IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 15, No. 6, December 2026, pp. 5083~5095
ISSN: 2252-8938, DOI: 10.11591/ijai.v15.i6.pp5083-5095

Towards child-friendly conversational artificial intelligence: system design of an analogy-based chatbot using OpenAI

Eka Stephani Sinambela, Tanisha Natalia Aritonang, Samuel Arta Raju Manik, Grace Angel Nainggolan, Marojahan Mula Timbul Sigiro, Frengki Simatupang, Istas Pratomo Manalu
Department of Computer Technology, Faculty of Vocational Studies, Institut Teknologi Del, Laguboti, Indonesia

Article Info

Article history:

Received Sep 15, 2025
Revised Aug 10, 2026
Accepted Aug 31, 2026

Keywords:

Analogy-based learning
Artificial intelligence
Chatbot
Curiosity
Retrieval-augmented generation

ABSTRACT

Curiosity plays a central role in children's learning by motivating active engagement in question–answer interactions. According to Piaget's theory of cognitive development, children aged 6–12 are in the concrete operational stage, in which logical reasoning is best supported through concrete representations and familiar analogies. Motivated by this perspective, this study proposes an analogy-based conversational approach to facilitate child-friendly knowledge transfer across multiple domains, including physics, mathematics, chemistry, biology, geography, sociology, economics, computer technology, and everyday-life contexts. This work presents CurioKids, a retrieval-augmented generation (RAG)-based chatbot that generates analogy-driven explanations tailored to children's cognitive characteristics. The system employs a multi-stage evaluation pipeline, beginning with large language model (LLM)-based screening to assess analogy quality, followed by expert review to ensure semantic validity and pedagogical appropriateness. The curated dataset is then evaluated using semantic similarity (BERTScore), response time, and task success rate (TSR). Experimental results show that CurioKids achieves high semantic alignment with reference answers and successfully completes 73% of user tasks. A comparative evaluation against a general-purpose chatbot (Gemini) further demonstrates that CurioKids produces more semantically faithful and coherent responses. Overall, this study contributes a structured design and evaluation framework for developing reliable and educational conversational agents for children.

This is an open access article under the CC BY-SA license.

Corresponding Author:

Eka Stephani Sinambela
Department of Computer Technology, Faculty of Vocational Studies, Institut Teknologi Del
Sisingamangaraja Road, Sitoluama, Laguboti, Toba Region, North Sumatera, Indonesia
Email: ekastephanisinambela@gmail.com

1. INTRODUCTION

Curiosity is a fundamental driver of children's learning, motivating them to explore, ask questions, and acquire new knowledge [1], [2]. Research shows that curiosity manifests as information-seeking behavior driven by intrinsic motivation and closely related to children's existing interests and questions [3]–[5]. Among its forms, epistemic curiosity plays a central role in intellectual engagement by prompting learners to identify knowledge gaps and seek explanations [6]. Epistemic curiosity strongly supports the question–answer process, as children actively ask questions to reduce uncertainty and deepen understanding. Such curiosity-driven interactions align with the use of ChatGPT in educational settings, where learners engage with artificial intelligence (AI)-generated responses to critically explore and understand concepts [7]. Prior studies distinguish between convergent and divergent questioning strategies, with divergent questioning

supporting exploratory learning processes. Such open-ended inquiry is closely associated with epistemic curiosity, which motivates learners to seek explanations and construct a deeper understanding of unfamiliar concepts [8], [9]. From a developmental perspective, Piaget's theory suggests that children aged 6–12 increasingly develop logical reasoning skills but still rely on concrete representations to comprehend abstract concepts, especially when supported by analogical explanations [10]. Analogical reasoning has long been recognized as an effective cognitive mechanism for knowledge transfer, enabling learners to map familiar concepts onto unfamiliar domains. Structure-mapping theory formalizes this process by emphasizing relational correspondences rather than surface attributes [11].

This theoretical foundation has influenced early AI research, including computational models of analogy designed to support flexible reasoning and learning [12]. Such findings underscore the potential of analogy-based explanations to support children's conceptual development. Recent advances in large language models (LLMs) have significantly enhanced the capabilities of conversational AI systems**,** enabling fluent, context-aware, and scalable natural language interactions. Transformer-based models, such as BERT, GPT-2, and later large-scale models like GPT-3 and GPT-4, demonstrate strong performance across a wide range of language tasks [13]–[17]. These developments have accelerated the adoption of chatbots in educational settings, to support personalized learning and knowledge access [18]. However, most existing conversational AI systems are designed for adult users and often lack developmental sensitivity, age appropriateness, and safety considerations for children [19], [20].

To address this gap, this paper introduces a child-friendly chatbot that answers children's questions using analogy-based explanations. Leveraging OpenAI's generative models, this study design a system architecture that integrates analogical reasoning, developmental appropriateness, and contextual relevance. This study investigates the following research question: How effectively can a chatbot generate, meaningful and appropriate analogy-based responses for children? This study evaluate the system using semantic similarity (BERTScore) [21], response relevance, and response latency to assess real-time feasibility. By combining analogy-driven design with quantitative evaluation, this work contributes to child-centered AI by demonstrating a scalable approach to supporting children's learning through explainable and engaging question–answer interactions.

2. METHOD

2.1. System overview

The proposed system is a conversational AI chatbot designed to support children's question–answer interactions through age-appropriate, analogy-based explanations. The architecture comprises four primary components: input processing, analogy generation, context awareness, and response formulation, organized as a sequential processing pipeline. The system utilizes OpenAI's GPT-based language model as the core engine for natural language understanding and generation. When a child submits a question, a prompt-engineering layer encodes contextual and pedagogical constraints to guide explicit analogical reasoning before querying the language model. The generated response is subsequently post-processed to ensure clarity and contextual relevance before delivery.

