Building a Quality Dataset Through Digital Literacy Enhancement: Developing Age-Appropriate AI to Support Student Learning

Ricky Cangniago[1] and Viny Christanti Mawardi[2]

[1,2] Universitas Tarumanagara, Jakarta, 11440, Indonesia
ricky.535220210@stu.untar.ac.id

Abstract. For artificial intelligence (AI) to work well in education, it needs high-quality datasets that accurately reflect how people interact in the classroom. This study describes a community service project at SDN 129 Rancasawo Bandung that aims to create a high-quality conversational dataset between teachers and students to help create age-appropriate educational AI. By systematically recording classroom conversations with professional audio equipment (Saramonic), we hope to improve digital literacy and make useful tools for AI development. The compilation of data captures authentic dialogues between educators and learners, which can be utilized to develop artificial intelligence frameworks aimed at facilitating the educational experiences of elementary school pupils in a manner that aligns with their individual learning preferences. Our work connects digital literacy programs with the technical development of long-lasting educational AI, making them two goals that are related to each other.

Keywords: Digital Literacy, Educational AI, Dataset Development, Innovation for the Future, Classroom Interaction.

1 Introduction

1.1 Background

Even though AI is becoming more a common in educational tools, making AI that really helps young students is still a big problem. The main problem is that there aren't enough good datasets [1]. Digital literacy, which means being able to use, understand, and make digital content well, is now a must-have skill in modern education [2]. The research team contends that the improvement of digital literacy and the development of AI datasets should not be considered distinct endeavors. Instead, putting them together makes for a strong, new way to improve education.

Indonesian elementary schools are facing a growing disconnect, while educational technology is advancing rapidly, the localized datasets needed to train effective AI are critically lacking [3]. This shortage directly undermines the performance of AI educational tools, as they often fail to grasp the local languages, culturally specific teaching methods, and the actual learning patterns of Indonesian students [4], [5].

© The Author(s) 2026
K. Keni et al. (eds.), *Proceedings of the fourteenth International Conference on Entrepreneurship and Business Management 2025 (ICEBM 2025)*, Advances in Economics, Business and Management Research 387,
https://doi.org/10.2991/978-94-6239-670-8_12

1.2 Problem Statement

The predominant reliance on Western datasets for educational AI tools creates several critical problems in the Indonesian context. First, these systems fail to capture the ambience of Indonesian language as it is spoken in the classroom, including local dialects and colloquial language. Second, they lack a fundamental understanding of the unique pedagogical approaches and styles of interaction between local teachers and students. This results in significant shortage of AI models trained on authentic elementary school conversations, a problem exacerbated by the minimal involvement of local communities in the data collection process itself.

1.3 Research Objectives

To address these challenges, the research has four main objectives. First, the project aims to develop a high-quality conversational dataset by systematically capturing teacher-student interactions at SDN 129 Rancasawo Bandung. Simultaneously, the project seeks to improve digital literacy among participating teachers and students through a participatory data collection model. These efforts collectively will create foundational resources for developing age-appropriate AI systems that support personalized learning. Finally, the research aims to establish best practices for ethical and effective dataset development in educational settings in Indonesia.

2 Literature Review

2.1 Digital Literacy in Education

The modern understanding of digital literation got evolved far beyond basic computer skills, now encompassing the crucial abilities to critically evaluate digital information and apply technology ethically [6]. Within the elementary education, these competencies serve as the cornerstone for 21st-century learning. Consequently, recent studies argue against teaching these skills in isolation, advocating instead for their direct integration into meaningful, hands-on technology use in the classroom.[6], [7].

2.2 AI in Educational Settings

Artificial intelligence (AI) in education to personalize learning, provide instant feedback, and reduce teachers administrative loads remains largely aspirational without its essential fuel in high-quality, representative training data [8]. This data challenge becomes even more acute when designing systems for young learners. For AI itself to be genuinely age appropriate, it cannot simply process information, it must be built upon data that captures the nuanced developmental stages, cognitive abilities, and emotional maturity unique to children [9].

2.3 Dataset Quality and Ethics

The creation of high-quality educational datasets presents a complex challenge that extends beyond technical metrics. It demands careful attention to factors like audio clarity, the diversity of interactions captured, and thorough contextual annotation. However, in a classroom setting, these technical goals must be carefully balanced against the non-negotiable ethical duty to protect student privacy and ensure a transparent informed consent process [10].

