



Evaluating FAQ Based Generative AI Chatbot Performance in Islamic Banking Customer Service

Nurul Izzah Ahlam¹, Mohammad Givi Efgivia²

^{1,2}Faculty of Industrial Technology and Informatics, Universitas Muhammadiyah Prof. Dr. Hamka, Indonesia

ARTICLE INFO	ABSTRACT
Published Online: 15 July 2026	This study designs and evaluates a web-based customer service chatbot for Islamic banking services. The chatbot integrates the Google Gemini API with a validated Frequently Asked Questions knowledge base to generate responses based on institutional service information. The system was developed using the Flask framework and evaluated through Black Box Testing, FAQ-based response evaluation, and chat log analytics. The evaluation used 310 validated FAQ records and 1,000 anonymized raw chat log records, which were filtered into 360 valid categorized records for interaction-pattern analysis. Based on 310 FAQ-based test questions, the chatbot produced 287 correct responses and 23 incorrect responses. The results show an accuracy of 92.58%, precision of 95.67%, recall of 96.63%, F1-score of 96.14%, success rate of 92.58%, and an average response time of 1.84 seconds. Chat log analysis further indicates that Mobile Banking was the most frequent service category requested by users. These findings suggest that combining Generative AI with a structured FAQ knowledge base can support fast and relevant customer service responses in Islamic banking.
Corresponding Author: Nurul Izzah Ahlam	
KEYWORDS: Artificial Intelligence, Chatbot, Digital Banking, Google Gemini API, FAQ Knowledge Base, Chat Log Analytics	

I. INTRODUCTION

Customer service is an important component of digital banking because it helps users obtain information, solve service-related problems, and understand banking procedures. In Islamic banking services, many customer inquiries are repetitive and information based, such as questions about mobile banking, complaints, products, transactions, and financing services. These recurring inquiries require service channels that can provide fast, consistent, and relevant answers without relying entirely on human operators. Digital banking service quality has been associated with customer satisfaction, loyalty, and continued use of digital platforms [1], [2]. Therefore, banks need service-support systems that can respond to repeated inquiries efficiently while maintaining information accuracy.

Chatbots have become one of the technologies used to support customer service automation. A chatbot enables users to submit questions through a digital interface and receive automated responses. In the banking sector, chatbots can help answer frequently asked questions, reduce waiting time, and support customer service staff by handling routine inquiries

[3], [4]. However, chatbot implementation in banking should not only be evaluated from user satisfaction or perceived usefulness. It is also necessary to examine whether chatbot responses are correct, relevant, and aligned with service information.

The development of Artificial Intelligence, particularly Generative AI and Large Language Models, has improved chatbot capability in generating natural language responses. Generative AI can produce more flexible and human-like answers based on user prompts and contextual information [5], [6]. One technology that can be integrated into chatbot applications is Google Gemini API, which allows a web-based system to submit user prompts and receive generated responses automatically [7]. However, relying only on generative output may create the risk of fluent but inaccurate responses. This issue is important in banking services because users need reliable and consistent information.

To reduce this risk, this study integrates Google Gemini API with a Frequently Asked Questions knowledge base. FAQ data are used as the main service information reference so that generated responses remain connected to validated banking

service content. In the proposed chatbot, FAQ records function not only as supporting knowledge for response generation but also as the benchmark for evaluating whether chatbot answers are correct or incorrect. FAQ-based chatbot systems can improve the consistency of information delivery because the responses are guided by predefined question-answer pairs [8], [9].

In addition to FAQ-based response evaluation, this study also applies chat log analytics. Chat logs contain records of user inquiries, chatbot responses, timestamps, interaction histories, response time, and service categories. These records can be analyzed to identify the most frequently requested service categories and determine which FAQ areas should be expanded. Previous studies suggest that conversation records and chatbot logs can provide useful evidence for understanding user behavior and improving chatbot performance [8], [10], [11]. Therefore, chat log analytics can support continuous improvement of the chatbot knowledge base.

Previous chatbot studies in banking and service contexts have commonly discussed user satisfaction, service quality, technology acceptance, and customer experience. Although these aspects are important, fewer studies evaluate chatbot performance by directly comparing generated responses with validated FAQ answers while also analyzing chat log records. This creates a research gap in the objective evaluation of chatbot correctness and interaction patterns, especially in Islamic banking services. To address this gap, this study evaluates a web-based customer service chatbot using FAQ-based response testing and chat log analytics.