2.2. Goal-question-metric approach

To evaluate the effectiveness of the proposed chatbot, we adopt the goal–question–metric (GQM) approach [22], which provides a systematic framework for aligning evaluation objectives with measurable indicators. The goal of this study is to analyze the design and application of an OpenAI-powered chatbot for supporting analogy-based, child-friendly question answering, with respect to explanation quality, answer relevance, and contextual awareness from the viewpoint of system designers and educational technology researchers, within the context of AI-based learning tools for children aged 3–12. To achieve this goal, this study formulate the following research question: How well does the chatbot generate analogy-based responses that are meaningful and appropriate for children? To answer the question, the following metrics have been chosen:

- BERTScore: measures semantic similarity between the chatbot's response and a reference answer, capturing response fluency and relevance. The score ranges from 0 to 1.
- Response time: the time the chatbot takes to generate a response, measured in seconds, to assess system responsiveness.
- Task success rate (TSR): measures the proportion of successfully addressed tasks by evaluating the preservation of key concepts and keywords from the input question in the generated response. TSR is calculated as (1).

$$TSR = \frac{Number\ of\ successfully\ completed\ tasks}{Total\ number\ of\ tasks} \times 100\% \quad (1)$$

2.3. Dataset and analogy-based question-answers format

To ensure ecological validity, we constructed a dataset comprising 1,300 frequently asked questions posed by children aged 6–12. These questions were collected from semi-structured interviews, educational websites, children's knowledge forums, and curated parenting platforms, reflecting real-world information-seeking behavior. The dataset spans a diverse range of domains, including physics, mathematics, chemistry, biology, geography, sociology, economics, computer technology, and everyday-life questions.

Rather than using the dataset for model training, it is employed exclusively as a retrieval resource within a retrieval-augmented generation (RAG) framework. This design choice preserves the general reasoning capabilities of the underlying LLM while ensuring that responses remain domain-relevant, age-appropriate, and pedagogically aligned with children's cognitive development. Insights derived from the dataset informed the design of the chatbot's response structure, which follows a consistent analogy-based educational format comprising four components: analogy, factual explanation, consequence, and conclusion. This structured format aims to ground abstract concepts in familiar experiences, thereby enhancing comprehension and supporting meaningful knowledge transfer. Table 1 presents the examples of the collected questions and their corresponding responses.

Table 1. Analogy question–answer example

Domain	Question	Factual explanation	Consequence	Conclusion
Mathematics	Why is 1/2 greater than 1/4?	The smaller the denominator of a fraction, the larger each part becomes. Since 1/2 represents a division into fewer parts, each portion is larger than that of 1/4.	This can be illustrated by cutting a cake into two parts, which results in larger pieces, compared to cutting the same cake into four smaller parts.	Therefore, because 1/2 involves fewer divisions, each resulting portion is larger.
Physics	Why does the sun feel hot?	The Sun emits energy because of nuclear reactions occurring in its core.	This process can be analogized to a massive, extremely hot campfire.	The immense amount of energy released causes the Sun to feel hot.
Biology	Why does the heart beat?	The heart beats to pump blood throughout the body.	This function can be compared to a water pump that distributes water to a house.	Through its rhythmic beating, the heart ensures the continuous circulation of blood.
Chemistry	Why can water put out fire?	Fire needs heat to stay alive. Water takes the heat away from the fire and makes it cooler.	It is like giving a cold drink to someone who feels very hot.	The fire goes out because the water reduces its temperature.
Computer Technology	Why do we use antivirus software?	Antivirus software detects and removes malicious programs to keep a computer system secure and functioning properly.	This process is analogous to a doctor who prescribes medicine to help the body recover and remain healthy.	Therefore, antivirus software protects computers from viruses and other malicious programs.

2.4. Research design

The chatbot's development employed an iterative, incremental agile development methodology, which provides flexibility to adapt to evolving requirements, enables rapid integration of user feedback, and supports continuous evaluation throughout the development cycle. Figure 1 describes the process that begins with a requirement analysis phase involving both literature review and field data. The team conducted direct observations of interactions among children aged 3–12 and interviewed both children and parents to understand common question patterns, thinking styles, and the challenges in explaining complex concepts. Subsequent stages included system design, implementation, and testing.

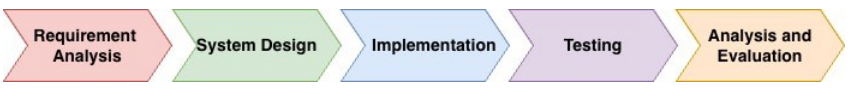

Figure 1. Research methodology

2.4.1. High level architecture

The system design phase involved defining the system architecture, modeling the question-answering process, designing a child-friendly user interface, and constructing the database structure.

The testing phase included both functional and non-functional testing, with a focus on performance. Additionally, answer quality was evaluated using metrics such as BERTScore, TSR, and response time to ensure the chatbot delivers relevant, accurate, timely, and pedagogically sound responses. The proposed system adopts a RAG architecture [23]–[25] to support analogy-based, child-friendly question answering, as illustrated in Figure 2. The design grounds LLM responses in a curated knowledge base, enabling contextually relevant and pedagogically structured explanations for children aged 6–12.

Figure 2 illustrates how the system operates through three main stages. First, during data ingestion, the system transforms curated question–answer content into semantic vector representations using OpenAI's text-embedding-3-small model and stores them in a vector database implemented with Supabase PGVector. Second, during query processing, the system converts a child's question into an embedding and retrieves relevant contextual information by matching it against stored vectors using cosine similarity. Third, during response generation, the system combines the retrieved context with the original query to construct a structured prompt for an OpenAI GPT model. The prompt constrains the model to generate responses in simplified Indonesian, following a predefined analogy-based educational structure consisting of analogy, factual explanation, consequence, and conclusion. This separation between retrieval and generation allows the system to update knowledge resources without retraining the language model, while ensuring consistent tone, style, and educational structure across responses.