2.4 Technopreneurship and the Strategic Value of Local Datasets in EdTech

become a primary engine of innovation within the education sector, giving rise to Educational Technology (EdTech) industry. However, the efficacy of EdTech solutions often hinges on their adaptation to local contexts. As identified in this study's Problem Statement , the reliance on Western datasets creates a significant "market gap" in Indonesia.

This gap is not merely an academic problem but also a distinct business opportunity for local entrepreneurs. In AI-driven business models, a high-quality, proprietary dataset is a core strategic asset. A dataset that successfully captures deep local nuances such as the code-switching between Bahasa Indonesia and regional dialects , along with culturally specific pedagogical styles becomes a competitive advantage that is difficult to replicate. This enables the development of superior AI educational products that are genuinely tailored to the domestic market, a void that global competitors cannot easily fill.

3 Methodology

3.1 Research Setting

This research was done at SDN 129 Rancasawo, Bandung, a public elementary school serving a diverse student population. The school was chosen as the research site for several key reasons. At first, school administration shows a much strong desire to participate in educational progress and innovation. Second, the classroom environment and area are representative of public schools in Indonesia in general, making the findings potentially more generalizable. Furthermore, there was a clear commitment from teachers to engage in digital literacy development, which was crucial to the participatory nature of this research. The last one, the school's comprehensive representation of grade levels (Grades 1-6) provided an ideal opportunity to capture differences in communication development across elementary school levels.

Table 1. Demographic characteristics of SDN 129 Rancasawo Bandung

Characteristic	Description
Total Student	Approximately 12 students
Grade Levels Covered	6 (Grade 1 through Grade 6)
Average Class Size	2 students per class

Teacher Participants	6 teachers (1 per grade level)
Location	Rancasawo, Bandung, West Java
School Type	Public Elementary School
Socioeconomic Context	Mixed (predominantly middle-income)

3.2 Data Collection Process

To ensure a high-quality audio, the research team utilized professional Saramonic recording equipment. The technical setup is designed for comprehensive classroom recording, incorporating two directional microphones to clearly capture the voices of both teachers and students. Audio is recorded using a digital recorder with noise-canceling capabilities, and the entire process is protected by a backup recording system or cloud drive to prevent data loss.

Data collection followed a structured three phases of protocol to ensure data quality and ethical compliance. The pre recording stage begins with obtaining and discussing about the administrative permission and approval from school, teachers, and parents. This was followed by an orientation session to explain the project's purpose and process, and finally, setting up the equipment in a non disruptive position to minimize disruption. During the recording phase, our team captured natural classroom interactions across various subjects and grade levels, documenting diverse teaching scenarios such as lectures and group work, while maintaining detailed metadata logs. After the recording process, involved securing the data with encryption, conducting an initial audio quality assessment, and preparing the files to be ready for transcription.

3.3 Digital Literacy Enhancement Component

In conjunction with the data collection process, the research team conducted workshops focused on digital literacy for the educators involved in the study. The subsequent sessions were meticulously designed to encompass several pivotal themes, such as comprehending artificial intelligence and its applicability in enhancing educational outcomes for a greater number of students, deliberations on the ethical ramifications associated with educational technology, foundational principles of data literacy and privacy consciousness, as well as the cultivation of practical competencies in utilizing a diverse array of educational technology tools. In a distinct engagement, students were acquainted with age-appropriate interpretations of the mechanisms by which technology derives insights from data and the significance of their proactive involvement in the development of superior educational instruments.

3.4 Data Analysis Framework

Upon acquisition, the audio data is subjected to processing through a carefully designed analytical framework that includes multiple essential stages of processing. The operation starts with audio preprocessing, which integrates the application of noise reduction methods, segmentation, and quality filtration to suitably arrange the unrefined audio

recordings. Subsequently, the data is subjected to transcription, wherein the spoken language is systematically converted into written text, with particular emphasis placed on the diverse variants of the Indonesian language. The produced text is consequently directed towards annotation, a detailed protocol that involves the exact labeling of speaker roles, forms of interaction, and specific educational environments. The last stage features an exacting quality assurance analysis aimed at corroborating the reliability of the transcriptions and the clarity of the annotations, consequently ensuring that the final dataset is valid for research purposes.

4 Preliminary Results

4.1 Dataset Characteristics

Initial data collection has yielded 6 video recordings representing all elementary grade levels (Grade 1-6) at SDN 129 Rancasawo Bandung. Table 2 presents the comprehensive breakdown of the collected dataset.

