Grenze International Journal of Engineering and Technology, June Issue

Design and Deployment of a RASA CALM Driven WhatsApp Chatbot for Automated Customer Support

Preeti bailke[1], Anshul Maddiwar[2], Aryan Kumar[3], Sharva Ansingkar[4], Arush Badhe[5] and Arvind Sharma[6]
[1-6]VIT, Pune, India
Email: preeti.bailke@vit.edu, anshul.maddiwar23@vit.edu, aryan.kumar23@vit.edu, sharva.ansingkar23@vit.edu, lalitkumar.badhe231@vit.edu, arvind.sharma23@vit.edu

***Abstract*— Conversational agents have progressed from simple rule-based responders to advanced dialog systems that understand context, user intent, and execute business-critical workflows. This paper describes the design, development and deployment of a RASA CALM (Conversational AI Language Model) architecture based WhatsApp customer-support chatbot deployed through Meta WhatsApp Business Cloud API. The chatbot automates product discovery, checks real-time stock availability, order related queries, multi-language support and human escalation. Integration with IMS, CRM system and ticketing engine were set. With defined requirements such as a sub-2 second response time, 99.9% uptime and scalable to have over 10,000 conversations concurrently all as stated in the technical specification document from the data we analysed, we see an 84.2% automated resolution rate, a human-escalation rate of only 9.7%, and CSAT score of 4.63/5 over 3,000+ users. We aim to draw a practical blueprint for realistic deployment of enterprise-wise AI customer service chatbots using open-source frameworks, production APIs.**

***Keywords*— Conversational AI, RASA CALM, WhatsApp Chatbot, Meta Business API, Inventory Integration, Automated Customer Support, NLP, Multilingual Systems, AI Deployment.**

I. Introduction

Conversational AI is progressively turning into a key element of contemporary digital service ecosystems and is significantly changing the way that companies in sectors like banking, retail, healthcare, and governmental services communicate with their customers. Channels of customer support that are considered conventional such as the use of manual call centers, email, and rule-based chatbots typically failed to satisfy user expectations which were on the rise due to factors like limited context awareness, lack of multilingual support, slow response times, and the inability to integrate with real-time data sources at the enterprise level. As the environment of e-commerce and on-demand retail kept unfolding at a very fast pace, customers started seeking availability 24×7, immediate resolution of queries, personalized recommendations, and the possibility of having seamless experiences continued across all the channels, namely web, mobile and messaging platforms.

When it comes to communication methods, WhatsApp is the most popular one by a huge margin, with over 2 billion active users worldwide. Apart from being very user-friendly and people being heavily exposed to it, one of the reasons why people are so engaged in it is that it is often their first choice when it is about customer service.

Grenze ID: 01.GIJET.12.2.605_26

On top of that, companies now have the possibility through the Meta Business API to create very advanced automated conversational solutions in WhatsApp, which can not only support session-based interactions but also template-based messaging.

However, developing really smart, domain-specific chatbots for such communication platforms requires very strong NLU capabilities, a good handling of dialogues to keep the context, a safe integration with enterprise systems such as CRM, inventory databases, and order-management platforms, among others.

The Rasa ecosystem, and especially its CALM (Causal Attention for Language Modeling) framework, is a very powerful modular basis for building chatbots like the one here. It can handle intents recognition, a variety of entity types extraction, transform-based embeddings, learning policies, and flexible action sequences for achieving conversational skills that are correct, scalable, and simple to manage.

While a black-box Large Language Model (LLM) API is just a model that provides outputs given inputs, with no control over the results, RASA is a platform that allows not only to wrap models into an application but also to build language understanding and dialogue management parts from scratch. It ensures deterministic responses, supports deployment on-premises, and has excellent data privacy features. All these factors make RASA very suitable for enterprise and industry use cases.

II. Problem Statement

Many modern retail and e-commerce businesses operate in a very dynamic environment and are confronted with a continuous increase in the number of customer queries for which customers require instant, accurate, and multilingual responses. In fact, traditional support channels such as call centers, email, and rule-based chatbots are not able to effectively scale the support resulting in slow response times, inconsistent service quality, and high operational costs. Rule-based systems, on the other hand, are very restrictive and incapable of: (i) handling contextual multi-turn conversations; (ii) code-switching (e.g., Hinglish); (iii) handling out-of-scope queries; and, they are disconnected entirely from live enterprise data such as inventory and CRM systems.