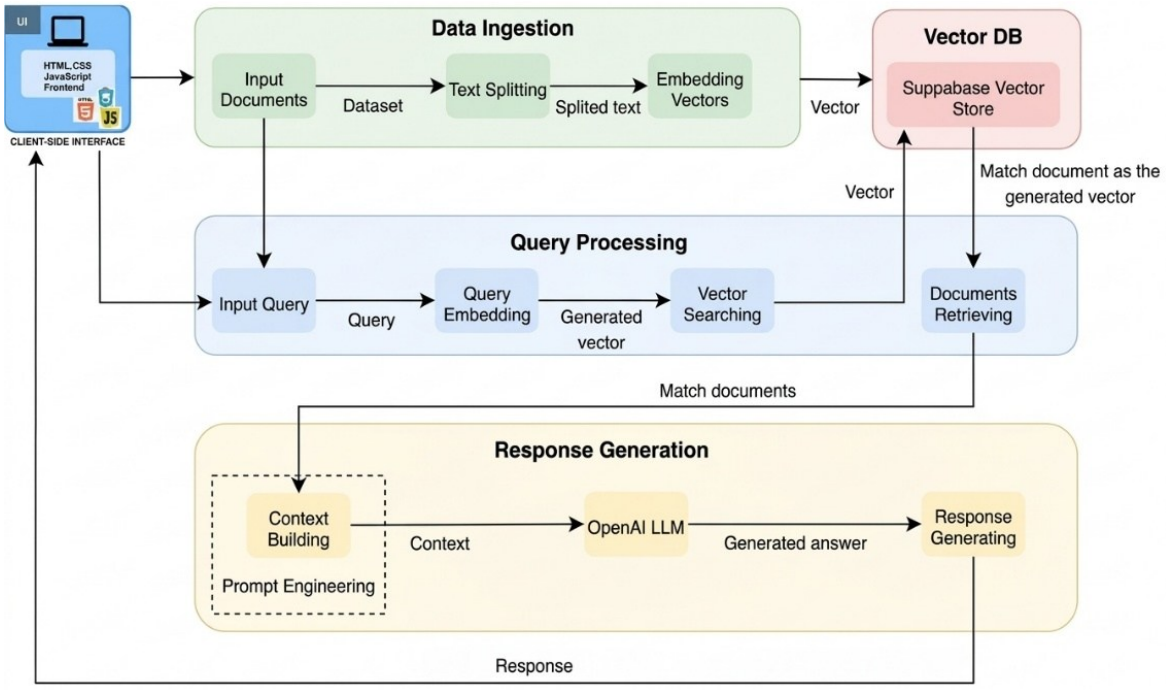

Figure 2. High level architecture chatbot-AI based

2.4.2. Analogy question–answer chatbot model

To operationalize the high-level architecture, the system follows a comprehensive workflow that coordinates real-time interactions among the user interface, backend services, and the vector database, as illustrated in Figure 3. The workflow initiates at the client-side interface, where the user submits a natural-language question. Upon receiving the query, the backend determines whether a session history is present. If an existing session is found, the system retrieves previous interactions to ensure conversational coherence; if not, it initializes a new session and stores the session data in both memory and persistent storage. Next, the backend conducts a semantic search on the Supabase vector database to retrieve contextually relevant knowledge chunks from the internal dataset. The retrieved content is combined with the current query and existing conversation context to form a structured input for the OpenAI language model. Once a response is generated, the system sends it to the frontend and logs the query-responses pair in the session history. If the input is identified as a question, it separately embedded and indexed to support interaction analysis without modifying the retrieval dataset.

