Vol. 6 No. 2, 2025, pp. 131-145 **Jurnal Pepadun**

AI VTuber Development Base on GPT-4 for Casual Chat Interaction on YouTube Live Streaming: Case Study KAIRA Channel

[1]Muhammad Fadhilah Ramadhani, [*2]Akmal Junaidi, [3]Ossy Dwi Endah Wulansari, [4]Rico Andrian, and [5]Favorisen Rosyking Lumbanraja

[1,2,3,4,5]Department of Computer Science, University of Lampung,
Jl. Prof. Dr. Ir. Sumantri Brojonegoro No. 1, Bandar Lampung, Lampung Province, Indonesia, 35145
e-mail: [1]ramadhanif154@gmail.com, [*2]**akmal.junaidi@fmipa.unila.ac.id**, [3]ossy.dwiendah@fmipa.unila.ac.id, [4]rico.andrian@fmipa.unila.ac.id, [5]favorisen.lumbanraja@fmipa.unila.ac.id

Abstract *- This study presents the development and evaluation of KAIRA, an Indonesian AI-based VTuber designed to engage in casual conversations during live streaming on YouTube. KAIRA integrates GPT-4 for language generation, ElevenLabs for voice synthesis, and VTube Studio for real-time avatar animation. The system was evaluated through both controlled and public testing sessions, focusing on interaction quality, contextual relevance, and character liveliness. Semantic similarity was evaluated using IndoBERT and IndoRoBERTa to measure how closely the system's responses matched user expectations. Additionally, the system employed a topic-filtering approach using cosine similarity to maintain KAIRA's conversational focus—in this case, specifically on gaming topics. The classifier's performance, assessed using a manually labeled dataset, achieved an accuracy of 50.33%, indicating frequent misclassifications. Despite these errors, the flexible design of the response logic allowed many off-topic messages to receive contextually suitable replies. This evaluation underscores the complexity of maintaining both topical consistency and conversational coherence in real-time interactions. Furthermore, the insights gained contribute to ongoing research on virtual character development and conversational AI, particularly within Indonesian-language applications.*

Keywords**:** *AI VTuber; GPT-4; Text-to-Speech; YouTube Live Chat; Cosine Similarity.*

1. INTRODUCTION

Virtual YouTubers, or more commonly known as VTubers have become a global phenomenon today, as they can involve interaction in live video streams using animated characters. Since the emergence and popularity of Kizuna Ai in 2016, VTubers have significantly achieved popularity on streaming platforms, at that time it was YouTube. [1]. The problem is that conventional VTubers rely heavily on human talent, which leads to several challenges related to scheduling, operating complex equipment, as well as the physical and mental strain that affects performance [2]. Maintaining a consistent character will also add to the level of difficulty, especially if the character is very different from the original personality and must be consistently upheld. Physical discomfort in operating the necessary devices can also exacerbate the problem [2][3]. Therefore, there is a need for alternative solutions that are more comfortable and practical.

Artificial intelligence offers a new approach to tackling various challenges in maintaining VTuber performance. With the support of advanced language models like GPT-4, virtual characters can now generate conversations that sound natural and resemble human conversation [4]. When combined with high-quality text-to-speech (TTS) services like ElevenLabs, this system can simulate interactions in real-time in a convincing manner [5][6][7]. One of the main advantages of AI-based VTubers is their resilience against human limitations, such as fatigue or mental stress, allowing them to maintain stable performance and sustained engagement [8][9].

One of the most well-known examples of this technology is Neuro-sama, an autonomous AI VTuber capable of interacting in real-time through natural language processing and voice synthesis. Neuro-sama demonstrates the potential of AI in building consistent and engaging virtual characters, as well as maintaining long-term viewer engagement, even surpassing conventional VTuber approaches in terms of content frequency and

 | doi: 10.23960/pepadun.v6i3.285

volume [8][10]. Moreover, the presence of AI-based VTubers opens up new spaces for digital storytelling, cultural expression, and various forms of other creative explorations [11][12][13].

This research presents KAIRA (Knowledge Artificial Intelligence Revolutionary Assistant), an AI-based VTuber designed to casually interact with audiences during live streaming sessions on YouTube. All conversations are generated and articulated by the AI system, then delivered through an animated avatar representation. The primary focus of this study is to explore the extent to which AI can take on roles typically performed by human VTubers in the context of casual conversation. Specifically, this research evaluates the generative AI's ability to maintain an engaging dialogue flow, produce speech that sounds natural, and present expressive and lively character animations.

2. RESEARCH METHODOLOGY

This study adopts a Research and Development (R&D) approach as the primary method for designing and evaluating AI-based VTuber systems. R&D is a methodology that aims not only to generate knowledge but also to produce tangible products or systems through iterative stages of design, development, testing, and refinement [14]. This approach is particularly suitable for technology development where the goal is to create functional prototypes that can be assessed in real or simulated environments.

According to Judijanto et al. (2024) in their book, the R&D model consists of several important stages, starting with initial research and ending with implementation [14]. These stages ensure that the product is designed based on real needs, improved through feedback, and ready for public application. The workflow is visualized in Figure 1.

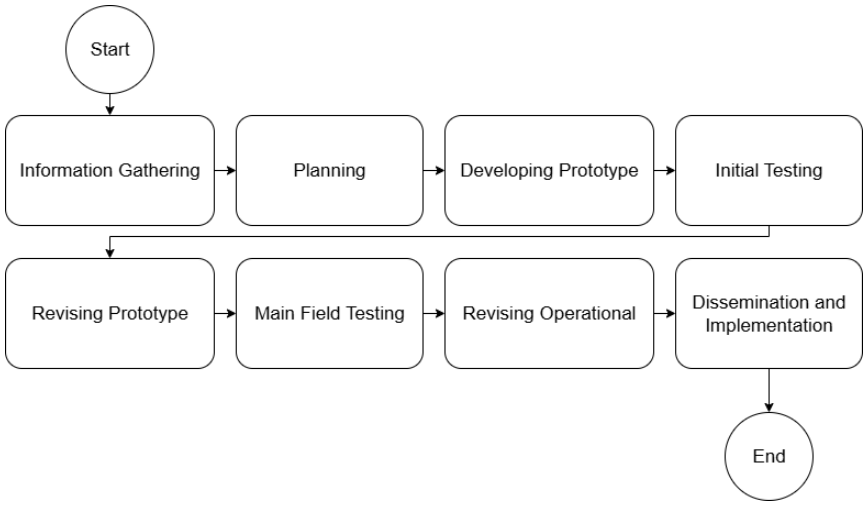

Figure 1. R&D method.