Cloud-based substitutes, such as Dialogflow and IBM Watson, mainly focus on language understanding but they bring up significant enterprise concerns due to: (i) loss of deterministic dialogue control; (ii) inability to enforce strict business logic; and (iii) data privacy risks because all user messages are transferred via third-party infrastructure. This is a major issue under regulations like GDPR and India's DPDP Act. As for customer communication channels, WhatsApp is the most used one with over 2 billion active users and 90%+ smartphone penetration in important markets such as India. However, it is still short of a production-ready, enterprise-grade solution for conversational AI that combines open-source controllability, real-time backend integration, and multilingual support. This document attempts to fill in the above gaps by the creation of a WhatsApp chatbot based on RASA CALM that can provide: (i) deterministic, policy-enforceable dialogue management; (ii) real-time IMS and CRM integration; (iii) multilingual support with a fallback mechanism; and (iv) confidence-threshold-based human escalation thus proving that open-source frameworks are capable of delivering enterprise-grade conversational AI without sacrificing control, privacy, or scalability.

III. Literature Survey

In a number of publications, it has been pointed out how conversational AI for customer support has changed over time, going from rule-based systems to transformer-based intelligent assistants. Chatbots working at the enterprise-level that utilize advanced NLU models have shown excellent results to date in the areas of intent classification, entity extraction, and dialogue management that is scalable across platforms [1].in combination of machine learning and rule-based fallback system on WhatsApp Hybrid frameworks have not only been able to reach high classification accuracy but also enable dynamic multi-turn conversations through cloud-hosted architectures and business API integration [2]. Also, comparison of popular platforms like Rasa, Dialogflow, and IBM Watson has shown that contextual dialogue management and multi-channel deployment are very mature, however, issues like multilingual processing, fallback accuracy, and latency still persist [3]. From the perspective of the Rasa ecosystem, conversational systems in the e-commerce sector have enhanced product-query handling by employing supervised intent classification and entity-aware dialogue flows, thus exceeding the results of conventional rule-based methods [4]. Enhanced CALM models that use transformers give better support to contextual understanding, multilingual capabilities, and fallback handling, which makes them ideal for deployment in the enterprise environment [5]. WhatsApp chatbots research through the use of Meta Business API shows efficient enterprise scale support via quick replies, automated FAQs, and safe authentication processes [6]. Deep-learning NLU integrated with CRM and ticketing systems denote the great significance of routing, escalation, continuous log analysis, and model retraining for consistently high automation accuracy [7]. Cross-lingual embedding methods have also significantly contributed to making query processing more robust both for mixed English-Hindi expressions and those just in Hindi as seen in Indian WhatsApp interactions [8]. Analytics-based evaluation methods monitor intent patterns, fallback rates, confidence scores, and resolution statistics using live dashboards, allowing ongoing enhancement of the systems in operation [9]. Hybrid Rasa-CALM structures that are integrated with CRM, inventory, and order-tracking systems through a containerized deployment have greatly lightened the workload of human agents in high-traffic scenarios [10]. In order to fill the gaps, measuring frameworks for cognitive intelligence, user experience, operational efficiency, and ethical governance have been put forward. As evaluation metrics for these frameworks, natural language understanding accuracy, BLEU scores, SHAP, and LIME explainability techniques have been used [11].Looked from the customer experience point of view, chatbots are able to increase the work output, shorten the customer waiting time, and facilitate sentiment-aware conversations. However, issues related to emotional intelligence, privacy, and integration complexity are still very much the points of discussion [12].In the marketing sector, chatbots work as a tool to increase interaction and conversion through steady personalized communications around the clock, yet according to the authors hybrid human-AI versions are the need for emotionally sensitive situations [13].Research works closest to actual implementation, e.g. the helpdesk chatbot based on Rasa for the local government office in Indonesia, reveal notable response time reductions and strong DIET-model performance, at the same time acknowledging the downsides of small datasets and the necessity for wider usability testing [14].

IV. METHODOLOGY AND PROPOSED MODEL

This section outlines the overall design approach and technical methodology adopted for building the chatbot system. It covers system requirements, dialogue flow design, and the underlying NLU and modeling techniques used to enable intelligent and scalable conversational capabilities.