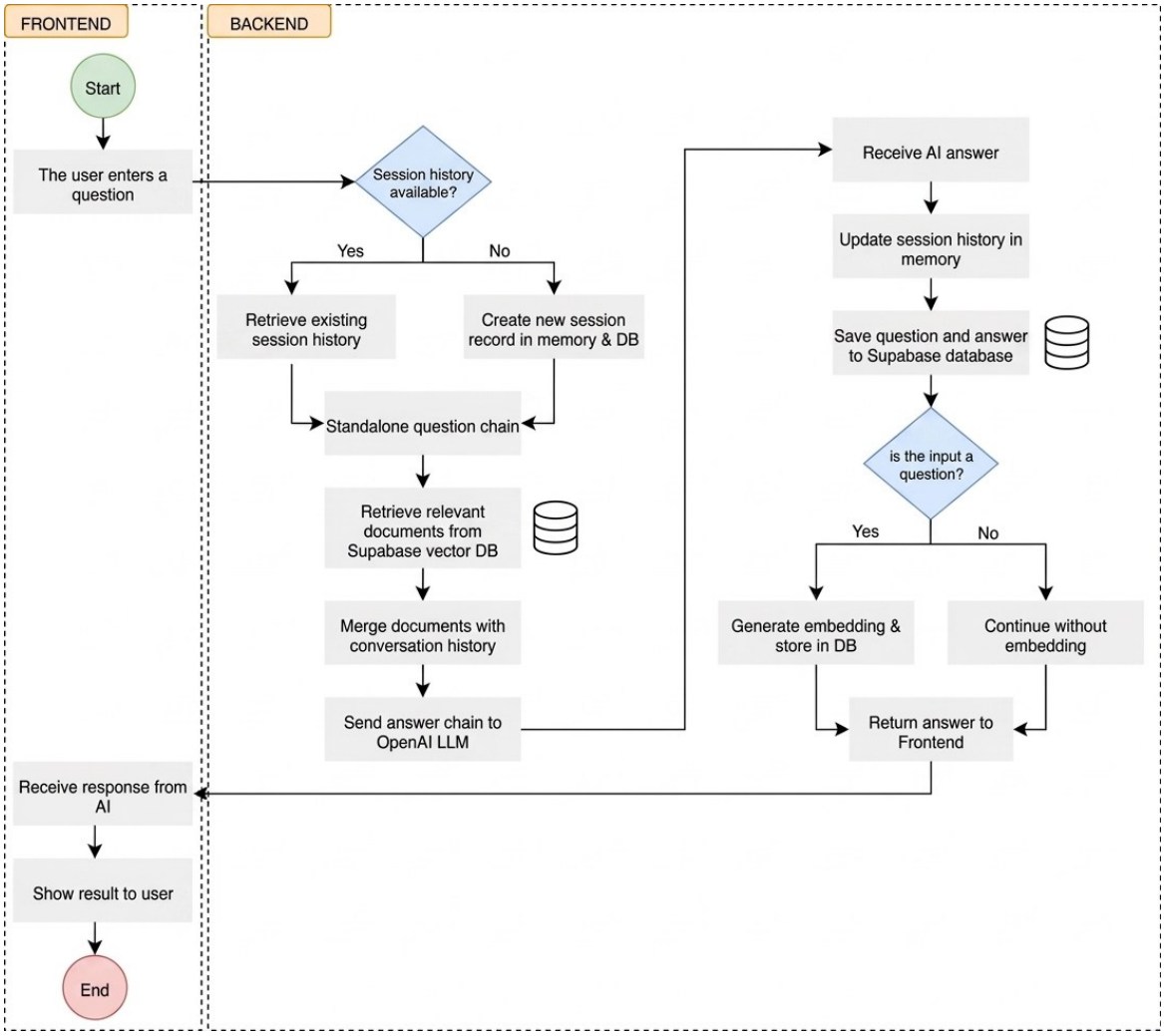

Figure 3. The flowchart of the AI-based models

This workflow enables context-aware, multi-turn dialogue while preserving the integrity of the retrieval corpus. Figure 4 present examples of the chatbot's responses, where Figure 4(a) to an internal prompt and Figure 4(b) to an external prompt. This design demonstrates how retrieval and generation complement each other in supporting analogy-based, child-friendly question answering.

2.4.3. Analogy question–answer quality control

Figure 5 illustrates the quality control process for analogy-based question–answer pairs generated by the chatbot. The method combines automated screening performed by an LLM with human review to ensure that the chatbot produces pedagogically sound, age-appropriate, and safe responses for children aged 6–12. The system enforces a structured response format consisting of an analogy, a factual explanation, a consequence, and a concluding statement. This structure promotes consistent explanation quality while helping children understand abstract concepts through familiar experiences.

In the first stage, the LLM conducts automated pre-screening using predefined criteria. The system checks factual correctness to detect misconceptions, evaluates sensitivity to identify potentially inappropriate content, and assesses overall explanation quality. The LLM assigns quality scores for clarity, familiarity, alignment with the user's question, cultural fit, and safety. Based on these indicators, the system classifies each response as pass, borderline, or rejected. In the second stage, human reviewers evaluate the responses using an expert judgment rubric. General reviewers assess clarity and age appropriateness, while domain experts examine scientific accuracy, misconception risk, and contextual safety. Reviewers score each response across five dimensions: scientific accuracy, age appropriateness, analogy clarity, misconception risk, and contextual safety. The review process determines final decisions (pass, borderline, or rejected) using

predefined scoring thresholds. The system subject borderline responses to targeted revision guided by the lowest-scoring criteria.



(a) (b)

Figure 4. CurioKids chatbot responses for evaluating adaptability and consistency: (a) internal prompt and (b) external prompt

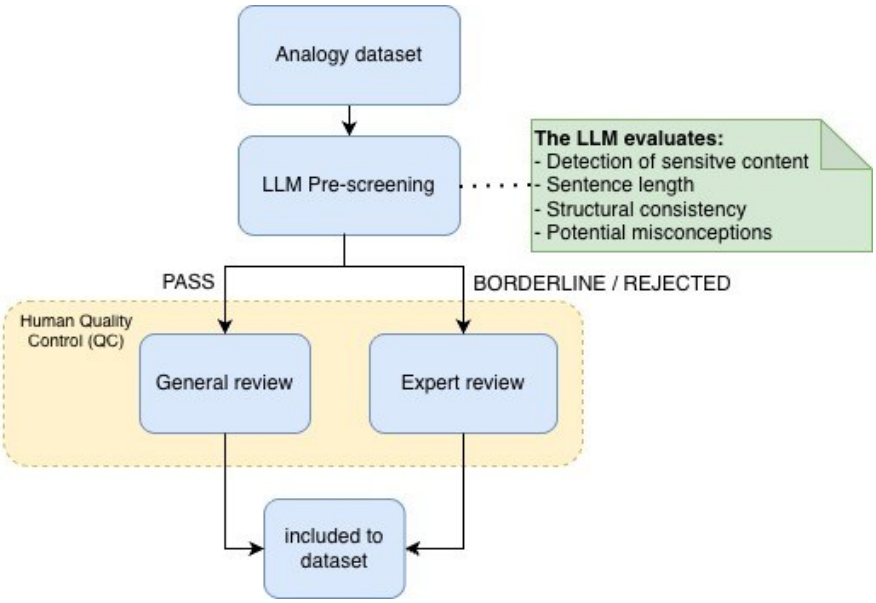

Figure 5. The analogy quality control

This quality control process aligns directly with the GQM framework adopted in this study. The quality control pipeline operationalizes the evaluation goal of supporting child-friendly, analogy-based question answering by enforcing pedagogical and safety constraints before quantitative assessment. This process also addresses the research question regarding the meaningfulness and appropriateness of generated analogies through both automated screening and expert human judgment. Moreover, the quality control mechanism complements the selected evaluation metrics. The system ensures semantic alignment and conceptual coverage before computing BERTScore and TSR, while response time captures system responsiveness during real-time interaction. By integrating quality assurance before quantitative evaluation, the study ensures that performance measurements reflect not only linguistic similarity and efficiency but also educational validity and child-centered design principles.

2.4.4. Implementation environment

The system operates on a cloud-based infrastructure to support scalable and reproducible deployments. This study implement the backend using Supabase, which provides PostgreSQL database services, authentication, and vector storage via the pgvector extension to enable RAG functionality. The backend runs on shared cloud resources, approximately two vCPUs and 8 GB of RAM, which are sufficient for vector retrieval and handling concurrent user requests. This study deploy the application on Vercel, which automatically scales system resources in response to user traffic. The OpenAI API is integrated for natural language processing and employing the text-embedding-3-small model for semantic retrieval and GPT-3.5 for response generation. Under typical usage conditions of approximately 1,000 conversations, the system incurs an estimated operational cost ranging from USD 5 to 30 per month. This study implement the software stack using JavaScript (ES6+) on a Node.js runtime with RESTful API architecture. Short-term conversational memory is managed using an in-memory cache with a 30-minute time-to-live (TTL), which stores session-level responses and contextual information. This design allows the system to reuse previously generated responses within the same session, reduce response latency, minimize redundant API calls, and lower overall token consumption.