This research was conducted as a case study of one Islamic banking institution in Indonesia. The institution name is not explicitly stated to maintain confidentiality and data privacy. The chatbot was developed using the Flask framework and integrated with Google Gemini API. The evaluation involved 310 validated FAQ records and 1,000 anonymized raw chat log records. After filtering incomplete, duplicate, and uncategorized records, 360 valid categorized chat logs were used for interaction-pattern analysis.

The novelty of this study lies in combining three elements: Google Gemini API as the generative response engine, FAQ knowledge base as the response reference, and chat log analytics as the basis for analyzing user inquiry patterns. The contribution of this study is an evaluation model that measures chatbot performance not only from functional testing but also from response correctness, evaluation metrics, response time, and user inquiry distribution. This approach provides technical performance evidence and practical insight into customer service demand in Islamic banking.



Figure 1: Conceptual Framework

II. LITERATURE REVIEW

A. Digital Banking Customer Service and Chatbot

Digital banking customer service refers to the process of helping users obtain banking information through digital channels. In this context, service quality is related to speed, reliability, accessibility, and the ability to provide correct information. Digital banking users expect services that are available, responsive, and easy to use. Previous studies have shown that electronic banking service quality can influence customer satisfaction and loyalty [12]. Therefore, customer service systems in digital banking should be able to support repeated information requests effectively.

Chatbots can support customer service by automating responses to common inquiries. In service settings, chatbots are used to answer routine questions, provide guidance, and reduce dependence on human operators for repetitive tasks [11], [14]. In this study, the chatbot is treated as a web-based question-answering system that receives user inquiries and returns service-related responses. The chatbot scope is limited to information-based customer service questions that can be answered using FAQ records. Therefore, the evaluation focuses on answer correctness, relevance, response time, and user question patterns rather than emotional or persuasive conversational ability.

B. Generative AI, Google Gemini API, and FAQ Knowledge Base

Generative AI refers to AI systems capable of producing new content, including text, based on user input and learned language patterns. The growth of Large Language Models has expanded the use of generative systems in chatbot applications because these models can process contextual prompts and generate coherent responses [15], [16]. Google Gemini API can be used as a response-generation component by connecting the model to a web-based application [7]. In this study, user questions are processed through a Flask application and sent to the Google Gemini API with FAQ context.

Nevertheless, generative responses need to be controlled by reliable domain information. In banking services, inaccurate or unsupported responses may cause misunderstanding. Therefore, this study uses an FAQ knowledge base as the main reference for response generation. The FAQ dataset contains validated question-answer pairs related to Islamic banking services. FAQ-based knowledge helps the chatbot provide more consistent and relevant answers because the system is guided by predefined service information [8], [10]. In the proposed system, the FAQ dataset has two roles: as the knowledge reference for chatbot responses and as the evaluation benchmark for determining response correctness.

C. Chat Log Analytics and Performance Evaluation

Chat log analytics refers to the analysis of stored chatbot interaction records. These records may include user inquiries, chatbot responses, timestamps, interaction history, categories, and response time. Chat log analysis can reveal which topics are most frequently asked and which categories require additional FAQ coverage [13], [12]. In this research, chat log analytics is used to identify dominant service categories and support improvement of the knowledge base. Chatbot performance evaluation requires measurable indicators. In this study, chatbot responses are compared with reference answers in the FAQ dataset. A response is considered correct when it answers the user question and is semantically consistent with the validated FAQ answer. The evaluation uses accuracy, precision, recall, F1-score, success rate, and response time. These metrics are commonly used to assess correctness, relevance, and efficiency of AI-based systems [15], [16], [17]. Using multiple metrics provides a more complete view of chatbot performance than relying on a single measurement.

III. METHOD

A. Research Method

This study used a Research and Development method because it involved designing, developing, and evaluating a web-based customer service chatbot. A descriptive quantitative approach was also applied to describe chatbot performance using measurable indicators, including response accuracy, precision, recall, F1-score, success rate, response time, and user inquiry distribution.

