

EVALUATION DATA ON SYSTEM PERFORMANCE, USABILITY, AND USER ACCEPTANCE OF EDGE-BASED SELF-REFLECTIVE RAG HEALTH ASSISTANT

Binsarniari Sinaga* ¹
Ngatindriatun Ngatindriatun ¹
Hanif Pandu Suhito ²
Sholikun Sholikun ²
Prahita Indriana Rantiasmi ²

¹Digital Business, Binus Business School, Binus University, Semarang, Indonesia, 2026
²Semarang City Health Office (Dinas Kesehatan Kota Semarang), Semarang, Indonesia, 2026

Corresponding Author's Email: binsarniari.sinaga@binus.ac.id

Keywords: Edge Computing; Gemma 4; Retrieval-Augmented Generation (RAG); Two-Stage PLS-SEM; System Usability Scale (SUS); Health Informatics Dataset

ABSTRACT

Public healthcare services in low-resource environments face significant data privacy risks and cost barriers in deploying cloud-based AI. To address this, Emi an offline, localized generative AI digital health assistant was developed for the Semarang City Health Office (Dinas Kesehatan Kota Semarang) to deliver secure Psychological First Aid (PFA) counseling and Non-Communicable Disease (PTM) health screening and education. This data article describes the technical performance and user usability evaluation datasets compiled during Emi’s development and deployment. The repository contains two primary datasets: technical interaction logs (N=115) mapping query categories (such as greeting, emergency, affective, general, and medical RAG) to system processing metrics (latency and accuracy), and user usability and acceptance survey logs (N=20) evaluating Semarang’s Gen-Z demographic across structural equation modeling (SEM) and System Usability Scale (SUS) constructs. These datasets provide empirical baselines for validating privacy-preserving, localized generative AI in municipal healthcare, serving as rare research data at the intersection of human-computer interaction, edge computing, and digital health informatics.

SPECIFICATIONS TABLE

Subject	Computer Science, Human-Computer Interaction, Health Informatics
Specific subject area	Evaluating the usability, user acceptance, and computing performance of a localized, quantized generative AI assistant for public health services in a low-resource environment.

Type of data	Tables, CSV files (technical logs and survey questionnaires)
Data collection	Data were collected via (1) automated system log recordings during technical testing (N=115 queries) and (2) a post-interaction questionnaire survey (N=20) with representative Generasi Z users in Semarang
Data source location	Semarang City, Central Java, Indonesia (Geographical coordinates: 6.9667 deg S, 110.4167 deg E)
Data accessibility	Repository name: Zenodo Data identification number: DOI 10.5281/zenodo.22249913 Direct URL: https://doi.org/ 10.5281/zenodo.22249913 [12]
Related research article	none

1. Value of the Data

- The datasets provide rare empirical evidence of local, privacy-preserving LLM deployment (Google Gemma 4 12B) in public healthcare under consumer hardware constraints ($\leq 8\text{GB}$ VRAM) [12].
- The interaction logs (N=115) provide detailed, category-specific benchmarking of latency and accuracy, serving as a baseline to optimize local retrieval strategies in RAG pipelines [12].
- The survey dataset (N=20) integrates both usability (SUS) and acceptance (TAM fused with Social-Trust) to validate behavioral modeling in human-agent interactions [3, 12].
- The qualitative feedback (n=11) captures UX friction points, such as computing latency and conversational consistency, helping explain statistical findings and adoption gaps in public AI services [12].

