talenta PUBLISHER

Data Science : Journal Of Computing And Applied Informatics

Journal homepage: https://jocai.usu.ac.id

Large Language Model (LLM)-Based Conversational Survey Design and Comparison with Web-Based Survey

Falana Rofako Hakam[*1], Nori Wilantika[2]

[1,2]*Department of Statistical Computing Politeknik Statistika STIS, Jakarta, 13330, Indonesia*

[*]Corresponding Author: 222112038@stis.ac.id

ARTICLE INFO

Article history:
Received 1 September 2025
Revised 30 January 2026
Accepted 12 June 2026
Available online 12 June 2026

E-ISSN: 2580-829X
P-ISSN: 2580-6769

How to cite:
Priyanka M. Sahu and Prof. D. M. Sable, "A Survey Paper on Targeted Advertising using Location – Based Behavioural Data and Social Data," International Journal of Engineering Research and, vol. V5, no. 11, Nov. 2016, doi: 10.17577/ijertv5is110238.

http://doi.org/10.26594/register.v6i1.idarticle

ABSTRACT

This study addresses the low response quality often observed in conventional web-based surveys due to respondent *satisficing*. This research developed a prototype conversational survey powered by a Large Language Model (LLM) that applies prompt engineering and Retrieval-Augmented Generation (RAG) to enable more natural survey interactions. A comparative experiment was conducted between the LLM-based conversational survey and a conventional web survey with 36 respondents whose characteristics were matched. The evaluation focused on response quality and user perceptions. Statistical analyses show that the LLM-based conversational survey significantly reduces satisficing, evidenced by a lower rate of item nonresponse (p-value = 0.0036) and longer per-item response times (p-value = 0.0001), indicating greater respondent engagement. From a user-experience perspective, the LLM-based conversational survey was rated as significantly more enjoyable (p-value = 0.023) and cognitively less demanding (p-value = 0.0002). This research concludes that LLM-based conversational surveys can simultaneously improve response quality and user experience.

Keyword: conversational survey, large language model (LLM), retrieval-augmented generation (RAG), response quality, user perceptions.

ABSTRAK

Penelitian ini bertujuan mengatasi rendahnya kualitas respons pada kuesioner berbasis web konvensional akibat perilaku *satisficing* pada responden. Sebagai solusi, dikembangkan prototipe kuesioner percakapan berbasis *Large Language Model* (LLM) yang menggunakan *prompt engineering* dan *Retrieval-Augmented Generation* (RAG) untuk menciptakan interaksi survei yang lebih alami. Sebuah eksperimen perbandingan dilakukan antara kuesioner percakapan berbasis LLM dan kuesioner berbasis web konvensional dengan 36 responden yang karakteristiknya telah disetarakan dengan evaluasi berfokus pada kualitas respons dan persepsi pengguna. Hasil analisis statistik menunjukkan kuesioner percakapan berbasis LLM secara signifikan lebih unggul dalam menekan *satisficing* yang dibuktikan dengan tingkat jawaban kosong (*item nonresponse*) yang lebih rendah (*p-value*=0.0036) dan waktu respons per item yang lebih panjang (*p-value*=0.0001) yang mengindikasikan keterlibatan responden lebih tinggi. Dari sisi pengalaman pengguna, platform kuesioner percakapan berbasis LLM juga terbukti secara signifikan lebih menyenangkan (*p-value*=0.023) dan lebih ringan secara kognitif (*p-value*=0.0002). Penelitian ini menyimpulkan bahwa kuesioner percakapan berbasis LLM dapat meningkatkan kualitas respons dan pengalaman pengguna secara bersamaan.

Keyword: kuesioner percakapan, *large language model* (LLM), *retrieval-augmented generation* (RAG), kualitas respons, persepsi pengguna

1. Introduction

In line with the need for high-quality data, web-based surveys have now become a primary choice for data collection due to the various advantages they offer. This method has grown rapidly since the emergence of the first graphical browsers such as NCSA Mosaic in 1992, followed by Netscape Navigator and Internet Explorer, up to the first research paper on it being published in 1996 [1]. Its main advantages include the speed of the

data collection process, which far surpasses traditional methods such as paper questionnaires [1], [2], significant savings in management and data collection costs [3], and its ability to reduce social bias on sensitive topics [4]. Various optimization efforts have also continued, such as research on the impact of Human-Computer Interaction (HCI) aspects through cursor tracking to identify response bias and improve data validity [5].

Although web-based surveys offer convenience and various advantages, the quality of response characteristics they produce remains a serious concern. Studies show that two main causes of low data quality in online surveys are satisficing behavior and item nonresponse [6], [7]. Satisficing arises when respondents answer hastily or do not fully understand the questions, usually because they want to finish the survey quickly or due to lack of motivation [8]. Item nonresponse refers to the number of survey questions left unanswered by respondents [6]. A study by Vriesema and Gehlbach (2021) [9] showed that about 30.36% of student respondents engaged in one or more forms of satisficing, with nonresponse (not answering one or several questions) reaching 24.99%, straight-lining (choosing the same answer in sequence without considering the content of the questions) at 5.38%, and early termination (stopping the survey before completion) at around 3.73%. In Indonesia, this phenomenon can be seen from the results of the 2020 Online Population Census conducted by Badan Pusat Statistik (BPS). Of approximately 270 million residents, only 51.36 million individuals (19.05%) and 13.63 million families (16.87% of total families) completed the census online [10], even though in the same year internet access had reached about 73.7% of the population [11]. These data indicate that access to technology does not automatically guarantee good response rates and active engagement in online or web-based surveys.

One root of this problem is the absence of direct interaction with an interviewer in web-based surveys. In face-to-face surveys, interviewers play an important role in clarifying confusing questions and maintaining respondent engagement, especially in long or complex surveys [12], [13]. The lack of this role in online surveys makes respondents more likely to lose focus, be distracted by other activities such as checking email or social media, and ultimately lose interest in completing the survey [8]. In addition, ensuring that respondents can provide optimal and accurate answers is an important step to improve survey data quality. Artino et al. (2022) [14] state that high-quality answers tend to come from respondents who are motivated to optimize the survey process, whereas answers from respondents who do not do so are suboptimal and yield less trustworthy data. Therefore, challenges such as confusion, loss of concentration, and lack of motivation in online surveys need to be addressed with new approaches capable of reintroducing interactive elements.