B. Research Object and System Architecture

The research object was a web-based customer service chatbot designed for an Islamic banking institution in Indonesia. The institution name was anonymized to protect confidentiality. The chatbot was developed using the Flask framework, while Google Gemini API served as the response-generation engine. The FAQ knowledge base was used as the primary reference for answering user inquiries. The response process started when a user submitted a question through the web interface. The Flask application processed the input and sent it to the Google Gemini API with relevant FAQ context. After the response was generated, the system displayed the answer to the user and stored the

interaction in the chat log database. The stored chat log was later used for evaluation and interaction-pattern analysis.

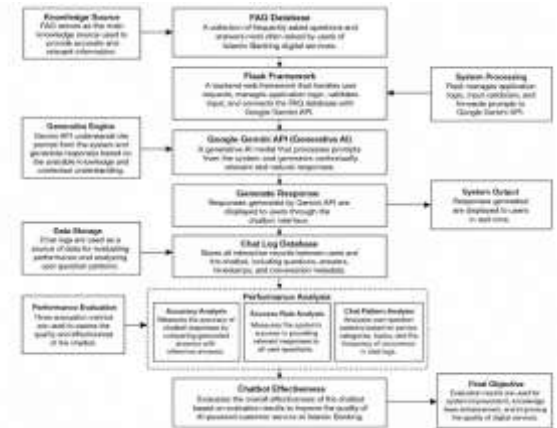


Figure 2: Research Framework

C. Research Dataset

This study used two datasets: an FAQ dataset and a chat log dataset. The FAQ dataset consisted of 310 validated question-answer pairs categorized into six service groups. The distribution was Mobile Banking 120 records, Complaints 75 records, Products 50 records, Transactions 35 records, Financing/Credit 20 records, and Others 10 records.

TABLE I. FAQ DATASET DISTRIBUTION

Category	Number of Records	Percentage (%)
Mobile Banking	120	38.71
Complaints	75	24.19
Products	50	16.13
Transactions	35	11.29
Financing/Credit	20	6.45
Others	10	3.23
Total	310	100

The chat log dataset was generated from user interactions with the chatbot during system operation. The recorded information included user inquiries, chatbot responses, timestamps, interaction history, and service categories. From 1,000 collected user interaction records, 360 valid and categorized records were retained for final analysis. The filtering process removed empty inquiries, incomplete responses, duplicate interactions, and records without category labels. All records were anonymized before analysis to protect user identity and sensitive institutional information.

D. System Evaluation

Black Box Testing was used to verify the functional performance of the chatbot without examining the internal source code. The tested functions included user input, prompt submission to Google Gemini API, response generation, chat log storage, and response display.

Response accuracy was evaluated using the complete FAQ dataset. Each FAQ question was submitted to the chatbot, and

the generated response was compared with the corresponding reference answer. Responses were classified into two classes: Correct and Incorrect. A response was labeled Correct when it addressed the inquiry and was semantically consistent with the FAQ reference answer. A response was labeled Incorrect when it was irrelevant, incomplete, or failed to answer the inquiry.

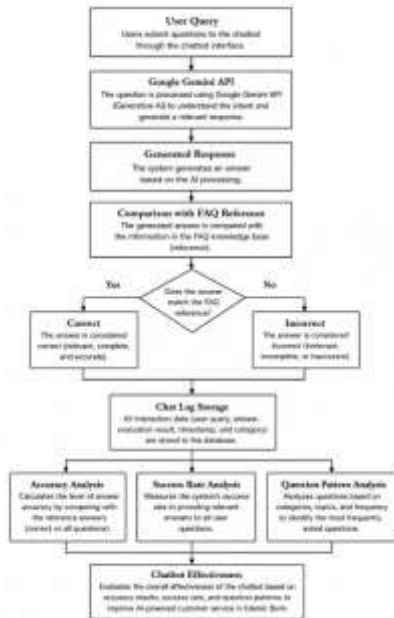


Figure 3: Research Evaluation Flow

E. Data Analysis

Chatbot performance was measured using accuracy, precision, recall, F1-score, success rate, and response time. Accuracy measured overall correctness. Precision measured the proportion of relevant predicted responses that were correct. Recall measured the chatbot ability to identify relevant responses. F1-score measured the balance between precision and recall. Success rate measured the proportion of successfully answered questions. Response time measured the average duration needed to generate an answer.