2. Background

Urban health in Semarang City faces a severe burden of Non-Communicable Diseases (PTM), with 47,127 new cases of hypertension and 22,304 new cases of type-2 diabetes mellitus recorded in 2024 [1]. At the national level, adolescents face a critical mental health crisis: one-third (15.5 million) of Indonesian adolescents experience mental health challenges, with depression affecting 2% of youth aged 15–24—the highest among all age groups—and 61% of these depressed youth contemplating suicide [2]. Furthermore, a severe treatment gap exists, as only 10.4% of depressed adolescents seek professional treatment [2]. These local and national health vulnerabilities are directly reflected in the field survey of 20 representative Gen-Z users in Semarang. Specifically, sleep duration logs reveal that 95% of respondents suffer from insufficient sleep, with 25% sleeping less than 5 hours and 70% sleeping only 5 to 7 hours per night. A significant communicative barrier is also documented: only 10% frequently share personal struggles, while 25% never open up to others, indicating widespread psychological isolation. Moreover, early-onset metabolic risks are already manifest: 20% of these young respondents report a personal history of hypertension or diabetes, and an alarming 25% remain completely unaware of their status. To address these overlapping physical and mental health challenges under local resource constraints, 'Emi' was developed as a localized, offline digital health assistant. Built on a 4-bit quantized

Google Gemma 4 12B model and Qdrant, Emi provides a cost-effective, secure, and privacy-preserving edge computing architecture to deliver secure Psychological First Aid (PFA) counseling and PTM screening directly to the community [12].

3. Data Description

The published repository comprises two primary datasets: emi interaction logs and user acceptance survey which provide a multi-dimensional technical and user-centric evaluation of Emi's localized deployment.

Table 3.1: Database Schema for emi interaction logs

Column Name	Data Type	Description
Query_ID	Integer	Unique identifier for each interaction log.
Category	Categorical	Class of prompt: SAPAAN (greeting), DARURAT (suicidal/acute), KONSULTASI (affective/empathy), UMUM (general), MEDIS (RAG queries).
Latency_Seconds	Float	Total processing time in seconds for local model inference.
Accuracy_Score	Float	Manual expert verification score of factual correctness (0.0 to 1.0).
RAGAS_Faithfulness	Float	Retrieval-Augmented Generation Assessment of factual consistency with source documents.

Data Insight (Table 3.1): The database schema maps five core parameters, balancing system-level metrics (inference latency) with response quality indices (accuracy and RAGAS faithfulness) [12]. This structured schema enables multi-criteria evaluation of edge-based LLM deployment trade-offs [12].

Table 3.2: Technical Performance and Accuracy Summary by Query Category (N=115)

Category	Sample Size (n)	Mean Accuracy (%)	Median Accuracy (%)	Mean Latency (s)
SAPAAN (Greeting)	30	97.50%	100.0%	13.20 s
DARURAT (Emergency PFA)	4	93.75%	100.0%	62.10 s
KONSULTASI (Affective/Empathy)	13	80.77%	75.0%	59.47 s
UMUM (General Local)	52	55.77%	62.5%	39.55 s
MEDIS (Medical RAG)	16	51.56%	37.5%	67.90 s
TOTAL / AVERAGE	115	70.22%	100.0%	39.66 s

Data Insight: Greeting (97.50%) and emergency (93.75%) queries achieve high accuracy, but medical RAG queries show a drop (51.56%) and the longest processing time (mean = 67.90 s) due to local GPU processing bottlenecks.

Table 3.3: Outer Model Evaluation and Construct Reliability (Stage 1 PLS-SEM)

Construct	Cronbach's Alpha	Composite Rel. (rho_a)	Composite Rel. (rho_c)	AVE	Evaluation Result
ANTHRO (Anthropomorphism)	0.870	0.874	0.903	0.608	Valid & Reliable (>0.50 / >0.70)
BI (Behavioral Intention)	0.867	0.886	0.905	0.661	Valid & Reliable (>0.50 / >0.70)
PEOU (Ease of Use)	0.863	0.892	0.897	0.594	Valid & Reliable (>0.50 / >0.70)
PU (Perceived Usefulness)	0.925	0.943	0.936	0.623	Valid & Reliable (>0.50 / >0.70)
SP (Social Presence)	0.961	0.963	0.970	0.865	Valid & Reliable (>0.50 / >0.70)
SUS (System Usability)	0.775	0.813	0.849	0.537	Valid & Reliable (>0.50 / >0.70)
TRUST (Trustworthiness)	0.730	0.810	0.844	0.649	Valid & Reliable (>0.50 / >0.70)
HOC-SocTrust* (Social-Trust Construct)	0.934	0.939	0.958	0.885	Valid & Reliable (HOC)*
HOC-Util* (Utilitarian Construct)	0.785	0.790	0.904	0.825	Valid & Reliable (HOC)*