Along with the development of digital technology, various innovations have begun to be applied, including the use of chatbots as interactive aids. As part of these innovations, chatbots have the potential to replace some of the roles of human interviewers. Chatbots function to simulate conversation with humans through interaction using natural language [15] and have been applied in various fields such as education, business, and health [16]. As virtual interviewers, chatbots can provide automated guidance and more personalized interaction to reduce respondent confusion and maintain their engagement. Chatbots also have anthropomorphic characteristics that make respondents feel as if they are talking to someone and more willing to disclose their personal insights [17]. In line with this, research by Celino and Re Calegari (2020) [18] states that conversational survey design is preferred by respondents compared to conventional web surveys (form-based) because it is considered more engaging, intuitive, and not boring.

Currently, Large Language Models (LLMs) are developing rapidly and open new opportunities to create sophisticated chatbots that can understand and respond to input in natural language. LLMs are a type of advanced neural network trained on very large text data so that they can process and generate human-like text by attending to grammatical structure and the meanings of words [19]. By using prompt engineering techniques and the Retrieval-Augmented Generation (RAG) approach, the chatbots built can provide responses that are relevant and feel more natural to users. Although chatbots have great potential to enhance interaction, the use of advanced LLM-based chatbots specifically designed to address response quality problems (satisficing and item nonresponse) in the context of online surveys in Indonesia remains very limited and has not been widely explored.

Therefore, this study fills that gap by building and testing a conversational survey prototype powered by an LLM. The main contributions of this study are threefold: (1) the development of a conversational survey prototype leveraging prompt engineering and Retrieval-Augmented Generation (RAG) to simulate natural interviewer interactions; (2) a comparative analysis of response patterns between the proposed conversational survey and a standard web survey using satisficing metrics; and (3) a comprehensive evaluation of user perceptions regarding ease of use, usefulness, enjoyment, security, and cognitive load.

The remainder of this paper is organized as follows: Section 2 describes the methodology used to develop the prototype and the experimental design. Section 3 presents the results of the system implementation,

performance evaluation, and the comparative analysis of user data. Finally, Section 4 concludes the study with a summary of findings and suggestions for future work.

2. Method

This study adopts a mixed-methods approach that integrates two main streams, namely the development of software to produce a technological artifact in the form of a survey application prototype, and an experimental design to conduct a comparative evaluation between the developed prototype and the conventional platform. The scope of the study encompasses a series of systematic stages, ranging from design, implementation, and technical evaluation to controlled experiments.

2.1. Prototype Development

The prototype application development process, encompassing both the LLM-based conversational survey and the conventional web-based survey, employs the Software Development Life Cycle (SDLC) methodology using a Modified Waterfall Model approach. This model was chosen because its structure is sequential and well-defined, yet it still provides flexibility to make revisions at each stage. This process is divided into four main phases as illustrated in Figure 1.

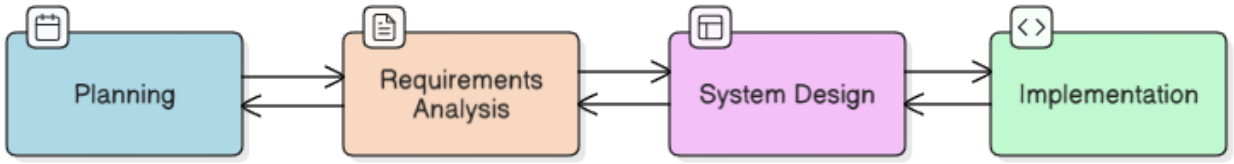

Figure 1. Prototype development process diagram.

2.1.1. Planning

This stage begins by clearly defining the scope and objectives of the project, as well as conducting an in-depth exploration of the technologies to be used. The main activities include selecting frameworks such as Next.js for the frontend and FastAPI for the backend, as well as determining MongoDB as the database.

2.1.2. Requirements Analysis

In this phase, an analysis of the conventional web-based survey system is conducted to identify its main weaknesses, particularly those related to the lack of interaction and the potential for satisficing. From this analysis, functional requirements are formulated such as intent classification and information extraction as well as non-functional requirements such as security and system performance.

2.1.3. System Design

This stage focuses on translating all the analyzed requirements into a comprehensive technical blueprint. The process includes designing a distributed system architecture, creating a flexible database schema, and developing the user interface (UI/UX) design through wireframes and mockups.

2.1.4. Implementation

This phase is the realization stage in which the entire system design is materialized into functional and testable code. The researchers carry out the coding process to build both application prototypes, namely the LLM-based conversational survey and the conventional web-based survey as a comparator. At the end of this stage, two software artifacts are produced that are ready for technical evaluation and comparative user testing.

2.2. Module Performance Evaluation

After both prototypes were completed, a technical evaluation was conducted on the functionality of the core modules of the LLM-based conversational questionnaire platform, as shown in Figure 2. This stage aims to validate the performance of each main application module before it is tested with end users.

