Boosting CRM Chatbot Solutions with Flash Attention and Probabilistic Inference

Conference Paper · June 2024

CITATIONS
0

READS
166

4 authors:

Ahmet Can Günay
Istanbul Kültür University
5 PUBLICATIONS **0** CITATIONS
SEE PROFILE

Muhammet Furkan Özara
Next4Biz
6 PUBLICATIONS **3** CITATIONS
SEE PROFILE

Akhan Akbulut
Istanbul Kültür University
78 PUBLICATIONS **1,125** CITATIONS
SEE PROFILE

Ahmet Erkan Çelik
Next4biz Information Technologies
14 PUBLICATIONS **3** CITATIONS
SEE PROFILE

Boosting CRM Chatbot Solutions with Flash Attention and Probabilistic Inference

Ahmet Can Günay[1][0009−0007−1145−2888], Muhammet Furkan Özara[1][0009−0006−2652−7179], Ahmet Erkan Çelik[1][0000−0001−5462−698X], and Akhan Akbulut[1,2][0000−0001−9789−5012]

[1] R&D Center of Next4biz, Sahrayicedit Mah, Pakpen Plaza, No:40/4, 34734 Istanbul, Turkey
{can.gunay,furkan.ozara,erkan.celik,akhan.akbulut}@next4biz.com
[2] Department of Computer Engineering, Istanbul Kültür University, 34536 Istanbul, Turkey
a.akbulut@iku.edu.tr

Abstract. In an effort to deliver exceptional customer service, organizations are increasingly recognising the critical role that advanced language models (LLMs) play in the integration of CRM systems. This article explores the forefront of chatbot technology, with a particular focus on the essential function they play in contemporary consumer relationship management. The ability of chatbots to provide personalized and contextually pertinent responses to customer inquiries has been significantly enhanced by the rapid development of LLMs, specifically in the areas of sentiment, intent, and context comprehension. Organizations have the opportunity to enhance customer contentment and optimize operations by leveraging the capabilities of these advanced AI systems. In addition, offline capabilities guarantee continuous support, thereby enhancing customer confidence and loyalty in a time when connectivity fluctuations continue to be a challenge. The present study introduces an innovative offline chatbot system that aims to overcome the limitations of traditional cloud-based counterparts. By efficiently integrating data from various sources, such as project documentation and social media, this chatbot is capable of operating independently and delivering assistance to users, even in locations with sporadic internet connectivity. The experimental outcomes illustrate that our newly developed chatbot model outperforms established benchmarks, as evidenced by its 4.1-second inference time, 90.2-point BLEU score, and 9.7-point WER score. These metrics underscore the model's effectiveness, precision, and timeliness in handling user inquiries and producing responses of exceptional quality. The experiments we conduct are designed to validate the effectiveness of our model in improving customer service experiences within CRM systems through comprehensive performance testing.

Keywords: Natural Language Processing · NLP · Chatbot · Flash Attention · Probabilistic Inference · Data ingestion.

1 Introduction

In the contemporary landscape of customer support, the fusion of artificial intelligence and chatbot technology has emerged as a pivotal force in elevating user experiences [1]. As organizations seek innovative solutions to provide efficient and seamless support, the limitations of conventional online chatbots in scenarios of unreliable or absent internet connectivity have become increasingly apparent. Addressing this challenge, introduces a cutting-edge offline chatbot system that not only transcends the constraints of connectivity but also redefines the scope and capabilities of autonomous customer support. The integration of artificial intelligence into customer service is no longer a luxury but a strategic necessity. Our offline chatbot stands at the forefront of this revolution, operating independently of real-time internet access and ensuring that users receive responsive assistance regardless of their connectivity status. While traditional online chatbots might falter in situations of network instability, our solution remains resilient, providing a consistent and reliable support experience.

This groundbreaking chatbot derives its strength from a multifaceted approach to data ingestion, drawing insights from diverse sources. By incorporating real-time information from dynamic social media platforms, the chatbot stays dynamically attuned to user sentiments, emerging trends, and evolving concerns. Simultaneously, the assimilation of data from project documents imparts a stable foundation of organizational knowledge, allowing the chatbot to navigate intricate product details, service nuances, and operational intricacies. This paper aims to unravel the intricacies of our offline chatbot system, delving into the architectural nuances that empower its autonomy. The meticulous process of data ingestion, from both social media channels and project documents, will be scrutinized to illustrate the comprehensive knowledge base upon which the chatbot relies. Additionally, the artificial intelligence algorithms governing the chatbot's decision-making processes will be explored, elucidating the sophistication that underpins its proficiency.