TABLE II. CHATBOT PERFORMANCE EVALUATION

Metric	Evaluation Objective	Formula
Accuracy	Measures the overall correctness of the chatbot system.	$(TP + TN) / (TP + TN + FP + FN)$
Precision	Measures the accuracy of responses predicted as relevant.	$TP / (TP + FP)$
Recall	Measures the chatbot's ability to identify all relevant responses.	$TP / (TP + FN)$
F1-score	Measures the balance between precision and recall.	$2 \times (Precision \times Recall) / (Precision + Recall)$

Metric	Evaluation Objective	Formula
Response Time	Measures the average time required for the chatbot to generate a response.	$\sum t_i / n$
Success Rate	Measures the overall success rate of the chatbot in providing relevant responses.	$\frac{\text{Number of Relevant Responses}}{\text{Total Number of Questions}}$

Chat log analysis was conducted by grouping valid interaction records based on service category. The frequency and percentage of each category were calculated to determine the most common user inquiry patterns.

IV. RESULT AND DISCUSSION

A. Functional Testing Results

Black Box Testing confirmed that the core chatbot functions worked according to system requirements. The tested features included user query input, Google Gemini API integration, response generation, chat log storage, and conversation history display. All tested features were successful and valid. This result shows that the chatbot was functionally ready for FAQ-based response evaluation and chat log analysis.

TABLE III. BLACK BOX TESTING RESULTS

No.	Feature Tested	Expected Result	Test Result	Status
1	User Query Input	The user query is successfully received by the system.	Successful	Valid
2	Google Gemini API Integration	The prompt is successfully transmitted to the Google Gemini API.	Successful	Valid
3	Response Generation	The generated response is successfully displayed to the user.	Successful	Valid
4	Chat Log Storage	Conversation records are successfully stored in the database.	Successful	Valid
5	Conversation History	Conversation history is successfully retrieved and displayed.	Successful	Valid

B. Response Accuracy Evaluation

The response accuracy evaluation used 310 FAQ-based test questions. Each chatbot response was compared with the corresponding FAQ reference answer. The chatbot produced 287 correct responses and 23 incorrect responses. Mobile Banking recorded 113 correct and 7 incorrect responses from 120 questions. Complaints recorded 69 correct and 6 incorrect responses from 75 questions. Products recorded 46 correct and 4 incorrect responses from 50 questions. Transactions recorded 32 correct and 3 incorrect responses from 35 questions. Financing recorded 18 correct and 2 incorrect responses from 20 questions. Others recorded 9 correct and 1 incorrect response from 10 questions.

TABLE IV. CHATBOT RESPONSE EVALUATION

Category	Number of Questions	Correct Responses	Incorrect Responses
Mobile Banking	120	113	7
Complaints	75	69	6
Products	50	46	4
Transactions	35	32	3
Financing	20	18	2
Others	10	9	1
Total	310	287	23

These results indicate that the chatbot answered most FAQ-based questions correctly. Mobile Banking had the largest number of incorrect responses because it also had the largest number of test questions. However, proportionally, Financing and Others had the highest error ratio, each reaching 10.00%. This suggests that categories with fewer FAQ records may require additional question variations to improve response consistency.

C. Confusion Matrix and Evaluation Metrics

The confusion matrix consisted of TP = 287, FN = 10, FP = 13, and TN = 0. The TN value was zero because the evaluation dataset did not include a separate group of intentionally irrelevant questions. Therefore, the matrix mainly represents response relevance within the FAQ-based evaluation scenario.

TABLE V. CONFUSION MATRIX

	Predicted Relevant	Predicted Irrelevant
Relevant	TP = 287	FN = 10
Irrelevant	FP = 13	TN = 0

Based on the confusion matrix, the chatbot achieved 92.58% accuracy, 95.67% precision, 96.63% recall, and 96.14% F1-score. The system also achieved a 92.58% success rate and an average response time of 1.84 seconds. These results indicate that the chatbot was able to generate relevant and correct

responses for most FAQ-based questions. The response time also shows that the system could provide answers quickly during testing. However, this result should be interpreted as an initial performance indication because the study did not yet test the system under high concurrent user traffic.