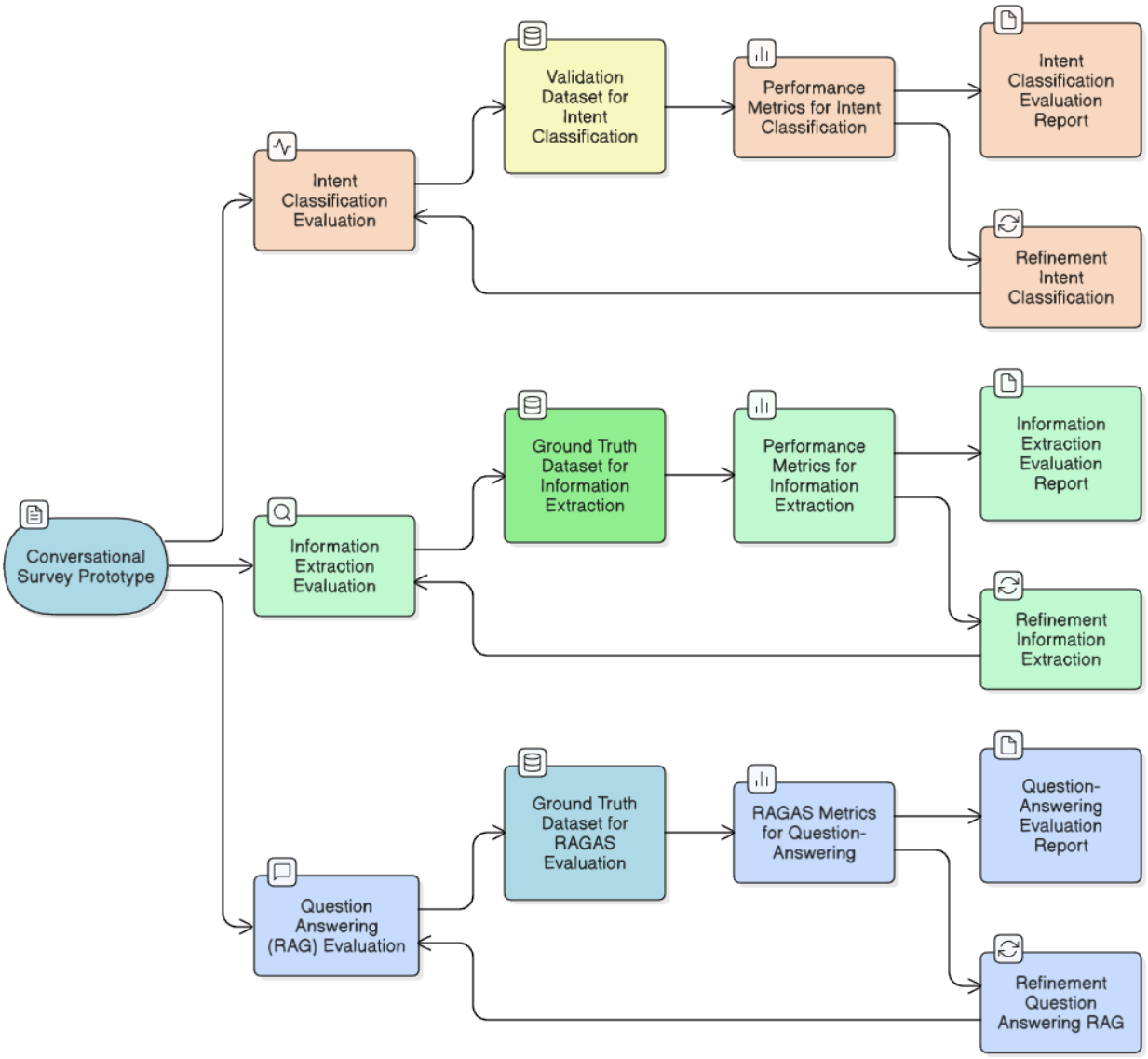

Figure 2. Module performance evaluation process diagram.

2.2.1. Evaluation of the Intent Classification Module

This module functions as the initial "brain" that interprets the intent of each user response and categorizes it as an answer, question, or other irrelevant input. This performance module is tested using standard classification metrics calculated based on the confusion matrix. The metrics used are:

- *Accuracy*
  This measure is used to measure the total proportion of correct predictions, calculated using Equation (1):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

- *Precision*
  This measure is used to measure the proportion of positive predictions that are truly positive, as shown in Equation (2):

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

- *Recall*
  This measure is used to measure the proportion of actual positive cases that were successfully identified, calculated using Equation (3):

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

- *F1-Score*
  This measure is the harmonic mean of Precision and Recall to provide a balancing score, as defined in Equation (4):

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

2.2.2. Evaluation of the Information Extraction Module

Once a response is identified as an answer, this module extracts specific information from the natural language text and transforms it into structured data. The module's accuracy is evaluated by comparing the extracted data from the user's response to the ground truth using an exact matching method. Accuracy is calculated using Equation (5):

$$Accuracy = \frac{Number\ of\ Matched\ Extractions}{Total\ Test\ Data} \quad (5)$$

2.2.3. Evaluation of the Question-Answering Module (RAG)

This module is activated when a user asks a question and acts as a virtual assistant providing clarification. It uses the Retrieval-Augmented Generation (RAG) framework, an approach that first retrieves relevant information from a knowledge base and then uses that information to generate factual answers. Answer quality is evaluated using the RAGAS framework with the following metrics:

- *Faithfulness*
  A metric used to assess the extent to which the generated answers are factually supported by the provided context, calculated as shown in Equation (6):

$$Faithfulness = \frac{Total\ claims\ in\ the\ answer}{Number\ of\ verified\ claims} \quad (6)$$

- *Answer Relevancy*
  A metric used to assess how relevant the generated answer is to the original question, defined in Equation (7):

$$Answer\ Relevancy = \frac{1}{k}\sum_{i=1}^{k} sim(q, q_i) \quad (7)$$

  where:
  - q is the original question.
  - q_i is the question generated from the answer.
  - k is the number of generated questions.
  - sim is the cosine similarity function.

- *Context Precision*
  This metric is used to measure the ratio of relevant information to irrelevant information (noise) in the retrieved context, calculated using Equation (8):

$$Context\ Precision = \frac{Number\ of\ relevant\ sentences\ in\ the\ context}{Total\ sentences\ in\ the\ context} \quad (8)$$

- *Context Recall*
  This metric is used to assess whether the retrieved context successfully covers all essential information required from the ground truth, as shown in Equation (9):

$$Context\ Recall = \frac{Number\ of\ ground\_truth\ sentences\ that\ can\ be\ attributed\ to\ the\ context}{Total\ sentences\ in\ the\ ground\_truth} \quad (9)$$

2.3. Comparative Testing and Data Analysis

This stage serves as the final phase, conducting a controlled experiment with a between-subjects design to compare the effectiveness of the two platforms, as described in Figure 3.

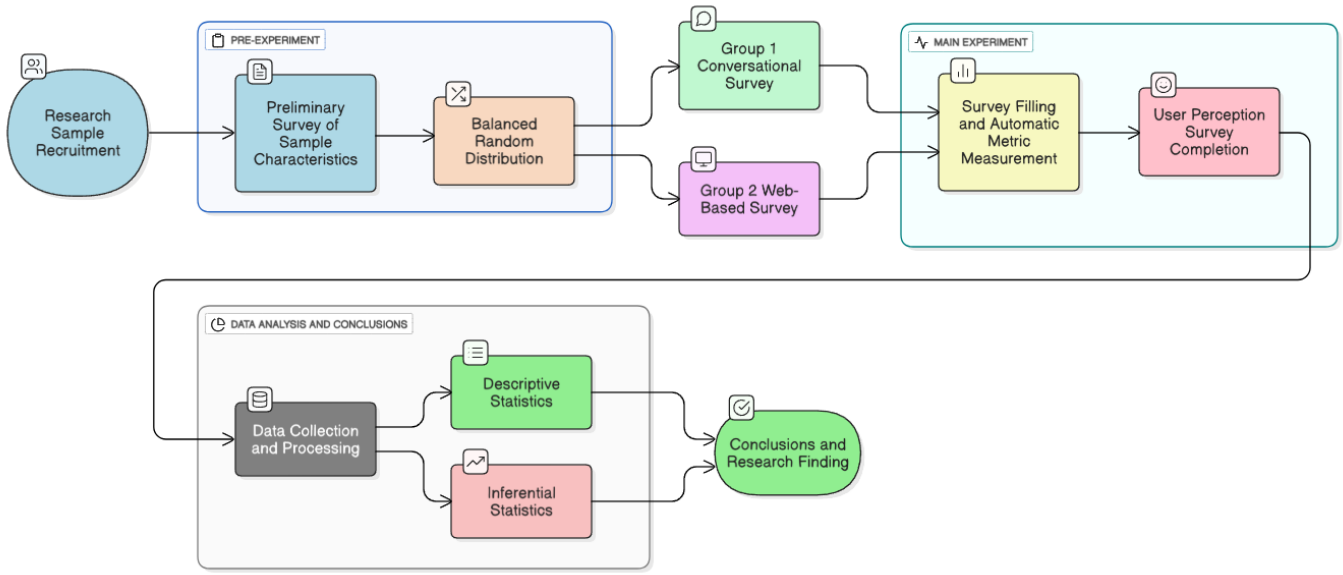

Figure 3. Comparative testing and data analysis process diagram.