2.1. Information Gathering and Planning

The first stage of the research focused on identifying essential requirements for creating an AI-powered VTuber capable of mimicking human-like interactions during live streaming sessions. Due to limited academic literature specifically on AI-driven VTubers, information was gathered primarily by observing human-operated VTubers, particularly during informal "freetalk" segments on YouTube. These observations showed that spontaneous, relevant responses and consistent portrayal of character personality greatly strengthen emotional connections between VTubers and viewers, differentiating them from static digital media. Based on these insights, the minimum necessary system requirements for an effective AI VTuber were identified as follows:

a) The ability to capture live chat messages in real-time from a streaming platform.

 | doi: 10.23960/pepadun.v6i3.285

b) A mechanism to generate contextually relevant and human-like responses.
c) A text-to-speech engine to convert responses into expressive vocal output.
d) A visual representation in the form of a live-animated avatar that reacts synchronously to the generated speech.

Next, the system was developed and named KAIRA (Knowledge Artificial Intelligence Revolutionary Assistant) following a modular approach. This modular structure allows each component to function independently while seamlessly integrating into the overall system workflow. KAIRA's architecture comprises four main modules:

a) Language Model (GPT-4): Responsible for generating contextual responses based on user input from YouTube live chat. GPT-4 was chosen due to its advanced capabilities in producing coherent, contextually relevant, and human-like responses. Its architecture supports nuanced understanding of informal language and diverse topics, making it particularly suitable for interactive live environments [15][16].
b) Text-to-Speech (TTS) Engine: ElevenLabs was used to convert text responses into natural, expressive speech output.
c) Live Chat Handler: A custom module designed to retrieve incoming messages from the live stream chat in real-time.
d) Avatar Visual Output: The character is rendered and animated in real-time using VTube Studio to deliver visual feedback alongside the audio response.

During the development stage, an evaluation strategy was also devised to assess the extent of KAIRA's performance in providing appropriate responses. The evaluation was carried out through two approaches: first, subjective assessments from users using a Likert scale questionnaire; and second, objective semantic-based analysis using the cosine similarity method. In addition, real-time topic filtering systems were also tested by utilizing semantic similarity-based classification techniques, using an Indonesian language corpus taken from Reddit. Testing was conducted on a manually labeled dataset. The classification results were then analyzed using a confusion matrix and supported by examples of real conversations to see how topic classification affects KAIRA's responses in actual situations.

2.1.1. Use Case Diagram

Use case diagrams are one of the important elements in the Unified Modeling Language (UML) that are used to clearly depict the functional requirements of a system while also showing how the system interacts with external parties [17]. In this diagram, each main feature offered by the system is represented as a use case, while actors depict external entities—such as users or other systems—that are involved in the interaction. The notations used are standard, such as an oval symbol for use cases, stick figures for actors, and connecting lines to indicate the relationships between the two [18]. In this study, use case diagrams are used to illustrate the features of KAIRA and how KAIRA operates during a live stream session, as shown in Figure 2. This diagram highlights the interaction between KAIRA and the audience, detailing key processes such as reading live chat input, generating appropriate responses, converting responses into speech, and delivering audio-visual output synchronized with avatar animations.

2.1.2. Likert-scale

The Likert scale is a psychometric tool widely used in quantitative research to measure attitudes, preferences, or user perceptions. In the Likert scale, participants are asked to indicate their level of agreement with a statement using a numerical scale, typically ranging from 1 (Strongly Disagree) to 5 (Strongly Agree) [19]. This approach effectively transforms qualitative opinions into numerical data, facilitating statistical analysis.

In this study, the Likert scale was chosen for its clear and sensitive measurement capabilities. This scale evaluates several aspects of KAIRA, such as the naturalness of speech, relevance of responses, character engagement, and overall user experience. Due to its simplicity and reliability, this method is used to evaluate technology systems [20], in this case assessing user experience with KAIRA.

 | doi: 10.23960/pepadun.v6i3.285

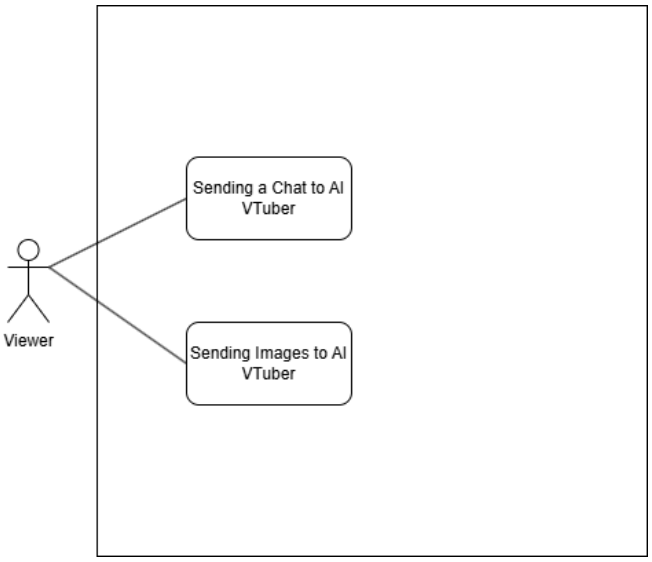

Figure 2. Use case diagram of KAIRA system.

2.1.3. Cosine Similarity

Cosine similarity is a method for measuring the extent to which two vectors are similar in vector space. Cosine similarity itself is widely used to measure semantic similarity between two texts, especially in Natural Language Processing (NLP). This method computes how similar two vectors are by measuring the angle between them, focusing on the direction of the vector rather than its magnitude [21]. Mathematically, cosine similarity is defined as:

$$sim(A, B) = \frac{A.B}{\|A\|\|B\|} \quad (1)$$

where A and B are sentence embeddings derived from textual input, and the result ranges from –1 to 1. A value closer to 1 indicates high semantic similarity.