In summary, an offline chatbot represents a transformative leap in the realm of customer support, seamlessly blending dynamic social media insights with the stability of project document knowledge. By doing so, it not only overcomes the limitations of online chatbots but sets a new standard for autonomous, versatile, and effective customer support, making it an invaluable asset for organizations seeking to enhance customer satisfaction and loyalty.

The remainder of this paper is structured as follows. Section 2 presents a comprehensive review of the current literature on chatbot usage, summarizing existing work, the current state of the field, and identifying gaps that this study aims to address. In Section 3, we detail the methodology for data collection, including an exhaustive description of how data from various sources, such as project documentation and social media, is integrated to ensure the chatbot's contextual relevance. Section 4 describes the incorporation of Flash Attention and the Probabilistic Inference Layer into the chatbot, explaining their roles and how they enhance the system's performance. Section 5 presents a comparative analysis of our chatbot against ChatGPT-3.5, Mistral-7b, and the original

LLama-2 7b LLM, evaluating performance metrics such as inference time, BLEU score, and WER score. Finally, Section 6 reviews the key findings, discusses the implications for improving customer service experiences, and suggests directions for future research in the field of advanced AI-driven CRM solutions.

2 Related Work

The Turing Test, developed by computer scientist and inventor Alan Turing in the 1950s, sought to ascertain if a machine could behave intelligently enough to be mistaken for a person [2]. This is where the idea of chatbots originated. The first chatbot, ELIZA, was developed in 1966 by Joseph Weizenbaum. Its purpose was to mimic a psychotherapist by reacting to user inputs with preprogrammed answers [3]. More sophisticated chatbots, like Parry and ALICE, were created in the 1980s and 1990s as a result of developments in machine learning and natural language processing. The emergence of mobile devices and messaging platforms in the early 2000s simplified the process for businesses to include chatbots into their customer support systems. Advances in artificial intelligence (AI), such as deep learning and natural language processing, have allowed chatbots (like IBM Watson Assistante and OpenAI's ChatGPT) to conduct more sophisticated and natural conversations with users. As a result, chatbots are now widely used across a variety of industries, including e-commerce, finance, and healthcare [4, 5].

Chatbots for e-commerce are AI-driven programs that mimic human communication to keep customers interested throughout the purchasing process [6]. They are employed by online retailers to enhance user experience, increase sales, and provide real-time answers to a variety of client questions. E-commerce chatbots can enhance customer experience and website functioning. FAQs, customer engagement, sales automation, post-sale assistance, feedback collection, and data collection are all capabilities they provide. AI-powered educational chatbots are programs created to include students in their educational process. In the realm of education, from elementary schools to colleges, they are growing in popularity. Chatbots for education have a number of applications. They can use a pre-planned conversational route to act as peer or instructional agents. A tailored learning method that meets the requirements of pupils is used by more than 25 percent of teachers [7]. In addition to other design concepts, additional chatbots make advantage of experience and collaborative learning [8]. To ensure the secure and efficient transmission of sensitive customer data, chatbot system leverages lightweight encryption techniques similar to those used in IoT devices and medical sensors, as demonstrated by the LWE algorithm [9], which balances security and performance in resource-constrained environments. To ensure a safe and secure user experience, our chatbot system integrates advanced filtering mechanisms akin to the agent-based approach used for detecting and classifying inappropriate video content, thereby preventing the dissemination of harmful materials to users [10].

AI-powered software called chatbots for HR is used to automate and optimize HR processes. In the HR domain, they are growing in popularity for anything from hiring to employee engagement [11]. By automating repetitive chores, HR chatbots may greatly enhance HR operations and free up HR experts to concentrate on strategic work like hiring, retaining employees, motivating them, developing their leadership, and fostering a positive company culture [12]. By incorporating advanced classification techniques similar to the HMM-based approach used for categorizing sports videos by color features, chatbot system can enhance user interactions through improved contextual understanding and response accuracy [13]. They are able to offer new hires and staff prompt assistance around-the-clock. Even while chatbots are becoming more and more common, many individuals still have ethical reservations about them. Studies on chatbots and human trust and emotion have started to surface. Considerable work has gone into developing algorithm-level solutions, which primarily address a subset of moral precepts including explainability, fairness, and privacy. Nevertheless, there aren't enough RAI governance and engineering studies to evaluate and reduce chatbot AI dangers in accordance with all AI ethical guidelines.