2.3.1. Pre-Experiment

This phase constitutes the preparatory stage prior to the implementation of the main experiment. A total of 36 respondents were recruited using purposive sampling to collect information on individual characteristics. Based on these data, the respondents were randomly allocated into two groups (18 for the LLM-based conversational survey and 18 for the conventional web-based survey) with compositions balanced by level of education and digital skills to enhance the study's internal validity.

2.3.2. Main Experiment

In this phase, each respondent group was asked to complete the questionnaire through the assigned platform. During the completion process, the system automatically recorded response quality metrics—based on research by Zarouali et al. (2024) and Sánchez-Fernández et al. (2012)—including Breakoff Rate, Response Time, Item Nonresponse, and Don't Know Response. After completing the survey, respondents filled out a brief questionnaire to measure user perceptions, with an evaluation framework adapted from Zarouali et al. (2024) that covers Ease-of-Use, Usefulness, Enjoyment, Security, and Cognitive Load.

2.3.3. Data Analysis

The final phase is the processing and interpretation of the collected data. The data were analyzed descriptively and inferentially, wherein, given the relatively small sample size (N = 36), non-parametric statistical tests were the primary choice. Fisher's Exact Test was applied to analyze categorical data (breakoff rate), whereas the Mann–Whitney U test was used for numerical data (such as response time and user perception scores), with the significance level (α) set at 0.05.

3. Result and Discussion

This section presents the results from the study's three main stages, namely the system design and implementation, the technical performance evaluation of the modules, and the comparative evaluation of user behavior and experience.

1.1. System Design and Implementation

The analysis of the current system is focused on the workflow of the conventional web-based questionnaire used by Badan Pusat Statistik (BPS), as illustrated in Figure 4. This process is static and linear, in which respondents receive invitations via SMS, access the link, and fill out the questionnaire independently (self-administered). Its main weakness is the lack of interaction, which potentially causes confusion and triggers satisficing behavior.

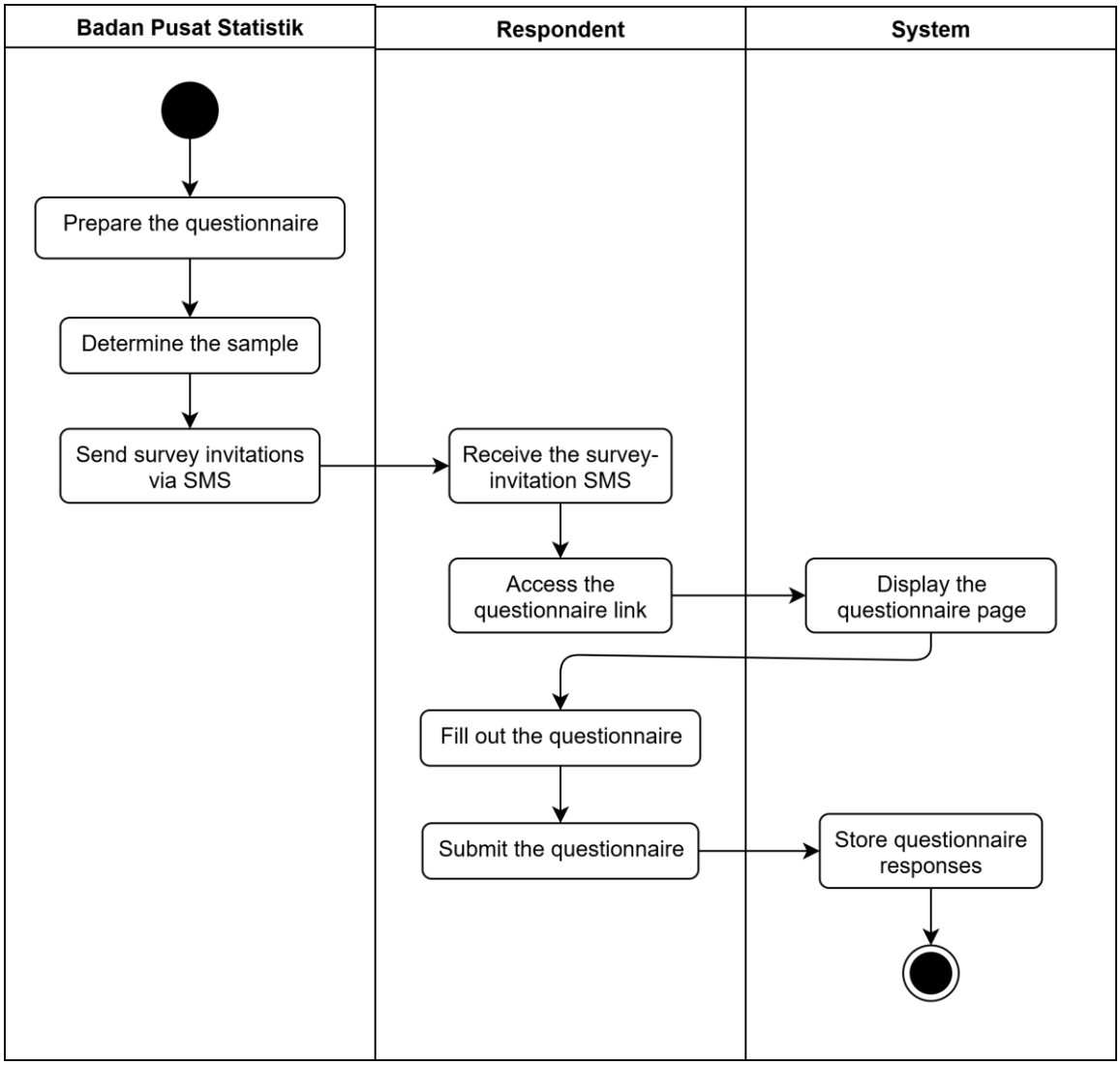

Figure 4. Current system business process.