In this study, cosine similarity is used to measure how close the meaning is between the responses given by the AI and the expected answers from participants during the trials. This evaluation will complement the Likert scale assessment by providing numerical insights into the coherence between input and output. Since cosine similarity works with numerical vectors, sentences must first be transformed into a vector representation called sentence embeddings. This embedding captures the semantic meaning of the entire sentence in a fixed-length numerical form, maintaining their relationships in high-dimensional vector space [21].

To generate accurate sentence embeddings suitable for Indonesian language interaction, this research uses two pretrained transformer models commonly used for Indonesian, namely IndoBERT [22] and IndoRoBERTa [23]. These two models can produce contextual embeddings that capture the linguistic nuances of the Indonesian language, making them highly suitable for semantic similarity tasks.

2.1.4. Language Models for Semantic Evaluation

Cosine similarity heavily relies on high-quality sentence embeddings—numeric vectors that accurately represent the meaning of the text. Therefore, for interactions in Indonesian, generating precise embeddings requires a language model specifically trained on local linguistic structures and vocabulary. This specialized training ensures that the embeddings reflect the subtle characteristics of the language, enhancing the reliability of semantic similarity analysis. To meet this need, this study utilizes two transformer-based models developed for the Indonesian language: IndoBERT and IndoRoBERTa. Both models have demonstrated their effectiveness in various NLP tasks, making them ideal for generating accurate sentence embeddings in semantic evaluation.

 | doi: 10.23960/pepadun.v6i3.285

IndoBERT is a monolingual language model specifically developed to address the limited availability of annotated data and linguistic resources for the Indonesian language based on BERT. This model is pretrained using Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) with a wide collection of Indonesian texts, including Wikipedia, news sources such as Kompas, Tempo, Liputan6, and the Indonesian Web Corpus (IdWaC) [22]. IndoBERT has two versions—base and large—which differ in the number of transformer layers and parameters. This flexibility allows it to be fine-tuned for various NLP tasks [24]. Training specifically targeted at the Indonesian language enables this model to represent local semantics more accurately than multilingual alternatives [25].

On the other hand, IndoRoBERTa is an adaptation of RoBERTa (Robustly Optimized BERT Pretraining Approach) specifically designed for the Indonesian language [26]. IndoRoBERTa was trained using OSCAR, a comprehensive open corpus derived from Indonesian web content [23]. Unlike IndoBERT, IndoRoBERTa eliminates Next Sentence Prediction and employs a dynamic masking approach during pretraining, which enhances its ability to capture context. With around 124 million parameters, IndoRoBERTa effectively models complex linguistic patterns, making it highly capable in various NLP tasks, including classification and semantic similarity.

By using both IndoBERT and IndoRoBERTa, this research ensured that cosine similarity calculations reflected a deeper understanding of Indonesian language structure and meaning. This dual-model approach was particularly beneficial for accurately evaluating the informal conversational language typical of live-stream chats. As a result, the integration of these two models significantly enhances the reliability of semantic similarity assessments in this study.

2.2. Developing Prototype

This stage focused on developing a functional prototype of KAIRA based on the established design plan. Each of the system's modules—including the language model, text-to-speech (TTS) engine, live chat handler, and visual avatar—was created and integrated into a cohesive real-time pipeline. This integration enabled continuous interactions between viewers and the AI VTuber without requiring human intervention.

2.2.1. System Architecture

The KAIRA system was implemented in Python using multiple asynchronous threads. These threads managed live chat input, generated responses via the OpenAI GPT-4 API, synthesized voices using the ElevenLabs TTS API, and synchronized avatar animations through VTube Studio. Message queues and timing control mechanisms were put in place to ensure natural turn-taking behavior during the delivery of chat responses. Each module operates concurrently, enabling the system to handle multiple chat requests simultaneously while maintaining a seamless real-time interaction experience. To illustrate the flow and structure of the system's operational pipeline, Figure 3 provides a schematic representation of KAIRA's overall architecture.

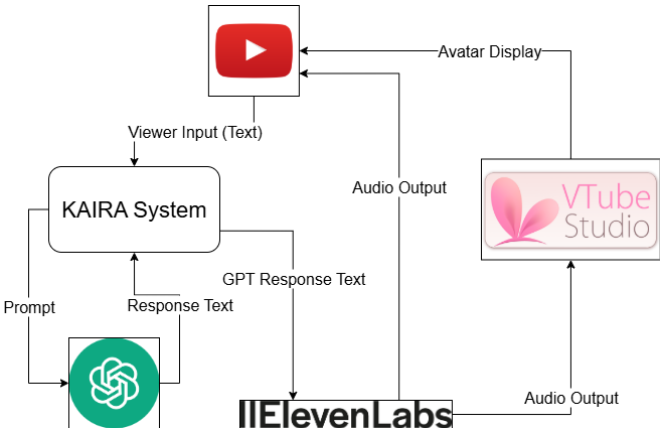

Figure 3. Overall architecture of the KAIRA system.

 | doi: 10.23960/pepadun.v6i3.285

2.2.2. Avatar Implementation

The virtual avatar used in this system originates from the VTube Studio community and is integrated via the VTube Studio software. To simulate a lifelike presence, a set of consistent idle animations, such as subtle head and body movements are predefined and run independently of the conversation. Meanwhile, only the mouth movement is dynamically triggered in real time based on synthesized speech output, serving as a visual cue that the character is speaking. Although the avatar does not track facial expressions or perform real-time motion capture, this combination of idle motion and reactive mouth animation helps maintain the illusion of a responsive and expressive character throughout the interaction. The simple flow of this visual response mechanism is illustrated in Figure 4.

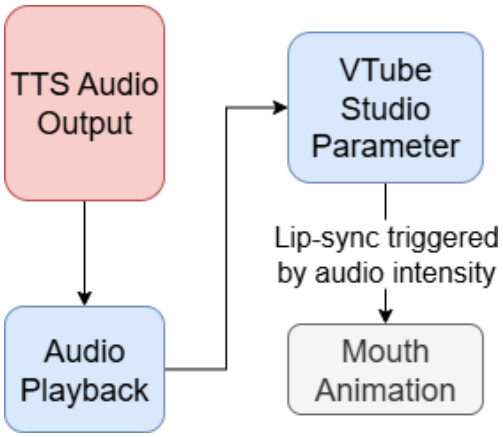

Figure 4. Avatar response flow based on audio triggers.