Table 2. Dataset composition from classroom recordings at SDN 129 Rancasawo Bandung

Grade Level	Duration (minutes)	Subject	Recording Date	Audio Quality
Grade 1	5.43	Indonesian	28th May 2025	High
Grade 2	4.45	Indonesian	28th May 2025	Medium
Grade 3	3.50	IPAS	28th May 2025	High
Grade 4	4.32	PPKN	28th May 2025	High
Grade 5	5.04	IPAS	28th May 2025	High
Grade 6	5.34	IPAS	28th May 2025	High
Total	29.46	-	-	-

The dataset demonstrates comprehensive coverage across all elementary grade levels, enabling analysis of age-specific communication patterns and pedagogical approaches. Each recording captures authentic teacher-student interactions using professional Saramonic audio equipment, ensuring high-quality data suitable for AI model training. The initial data collection resulted in a total of 29.46 minutes of authentic classroom dialogue, distributed across all six elementary grade levels.

4.2 Recording Session Overview

The data collection endeavor effectively documented genuine classroom interactions across all six elementary grade levels at SDN 129 Rancasawo Bandung. All recording sessions transpired during standard instructional hours to guarantee that the interactions captured accurately depict unaltered classroom dynamics devoid of artificial interventions.

In order to promote consistency throughout all instructional sessions, a uniform method of documentation was adopted. A comprehensive lesson was meticulously documented for each educational grade, encompassing grades 1 through 6, and the sessions

were conducted with minimal disruption to conventional pedagogical practices. Furthermore, the Saramonic audio equipment was judiciously positioned to proficiently capture the vocalizations of both educators and students, while teachers were prompted to execute their lessons in a manner congruent with their established methodologies to preserve the authenticity of the educational interactions.

This methodological framework guarantees that the dataset authentically reflects Indonesian elementary classroom environments, encapsulating genuine pedagogical strategies and student-teacher interactions that are crucial for the development of contextually relevant AI systems.

4.3 Initial Observations from Recordings

A preliminary review of the recorded materials reveals several key characteristics valuable for AI development, particularly in language use patterns and classroom interaction types. Linguistically, we observed frequent and natural code-switching between Bahasa Indonesia and the regional dialect, alongside age-appropriate vocabulary and varied sentence structures. The dataset also successfully captured a rich diversity of interaction types, including direct instruction, question-and-answer exchanges, active student participation, classroom management dialogue, and teacher feedback.

4.4 Technical Quality Assessment

To ensure the dataset has adequate quality for training AI models, a technical quality assessment was performed. The primary focus of this assessment was on the transcribable clarity of the audio and the overall data integrity. The use of professional recording equipment (Saramonic Blink 500 B2) served as the primary foundation for achieving the targeted quality standards.

The technical specifications of the recording process and the results of the qualitative assessment of the collected data are summarized in the table below.

Table 3. Technical quality assessment of the recorded dataset

Technical Parameter	Specification / Assessment
Recording Equipment	Saramonic Blink 500 B2 Wireless Microphone System
Audio File Format	WAV
Audio Settings	48kHz, 24-bit
Microphone Configuration	Dual-channel system: 1 transmitter for the teacher, 1 transmitter for the student area
Qualitative Audio Assessment	Speech (between teacher and students) is clear and distinctly audible
Background Noise Level	Low and non-intrusive to the main dialogue
Data Integrity	No data loss or file corruption detected

The combination of a lossless audio format (WAV) and high auditory clarity ensures that this dataset is highly adequate for an accurate transcription process. Accurate transcription is a fundamental step before the data can be effectively used to train speech

recognition and NLP models. Therefore, this technical quality validates the dataset's suitability for the purpose of developing educational AI within the Indonesian context.

During the data acquisition process, the project team successfully identified and addressed several practical challenges to maintain quality. To mitigate background noise, microphones were strategically placed near the primary sound sources, and the classroom was conditioned to minimize external disturbances. Regarding student privacy concerns, a comprehensive informed consent procedure was implemented prior to recording, with a clear protocol for anonymizing the data during the transcription phase. Finally, to maintain natural participant behavior, we conducted an acclimatization period where the equipment was left in the classroom before recording began, allowing everyone to become accustomed to its presence.