3. RESULTS AND DISCUSSION

3.1. Results

Following the methodology described in the previous section, this part presents the evaluation and discussion of the CurioKids chatbot in a structured, multi-stage manner. As outlined in the system pipeline, the review begins with an LLM-based screening of all generated analogy-based question–answer pairs, which are categorized into pass, borderline, and reject based on predefined analogy quality indicators. Items classified as borderline and reject are then revised by human reviewers, including general reviewers and external expert reviewers, using a rubric. This process yields in a curated final dataset that aligns with the quality control strategy outlined in the methodology. The finalized dataset is then evaluated quantitatively using three performance metrics: BERTScore to measure semantic alignment between generated and reference answers, response time to assess system efficiency, and TSR to evaluate task completion effectiveness. Finally, to examine the robustness of the proposed system, CurioKids is compared with Gemini, using BERTScore as the primary metric.

The following subsections present the results of each evaluation stage and discuss their implications in Table 2. The screening outcomes support the suitability of LLM-based review as a scalable first-stage quality control mechanism. The relatively high proportion of pass and borderline items (82.39% combined) suggests that most generated analogies are either directly usable or correctable through targeted revision, reducing the burden on human reviewers.

Table 2. LLM-based review as a first stage quality control mechanism

Question category	Number of question	Percentage (%)
Pass	589	45.31
Borderline	482	37.08
Reject	229	17.62
Total	1,300	100

Table 3 illustrates the percentage for each question category across the diverse domains. The strong performance in domains such as biology, physics, mathematics, and economics can be attributed to well-established concepts and canonical explanations, which facilitate stable analogy construction. These domains often allow analogies to be grounded in everyday physical or numerical experiences, making them more accessible for children.

Table 3. The percentage for each question category across the diverse domains

Question domain	Pass (%)	Borderline (%)	Reject (%)
Chemistry	65	26	9
Biology	87	6	7
Physics	80	19	1
Mathematics	79	14	7
Economy	93	4	3
Sociology	40	47	13
Geography	67	32	1
Computer technology	16	66	18

Conversely, the lower pass rates in computer technology and sociology highlight important limitations. In computer-related questions, abstract notions (e.g., algorithms, data, networks) and adult-centric metaphors may increase the risk of unclear or developmentally inappropriate analogies. In sociology, the dependence on cultural context and social norms increases sensitivity and ambiguity, leading to more borderline classifications rather than outright rejection. Notably, the proportion of rejected items remains relatively low across most domains, indicating that the LLM rarely produces fundamentally unsafe or misleading analogies. Instead, most quality issues fall into the borderline category, validating the design decision to route such cases to expert review rather than discard them. These findings justify the proposed hybrid evaluation pipeline, where LLM-based screening ensures feasibility at scale. At the same time, human and expert reviews provide pedagogical validation and contextual refinement for child-oriented applications.

This study present the evaluation results using descriptive statistics and boxplot visualizations. Table 4 summarizes the first two metrics: semantic quality (BERTScore F1) and efficiency (response time in seconds). For BERTScore and response time, this study report the mean, median, standard deviation (σ), minimum, maximum, and confidence interval (CI). The chatbot demonstrates consistently high semantic quality. It achieves a mean BERTScore F1 of 0.97, a median of 1.00, and a σ of 0.06. Most scores fall between 0.96 and 0.99, with a minimum of 0.76 and a maximum of 1.00.

Table 4. Summary statistics for BERTScore F1 across 1,300 CurioKids Bot-generated responses

Metric	Mean	Median	σ	Minimum	Maximum	CI
BERTScore (F1)	0.97	1.00	0.06	0.76	1.00	0.97
Response Time (s)	4.61	4.41	1.15	3.51	9.95	4.53

Figure 6 shows the boxplot distribution of BERTScore F1 and response time. Figure 6(a) shows the distribution of BERTScore F1, confirming a symmetric distribution with no significant outliers. These results indicate that the chatbot reliably generates semantically accurate responses, with the majority of predictions achieving high performance. The chatbot averages a response time of 4.61 seconds (median = 4.41, σ = 1.15, CI ≈ 4.53). Most responses fall within 4–5 seconds, although several outliers exceed 7 seconds, reaching a maximum of 10 seconds. Figure 6(b) highlights this distribution, showing generally stable performance with occasional delays. While the system is efficient for typical interactions, these outliers suggest the need to optimize real-time response handling. This study measure the third metric TSR, to evaluate task completion. The chatbot completes 73% of tasks successfully, indicating that nearly three out of four user tasks reach completion. The remaining 27% reflect instances in which the chatbot fails to guide users thoroughly. Table 5 presents the semantic alignment comparison between CurioKids and the Gemini chatbot using BERTScore-based metrics. CurioKids consistently outperforms Gemini across all evaluated dimensions. Specifically, CurioKids achieves a higher average BERTScore (F1 = 0.89) than Gemini (F1 = 0.71), indicating more substantial semantic alignment with the reference answers and representing an improvement of 0.17 points in the F1 score.

In terms of faithfulness-related metrics, CurioKids also demonstrates superior performance. Its average precision (0.88) exceeds Gemini's precision (0.69), suggesting that CurioKids produces responses with fewer semantically irrelevant or unsupported elements. Similarly, CurioKids achieves a higher average recall (0.89) than Gemini (0.73), indicating greater coverage of the reference content. The lower precision (–0.18) and recall (–0.16) observed for Gemini imply that its responses are comparatively less accurate and less comprehensive. This study evaluate the models using structured JSON input files containing question–prediction–ground truth triplets. For each prompt, we compute BERTScore, precision, and recall with the bert-base-multilingual-cased model from HuggingFace to ensure scalable, reproducible, and objective assessment across all test cases. Based on these results, the following discussion examines

how task-specific design choices and analogy-driven scaffolding account for the observed performance differences between CurioKids and a general-purpose chatbot.