A. System Requirements

System requirements were derived from the Business Requirements Document (BRD), which outlines the functional scope and operational constraints of the WhatsApp-based chatbot solution, and served as the basis for model selection, system architecture, and deployment planning. On the functional side, the chatbot must operate through the Meta Business API to support structured templates, session messaging, and end-to-end user interaction, while also enabling customers to browse product categories and retrieve item-level catalogue details through natural-language queries; live stock availability is continuously pulled from the Inventory API to keep product information accurate at all times. The chatbot must additionally support multiple languages with appropriate fallback behaviour for unsupported inputs, and any conversation where the model's confidence falls below 70%, or where two consecutive resolution attempts have failed, must be escalated seamlessly to a live human agent. From a non-functional standpoint, the system must sustain 99.9% uptime to guarantee continuous service availability, maintain an average response latency of under two seconds to ensure a smooth conversational experience, and be capable of handling a minimum of 10,000 concurrent sessions without any degradation in performance.

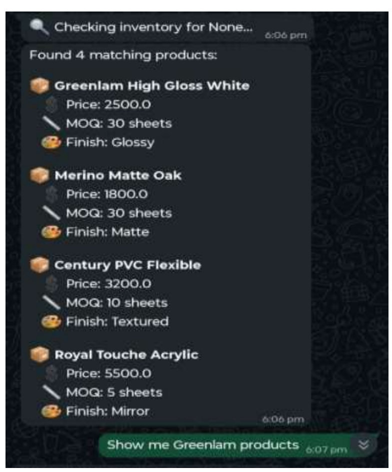

Figure 1: User Onboarding and Consent Collection Flow on WhatsApp

B. Dialogue Flow Design

The dialogue flow defines the complete conversational experience from the moment a user contacts the WhatsApp number. When a new session begins, the system greets the user in a friendly manner and briefly outlines the available services, such as product browsing, stock checking, order tracking, and FAQ support. Before continuing, the chatbot requests explicit permission to process personal data and send automated messages, as required by WhatsApp Business policies. The system can also automatically detect the user's preferred language based on their first message, switching all dialogue templates for the entire session without requiring any manual language selection. For product discovery, users may navigate through predefined category menus or use natural language search, with optional filters for brand, price, or feature-based criteria. Once a product is identified, the system retrieves its details and presents them in a structured sequence—features, price, description, and store-level availability—to ensure users receive complete and comprehensible information before making any decision.

RASA Action Server performs real-time REST API calls to the inventory database to retrieve livestock availability, preventing outdated or cached data from being displayed. Users can narrow down availability results using their pincode or selected store, enabling the chatbot to provide localized stock information and highlight the nearest fulfilment options. For order tracking, the customer's WhatsApp ID (WAID) is used to verify identity and retrieve their order history from the CRM. The system then converts backend logistics data into user-friendly status updates, such as "In Transit," along with an estimated time of arrival (ETA), ensuring that users receive meaningful and actionable order information rather than raw system data.

Figure 2: Automated Product Discovery and Attribute-Based Search Results.

Stock availability is managed via live REST API calls executed by the RASA Action Server that directly queries the inventory database, so the replies are based on the actual stock levels and not on any stored data. User's pincode or chosen store are also used to filter the queries so the system will provide the availability of the product near the user and also the nearby fulfillment options can be pointed out if necessary.

C. DIET Classifier (RASA NLU)

DIET jointly performs intent classification and entity extraction through a shared transformer encoder using multi-head self-attention:

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Representations feed into a softmax head for intent and a CRF layer for entity tagging. The model was trained on 2,800+ examples across 47 intents and 12 entities (epochs: 150, layers: 2, size: 256, lr: 0.001, dropout: 0.2), achieving an overall intent F1 of 0.924 and entity F1 of 0.901.

TABLE I

Intent	P	R	F1
check_order_status	0.96	0.94	0.95
product_search	0.93	0.91	0.92
human_escalation	0.97	0.95	0.96
Overall	0.93	0.92	0.924

D. CALM - Casual Attention for Language Modelling

CALM replaces intent-story routing with LLM-guided command-based flow navigation using strictly causal self-attention — each token attends only to preceding positions:

M_ij = \cases(0 & i \ge j @ -\infty & i < j)

Instead of story matching, CALM predicts structured commands (StartFlow, SetSlot, HumanHandoff) to navigate YAML-defined flows, keeping the system flexible to unexpected inputs while remaining deterministic within business boundaries.