TABLE VI. EVALUATION METRICS RESULTS

Metric	Value
Accuracy	92.58
Precision	95.67
Recall	96.63
F1-score	96.14
Success Rate	92.58
Average Response Time	1.84 seconds

D. Chat Log Analysis

Chat log analysis was conducted using 360 valid categorized interaction records. The results show that Mobile Banking was the most frequent category, with 145 inquiries or 40.28% of the analyzed records. Complaints ranked second with 90 inquiries or 25.00%. Products had 55 inquiries or 15.28%, Transactions had 40 inquiries or 11.11%, Financing had 20 inquiries or 5.56%, and Others had 10 inquiries or 2.78%.

TABLE VII. DISTRIBUTION OF USER QUERIES

Category	Frequency	Percentage (%)
Mobile Banking	145	40.28
Complaints	90	25.00
Products	55	15.28
Transactions	40	11.11
Financing	20	5.56
Others	10	2.78
Total	360	100.00

The dominance of Mobile Banking indicates that users mostly interacted with the chatbot to ask about practical digital banking issues. These issues may include login problems, transaction status, application access, and mobile banking-related complaints. This finding suggests that Mobile Banking should become the main priority for FAQ expansion. Complaint-related topics should also be improved because they represent the second most frequent category.

E. Discussion

The evaluation results show that integrating Google Gemini API with a structured FAQ knowledge base can support information-based customer service in Islamic banking. The chatbot correctly answered 287 of 310 FAQ-based questions, which indicates strong performance in controlled FAQ testing. The high precision, recall, and F1-score values suggest that most responses classified as relevant were appropriate and that the chatbot was able to identify most relevant FAQ-based inquiries.

However, the 23 incorrect responses show that the system still requires improvement. Incorrect responses may occur when user questions use different wording, require more specific context, or are related to categories with fewer FAQ variations. The error ratio in Financing and Others indicates that limited FAQ coverage can affect response consistency. Therefore, improving FAQ diversity is important for future development.

Chat log analysis provides practical insight into user service needs. Since Mobile Banking was the most frequent category, chatbot improvement should prioritize mobile banking topics. This result also shows that chat log analytics can support continuous knowledge base refinement. Rather than relying only on static FAQ content, the institution can use actual interaction patterns to decide which service topics should be expanded.

Overall, the proposed chatbot can support repeated and FAQ-based inquiries in Islamic banking customer service. Nevertheless, this study has several limitations. The evaluation used FAQ-based questions and categorized chat logs from a single chatbot implementation. The system was not evaluated through direct customer satisfaction surveys, complex multi-turn conversations, or large-scale concurrent user testing. Future research may implement Retrieval-Augmented Generation, vector databases, semantic similarity evaluation, and comparison with other Large Language Models.

V. CONCLUSION

This study developed and evaluated a web-based customer service chatbot for Islamic banking services by integrating Google Gemini API with a validated FAQ knowledge base. The chatbot was designed to support repeated service inquiries and provide responses based on available service information. The system was implemented using the Flask framework and evaluated using Black Box Testing, FAQ-based response evaluation, performance metrics, and chat log analytics.

The evaluation using 310 validated FAQ records showed that the chatbot generated 287 correct responses and 23 incorrect responses. The system achieved 92.58% accuracy, 95.67% precision, 96.63% recall, 96.14% F1-score, 92.58% success rate, and an average response time of 1.84 seconds. Chat log analysis showed that Mobile Banking was the most dominant service category, with 145 inquiries or 40.28% of 360 valid chat log records.

These findings indicate that combining Generative AI with an FAQ knowledge base can support fast and relevant digital customer service responses. The study also shows that chat log analytics can help identify user information needs and guide future FAQ improvement. Future studies may extend the system using Retrieval-Augmented Generation architecture, larger datasets, semantic evaluation metrics, direct user satisfaction measurement, and comparative testing with other Large Language Models.