3 Methodology

Offline chatbot system uses information from a wide range of sources, including technical documents produced by the firm and films available on other platforms. An overview of the data types and methods used for data collecting and processing by the chatbot is given in this section.

3.1 Video Data

Using movies as a rich source of information, the chatbot system captures and distributes information through both visual and auditory input. OpenAI's Whisper infrastructure is used to extract data from videos [14]. Whisper makes it easy to extract and process important information from videos so the chatbot can easily access and analyze relevant information. This requires extracting spoken speech, visual signals, and underlying context from films.

3.2 Technical Documents

Apart from videos, the chatbot system gets information from a ton of technical documentation that the organization has written. These papers come in a variety of forms, such as Word documents, Excel spreadsheets, PDF files, and other proprietary formats. The variety of document kinds is a reflection of the complexity of operational data and organizational knowledge.

3.3 Data Acquisition Methodologies

Acquiring data from technical documentation and audiovisual sources requires methodical procedures specific to each type of data.

Video Data Acquisition The tremendous visual and auditory richness that video data offers may considerably enhance user interactions for the chatbot system.

Utilization of OpenAI's Whisper Infrastructure

- The advanced Whisper infrastructure from OpenAI is made to extract insights from video material. It uses sophisticated algorithms to interpret the data encoded in movies, such as sentiment analysis, object identification, and speech recognition.
- Speech recognition software converts spoken speech to text so that the chatbot can comprehend and analyze spoken communication.
- Algorithms for object identification detect visual components in the video, including people, objects, or sceneries, and provide further context to improve the chatbot's comprehension.
- Emotions analysis algorithms examine the tone of the text from an emotional perspective, allowing the chatbot to determine the user's emotions and react appropriately, providing comfort or empathy as needed.

Technical Document Acquisition Technical documents comprise a broad spectrum of items produced by the organization, such as project paperwork, specifications, and manuals.

Diverse Document Formats

- Word documents, Excel spreadsheets, PDF files, and proprietary formats exclusive to certain software or tools are just a few of the formats in which technical documentation can be found.
- The extraction and processing of data is posed with different problems by different document formats. Word documents, for instance, could include structured text, whereas PDF files might have non-searchable text or scanned photos.

Data Extraction Methodologies

- Optical character recognition (OCR), natural language processing (NLP), and text mining techniques are used in conjunction to extract data from technical papers.
- Text mining methods help extract pertinent information by analyzing document content to find important phrases, concepts, and trends.
- To help the chatbot comprehend the information being delivered, natural language processing (NLP) algorithms examine the textual content of documents to extract meaning and context.
- By utilizing OCR technology, non-searchable text found in PDF files or scanned pictures may be transformed into machine-readable text that can be processed and analyzed further.

3.4 Data Preprocessing and Integration

Prior to being incorporated into the chatbot system, data is preprocessed and standardized to ensure consistency and usefulness across various sources. This means standardizing data formats, removing noise and unnecessary information, and compiling data into a unified knowledge base. The integrated data store serves as the chatbot's knowledge base, enabling it to respond to user questions with accuracy and appropriateness for the given context shown on Fig. 1.

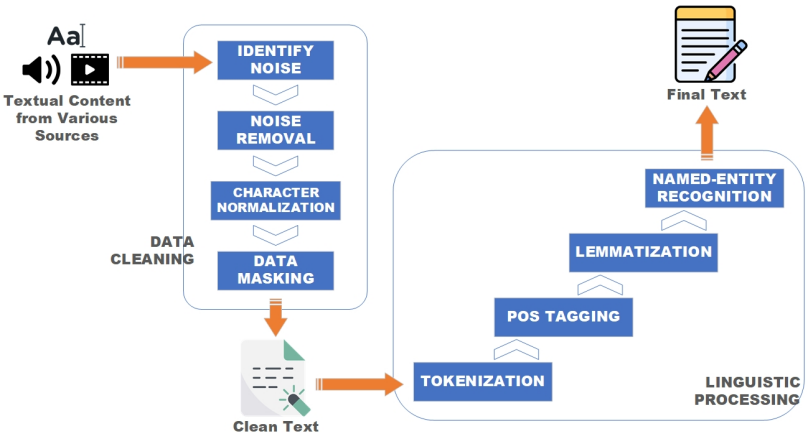

Fig. 1. The pipeline created until the text data is collected from different sources and ready.