To address these weaknesses, an LLM-based conversational survey system was designed to simulate the role of an interactive interviewer. The proposed system has two main modes:

- Survey Mode

The process of completing the questionnaire is transformed into a dynamic dialogue, as shown in Figure 5. Each user response has its intent classified. If it is an answer, the system extracts the information; if it is a question, the system activates the Question-Answering module to provide clarification or definitions.

- Question-and-Answer Mode

A simpler flow in which users can ask questions directly without being tied to a particular survey question, and the system provides answers based on the existing knowledge base using the RAG pipeline illustrated in Figure 6.

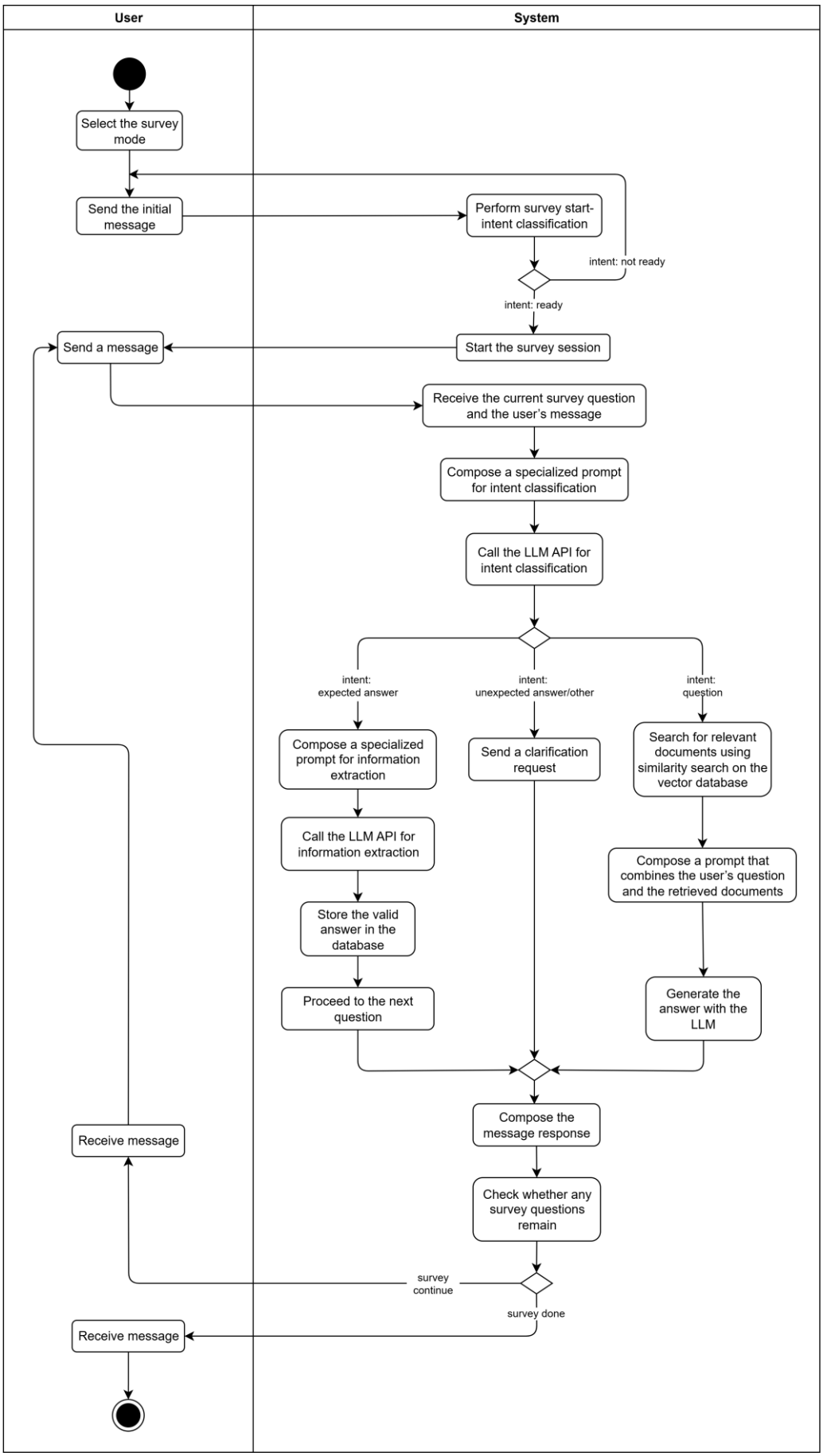

Figure 5. Proposed survey-mode system business process.

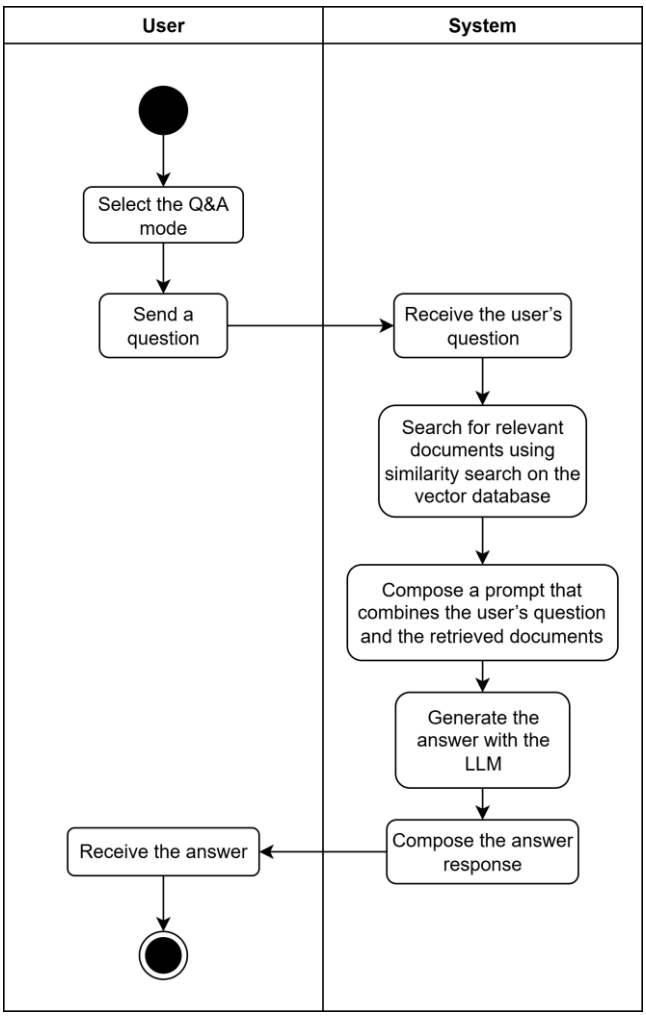

Figure 6. Proposed Q&A-mode system business process.