REFERENCES

1. C. Ledro, A. Nosella, and A. Vinelli, “Artificial intelligence in customer relationship management: literature review and future research directions,” Dec. 19, 2022, *Emerald Group Holdings Ltd.* doi: 10.1108/JBIM-07-2021-0332.
2. S. G. Ayinaddis, B. A. Taye, and B. G. Yirsaw, “Examining the effect of electronic banking service quality on customer satisfaction and loyalty: an implication for technological innovation,” *J. Innov. Entrep.*, vol. 12, no. 1, Dec. 2023, doi: 10.1186/s13731-023-00287-y.
3. D. Leocádio, L. Guedes, J. Oliveira, J. Reis, and N. Melão, “Customer Service with AI-Powered Human-Robot Collaboration (HRC): A Literature Review,” in *Procedia Computer Science*, Elsevier B.V., 2024, pp. 1222–1232. doi: 10.1016/j.procs.2024.01.120.
4. L. Jenneboer, C. Herrando, and E. Constantinides, “The Impact of Chatbots on Customer Loyalty: A Systematic Literature Review,” Mar. 01, 2022, *MDPI*. doi: 10.3390/jtaer17010011.
5. L. Banh and G. Strobel, “Generative artificial intelligence,” *Electronic Markets*, vol. 33, no. 1, Dec. 2023, doi: 10.1007/s12525-023-00680-1.
6. Z. Bin Akhtar, “Unveiling the evolution of generative AI (GAI): a comprehensive and investigative analysis toward LLM models (2021–2024) and beyond,” *Journal of Electrical Systems and Information Technology*, vol. 11, no. 1, Jun. 2024, doi: 10.1186/s43067-024-00145-1.
7. T. R. McIntosh *et al.*, “From Google Gemini to OpenAI Q* (Q-Star): A Survey on Reshaping the Generative Artificial Intelligence (AI) Research Landscape,” Feb. 01, 2025, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/technologies13020051.
8. L. A. L. F. da Costa *et al.*, “Advancing Chatbot Conversations: A Review of Knowledge Update Approaches,” *Journal of the Brazilian Computer Society*, vol. 30, no. 1, pp. 55–68, Mar. 2024, doi: 10.5753/jbcs.2024.2882.
9. H. Ho, S. M. Han, J. Cha, and L. Pham, “Mobile Banking Customer Satisfaction and Loyalty: The Roles of Technology Readiness,” *Journal of Risk and Financial Management*, vol. 18, no. 7, Jul. 2025, doi: 10.3390/jrfm18070403.
10. K. W. Church, J. Sun, R. Yue, P. Vickers, W. Saba, and R. Chandrasekar, “Emerging trends: A gentle introduction to RAG,” *Nat. Lang. Eng.*, vol. 30, no. 4, pp. 870–881, Jul. 2024, doi: 10.1017/S1351324924000044.
11. K. Kuligowska and M. Stanusch, “Commercial chatbot monitoring: Approaches focused on automated conversation analysis,” *Humanities &*

Social Sciences Reviews, vol. 12, no. 2, pp. 54–60, Sep. 2024, doi: 10.18510/hssr.2024.1227.

12. S. Woollaston, B. Flanagan, P. Ocheja, Y. Toyokawa, and H. Ogata, “ARCHIE: Exploring Language Learner Behaviors in LLM Chatbot-Supported Active Reading Log Data with Epistemic Network Analysis,” in *15th International Conference on Learning Analytics and Knowledge, LAK 2025*, Association for Computing Machinery, Inc, Mar. 2025, pp. 642–654. doi: 10.1145/3706468.3706556.
13. G. Caldarini, S. Jaf, and K. McGarry, “A Literature Survey of Recent Advances in Chatbots,” *Information (Switzerland)*, vol. 13, no. 1, Jan. 2022, doi: 10.3390/info13010041.
14. S. U. Singh and A. S. Namin, “A survey on chatbots and large language models: Testing and evaluation techniques,” *Natural Language Processing Journal*, vol. 10, Mar. 2025, doi: 10.1016/j.nlp.2025.100128.
15. M. Abbasian *et al.*, “Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative AI,” *NPJ Digit. Med.*, vol. 7, no. 1, Dec. 2024, doi: 10.1038/s41746-024-01074-z.
16. T. K. Q. Nguyen, “E-Banking Service Quality and Customer Satisfaction with Moderator Factor,” *Emerging Science Journal*, vol. 9, no. 2, pp. 725–740, Apr. 2025, doi: 10.28991/ESJ-2025-09-02-012.
17. T. B. Mashigo, W. Dinath, and S. Khumalo, “Chatbot evaluation for effectiveness in customer query resolution,” *S. Afr. J. Inf. Manag.*, vol. 27, no. 1, Sep. 2025, doi: 10.4102/sajim.v27i1.1963.