2.2.3. Text-to-Speech Pipeline

Text responses generated by GPT-4 are converted into speech using ElevenLabs, selected due to its expressive and natural-sounding voice synthesis [7][27], although it has not yet fully matched the flexibility of human voice actors [28]. The audio output is generated in real-time and synchronized with the avatar to produce cohesive audiovisual interactions. Additional audio post-processing using tools like FFmpeg is also employed to ensure consistent quality and clarity of speech outputs. The integration of these processes guarantees the avatar's speech feels natural and engaging for viewers. A simplified overview of the text-to-speech processing pipeline is presented in Figure 5.

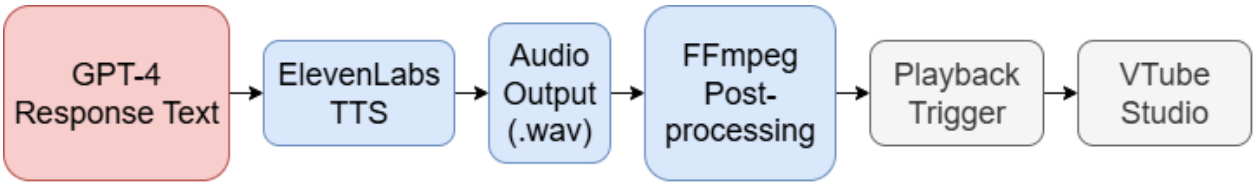

Figure 5. Text-to-speech processing pipeline in KAIRA.

2.2.4. Chat Interaction Flow

The chat processing workflow was set up to capture YouTube Live Chat messages in real-time using a Python library, eliminating dependence on YouTube's official API. Incoming messages were integrated directly into GPT-4 prompts alongside previous conversation history, preserving context and continuity. This enabled KAIRA to produce accurate and contextually suitable responses, making interactions more natural. Figure 6 visually details the user input management and processing flow during live streams.

 | doi: 10.23960/pepadun.v6i3.285

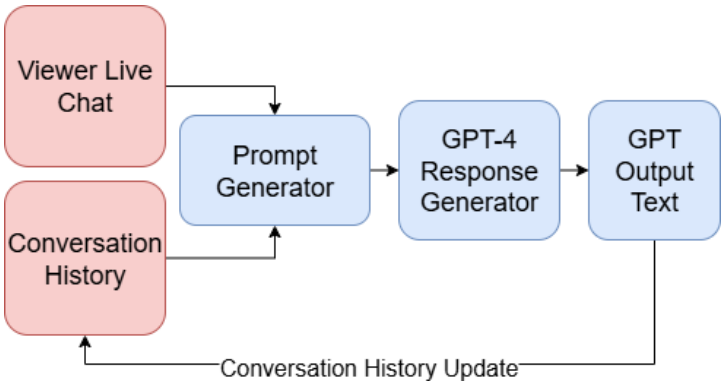

Figure 6. Live chat interaction flow from input to response.

2.3. Initial Testing

The initial testing phase aimed to verify the basic functionality, conversational quality, and consistency of KAIRA's character before moving to public deployment. This stage involved a controlled evaluation with a small participant group interacting with the AI VTuber in a semi-structured setting. Participants submitted various types of messages, such as greetings, personal inquiries, casual questions, trending topics, and hypothetical suggestions. These categories were designed to simulate typical freetalk scenarios observed in human VTuber streams. The goal was to evaluate how effectively the system could interpret user intent and respond in a manner that felt natural, relevant, and aligned with the character's intended personality. To evaluate the prototype, a Likert scale was used to assess participants' opinions about the KAIRA responses given the theme.

In parallel, semantic evaluation was conducted using cosine similarity analysis. For each participant input, an expected (ideal) response was defined, and the actual AI-generated reply was compared using sentence embeddings produced by IndoBERT and IndoRoBERTa models. This enabled quantitative assessment of semantic alignment. The themes, questions, and expectations are listed in Table 1.

Table 1. Sample themes, questions, and expectations.

Themes	**Questions**	**Expectations**
Greetings or Salutation	Hello KAIRA, I'm a new viewer *Halo kaira aku penonton baru*	Hello, thanks for stopping by to watch *Halo terimakasih udah mampir nonton*
Popular Trends	Do you know about the brainrot anomaly trend? *Kamu tau tentang tren anomali brainrot?*	A trend of creatures made by AI with strange names *Tren makhluk yang dibuat sama AI terus pake nama-nama yang aneh*
Activities During Holidays	Tomorrow's a holiday, where should I go? *Besok liburan nih, kemana enaknya?*	Somewhere that refreshes the body, mind, and eyes *Ke tempat-tempat yang bisa nyegerin badan, pikiran, sama mata*
Personal Questions	Where do you live? *Rumah kamu dimana?*	I live in an imaginary world *Rumah ku di dunia imajinasi*

 | doi: 10.23960/pepadun.v6i3.285

Themes	Questions	Expectations
Suggestions and Recommendations	Recommend me some songs that are relaxing! *Rekomendasiin lagu yang bikin adem dong!*	Try "Blue" by Yung Kai *Cobain Blue - Yung Kai*
Random Questions	Are you team stirred or unstirred porridge? *Kamu tim bubur diaduk atau atau ga diaduk?*	I prefer it stirred, it blends the flavors evenly *Aku tim bubur diaduk soalnya rasanya tercampur merata*
Daily Activities	What do you usually do after streaming? *Apa yang kamu biasa lakukan setelah melakukan streaming?*	I usually take a break to refresh my mind and body *Aku bakal istirahat dari semua kegiatan sambil refresh otak biar nggak cape*

2.4. Revising Prototype

This stage was carried out after the initial testing phase to refine the system based on collected feedback and observed interactions. The revision involved reviewing the conversational flow, evaluating response timing, and reassessing the alignment between the system's components. Adjustments were made iteratively throughout the refinement phase. These changes aimed to ensure the system aligned more closely with the intended design objectives and functional requirements.

By fixing any potential problems discovered during the preliminary testing, this step's primary goal was to prepare the system for a more thorough evaluation. To guarantee uniformity and enable integration, all modifications were made in exact alignment with the modular structure of the system. This method allowed improvements while preserving the general structure's integrity.