The user interface (UI) is designed to create an intuitive and engaging experience that differs from conventional form-based online surveys. The final implementation adopts a conversational format with chat bubbles, a modern color palette, and additional functionality such as viewing survey progress and editing answers, as shown in Figures 7, 8, and 9.

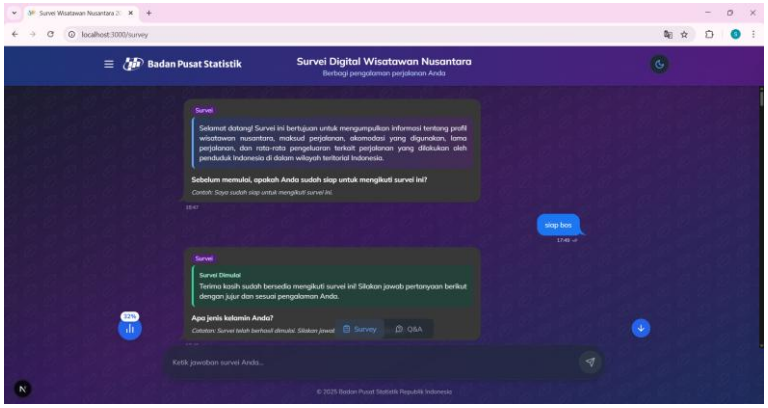

Figure 7. Main survey page.

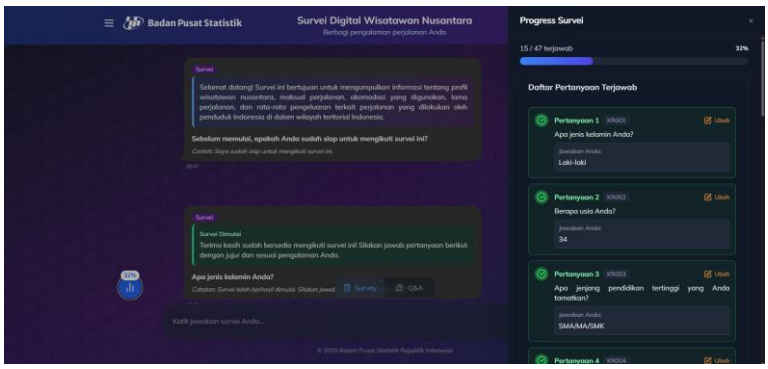

Figure 8. Detailed answered questions progress page.

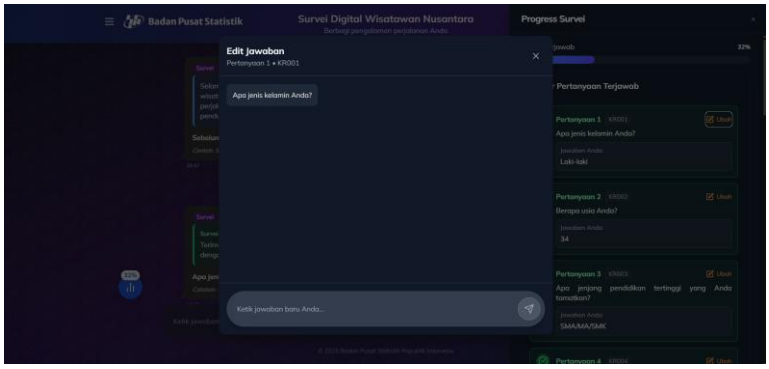

Figure 9. Answer edit page.

The Question-Answering module is implemented using an enhanced RAG workflow as shown in Figure 10. This process consists of two stages:

- *Indexing*

This stage is the preparation phase in which the system's knowledge base is constructed. The process begins by loading (Load) source documents, such as survey manuals, into the system. Next, these documents are split (Split) into smaller text chunks based on paragraphs to preserve contextual integrity. Each text chunk is then converted into numerical vectors (Embed) using an embedding model that represents its semantic meaning. Finally, all these vectors are stored (Store) in a vector database (MongoDB Atlas) so that they are ready for similarity-based retrieval.

- *Retrieval & Generation*

This process is executed when the user submits a question. This stage begins with rewriting the query (Rewrite Queries), in which the user's original question is transformed by the LLM into four variations to broaden the search scope. These four queries are then used to retrieve documents (Multi-Query Retrieve) in parallel from the vector database, and the results are merged. To ensure that the retrieved documents are not only relevant but also diverse, the Maximal Marginal Relevance (MMR) method is used. After that, a reranking process (Rerank) is carried out, in which the LLM re-evaluates all the retrieved documents and orders them based on the highest relevance to the original question. The most relevant documents are then combined with the user's question into a structured prompt, which is ultimately sent to the LLM to generate (Generate) a final answer that is factual and contextual.

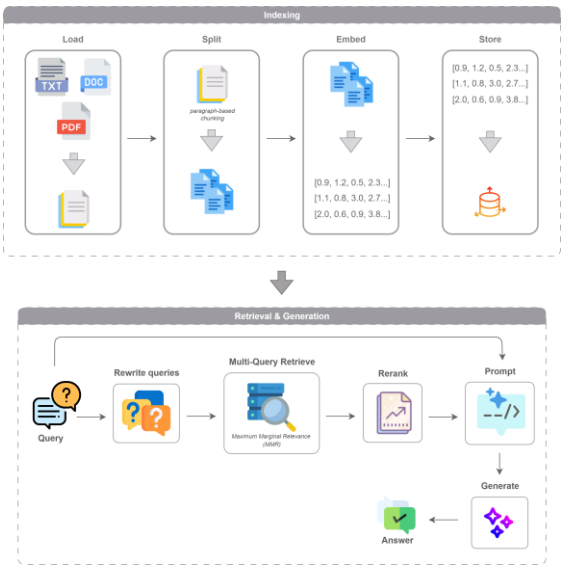

Figure 10. Workflow for enhancing the RAG pipeline.

1.2. System Performance Evaluation

1.2.1. Intent Classification Module

Evaluated using 120 test samples, this module achieved an accuracy of 91.67%. The confusion matrix in Figure 11 shows that the model is able to distinguish well between expected answers, questions, and other categories, with a low misclassification rate.

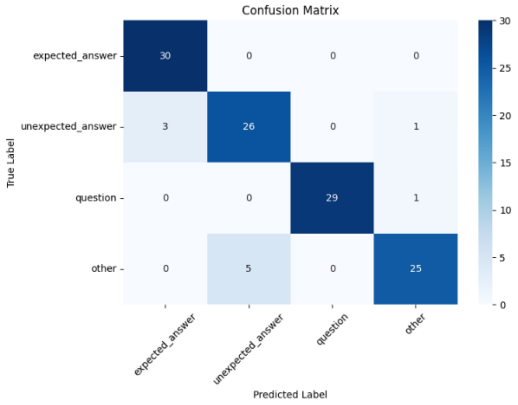

Figure 11. Confusion matrix of the intent classification module.