To sum up, the data description gives a general overview of the many information sources that our offline chatbot system draws from, along with the methods for gathering, preparing, and integrating data. The chatbot system may provide educated and efficient customer care by using technical documents and videos to obtain a deep and nuanced understanding of organizational expertise.

4 Text Generation Model Description

The offline chatbot system works by combining ready-to-use texts from various sources with the state-of-the-art 7b parameter model of LLama-2 [15], as described in the above sections. This section explores the complexity of integrating models, identifies new approaches used to improve performance, and provides a comprehensive performance analysis comparing the created model to industry standards.

4.1 Model Integration

The fundamental component of the chatbot system is the smooth integration of the LLama-2 7b parameter model with cutting-edge methods designed to im-

prove the accuracy of information retrieval and inference. To be more precise, our model combines two innovative approaches: Probabilistic Inference Layer Integration in LLMs for Accurate Information Retrieval [16] and the Flash Attention [17] technique.

Flash Attention Method Due to its quadratic time and memory complexity, the self-attention mechanism significantly impedes the scaling of the transformer design. Advancements in accelerator technology in recent times have mostly concentrated on increasing computing capacity rather than memory and data transmission across devices [17]. This causes a memory bottleneck in the attention process. An attention approach called Flash Attention is utilized to lessen this issue and scale transformer-based models more effectively, allowing for quicker inference and training.

High Bandwidth Memory (HBM) is used by the standard attention mechanism to store, read, and write keys, queries, and values. While SRAM has less memory but processes information more quickly, HBM has more memory but processes information more slowly. The cost of loading and writing keys, queries, and values from HBM is considerable with the typical attention implementation. It loads the values, queries, and keys from HBM into the GPU's on-chip SRAM, executes one attention step, writes the result back to HBM, and repeats the process for each and every attention step. Rather, Flash Attention loads the keys, queries, and values all at once, combines the attention mechanism's activities, and publishes the results back.

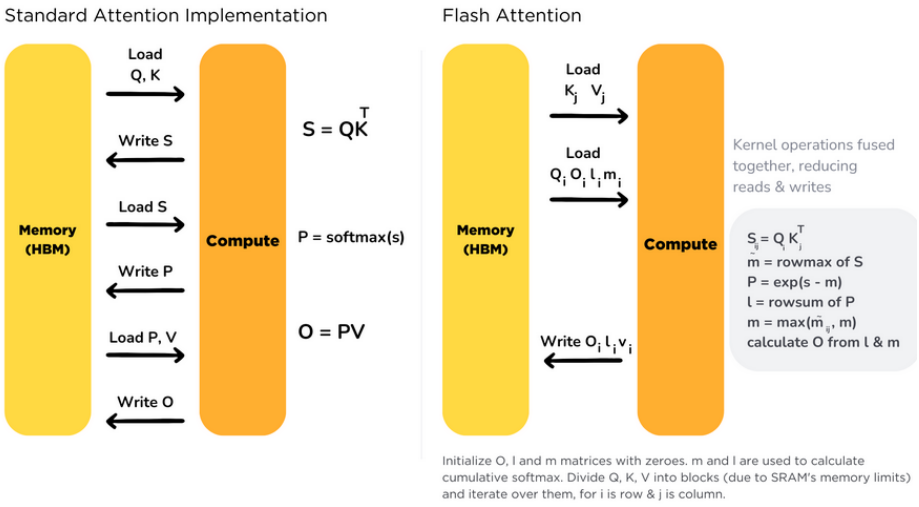

Fig. 2. Flash Attention working principle [17]