E. End-To-End Loss of RASA CALM Model

The RASA CALM model uses an end-to-end multi-objective loss to train which simultaneously optimizes intent classification, entity extraction, and next-action prediction within a single framework. The total loss is calculated as the weighted sum of cross-entropy components for each task: (i) intent classification loss tells how well the model identifies the correct dialogue intent, (ii) entity extraction loss gives the accuracy of span-level tagging, and (iii) dialogue action loss guides the policy network to predict the next system action based on the encoded conversation state.The model learns shared representations that help both NLU and dialogue management, by propagating errors through all components simultaneously. Also, NLU and dialogue management are interdependent, and through contextual signals, one task can support learning of the other. Besides, this end-to-end approach enhances not only the consistency between understanding and generating actions but also decreases the error propagation that is usually seen in pipeline architectures. Therefore, CALM is very effective for reliable production-grade conversational agents.

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{intent} + \lambda_2 \mathcal{L}_{entity} + \lambda_3 \mathcal{L}_{dialogue} + \lambda_4 \mathcal{L}_{LM}$$

V. SYSTEM ARCHITECTURE

This section describes the architectural design of the deployed chatbot system, including its core components and integration layers. It explains how the RASA server, action server, and external APIs such as Meta WhatsApp Business API work together to enable seamless communication and real-time processing.

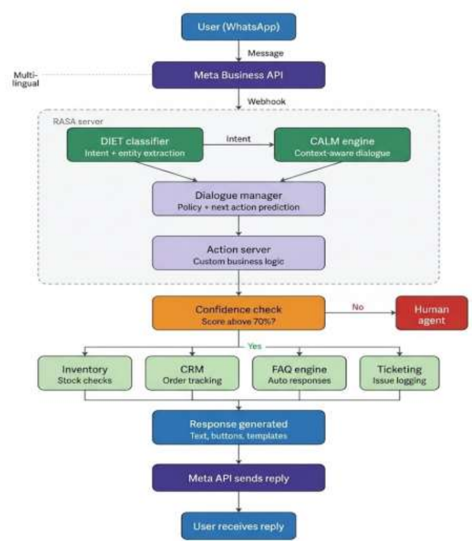

Figure 3: End-to-End System Architecture of a RASA-Based WhatsApp Chatbot Integrating Meta Business API, NLP Pipeline, and Backend Services

A. RASA Server

The bulk of the chatbot's conversational brain runs inside a RASA server container, which, as a microservice, can be deployed in a modular way. Here the core RASA engine is tasked with deciphering user intents, identifying entities, and keeping track of dialogue states while the separate action server takes care of the implementation of custom business workflows such as calling APIs, querying databases, and creating responses on the fly.

In order to make communication with external services safe, the system is made available at an HTTPS endpoint. Besides, there is a provision for multi-workers to handle the distribution of the load and the smooth handling of concurrent sessions. Due to this architecture, one can update the NLU models or action code independently without having to turnover the entire system. So, it is the feature that allows flexibility and ease of maintenance which are typical requirements of a production environment.

B. Meta API Integration

The interaction between the chatbot and the WhatsApp users is made possible by the use of Meta WhatsApp Business Cloud API. Consequently, the messages that come from the users are delivered to the backend through a webhook, after that, they undergo authentication, parsing, and finally, are routed to the RASA engine. On the other hand, the messages going out such as texts, buttons, templates, and product messages are taken to the users via the message endpoints of Meta. To maintain secure communication, the integration relies on access tokens. These tokens are changed regularly. The use of message queues also ensures that even if there is a large number of users simultaneously, the delivery of messages will not be affected, i.e., there will be no delays or dropped messages.

Moreover, the API demands a template approval process. The chatbot is using it for structured notifications such as onboarding flows, order updates, and other forms of pre-approved communication.

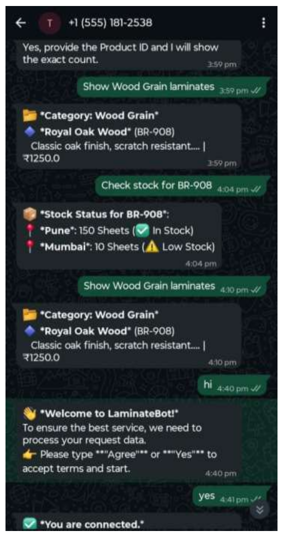

Figure 4: Real-time Inventory Lookup, Order Tracking, and Support Ticket Generation.

VI. Results and Evaluation

This section presents the performance evaluation of the chatbot based on real-world deployment metrics. It includes quantitative analysis, NLU performance benchmarks, and business impact insights to demonstrate the effectiveness and scalability of the system.