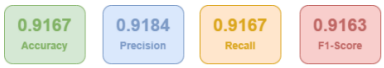

Figure 12. Evaluation results of the intent classification module.

1.2.2. Information Extraction Module

The Information Extraction module, which is responsible for transforming natural language responses into structured data, was tested on 30 data samples and demonstrated a perfect accuracy of 100%, as summarized in Figure 13. This proves its capability to extract and standardize various types of input without any errors.

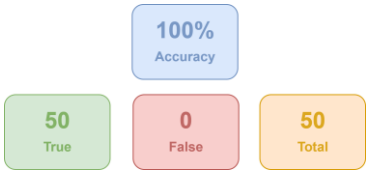

Figure 13. Evaluation results of the information extraction module.

1.2.3. Question-Answering Module (RAG)

For the Question-Answering module, the evaluation using the RAGAS framework as shown in Figure 14 indicates a very strong retrieval capability with a Context Precision of 93% and a Context Recall of 98%. This means the system is highly effective at finding relevant information. However, at the generation stage, the Faithfulness (84%) and Answer Relevancy (78%) scores indicate that there is still area for improvement in ensuring that the generated answers are fully factual and relevant.

Figure 14. Evaluation results of the question-answering module.

1.3. User Comparative Evaluation

A comparative analysis between the LLM-based conversational survey group (A) and the conventional web-based survey group (B) shows significant differences in several aspects, as summarized in Table 1.

Table 1. Summary of statistical tests.

Evaluation Domain	Variable Tested (unit)	*P-value*	Decision (α=0,05)	*Mean/Proportion** A	*Mean/Proportion** B	Direction*
Response Quality	*is_breakoff (%)*	0.4857	Fail to Reject H₀	11.00	0.00	A = B
	avg_response_time (ms)	0.0001	Reject H₀	43462.88	22392.62	A > B
	item_nonresponse (%)	0.0036	Reject H₀	1.00	3.00	A < B
	dont_know_response (%)	0.2879	Fail to Reject H₀	2.00	5.00	A = B
User Perception	*ease_of_use (unit)*	0.3074	Fail to Reject H₀	6.11	6.06	A = B
	usefulness (unit)	0.3814	Fail to Reject H₀	6.39	5.89	A = B
	enjoyment (unit)	0.0230	Reject H₀	6.61	5.72	A > B
	data_security (unit)	0.5247	Fail to Reject H₀	5.61	5.78	A = B
	privacy_safety (unit)	0.0784	Fail to Reject H₀	4.67	5.56	A = B
	cognitive_load (unit)	0.0002	Reject H₀	2.17	5.72	A < B

Description:

- *A = LLM-Based Conversational Survey*
- *B = Conventional Web-Based Survey*

Based on the results of the statistical tests, there are significant differences between the LLM-based conversational survey and the conventional web-based survey in several aspects tested. The LLM-based conversational survey produced a lower average rate of blank answers (item nonresponse), namely only 1% compared to 3% in the conventional web-based survey (p-value = 0.0131), and a longer average completion time per item (response time), namely 43.5 seconds compared to 22.4 seconds in the conventional web-based survey (p-value = 0.0001). This longer response duration may reflect deeper engagement and consideration of answers, although it does not rule out the possibility of an influence from system latency in the LLM-based conversational survey platform. In terms of user perception, the LLM-based conversational questionnaire is significantly superior in two aspects, namely a higher perception of enjoyment with an average score of 6.61 versus 5.72 (p-value = 0.0230) and a lower perception of cognitive load with an average score of 2.17, which is far below the conventional web-based questionnaire with an average score of 5.72 (p-value = 0.0002). Meanwhile, no statistically significant differences were found between the two platforms in terms of survey completion rate (breakoff rate), the rate of "don't know" answers (don't know response), and user perceptions regarding ease of use, usefulness, and security.

Although the LLM-based conversational survey excels in the aspects of enjoyment and cognitive load, the perception of security becomes the weakest point that is important to improve. Specifically, privacy safety in the LLM-based conversational survey records the lowest score with an average score of 4.67 among all positive perception variables, which is slightly lower than the conventional web-based survey that obtains an average score of 5.56 on the same aspect. This low score most likely reflects respondents' concerns that their conversation history is not fully secure and could potentially be accessed by other parties, thereby reducing the sense of privacy safety. This is reinforced by the data security score, which is also relatively low with an average score of 5.61 and not far from the conventional web-based survey that obtains an average score of 5.78, indicating that security issues are a common concern on both platforms. Therefore, improvements in privacy guarantees and data security are needed, especially on the LLM-based conversational survey platform.

Table 2. Analysis of feedback from respondents in the LLM-based conversational survey group.

Aspect	Feedback Summary
Speed / Performance	Although some users judged the questionnaire responses to be fairly fast, complaints arose about loading time between questions and the chatbot's

	response speed, which tended to slow down when users asked follow-up questions or required clarification.
Question Relevance	Feedback on the relevance of the questions was very positive because users appreciated word choices that were comfortable to read and questions that were detailed, thereby creating an engaging experience in completing an online survey. However, some feedback suggested adjustments for questions that were insufficiently specific or used unfamiliar terms so as to improve the clarity of the question's intent.
Questionnaire Flow	A great deal of very positive feedback emerged for the Questionnaire Flow aspect because users appreciated an interactive conversational survey model that resembles chatting, feels comfortable, and is easy to access, and is supported by good AI functions with natural word choices.
Visual Design	User feedback on the Appearance/Visual Design was largely positive, driven by a modern, attractive, and aesthetic chat interface, as well as appropriate and consistent color choices.
General	User feedback was very positive because they considered that the chatbot conversational format on this survey platform created a new, engaging, and easy survey-completion experience that felt personal, like sharing together.

Based on Table 2, the sentiments expressed are predominantly positive. Respondents appreciated the interactive questionnaire flow, the relevance and appropriate wording of the questions, as well as the modern and comfortable visual design. Together, these aspects create a survey completion experience that is considered engaging, easy to carry out, and personal. The only note for improvement that emerged relates to technical performance, particularly the chatbot's response speed and the loading time between questions.