Data Insight (Table 3.3): Following Sarstedt et al. (2019), all seven lower-order constructs demonstrate excellent reliability (alpha, rho_a, rho_c > 0.70) and convergent validity (AVE > 0.50) [3, 11]. The calculated higher-order constructs (HOC-SocTrust and HOC-Util*) also show strong convergent validity (AVE >= 0.825) and reliability, establishing model robustness [11].*

Table 3.4: Structural Model Path Analysis Results (Stage 2 PLS-SEM)

Hypothesis	Path / Relationship	Path Coeff. (β)	Std. Error	t-Value	p-Value	f ² Effect Size	Decision
H3a	HOC-SocTrust -> BI	0.981	0.166	5.895	0.000	1.564 (Large)	ACCEPTED
H3b	HOC-Util -> BI	-0.081	0.172	0.472	0.319	0.000 (None)	REJECTED
H3c	BI -> SUS	0.841	0.107	7.882	0.000	2.422 (Large)	ACCEPTED
H3d (Med.)	HOC-SocTrust -> BI -> SUS	0.825	0.178	4.633	0.000	-	SIGNIFICANT MEDIATION
H3e (Med.)	HOC-Util -> BI -> SUS	-0.068	0.144	0.474	0.318	-	NOT SIGNIFICANT

Data Insight (Table 3.4): Under the two-stage PLS-SEM reporting standard of Sarstedt et al. (2019), Social-Trust (HOC-SocTrust) exhibits a dominant, significant direct effect on Behavioral Intention (beta = 0.981, p < 0.001, f² = 1.564), while utilitarian value is non-significant (beta = -0.081, p = 0.319) [11]. Behavioral Intention strongly translates to System

Usability ($\beta = 0.841$, $p < 0.001$, $f^2 = 2.422$) [11]. The model shows high explanatory power ($R^2 BI = 0.833$, $R^2 SUS = 0.708$), conforming to Hair et al. (2021) and Sarstedt et al. (2019) guidelines [3, 11].

4. Experimental Design, Materials and Methods

Using the Design Science Research (DSR) framework [8], the 'Emi' health assistant was developed and evaluated through four key stages:

1. Hardware & Model Quantization Setup: Google Gemma 4 12B [10] was quantized to 4-bit GGUF via Quantization-Aware Training (QAT) [5], reducing model size to ~3.5 GB for offline deployment on standard consumer-grade GPUs (≤ 8 GB VRAM) in local public health workstations.

2. Retrieval-Augmented Generation (RAG) Integration: A local instance of Qdrant Vector Database [7] indexed the official 'Profil Kesehatan Kota Semarang 2024' [1] and UNICEF's 'Indonesia Adolescent Health Profile 2024' [2]. A self-reflective retrieval pipeline indexed document chunks to fetch relevant health data and eliminate hallucinations.

3. On-Premise Deployment & Testing: A secure web-based chat interface was deployed locally [7]. Representative users entered queries, which were logged (N=115 interactions) to analyze processing latency, expert-verified accuracy, and RAGAS scores [12].

4. Usability & Psychometric Evaluation: N=20 Gen-Z users completed post-test surveys covering SUS, TAM [4], and Social-Trust indicators. Path relationships were modeled via Two-Stage PLS-SEM in SmartPLS 4 [3]. SUS ratings were calculated using standard formulas, while qualitative comments (n=11) were coded using thematic analysis [9].

5. Limitations

The primary limitation is the small sample size for user acceptance testing (N=20) and qualitative feedback (n=11), focusing exclusively on Gen-Z respondents in Semarang City. Additionally, the technical log dataset (N=115) represents evaluations conducted under controlled hardware conditions (single-node edge workstation with ≤ 8 GB VRAM) [12]. External network variations or multi-device deployments are not represented.