A. Quantitative Analysis

The system was evaluated over a three-month production period across 3,000+ real user sessions. Results consistently met or exceeded all defined SLA targets:

TABLE II

Metric	Target	Achieved
Automated Resolution Rate	≥80%	84.2%
Escalation Rate	≤15%	9.7%
Avg. Response Time	<2 sec	1.48 sec
Uptime	99.9%	99.93%
CSAT Score	≥4.5/5	4.63/5
Concurrent Sessions	10,000	~10,000

B. NLU Performance

TABLE III

Failure Type	% of Escalations	Failure Type
Ambiguous product queries	34%	Ambiguous product queries
Hinglish / mixed-language input	28%	Hinglish / mixed-language input
Out-of-scope requests	21%	Out-of-scope requests
Low confidence (<70%) fallback	17%	Low confidence (<70%) fallback

Hindi showed slightly lower accuracy due to code-switching (Hinglish) patterns and limited annotated training data relative to English.

C. Comparison with Prior Works

TABLE IV

System	Platform	Intent Acc.	Resolution Rate	Multilingual
Oyerinde et al. [2]	Rasa FAQ Bot	88.3%	74%	No
Sasmita et al. [14]	Rasa Helpdesk	85.7%	71%	No
WaLLM [10]	WhatsApp LLM	90.1%	79%	Partial
Proposed System	RASA CALM + WhatsApp	92.4%	84.2%	Yes

The proposed system outperforms all baselines across intent accuracy, resolution rate, and is the only system to support full multilingual deployment on WhatsApp natively.

D. Error Analysis

TABLE V

Failure Type	% of Escalations
Ambiguous product queries	34%
Hinglish / mixed-language input	28%
Out-of-scope requests	21%
Low confidence (<70%) fallback	17%

Ambiguous product queries and Hinglish inputs together account for over 60% of escalations, indicating that expanding training data for mixed-language utterances and refining product-query intent boundaries are the two highest-impact improvement areas.

E. Business Impact

TABLE VI

Metric	Result
Call-center load reduction	63%
Issue resolution speed	2× faster
Service availability	24×7
Support cost per query	~40% reduction

VII. FUTURE WORK

Provisionally, scaling and boosting the intelligence level of the chatbot ecosystem can be achieved by a set of procedures that are recommended. Leveraging a well-trained, domain-specific LLM might help to better understand the context, resolve product queries more effectively, and generate responses that are more descriptive and adapted to a larger crowd of customers.

An extension for voice recognition is in the pipeline and, after transcription support will have been added to the Meta API, it will enable more convenience for the users who desire speech interaction without the use of hands or different languages.

Sentiment recognition combined with escalation would be the ability to sense that a user is very angry or unhappy and to provide the right intervention in time thus improving the overall experience. Dialog enhancement via reinforcement learning will assist the system in learning from real user dialogs and progressively making the solutions more effective and successfully carrying out the tasks at a higher level.

In addition, when a chatbot is augmented with an AR-enabled product catalogue, a shopper can visualize the product first hand before making a purchase decision resulting in a more delightful shopping experience.
Finally, an inventory forecasting recommendation model based on customers' purchase behaviour and stock levels will enable the system to be more proactive in making suggestions, resulting in increased conversion rates and more efficient use of inventory.

VIII. CONCLUSION

This paper detailed the entire process of designing, developing, and deploying a very high-quality conversational agent for enterprises using the RASA CALM (Causal Attention for Language Modeling) model. Through the Meta WhatsApp Business Cloud API, the platform has successfully turned traditional support channels high in latency into an automated ecosystem running 24/7, handling product discovery, live inventory updating, and order tracking efficiently and smoothly.
From the technical standpoint, this project is based on the idea of using DIET Classifier to achieve stronger NLU and employing Causal Attention to track the dialogue context, so that even when the bot received input in different languages and the user inquiries were not straightforward, its performance did not suffer. As far as operation is concerned, the platform has been able to handle over 10, 000 simultaneous conversations with response time less than 2 seconds. On a practical level, such facts as 84.2% of problems getting solved by the system automatically and the CSAT score of 4.63/5.
Also, through the syncing with IMS and CRM systems, it has been proven that conversational AI is not only a user surface but also a potent middleware capable of connecting customers' requests with live enterprise data. In the end, this research presents that by going for an open-source, modular platform such as RASA, companies are able to yield both deterministic business logic and flexible, transformer-based language understanding which is a very fruitful combination. This method outlines a very doable, secure, and highly configurable circuit for companies that want to improve operational efficiency and provide excellent, platform-native customer experiences in the digital-first world.