5 Discussion

5.1 Implications for Age-Appropriate AI Development

The dataset developed through this research offers a holistic framework for the enhancement of artificial intelligence systems that are authentically adapted to the context of Indonesian elementary education. To begin with, the documented interactions highlight important linguistic and communicative patterns, including the intrinsic occurrence of code-switching between the Indonesian language and various regional dialects, the use of simplified terminology to clarify intricate concepts, and culturally significant pedagogical analogies. A comprehensive grasp of this linguistic framework is necessary for deciphering the educational landscape, which ultimately facilitates the design of an AI that can proficiently address multiple educational situations, whether that involves clarifying concepts, originating guiding questions, or supplying motivational support. Ultimately, this reaches a point of true developmental relevance, as the dataset comprises the age-specific communicative patterns that facilitate AI systems in calibrating their responses to the varied cognitive and emotional developmental stages of elementary school students.

5.2 Digital Literacy as Innovation Driver

This research demonstrates that improving digital literacy can be integrated with technological innovation, rather than treated as a separate objective. By directly involving the education community in the dataset development process, we achieved several interrelated outcomes. This participatory model fostered the development of practical digital literacy through meaningful engagement, which in turn fostered community ownership of the resulting technology initiatives. Furthermore, this approach ensured an ethical data collection process and fully respected participants' rights to participate. In overall, these elements form a sustainable innovation model that builds local capacity, empowering communities as active partners in their own technological advancement.

5.3 Future Implications

The methodology established in this research can be scaled to other educational contexts, creating a comprehensive dataset representing diverse Indonesian educational settings. This foundation supports the development of AI systems that truly understand and serve local educational needs rather than imposing foreign models.

5.4 Implications for Educational Entrepreneurship

This research offers practical implications for entrepreneurs in the educational field. First, the methodology employed in this project can be adopted as a model for lean startup practices in EdTech product development. The participatory data collection process, involving a direct partnership with SDN 129 Rancasawo, serves as a profound form of customer discovery and co-creation. By engaging end-users (teachers and students) from the outset, this approach not only ensures data authenticity but also validates market acceptance and builds relationships with potential early adopters, which is critical for de-risking product failure.

Second, the resulting dataset, while still at an initial scale, functions as a proof-of-concept that validates a market need. It demonstrates a tangible problem (the failure of global AI to understand local context) that is solvable with technology. For an entrepreneur, this dataset represents the foundation for developing a Minimum Viable Product (MVP) an AI tool specifically trained to serve the Indonesian educational context.

6 Conclusions and Recommendations

6.1 Conclusions

The work in this project pioneers an approach where building a quality dataset is not just a technical task, but also a direct method for enhancing digital literacy within the community. By integrating technological innovation into a community service (PKM) framework, we created a symbiotic relationship: our research gained invaluable authentic data, and in return, the community developed critical digital skills.

The initiative's key achievement is comprehensive. The project successfully developed an ethical and effective educational data collection protocol, facilitating the creation of a high-quality conversational dataset representing authentic classroom interactions in Indonesia Through this participatory process, we also achieved a significant enhancement of digital literacy among participating teachers and students. Overall, these outcomes provide a foundational framework for developing age-appropriate AI systems that are based on local educational contexts.

6.2 Recommendations

Based on the research and studies conducted, there are a few conclusion, a series of recommendations are proposed to enchanced the strenghtness and more ethical educational AI ecosystem in Indonesia.

For the educational institutions, a crucial transformation towards a more participatory paradigm of innovation is advocated. This transformation commences with the adaptation and implementation of a methodology that actively engages educators and learners, thereby converting them from passive recipients into dynamic co-creators. To enable the transition itself, it is recommended that educational institutions allocate resources towards effective digital literacy initiatives, ensuring that users possess both theoretical understanding and a focus on practical and direct involvement. All of these initiatives must lie the establishment of clear ethical protocols regulating the gathering and application of data, ensuring that innovation may thrive within a safe and accountable context.

For the policymakers, the priority should be to build a national framework that supports these grassroots initiatives. A recommended focus is on policies that build local capacity in educational technology, thereby reducing the nation's reliance on ill-suited foreign models. This includes creating robust legal and ethical frameworks for the safe development of AI in education and actively promoting strategic partnerships between educational institutions and technology developers. A few of collaborations are essential to accelerate the journey from research to the deployment of practical, locally relevant tools in classrooms.

For the future research, this study lays the foundation for a clear and critical research agenda. The next step is to expand data collection efforts to encompass broader geographic or regional and socioeconomic contexts, which is crucial for building a truly nationally representative dataset. The next phase should focus on leveraging the dataset to develop and rigorously test new AI models to validate their practical utility. Moving forward, longitudinal studies are highly recommended to measure the tangible impact of these AI-assisted tools on student learning outcomes across grades in Indonesia.