2.5. Main Field Testing

This phase served as a public beta test, deploying the AI VTuber KAIRA on a live YouTube broadcast to assess its performance under realistic conditions. Unlike the controlled environment of the initial testing, which included both subjective and objective evaluations, this phase relied exclusively on subjective participant feedback. After the livestream, voluntary participants completed a Likert-scale questionnaire via Google Forms, evaluating the system on several key indicators, such as:

a) The clarity of its answers to questions,
b) Its understanding of question intent,
c) The fluency and naturalness of its responses,
d) The appropriateness of the chat context,
e) The speed of its responses,
f) Comfort with its language style,
g) The impression of KAIRA as a lifelike character, and
h) The overall interaction experience.

Participants were also encouraged to provide open-ended feedback to identify areas for improvement from an audience perspective. These open-ended responses supplemented the numerical ratings, providing valuable qualitative insight into KAIRA's performance in an unsupervised public setting. The combination of qualitative and quantitative data helped to confirm the system's effectiveness in real-world scenarios, serving as a benchmark for user experience and engagement.

© 2025 The Authors | doi: 10.23960/pepadun.v6i3.285

2.6. Revising Operational

The next phase concentrated on refining the AI VTuber system utilizing the public feedback and insights gleaned from the primary field test. Though the primary aim was to assess real performance, the coordinated Likert-scale ratings and open-ended comments from volunteers pointed out areas for development. Rather than on large structural changes, those made during this phase focused on particular improvements to the quality of communication and general performance. Better personality alignment, better quick creation, and minor timing adjustments to encourage more natural interactions comprised these enhancements. Gradual changes were made to improve KAIRA's reactions with the intended personality traits by adjusting prompts and behavior settings.

These operational changes assured that the ultimate iteration of KAIRA not only passed technical dependability but also closely matched consumer expectations about character authenticity and participation. Following the modular design ideas laid out earlier, the development process let precise updates without changing the general structure of the system.

2.7. Dissemination and Implementation

The final stage focused on disseminating the completed AI VTuber system and documenting the research process and outcomes. Rather than deploying the system widely for continuous use, dissemination involved compiling the detailed development process, test results, and methodological insights into a formal research paper intended for academic publication.

As part of the dissemination, comprehensive documentation was created to clearly capture both the technical and evaluative aspects of KAIRA. This documentation serves as a resource for future research and promotes transparency, reproducibility, and a broader understanding of AI-driven virtual character interactions. By detailing the technical architecture, evaluation strategies, and system performance, the report provides a solid foundation for future adaptation and innovation within both academic and developer communities.

3. RESULTS AND DISCUSSION

This section presents the findings from each research phase. The results are analyzed in terms of KAIRA's ability to deliver timely, contextually suitable, and character-consistent interactions during livestreams. Both subjective Likert-scale feedback and objective semantic similarity assessments are examined to evaluate the system's effectiveness. Additionally, key insights, implications, and limitations observed throughout the testing are summarized and discussed.

3.1. Initial Testing Results

3.1.1. Likert Score

To assess KAIRA's conversational capability during initial testing, participants evaluated interactions using a 5-point Likert scale (1 = Strongly Disagree; 5 = Strongly Agree). Evaluations were divided into seven thematic categories: Greetings, Popular Trends, Holiday Activities, Personal Questions, Suggestions and Recommendations, Random Questions, and Daily Activities. Within each theme, participants rated how effectively KAIRA's responses matched their expectations. The summarized results for each theme, reflecting user perceptions of KAIRA's conversational quality, are presented in Table 2.

Table 2. Average likert scores by theme during initial testing.

Theme	Mean Score	Std. Deviation
Greetings or Salutation	4.17	±1.17
Popular Trends	3.67	±1.51
Holiday Activities	3.83	±0.98

 | doi: 10.23960/pepadun.v6i3.285

Theme	Mean Score	Std. Deviation
Personal Questions	3.67	±1.03
Suggestions and Recommendations	4.50	±0.55
Random Questions	4.00	±1.10
Daily Activities	4.17	±1.17

Overall, participants positively rated KAIRA's conversational interactions, notably praising its ability to provide relevant recommendations, shown by the highest average score in "Suggestions and Recommendations" (Mean = 4.50). Conversely, "Popular Trends" received the lowest average score (Mean = 3.67) and greater response variability (SD = ±1.51), indicating challenges in consistently addressing emerging or niche topics. These variations highlight areas for improvement in KAIRA's ability to interpret and respond to evolving conversational contexts.

3.1.2. Semantic Similarity

To complement subjective ratings, an objective semantic similarity analysis was conducted to measure how closely KAIRA's responses matched participant expectations. This analysis employed cosine similarity, a numerical metric quantifying the semantic closeness between two sentences. Sentence embeddings were generated using two pretrained Indonesian transformer models, IndoBERT and IndoRoBERTa, chosen specifically for their strength in capturing nuanced conversational meanings in Indonesian contexts.

Each AI-generated response was compared with the ideal participant-expected response, and cosine similarity scores were computed. Scores ranged from 0 to 1, with higher scores indicating stronger semantic alignment. A threshold of 0.80 was used to denote strong similarity, following recommendations by Zhou et al. (2022) [29]. The complete similarity results across themes and models are summarized in Table 3.

Table 3. Comparison of cosine similarity scores using IndoBERT and IndoRoBERTa across different question themes.

Theme	Model	Q1	Q2	Q3	Q4	Q5	Q6
Greetings or Salutation	IndoBERT	0.68	0.72	0.57	0.46	0.52	0.50
	IndoRoBERTa	0.87	0.66	0.54	0.79	0.67	0.44
Popular Trends	IndoBERT	0.62	0.67	0.67	0.54	0.53	0.46
	IndoRoBERTa	0.55	0.57	0.60	0.39	0.45	0.45
Holiday Activities	IndoBERT	0.71	0.62	0.55	0.65	0.61	0.47
	IndoRoBERTa	0.66	0.48	0.39	0.46	0.60	0.51
Personal Questions	IndoBERT	0.46	0.60	0.75	0.37	0.71	0.70
	IndoRoBERTa	0.36	0.44	0.62	0.36	0.64	0.53
Suggestions and Recommendations	IndoBERT	0.46	0.53	0.74	0.68	0.50	0.49
	IndoRoBERTa	0.39	0.56	0.57	0.52	0.39	0.44
Random Questions	IndoBERT	0.69	0.61	0.82	0.68	0.64	0.63
	IndoRoBERTa	0.42	0.50	0.52	0.40	0.39	0.51
Daily Activities	IndoBERT	0.62	0.63	0.71	0.50	0.62	0.60
	IndoRoBERTa	0.54	0.63	0.61	0.44	0.50	0.58

Note: Q1–Q6 refer to question numbers grouped under each theme in Table 1.