The LLama-2 model's attention processes are revolutionized by the Flash Attention technique, making it possible to digest textual input quickly and effectively. This innovative attention method significantly enhances the model's ability to process and understand large volumes of text, leading to faster and more accurate comprehension of user inquiries. By optimizing the attention mechanism, Flash Attention reduces the computational complexity and latency typically associated with traditional attention methods. This improvement enables the chatbot to contextualize user queries with greater speed and precision, ultimately resulting in more accurate and responsive interactions. The integration of Flash Attention not only accelerates the model's processing capabilities but also improves its overall efficiency in handling diverse and complex inquiries. This is particularly beneficial in scenarios where rapid and reliable customer support is essential. The enhanced processing power allows the chatbot to deliver high-quality responses in real-time, even in demanding environments with high query volumes. To illustrate the performance gains achieved through Flash Attention, we present the forward-speedup and Average Inference Time (seconds) graph in the model created using this method. The graph demonstrates the significant reduction in inference time, showcasing the superior speed and efficiency of the Flash Attention-enabled LLama-2 model compared to its predecessors. These advancements underscore the transformative impact of Flash Attention on the LLama-2 model's capabilities, making it a highly effective solution for modern CRM chatbot applications.

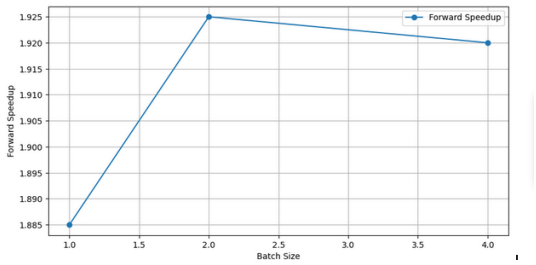

Fig. 3. LLama-2 forward speedup graph while using Flash Attention Method

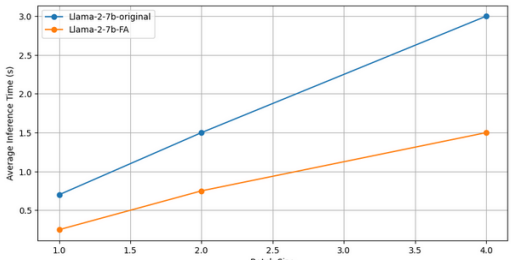

Fig. 4. LLama-2 inference time graph while using Flash Attention Method

Probabilistic Inference Layer Integration In order to solve the crucial problem of accurate and dependable information retrieval in natural language processing, this section provides an innovative integration of the Probabilistic Inference Layer (PIL) into the Llama-2 Large Language Model (LLM). Significant gains in bias reduction, context understanding, and knowledge retrieval accuracy are demonstrated by the upgraded Llama-2 Master's PIL[7]. By utilizing Bayesian networks, intricate mathematical structures like matrix calculus, and concepts akin to Bernoulli and Lorentz transformations, PIL's model enhances its accuracy and dependability in processing data. The study's findings demonstrated that the model performed better on a number of tasks, particularly when it came to separating context from purpose, minimizing biases, and managing intricate logical processes. Even with its high computing requirements and the continuous difficulty in totally removing biases, the PIL integration sets a new standard for LLMs and creates opportunities for more study. By highlighting the potential of probabilistic approaches to improve the capabilities of generative AI models, this work advances the discipline and opens the door to the development of more complex and dependable AI-driven information systems. Below Fig. 5. depitcs the flow of this structure.

Information about the structure here is listed below.

1. **Data Preprocessing:** Input data is preprocessed to extract important components and context, which are necessary for the PIL to evaluate information correctly.
2. **Probabilistic Assessment:** Following processing, the PIL analyzes the incoming data and applies Bayesian networks to determine the probability of accurate information.
3. **Feedback Integration:** By integrating the feedback from the generic LLama-2 LLM with the computed probabilities, the model's final output is better contextually relevant and accurate.
4. **Output Generation:** The LLama-2 LLM develops its answer using the integrated probabilistic evaluations, prioritizing the data that is judged to be the most correct.

This improved integration is shown in Figure 5, which also highlights the bidirectional flow and the PIL's critical function in enhancing the capabilities of the generic Llama-2 LLM. Comparisons of the new model structure created with the base model in terms of "Accuracy Improvement in Retrieving Factual Information", "Enhanced Ability to Discern Context and Intent" and "Reduction in Biased or Outdated Information" are shown in the Fig. 6.