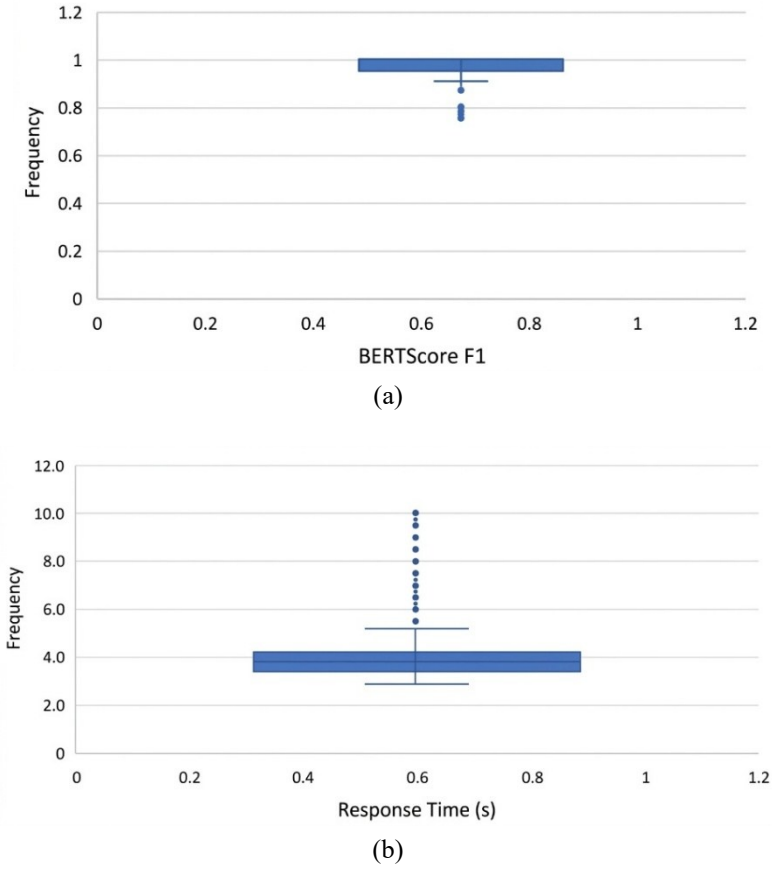

Figure 6. Boxplot distributions for the chatbot application: (a) BERTScore F1 and (b) response time

Table 5. Summary statistics for CurioKids chatbot compares to Gemini

Model	Avg. BERTScore (F1)	Avg. precision	Avg. recall
CurioKids	0.89	0.88	0.89
Gemini	0.71	0.69	0.73

3.2. Discussion

The evaluation results indicate that CurioKids achieves strong semantic alignment and linguistic quality, as reflected by consistently high BERTScore values. The analogy-based explanation strategy and domain-constrained prompting effectively support coherent, contextually relevant responses, which are essential for child-oriented educational chatbots. These findings align with prior studies showing that high semantic similarity contributes to perceived naturalness and user trust [26], [27]. Despite this strong semantic performance, the TSR of 73% reveals a gap between linguistic quality and functional effectiveness. A substantial portion of interactions fail to fully satisfy user intent, likely due to limitations in intent recognition, knowledge coverage, or dialogue management.

Similar mismatches between semantic quality and task completion have been reported in earlier chatbot studies [28], [29], underscoring the need for multi-metric evaluation. Response time analysis highlights further trade-offs. While the average response time of approximately 4.6 seconds remains acceptable, it exceeds the recommended threshold for optimal conversational interaction, and occasional delays may affect user experience [30]–[32]. This motivates future optimizations such as caching, asynchronous processing, and inference acceleration. Comparative results with Gemini further emphasize the benefits of task-specific design. CurioKids consistently outperforms the general-purpose chatbot in BERTScore, precision, and recall, indicating more faithful and complete alignment with child-oriented,

analogy-based references. These findings demonstrate that pedagogically informed constraints and analogy-driven scaffolding offer clear advantages for educational question–answer systems, supporting the effectiveness of the proposed hybrid evaluation pipeline. Overall, these findings show that CurioKids delivers strong semantic and pedagogical performance through its analogy-driven, task-specific design, while revealing remaining challenges in task completion and response efficiency. This aligns with previous observations in AI evaluation literature, which suggest that language generation quality alone does not ensure functional success or operational speed [33], [34]. This balance of strengths and limitations informs the concluding discussion on the system's contributions, constraints, and directions for future improvement.

4. CONCLUSION

This study demonstrates that CurioKids represents a viable step toward child-centered conversational AI by integrating analogy-driven explanations with task-specific design and structured quality control. The evaluation shows that CurioKids consistently delivers semantically aligned and pedagogically appropriate responses, achieving a BERTScore F1 score 0.18 higher than a general-purpose chatbot. These results underline the importance of domain adaptation and educational scaffolding in supporting children's exploratory question answering. Beyond linguistic quality, the findings also reveal key challenges that shape future development. A TSR of 73% and occasional response delays indicate that effective educational dialogue requires not only fluent language generation but also robust intent understanding and efficient system behavior. These limitations highlight the need to advance conversational AI for children beyond surface-level coherence toward deeper functional reliability. Looking forward, future work will focus on strengthening intent recognition and dialogue management, optimizing response efficiency, and maintaining high semantic and pedagogical standards. More broadly, this work contributes a scalable evaluation framework that combines LLM-based screening with human and expert review, offering a practical pathway for developing safe, faithful, and effective analogy-based educational chatbots. Together, these contributions position CurioKids as a foundation for next-generation child-friendly AI systems that support meaningful learning through natural and engaging interaction.

ACKNOWLEDGMENTS

The authors would like to express gratitude to Institut Teknologi Del for the support provided during the conduct of this research.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Eka Stephani Sinambela	✓	✓	✓	✓	✓	✓	✓	✓	✓			✓	✓	✓
Tanisha Natalia Aritonang	✓	✓	✓	✓	✓	✓		✓	✓		✓	✓	✓	
Samuel Arta Raju Manik	✓	✓	✓	✓	✓	✓		✓	✓		✓			
Grace Angel Nainggolan	✓	✓	✓	✓	✓	✓		✓	✓		✓			
Marojahan Mula Timbul Sigiro	✓			✓	✓	✓				✓				
Frengki Simatupang	✓			✓	✓	✓				✓				
Istas Pratomo Manalu				✓		✓				✓				

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are stored locally and are available from the corresponding author, [ESS], upon reasonable request.

REFERENCES

[1] J. Litman, T. Hutchins, and R. Russon, "Epistemic curiosity, feeling-of-knowing, and exploratory behaviour," *Cognition and Emotion*, vol. 19, no. 4, pp. 559–582, Jun. 2005, doi: 10.1080/02699930441000427.