Across all thematic categories, IndoBERT consistently achieved higher and more stable similarity scores compared to IndoRoBERTa, indicating superior semantic alignment with participant expectations. These findings suggest IndoBERT is more effective at modeling conversational semantics in casual, Indonesian-language contexts, particularly in informal interactions typical of livestream chats. However, in several themes, IndoRoBERTa exhibited comparable results, highlighting that both models may offer complementary strengths depending on sentence structure and topic domain.

 | doi: 10.23960/pepadun.v6i3.285

In general, the majority of similarity scores fell within the moderate range of 0.50–0.70, indicating that while KAIRA's responses were reasonably aligned with expected meanings, perfect semantic matching was not consistently achieved. This reflects the inherent difficulty of replicating nuanced human intent in open-domain conversations, even with sophisticated transformer-based models. These results underscore the importance of continued refinement in model fine-tuning and prompt design to enhance response relevance in real-time interactions.

3.1.3. Real-Time Topic Filtering Performance in Live Chat

Following the offline evaluation, the semantic topic filtering system was integrated into KAIRA's real-time chat pipeline to constrain interactions to gaming-related topics. This mechanism employed cosine similarity to compare each incoming user message with an Indonesian-language corpus constructed from Reddit discussions. Messages exceeding a similarity threshold of 0.7 were classified as game-related and routed to GPT-4 for response generation, while others received a polite refusal. To assess the effectiveness of this filtering mechanism, a manually labeled evaluation dataset was tested, yielding a classification accuracy of 50.33%. A confusion matrix of the classification results is shown in Figure 7, illustrating the distribution of correct and incorrect predictions across the game and non-game categories.

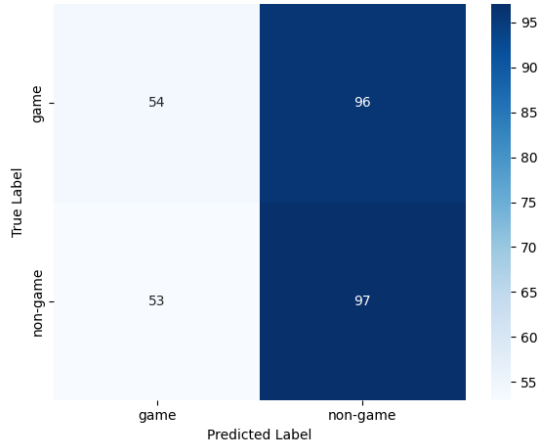

Figure 7. Confusion matrix of semantic topic classification.

While the confusion matrix provides an overview of classification accuracy, it does not reflect how those predictions influence the system's actual response behavior. In practice, KAIRA's routing logic does not enforce strict blocking based on classification results. Consequently, misclassified inputs are sometimes forwarded to GPT-4 and still receive responses regardless of their assigned labels. For instance, a game-related message mentioning a character or franchise may be misclassified as non-game, yet KAIRA proceeds to generate a relevant response. Conversely, some non-game inputs are mistakenly labeled as game-related and thus also receive full replies. These inconsistencies arise due to the limited scope of the Reddit-based corpus, which causes the system to misinterpret certain contexts beyond its training coverage. To further illustrate these behavioral anomalies, Table 4 presents selected samples from KAIRA's interaction logs. Each row shows how a particular message was labeled, how it was handled by the system, and whether the response aligned with the intended topic scope.

Table 4. Examples of classification and actual system response behavior.

User Input	True Topic	System Label	Response Behavior
Who do you think is the best hero in Mobile Legends? *Hero terbaik di Mobile Legends menurut kamu siapa?*	Game	Game	Answered (Game-related)

 | doi: 10.23960/pepadun.v6i3.285

User Input	True Topic	System Label	Response Behavior
Do you think ZZZ will be as popular as Honkai? *Menurut kamu, ZZZ itu bakal sepopuler Honkai gak?*	Game	Non-game	Answered (Game-related)
Do you think money can buy happiness? *Menurutmu, uang bisa beli kebahagiaan gak?*	Non-game	Non-game	Refused (Polite rejection)
I don't know why people like durian. *Aku ga ngerti kenapa orang suka durian.*	Non-game	Game	Answered (Non-game related)

The relatively low classification accuracy of 50.33% highlights the limitations of relying on a Reddit-derived corpus originally written in English and later translated into Indonesian. This translation process likely reduced the naturalness of linguistic expressions, making them less representative of authentic Indonesian conversations. As a result, many informal slang terms, code-switching patterns, and ambiguous contexts were not adequately captured, which weakened the model's ability to consistently distinguish gaming from non-gaming topics. To improve this performance, future work should focus on building more representative datasets directly in Indonesian, such as annotated chat logs from YouTube, Discord, or local gaming forums. In addition, hybrid approaches could be adopted by combining semantic similarity with keyword-based heuristics or by fine-tuning transformer models such as IndoBERT and IndoRoBERTa on manually labeled game-related data. Threshold tuning and the use of alternative evaluation metrics (e.g., precision, recall, and F1-score) may also help achieve a better balance between sensitivity and specificity. More advanced techniques, including few-shot or zero-shot classification with large language models, could further enhance robustness while maintaining the lightweight design required for real-time filtering.

3.2. Main Field-Testing Result

The main field testing was conducted via a public livestream session with 16 participants interacting freely with KAIRA on YouTube Live Chat. No predefined topics were set, allowing natural conversation within the scope of casual interaction to observe KAIRA's real-world performance. After the session, participants rated KAIRA using an 8-item Likert-scale questionnaire (1–5), summarized in Table 5.

Table 5. Mean and standard deviation of public testing evaluation.