6. Ethics Statement

The authors confirm that the current work does not involve high-risk clinical interventions or invasive animal experiments. Informed consent was obtained from all N=20 human subjects who participated in the usability and acceptance testing [12]. All data collected from users have been fully anonymized to preserve participant privacy, adhering to the ethical standards for publication in Data in Brief [12].

7. CRediT Author Statement

- Binsarniari Sinaga: Conceptualization, Methodology, Software, Investigation, Data Curation, Formal Analysis, Writing - Original Draft, Writing - Review & Editing.
- Ngatindriatun: Supervision
- Hanif Pandu Suhito: Supervision, Resources, Validation
- Sholikun: Supervision, Resources, Validation
- Prahita Indriana Ranasmi: Supervision, Resources

8. Acknowledgements

We express our gratitude to the Semarang City Health Office (Dinas Kesehatan Kota Semarang) for providing the official data sources used in this project [1]. We also thank the Digital Business Program of Binus University Semarang for the technical resources and administrative support. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

9. Declaration of Competing Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

10. References

- [1] Dinas Kesehatan Kota Semarang, Profil Kesehatan Kota Semarang Tahun 2024, Pemerintah Kota Semarang, 2024.
- [2] UNICEF Indonesia, Indonesia Adolescent Health Profile 2024, United Nations Children's Fund, 2024.
- [3] J. F. Hair, G. T. M. Hult, C. M. Ringle, M. Sarstedt, N. P. Danks, S. Ray, Partial Least Squares Structural Equation Modeling (PLS-SEM) Using R: A Workbook, Springer, 2021. <https://doi.org/10.1007/978-3-030-80519-7>
- [4] F. D. Davis, A. Granić, The Technology Acceptance Model: 30 Years of TAM, Springer Nature Switzerland, 2024. <https://doi.org/10.1007/978-3-030-45274-2>
- [5] V. Egiazarian, A. Panferov, D. Kuznedelev, E. Frantar, A. Babenko, D. Alistarh, Extreme Compression of Large Language Models via Additive Quantization, arXiv preprint arXiv:2401.06118 (2024). <https://doi.org/10.48550/arXiv.2401.06118>
- [6] N. Epley, A. Waytz, J. T. Cacioppo, On seeing human: A three-factor theory of anthropomorphism, Psychological Review 114 (2007) 864-886. <https://doi.org/10.1037/0033-295X.114.4.864>
- [7] K. Edemacu, X. Wu, Privacy Preserving Prompt Engineering: A Survey, ACM Computing Surveys 57 (2025) 1-36. <https://doi.org/10.1145/3729219>
- [8] K. Peffers, T. Tuunanen, C. E. Gengler, M. Rossi, W. Hui, The Design Science Research Process: A Model for Producing and Presenting Information Systems Research, Journal of Management Information Systems 24 (2007) 45-77. <https://doi.org/10.2753/MIS0742-1222240302>
- [9] V. Braun, V. Clarke, Using thematic analysis in psychology, Qualitative Research in Psychology 3 (2006) 77-101. <https://doi.org/10.1191/1478088706qp063oa>
- [10] Gemma Team, Gemma 4 Technical Report, arXiv preprint arXiv:2607.02770 (2026). <https://doi.org/10.48550/arXiv.2607.02770>
- [11] M. Sarstedt, J. F. Hair Jr, J.-H. Cheah, J.-M. Becker, C. M. Ringle, How to Specify, Estimate, and Validate Higher-Order Constructs in PLS-SEM, Australasian Marketing Journal 27 (2019) 197-211. <https://doi.org/10.1016/j.ausmj.2019.05.003>
- [12] B. Sinaga, N. Ngatindriatun, . hanif . pandu . suhito, S. Sholikun, and P. Indriana Rianasmi, "Dataset on System Performance, Usability, and User Acceptance of Edge-Based Self-Reflective RAG Health Assistant". Zenodo, Sep. 02, 2026. doi: 10.5281/zenodo.22249913.