1.4. Benefits and Limitations of the Proposed Methodology

The results of this study highlight several key benefits of the proposed LLM-based conversational survey. Primarily, the methodology successfully reduces satisficing behavior, evidenced by significantly lower item nonresponse rates. This suggests that the interactive, chat-like format encourages respondents to be more thorough. Furthermore, the significant reduction in cognitive load and increased enjoyment indicates that the conversational interface makes the survey process mentally easier and more engaging than filling out static forms. The RAG capability also provides a benefit that conventional surveys lack: the ability to provide instant, context-aware clarification for respondents who may be confused by specific questions.

However, there are limitations to this methodology. The most notable limitation is the increased response time. While this can indicate deeper thought, it may also be attributed to the latency inherent in LLM processing and text generation, which could potentially frustrate users in a time-sensitive context. Additionally, the study revealed a limitation regarding user trust; the conversational nature of the AI led to lower perceptions of privacy safety. Users may feel that "chatting" with an AI is less secure than a standard form, or they may fear their data is being processed by third-party models. Finally, the implementation of such a system is technically more complex and resource-intensive compared to standard web forms, requiring robust backend architecture and API management.

4. Conclusion

This study shows that the Large Language Model (LLM)-based conversational survey prototype was successfully developed with good technical performance in each core module; the application of prompt engineering techniques and Retrieval-Augmented Generation (RAG) effectively shaped an interactive survey flow, supported natural language interaction, and produced relevant responses. Empirically, this platform proved superior to conventional web-based survey in reducing satisficing behavior, as reflected in a lower rate of blank answers (item nonresponse) and a longer response time per item—an indication of higher respondent cognitive engagement. From the user experience perspective, the LLM-based conversational platform was rated as more enjoyable while also imposing a lower cognitive load, affirming its potential not only to improve survey data quality but also to transform the survey completion process into a more positive and less burdensome experience for respondents.

For future works, this study suggests several avenues for improvement. First, addressing the privacy concerns identified in the user perception evaluation is crucial; future iterations should explore privacy-preserving LLM techniques or on-premise model deployment to enhance user trust. Second, technical optimization is needed to reduce system latency, ensuring that the longer response times are solely due to cognitive engagement rather than technical delays. Finally, testing the platform on a larger, more diverse population would validate the generalizability of these findings beyond the current sample.

References

[1] M. P. Couper and P. V. Miller, "Web Survey Methods: Introduction," *Public Opin. Q.*, vol. 72, no. 5, pp. 831–835, Dec. 2008, doi: 10.1093/poq/nfn066.

[2] H. L. Ball, "Conducting Online Surveys," *J. Hum. Lact.*, vol. 35, no. 3, pp. 413–417, Aug. 2019, doi: 10.1177/0890334419848734.

[3] S. Lefever, M. Dal, and Á. Matthíasdóttir, "Online data collection in academic research: advantages and limitations," *Br. J. Educ. Technol.*, vol. 38, no. 4, pp. 574–582, Jul. 2007, doi: 10.1111/j.1467-8535.2006.00638.x.

[4] N. Berzelak and V. Vehovar, "Mode effects on socially desirable responding in web surveys compared to face-to-face and telephone surveys," *Adv. Methodol. Stat.*, vol. 15, no. 2, Jul. 2018, doi: 10.51936/lrkv4884.

[5] J. L. Jenkins, J. S. Valacich, and P. Williams, "Human-computer interaction movement indicators of response biases in online surveys," 2017.

[6] D. Heerwegh and G. Loosveldt, "Face-to-Face versus Web Surveying in a High-Internet-Coverage Population: Differences in Response Quality," *Public Opin. Q.*, vol. 72, no. 5, pp. 836–846, Dec. 2008, doi: 10.1093/poq/nfn045.

[7] J. A. Krosnick, "Response strategies for coping with the cognitive demands of attitude measures in surveys," *Appl. Cogn. Psychol.*, vol. 5, no. 3, pp. 213–236, May 1991, doi: 10.1002/acp.2350050305.

[8] M. Callegaro, K. L. Manfreda, and V. Vehovar, *Web Survey Methodology*. 1 Oliver's Yard, 55 City Road London EC1Y 1SP: SAGE Publications Ltd, 2015. doi: 10.4135/9781529799651.

[9] C. C. Vriesema and H. Gehlbach, "Assessing Survey Satisficing: The Impact of Unmotivated Questionnaire Responding on Data Quality," *Educ. Res.*, vol. 50, no. 9, pp. 618–627, Dec. 2021, doi: 10.3102/0013189X211040054.

[10] BPS, "Introducing Administrative Data to the Census in Indonesia," 2022.

[11] APJII, "Laporan Survei Penetrasi & Profil Perilaku Pengguna Internet Indonesia 2019–2020 (Q2)," Asosiasi Penyelenggara Jasa Internet Indonesia, Jakarta, 2020. [Online]. Available: https://survei.apjii.or.id/

[12] C. F. Cannell, L. Oksenberg, and J. M. Converse, "Striving for Response Accuracy: Experiments in New Interviewing Techniques," *J. Mark. Res.*, vol. 14, no. 3, p. 306, Aug. 1977, doi: 10.2307/3150768.

[13] Y. P. Ongena, "Interviewer and Respondent Interaction in Survey Interviews," Vrije Universiteit, Netherlands, 2005.

[14] A. R. Artino, Q. R. Youmans, and M. G. Tuck, "Getting the Most Out of Surveys: Optimizing Respondent Motivation," *J. Grad. Med. Educ.*, vol. 14, no. 6, pp. 629–633, Dec. 2022, doi: 10.4300/JGME-D-22-00722.1.

[15] B. Abu Shawar and E. Atwell, "Chatbots: Are they Really Useful?," *J. Lang. Technol. Comput. Linguist.*, vol. 22, no. 1, pp. 29–49, Jul. 2007, doi: 10.21248/jlcl.22.2007.88.

[16] M. Jovanovic, M. Baez, and F. Casati, "Chatbots as Conversational Healthcare Services," *IEEE Internet Comput.*, vol. 25, no. 3, pp. 44–51, May 2021, doi: 10.1109/MIC.2020.3037151.

[17] E. Tallyn, H. Fried, R. Gianni, A. Isard, and C. Speed, "The Ethnobot," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA: ACM, Apr. 2018, pp. 1–13. doi: 10.1145/3173574.3174178.

[18] I. Celino and G. Re Calegari, "Submitting surveys via a conversational interface: An evaluation of user acceptance and approach effectiveness," *Int. J. Hum. Comput. Stud.*, vol. 139, p. 102410, Jul. 2020, doi: 10.1016/j.ijhcs.2020.102410.

[19] L. F. Bouchard, L. Peters, and T. AI, *Building LLMs for Production: Enhancing LLM Abilities and Reliability with Prompting, Fine-tuning, and RAG*. Towards AI, 2024. [Online]. Available: https://books.google.co.id/books?id=siLP0AEACAAJ