Evaluation Indicator	Mean	Std. Deviation
Ability to answer questions clearly	4.56	±0.51
Understanding of question intent	4.44	±0.73
Fluency and naturalness of responses	4.12	±0.81
Contextual relevance of the answers	4.38	±0.62
Response speed in live interaction	3.44	±0.81
Conversational tone and linguistic comfort	4.31	±0.70
Impression of KAIRA as a lifelike character	3.44	±1.15
Overall interaction experience	4.19	±0.66

Overall results show positive user perception across most indicators. KAIRA scored highest in "Ability to answer questions clearly" (4.56) and "Understanding of question intent" (4.44), showing strong language comprehension and contextual relevance. Meanwhile, lower ratings were observed in "Response speed" and "Impression as a lifelike character" (both 3.44), with notably high standard deviation in the latter (±1.15), suggesting mixed impressions regarding avatar realism.

The perception of slow response may stem from two technical factors. First, KAIRA queues incoming messages during TTS playback to prevent audio overlap, which can delay response initiation. Second, YouTube's livestream latency (5–10 seconds) introduces a temporal gap between system response and user

 | doi: 10.23960/pepadun.v6i3.285

perception, making replies feel slower than they are. Despite this, system logs show KAIRA's actual response time ranges from 3 to 6 seconds, aligning with Nielsen's (2023) [30] category of "tolerable" responsiveness (1–10 seconds). This suggests user-reported delay is more likely caused by external factors like streaming lag and audio buffering than internal inefficiency. KAIRA's average delay remains acceptable for interactive use, though not perceived as instant.

Furthermore, Shi and Deng (2024) recommend a 1–3 second response time for fluid interaction [31], while Wang and Lo (2025) warn that delays above 5 seconds reduce satisfaction, especially among younger users. [32]. YouTube's own recommendation for "ultra-low latency" streaming defines < 5 seconds as the benchmark for enabling real-time audience interaction. Therefore, KAIRA's performance, while not optimal, can still be considered within the upper bound of acceptable latency for AI-driven livestream systems.

To mitigate these delay issues, several optimization strategies can be implemented in future iterations. Streaming text-to-speech output in smaller chunks, rather than waiting for full-sentence synthesis, could reduce perceived lag. Shortening or restructuring prompts may also lower GPT-4's generation time without sacrificing coherence. In addition, employing asynchronous buffering—such as displaying a placeholder message or triggering a non-verbal avatar gesture while the system prepares its reply—can help maintain conversational flow. On the platform side, enabling YouTube's ultra-low latency mode or experimenting with alternative streaming solutions may further reduce end-to-end delay.

4. CONCLUSIONS

This study demonstrates the feasibility of developing an Indonesian-speaking AI VTuber capable of maintaining casual real-time interactions during live YouTube broadcasts. The integration of GPT-4, ElevenLabs, and VTube Studio enabled KAIRA to deliver coherent, relevant, and naturally voiced responses. Evaluation through Likert-scale questionnaires and semantic similarity analysis confirmed overall positive conversational quality. A topic filtering system based on cosine similarity was also introduced to keep discussions within the gaming domain. However, classification accuracy remained limited at 50.33%, with confusion matrix analysis revealing frequent false positives and false negatives. Despite this, KAIRA's routing logic allowed it to respond appropriately even when classification results were incorrect. These results suggest that while semantic filtering provides a lightweight topic control method, its effectiveness depends heavily on corpus quality and system design. Future work should explore richer corpora, threshold tuning, or hybrid filtering techniques to enhance topical precision without compromising natural interaction flow.

LITERATURE

[1] D. R. Puspitaningrum and A. Prasetio, "Fenomena 'Virtual Youtuber' Kizuna Ai di Kalangan Penggemar Budaya Populer Jepang di Indonesia," *Mediator: Jurnal Komunikasi*, vol. 12, no. 2, Dec. 2019, doi: 10.29313/mediator.v12i2.4758.

[2] D. Kim, S. Lee, Y. Jun, Y. Shin, and J. Lee, "VTuber's Atelier: The Design Space, Challenges, and Opportunities for VTubing," in *Conference on Human Factors in Computing Systems - Proceedings,* Association for Computing Machinery, Apr. 2025. doi: 10.1145/3706598.3714107.

[3] W. A. Hamilton, O. Garretson, and A. Kerne, "Streaming on twitch: Fostering participatory communities of play within live mixed media," in *Conference on Human Factors in Computing Systems - Proceedings*, Association for Computing Machinery, 2014, pp. 1315–1324. doi: 10.1145/2556288.2557048.

[4] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language Models are Unsupervised Multitask Learners," 2023. [Online]. Available: https://github.com/codelucas/newspaper [Accessed: Sep. 6, 2024]

 | doi: 10.23960/pepadun.v6i3.285

[5] S. Oyucu, "A Novel End-to-End Turkish Text-to-Speech (TTS) System via Deep Learning," *Electronics (Switzerland)*, vol. 12, no. 8, Apr. 2023, doi: 10.3390/electronics12081900.

[6] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language Models are Few-Shot Learners," *arXiv preprint*, arXiv:2005.14165, Jul. 2020. [Online]. Available: https://arxiv.org/abs/2005.14165 [Accessed: Sep. 6, 2024].

[7] H.-W. Chen, Endoscopic Endonasal Skull Base Surgery For Pituitary Lesions: An AI-Assisted Creative Workflow To Develop An Animated Educational Resource For Patients And Physicians, *Master's thesis*, University of Toronto, Toronto, Canada, 2023.

[8] N. Amato, B. De Carolis, F. De Gioia, M. N. Venezia, G. Palestra, and C. Loglisci, "Can an AI-driven VTuber engage People? The KawAIi Case Study," in *Joint Proceedings of the ACM IUI Workshops 2024*, Greenville, SC, USA, Mar. 18–21, 2024.

[9] M. Gerlich, "The Power of Virtual Influencers: Impact on Consumer Behaviour and Attitudes in the Age of AI," *Adm Sci*, vol. 13, no. 8, Aug. 2023, doi: 10.3390/admsci13080178.

[10] Ç. Ö. Güzel, "The Autthenticity of AI Influencers in Marketing," in *Understanding Generative AI in a Cultural Context*, 2024, pp. 247–274. doi: 10.4018/979-8-3693-7235-7.ch010.