[2] H. Hunaepi, I. K. Suma, and I. W. Subagia, "Curiosity in science learning: a systematic literature review," *International Journal of Essential Competencies in Education*, vol. 3, no. 1, pp. 77–105, Jun. 2024, doi: 10.36312/ijece.v3i1.1918.

[3] J. Jirout and D. Klahr, "Children's scientific curiosity: In search of an operational definition of an elusive concept," *Developmental Review*, vol. 32, no. 2, pp. 125–160, Jun. 2012, doi: 10.1016/j.dr.2012.04.002.

[4] V. Ajuwon, T. Monteiro, A. K. Schnell, and N. S. Clayton, "To know or not to know? Curiosity and the value of prospective information in animals," *Learning and Behavior*, vol. 53, no. 1, pp. 114–127, Mar. 2025, doi: 10.3758/s13420-024-00647-y.

[5] J. J. Jirout, N. S. Evans, and L. K. Son, "Curiosity in children across ages and contexts," *Nature Reviews Psychology*, vol. 3, no. 9, pp. 622–635, Aug. 2024, doi: 10.1038/s44159-024-00346-5.

[6] A. Modirshanechi, K. Kondrakiewicz, W. Gerstner, and S. Haesler, "Curiosity-driven exploration: foundations in neuroscience and computational modeling," *Trends in Neurosciences*, vol. 46, no. 12, pp. 1054–1066, Dec. 2023, doi: 10.1016/j.tins.2023.10.002.

[7] P. Bitzenbauer, "ChatGPT in physics education: A pilot study on easy-to-implement activities," *Contemporary Educational Technology*, vol. 15, no. 3, Jul. 2023, doi: 10.30935/cedtech/13176.

[8] S. Mukaromah and H. J. Yulianto, "Classroom interaction in the teaching of English at a private high school in Indonesia," *English Education Journal*, vol. 13, no. 1, pp. 105–117, Mar. 2023, doi: 10.15294/eej.v13i1.71936.

[9] M. Alaimi, E. Law, K. D. Pantasdo, P.-Y. Oudeyer, and H. Sauzeon, "Pedagogical agents for fostering question-asking skills in children," in *2020 CHI Conference on Human Factors in Computing Systems*, New York, United States: ACM, Apr. 2020, pp. 1–13, doi: 10.1145/3313831.3376776.

[10] M. Cerovac and T. Keane, "Early insights into Piaget's cognitive development model through the lens of the Technologies curriculum," *International Journal of Technology and Design Education*, vol. 35, no. 1, pp. 61–81, Mar. 2025, doi: 10.1007/s10798-024-09906-5.

[11] X. Hu, S. Storks, R. Lewis, and J. Chai, "In-context analogical reasoning with pre-trained language models," in *61st Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, United States: Association for Computational Linguistics, 2023, pp. 1953–1969, doi: 10.18653/v1/2023.acl-long.109.

[12] K. D. Forbus, K. Chen, W. Xu, and M. Usher, "Hybrid primal sketch: Combining analogy, qualitative representations, and computer vision for scene understanding," *Cognitive Systems Research*, vol. 93, Oct. 2025, doi: 10.1016/j.cogsys.2025.101390.

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Stroudsburg, United States: Association for Computational Linguistics, 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423.

[14] S. M. S. Mohammadabadi, B. C. Kara, C. Eyupoglu, C. Uzay, M. S. Tosun, and O. Karakuş, "A survey of large language models: evolution, architectures, adaptation, benchmarking, applications, challenges, and societal implications," *Electronics*, vol. 14, no. 18, Sep. 2025, doi: 10.3390/electronics14183580.

[15] O. Shliazhko, A. Fenogenova, M. Tikhonova, A. Kozlova, V. Mikhailov, and T. Shavrina, "mGPT: few-shot learners go multilingual," *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 58–79, Jan. 2024, doi: 10.1162/tacl_a_00633.

[16] J. Gallifant *et al.*, "Peer review of GPT-4 technical report and systems card," *PLOS Digital Health*, vol. 3, no. 1, Jan. 2024, doi: 10.1371/journal.pdig.0000417.

[17] Z. Yao, B. Schloss, and S. Selvaraj, "Improving summarization with human edits," in *2023 Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, United States: Association for Computational Linguistics, 2023, pp. 2604–2620, doi: 10.18653/v1/2023.emnlp-main.158.

[18] D.-L. Chen, K. Aaltonen, H. Lampela, and J. Kujala, "The design and implementation of an educational Chatbot with personalized adaptive learning features for project management training," *Technology, Knowledge and Learning*, vol. 30, no. 2, pp. 1047–1072, Jun. 2025, doi: 10.1007/s10758-024-09807-5.

[19] H.-T. Ho, D.-T. Ly, and L. V. Nguyen, "Mitigating hallucinations in large language models for educational application," in *2024 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia)*, IEEE, Nov. 2024, pp. 1–4, doi: 10.1109/ICCE-Asia63397.2024.10773965.

[20] Y. Yu, T. Sharma, M. Hu, J. Wang, and Y. Wang, "Exploring parent-child perceptions on safety in generative AI: concerns, mitigation strategies, and design implications," in *2025 IEEE Symposium on Security and Privacy (SP)*, IEEE, May 2025, pp. 2735–2752, doi: 10.1109/SP61157.2025.00090.

[21] L. M. -Hoz, S. Y. -Luis, J. H. Cejudo, D. G. -Reina, and E. F. Álvarez, "A comparative study of BERT-based models for teacher classification in physical education," *Electronics*, vol. 14, no. 19, Sep. 2025, doi: 10.3390/electronics14193849.

[22] R. M. Carvalho, A. B. Marques, R. Andrade, A. Sousa, R. Dias, and G. Campagnoli, "An experience of using the GQM approach in a remote environment to define requirements metrics," in *XXI Brazilian Symposium on Software Quality*, New York, United States: ACM, Nov. 2022, pp. 1–10, doi: 10.1145/3571473.3571491.

[23] W. Fan *et al.*, "A survey on RAG meeting LLMs: towards retrieval-augmented large language models," in *30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, New York, United States: ACM, Aug. 2024, pp. 6491–6501, doi: 10.1145/3637528.3671470.