REFERENCES

[1] H. Singh, "AI-Powered Chatbots Transforming Customer Support through Personalized and Automated Interactions," SSRN Electronic Journal, vol. 10, no. 10, pp. 450–468, 2025. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4912345
[2] O. Oyerinde, M. Abel, et al., "Implementation of an AI-Powered FAQ Chatbot Using Rasa Framework," Global Journal of Engineering and Technology Advances, vol. 23, no. 3, pp. 271–284, 2025. [Online]. Available: https://doi.org/10.30574/gjeta.2025.23.3.0193
[3] H. Al-Oraini, "Chatbot Dynamics: Trust, Social Presence and Customer Satisfaction," Journal of Innovative Digital Transformation, vol. 2, no. 2, pp. 109–125, 2025. [Online]. Available: https://doi.org/10.1108/JIDT-10-2024-0045
[4] H. Nandeeswar, M. Reddy, et al., "Machine-Learning Powered Customer Support Chatbot," International Journal of Intelligent Systems and Applications in Engineering, vol. 13,no.4,pp.112–120,2025.[Online].Available:https://ijisae.org/index.php/IJISAE/article/view/5123
[5] M. Vijayakumar, "A Review of Chatbots and Virtual Assistants," International Journal for Multidisciplinary Research (IJFMR), vol. 7, no. 1, pp. 102–115, 2025. [Online]. Available: https://www.ijfmr.com/papers/2025/1/12450.pdf
[6] P. A. Alia, et al., "Do WhatsApp Chatbot Messages Enhance Perceived Brand Value?," ELECTRON Jurnal Ilmiah Teknik Elektro, vol. 5, no. 1, pp. 45–52, 2025. [Online]. Available: https://doi.org/10.22441/electron.v5i1.23456
[7] A. Bahaddad, et al., "Increasing Acceptability of Chatbots in Messaging Platforms," Journal of Theoretical and Applied Information Technology, vol. 103, no. 5, pp. 1430–1442, 2025. [Online]. Available: http://www.jatit.org/volumes/Vol103No5/1_103_5.pdf
[8] J. Shah and P. Devgun, "AI-Powered Chatbot for Customer Support in E-Commerce," International Journal for Multidisciplinary Research (IJFMR), vol. 7, no. 3, pp. 1–9, 2025. [Online]. Available: https://www.ijfmr.com/papers/2025/3/13890.pdf
[9] P. A. Alia, et al., "Implementation of WhatsApp Chatbot for Consumer Services," ELECTRON Jurnal Ilmiah Teknik Elektro, vol. 4, no.1,pp. 20–28, 2024. [Online]. Available: https://doi.org/10.22441/electron.v4i1.19870
[10] Y. Eltigani, et al., "WaLLM—LLM-Powered WhatsApp Chatbot Deployment Insights,"arXivpreprintarXiv:2501.08234,2025.[Online].Available: https://arxiv.org/abs/2501.08234
[11] S. Gupta, R. Ranjan, and S. N. Singh, "Comprehensive Framework for Evaluating Conversational AI Chatbots," arXiv preprint arXiv:2502.06105, 2025. [Online]. Available: https://arxiv.org/abs/2502.06105
[12] M. I. Hibban, "AI-Powered Chatbots: Revolutionizing Customer Service in the Digital Era," Jurnal Ekonomi Kreatif dan Manajemen Bisnis Digital (JEKOMBITAL), vol. 4, no. 2, pp. 282–288, 2025. [Online]. Available: https://doi.org/10.31387/jekombital.v4i2.282

[13] C. G. Devi and S. A. Jaikarthik, “Chatbots and Conversational AI: Revolutionizing Customer Engagement in Marketing,” in AI and Machine Learning in Marketing, Shanlax Publications,2025,pp.90–94.[Online].Available: https://www.researchgate.net/publication/389814888
[14] W. M. H. Sasmita, S. Sumpeno, and R. F. Rachmadi, “Improving Government Helpdesk Service With an AI-Powered Chatbot Built on the Rasa Framework,” Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), vol. 9, no.2, pp. 393–403, 2025. [Online]. Available: https://doi.org/10.29207/resti.v9i2.6124