[11] C.-M. Lee, "The Key Factors Affecting Audience's Support for VTubers," 2024. [Online]. Available: https://www.researchgate.net/publication/384159907 [Accessed: Sep. 30, 2024]

[12] H. Hermawan, P. Subarkah, A. T. Utomo, F. Ilham, and D. I. S. Saputra, "VTuber Personas In Digital Wayang: A Review Of Innovative Cultural Promotion For Indonesian Heritage," *Jurnal Pilar Nusa Mandiri*, vol. 20, no. 2, pp. 165–175, Sep. 2024, doi: 10.33480/pilar.v20i2.5921.

[13] M. T. Tang, V. L. Zhu, and V. Popescu, "Alterecho: Loose avatar-streamer coupling for expressive VTubing," in *Proceedings - 2021 IEEE International Symposium on Mixed and Augmented Reality, ISMAR 2021*, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 128–137. doi: 10.1109/ISMAR52148.2021.00027.

[14] L. Judijanto, A. I. Puspitasari, M. Rahmawati, M. S. Mahendra, and A. S. Nurhidayat, *Metodologi Research And Development (Teori dan Penerapan Metodologi RnD)*. Jakarta: PT. Sonpedia Publishing Indonesia, 2024. [Online]. Available: https://www.researchgate.net/publication/381290945 [Accessed: May. 2, 2025].

[15] M. Van Poucke, "ChatGPT, the perfect virtual teaching assistant? Ideological bias in learner-chatbot interactions," *Comput Compos*, vol. 73, Sep. 2024, doi: 10.1016/j.compcom.2024.102871.

[16] M. Aljanabi, M. Ghazi, A. H. Ali, S. A. Abed, and C. Gpt, "ChatGpt: Open Possibilities," 2023, *College of Education, Al-Iraqia University*. doi: 10.52866/20ijcsm.2023.01.01.0018.

[17] B. Alturas, "Connection between UML use case diagrams and UML class diagrams: a matrix proposal," *International Journal of Computer Applications in Technology*, vol. 72, no. 3, pp. 161–168, 2023, doi: 10.1504/IJCAT.2023.133294.

[18] E. Aquino, P. de Saqui-Sannes, and R. A. Vingerhoeds, "A Methodological Assistant for Use Case Diagrams," in *International Conference on Model-Driven Engineering and Software Development*, Science and Technology Publications, Lda, 2020, pp. 227–236. doi: 10.5220/0008938002270236.

[19] A. Joshi, S. Kale, S. Chandel, and D. Pal, "Likert Scale: Explored and Explained," *Br J Appl Sci Technol*, vol. 7, no. 4, pp. 396–403, Jan. 2015, doi: 10.9734/bjast/2015/14975.

[20] M. Koo and S.-W. Yang, "Likert-Type Scale," *Encyclopedia MDPI*, vol. 5, no. 18, Feb. 2025, doi: 10.3390/encyclopedia5010018.

[21] A. F. Hidayat, "Evaluasi Keandalan Cosine Similarity dalam Mendeteksi Plagiarisme Kode Program," unpublished student paper, IF2123 Aljabar Geometri, Institut Teknologi Bandung, 2024.

[22] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," *arXiv preprint*, arXiv:2011.00677, Nov. 2020. [Online]. Available: https://arxiv.org/abs/2011.00677 [Accessed: Jun. 15, 2025].

[23] M. R. Faisal, K. E. Fitriani, M. I. Mazdadi, F. Indriani, D. T. Nugrahadi, and S. E. Prastya, "Enhancing Natural Disaster Monitoring: A Deep Learning Approach to Social Media Analysis Using Indonesian BERT Variants," *Indonesian Journal of Electronics, Electromedical Engineering, and Medical Informatics* , vol. 7, no. 1, pp. 77–89, 2025, doi: 10.35882/ijeeemi.v7i1.38.

[24] L. Geni, E. Yulianti, and D. I. Sensuse, "Sentiment Analysis of Tweets Before the 2024 Elections in Indonesia Using Bert Language Models," *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika*, vol. 9, no. 3, pp. 746–757, Aug. 2023, doi: 10.26555/jiteki.v9i3.26490.

[25] E. Yulianti and N. K. Nissa, "ABSA of Indonesian customer reviews using IndoBERT: single-sentence and sentence-pair classification approaches," *Bulletin of Electrical Engineering and Informatics*, vol. 13, no. 5, pp. 3579–3589, Oct. 2024, doi: 10.11591/eei.v13i5.8032.

[26] Y. Sagama and A. Alamsyah, "Multi-Label Classification of Indonesian Online Toxicity using BERT and RoBERTa," in *Proceedings of the 2023 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology, IAICT 2023*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 143–149. doi: 10.1109/IAICT59002.2023.10205892.

[27] I. Ramli, N. Jamil, N. Seman, and N. Ardi, "An Improved Syllabification for a Better Malay Language Text-to-Speech Synthesis (TTS)," in *Procedia Computer Science*, Elsevier B.V., 2015, pp. 417–424. doi: 10.1016/j.procs.2015.12.280.

[28] R. A. F. Dewatri, A. Z. Al Aqthar, H. Pradana, B. Anugerah, and W. H. Nurcahyo, "Potential Tools to Support Learning: OpenAI and Elevenlabs Integration," *ODELIA: Southeast Asia Journal on Open Distance Learning*, vol. 01, no. 02, pp. 59–69, 2023.

[29] K. Zhou, K. Ethayarajh, D. Card, and D. Jurafsky, "Problems with Cosine as a Measure of Embedding Similarity for High Frequency Words," *arXiv preprint*, arXiv:2205.05092, May 2022. [Online]. Available: https://arxiv.org/abs/2205.05092 [Accessed: May. 25, 2025].

[30] J. Nielsen, "The Need for Speed in AI," *UX Tigers: Fearless Usability*, Aug. 2, 2023. [Online]. Available: https://www.uxtigers.com/post/ai-response-time [Accessed: Jun. 17, 2025].

[31] Y. Shi and B. Deng, "Finding the sweet spot: Exploring the optimal communication delay for AI feedback tools," *Inf Process Manag*, vol. 61, no. 2, Mar. 2024.

[32] Y. L. Wang and C. W. Lo, "The effects of response time on older and young adults' interaction experience with Chatbot," *BMC Psychol*, vol. 13, no. 1, Dec. 2025, doi: 10.1186/s40359-025-02459-9.