[24] Z. Li, Z. Wang, W. Wang, K. Hung, H. Xie, and F. L. Wang, "Retrieval-augmented generation for educational application: a systematic survey," *Computers and Education: Artificial Intelligence*, vol. 8, Jun. 2025, doi: 10.1016/j.caeai.2025.100417.
[25] M. Murtiyoso, I. Tahyudin, and B. Berlilana, "A systematic review of retrieval-augmented generation for enhancing domain-specific knowledge in large language models," *Sinkron*, vol. 9, no. 2, pp. 969–977, Jun. 2025, doi: 10.33395/sinkron.v9i2.14824.
[26] Z. Yi *et al.*, "A survey on recent advances in LLM-based multi-turn dialogue systems," *ACM Computing Surveys*, vol. 58, no. 6, pp. 1–38, Apr. 2026, doi: 10.1145/3771090.
[27] F. Sufi, "Just-in-time news: an AI Chatbot for the modern information age," *AI*, vol. 6, no. 2, Jan. 2025, doi: 10.3390/ai6020022.
[28] A. Testoni and R. Bernardi, "The interplay of task success and dialogue quality: an in-depth evaluation in task-oriented visual dialogues," in *16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, Stroudsburg, United States: Association for Computational Linguistics, Apr. 2021, pp. 2071–2082, doi: 10.18653/v1/2021.eacl-main.178.
[29] R. Fellows, H. Ihshaish, S. Battle, C. Haines, P. Mayhew, and J. I. Deza, "Task-oriented dialogue systems: performance vs quality-optima, a review," in *Computer Science and Information Technology*, Academy and Industry Research Collaboration Center (AIRCC), Jul. 2022, pp. 69–87, doi: 10.5121/csit.2022.121306.
[30] M. Maslych *et al.*, "Mitigating response delays in free-form conversations with LLM-powered intelligent virtual agents," in *7th ACM Conference on Conversational User Interfaces*, New York, United States: ACM, Jul. 2025, pp. 1–15, doi: 10.1145/3719160.3736636.
[31] Y. Shi and B. Deng, "Finding the sweet spot: exploring the optimal communication delay for AI feedback tools," *Information Processing & Management*, vol. 61, no. 2, Mar. 2024, doi: 10.1016/j.ipm.2023.103572.
[32] R. Lumbantoruan, A. M. Sinaga, M. P. Hutagalung, P. C. N. Parapat, and M. T. Simanjuntak, "Evaluating LLM-based institutional information chatbot responses using a preliminary human-scored analytic rubric and automatic metrics," *Elinvo (Electronics, Informatics, and Vocational Education)*, vol. 11, no. 1, pp. 85–99, May 2026, doi: 10.21831/elinvo.v11i1.91571.
[33] A. Braggaar, C. Liebrecht, E. V. Miltenburg, and E. Krahmer, "Evaluating task-oriented dialogue systems: a systematic review of measures, constructs and their operationalisations," *Northern European Journal of Language Technology*, vol. 12, no. 1, Mar. 2026, doi: 10.3384/nejlt.2000-1533.2026.5940.
[34] C.-W. Liu, R. Lowe, I. Serban, M. Noseworthy, L. Charlin, and J. Pineau, "How not to evaluate your dialogue system: an empirical study of unsupervised evaluation metrics for dialogue response generation," in *2016 Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, United States: Association for Computational Linguistics, 2016, pp. 2122–2132, doi: 10.18653/v1/D16-1230.

BIOGRAPHIES OF AUTHORS



Eka Stephani Sinambela received the degree in Computer and Network Engineering from Institut Teknologi Bandung, in 2013 and the M.Sc. degrees in Computer Science from VU University of Amsterdam, in 2017. Currently, she is lecturer and research member at the Department of Computer Technology at Institut Teknologi Del since 2018. The research interests include computer network, cyber security, and software defined network. She can be contacted at email: eka@del.ac.id.



Tanisha Natalia Aritonang is currently studying Computer Technology at Institut Teknologi Del, Indonesia. Her academic interests include artificial intelligence, computer networks, and cloud computing. She is actively involved in lab-based projects and coursework, aiming to contribute to the development of secure and efficient computing systems. She can be contacted at email: tanishaartng@gmail.com.



Samuel Arta Raju Manik is a Computer Technology student at the Institut Teknologi Del, Laguboti, North Sumatra, Indonesia. He is currently studying Computer Technology, which includes an in-depth understanding of hardware systems, software, computer networks, and the application of information technology to support the development of the digital world. He can be contacted at email: samuelarta7@gmail.com.



Grace Angel Nainggolan is a student of the Computer Technology program at Institut Teknologi Del, Indonesia. Her academic interests focus on software development with a specialization in front-end development, where she actively explores creating intuitive and responsive user interfaces. In addition, she has an interest in computer networks, particularly in the areas of design and network security. She can be contacted at email: grace.angell2805@gmail.com.



Marojahan Mula Timbul Sigiro received his bachelor's degree in Informatics Engineering from Institut Teknologi Bandung, Indonesia, and his master's degree in Computer Science from Delft University of Technology, the Netherlands. He is a lecturer at the diploma-3 program in Computer Technology, Institut Teknologi Del, Indonesia. His teaching and research interests include distributed computing, distributed applications, internet of things (IoT), and web applications. He is actively engaged in curriculum development and applied research in modern computing technologies. He can be contacted at email: marojahan@del.ac.id.



Frengki Simatupang received the degree in Informatics Engineering from Institut Teknologi Del, in 2017 and the M.T. degrees in electrical engineering from Institut Teknologi Bandung, in 2023. Currently, he is lecturer and research member at the Department of Computer Technology at Institut Teknologi Del since 2020. His research interests include computer network, cyber security, embedded system, internet of things, and digital forensics. He can be contacted at email: frengki.simatupang@del.ac.id.



Istas Pratomo Manalu since 2018, he have been a lecturer at the Institut Teknologi Del, specializing in the D3-Computer Technology Study Program. His primary research interests focus on the internet of things (IoT), computer networks, and LoRa technology. He have actively contributed to advancing these fields and have had the privilege of publishing my work in both national and international journals. He can be contacted at email: istas.manalu@del.ac.id.