For Technopreneurs and EdTech Startups, this study highlights a clear market opportunity. Entrepreneurs are encouraged to move beyond merely adapting foreign AI models and instead invest in developing fundamentally localized solutions. The participatory methodology presented herein can be adopted as a best practice for building proprietary datasets. It is these datasets that will serve as the key strategic asset for creating a sustainable competitive advantage. We recommend that startups form direct partnerships with educational institutions mirroring this project's community service (PKM) framework to act as "living labs" for co-creating and validating technology that authentically answers local educational needs.

6.3 Limitations

This study having several limitations that should be considered when interpreting its findings. At first, the geographic scope was limited to a single school in Bandung, which may affect the generalizability of the results to the other educational contexts in Indonesia. Additionally, time constraints affected the total of data that been collected, limiting the width of classroom scenarios captured in the dataset. Furthermore, while professional equipment was used, technical limitations may have prevented the capture of all subtle nuances of the classroom interactions, such as non-verbal cues. Finally, the presence of the research team and recording equipment could have introduced potential

observer effects, possibly influencing participant behavior despite efforts to minimize disruption.

Acknowledgments. This research was conducted as part of community service activities (PKM) at SDN 129 Rancasawo Bandung. We extend our gratitude to the school principal, participating teachers, students, and parents for their valuable contribution. We also thank Tarumanagara University for supporting this research initiative.

References

1. M. Grizioti and M.-S. Nikolaou, "'Playing, Moving and Designing with Data: Exploring Young Students'' Data Literacy Skills in Embodied Classification Games",'" in *Proceedings of the 3rd International Conference of the ACM Greek SIGCHI Chapter*, New York, NY, USA: ACM, Sep. 2025, pp. 197–202. doi: 10.1145/3749012.3749080.
2. R. Kurniawan and A. W. Kuncoro, "The Influence of Digital Competence, Digital Literacy, and Emotional Intelligence on Teacher Performance (Case Study at Senior High School 14, Tangerang City)," *Jurnal Multidisiplin Sahombu*, vol. 5, 2025, doi: 10.58471/jms.v5i06.
3. D. Mertkan Gezgin, "Examination of Teachers' Digital Literacy Levels in The Era of Digitalization in Education." [Online]. Available: https://www.ubakkongre.com/efes
4. E. Shalomita Hana, "Beyond Inevitability: AI Tutoring and Educational Equity in Indonesia", doi: 10.1007/s40593.
5. D. Wong-A-Foe, "Navigating the Implications of AI in Indonesian Education: Tutors, Governance, and Ethical Perspectives," in *Communications in Computer and Information Science*, Springer Science and Business Media Deutschland GmbH, 2023, pp. 349–360. doi: 10.1007/978-981-99-7969-1_26.
6. S. Papadakis, "COMPUTATIONAL THINKING BEYOND CODING IN EARLY CHILDHOOD EDUCATION: REFRAMING CS EDUCATION FOR THE AGE OF AI," 2025, *Scientia Socialis Ltd*. doi: 10.33225/jbse/25.24.592.
7. F. S. Hapsari, M. Farid, M. Ahyar, A. Ahmad, and M. Yusup, "INTEGRATING DIGITAL LITERACY FOR ELEMENTARY SCHOOL LEARNING: A COMMUNITY SERVICE PROGRAM."
8. J. Jr. G. Adil, "AI in Education: A Systematic Literature Review of Emerging Trends, Benefits, and Challenges," *Seminars in Medical Writing and Education*, vol. 4, p. 795, Sep. 2025, doi: 10.56294/mw2025795.
9. A. I. Amjad, S. Aslam, and Z. A. Sial, "Beyond borders: Examining bullying, social networks, and adolescents mental health in developing regions," 2024, *Frontiers Media SA*. doi: 10.3389/feduc.2024.1431606.
10. A. Langer, P. J. Marshall, and S. Levy-Tzedek, "Ethical considerations in child-robot interactions," Aug. 01, 2023, *Elsevier Ltd*. doi: 10.1016/j.neubiorev.2023.105230.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution- NonCommercial - NoDerivatives 4.0 International License (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits any noncommercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if you modified the licensed material. You do not have permission under this license to share adapted material derived from this chapter or parts of it.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.