5 Results

To evaluate the effectiveness of the model, a comprehensive performance test was conducted to compare the built model with three existing benchmarks: ChatGPT-3.5, Mistral-7b, and the original LLama-2 7b LLM models. Word Error Rate (WER), BLEU score, and inference time are examples of evaluation

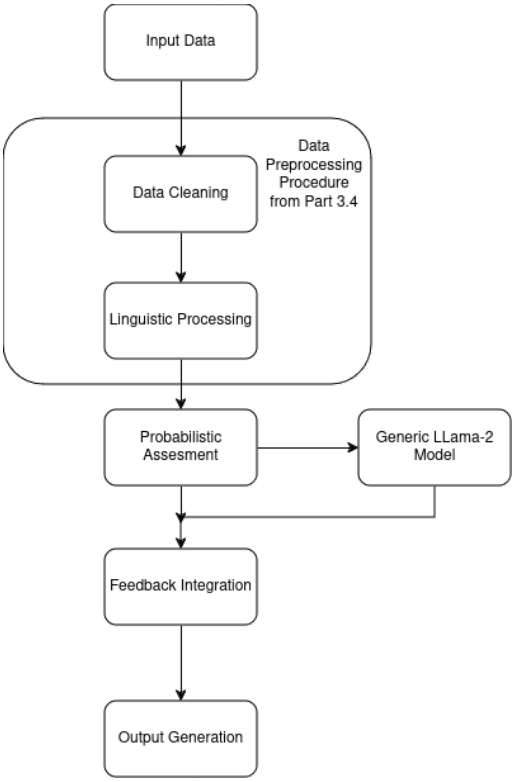

Fig. 5. Integration process of PIL into LLama-2

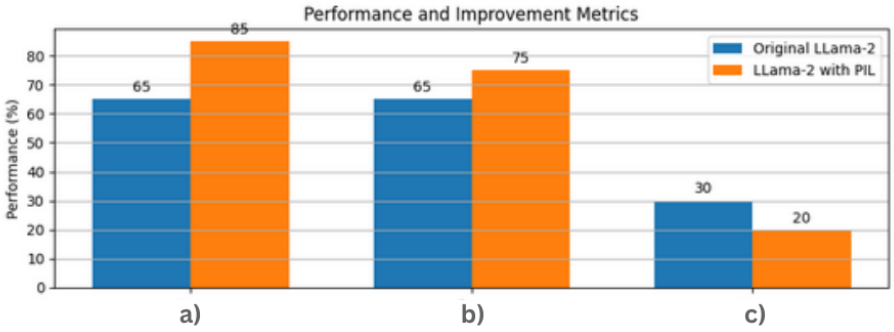

Fig. 6. Performance and improvement metrics of original LLama-2 and LLama-2 with PIL a)Accuracy Improvement in Retrieving Factual Information b) Enhanced Ability to Discem Context and Intent c) Reduction in Biased or Outdated Information)

measures. The tests performed here were performed on the corpus created with the data we obtained previously.

The table shows that our newly designed chatbot model performs better than existing 7b parameter models in terms of WER, BLEU score, and inference time. It retains superior accuracy and responsiveness while processing user requests and generating answers more quickly. The higher BLEU score of the model, which denotes a higher degree of similarity between its responses and the reference texts and, consequently, higher-quality replies, clearly shows how superior it is. Moreover, its WER is lower, suggesting fewer mispredicted words and improved response production accuracy. Our overall goal is to deliver the fastest feasible inference time while still offering the most precise and responsive chatbot experience.

Table 1. Comprehensive performance test results

Model	Inference Time (sn)	BLEU Score	WER Score
ChatGPT-3.5	3.5	91.8	10.2
Mistral 7-b	5.1	85.5	12.5
LLama-2 7b (Original)	4.8	86.4	11.8
LLama-2-based Proposed Model	**4.1**	**90.2**	**9.7**

ChatGPT-3.5, a SaaS solution with 175 billion parameters, demonstrates the best overall performance with the lowest inference time (3.5 seconds) and the highest BLEU score (91.8), although it has a slightly higher WER score (10.2) compared to our proposed model. Our LLama-2-based proposed model stands out among the on-premises solutions. With an inference time of 4.1 seconds, it is faster than both the Mistral 7b (5.1 seconds) and the original LLama-2 7b (4.8 seconds). Additionally, it achieves a BLEU score of 90.2, which is significantly higher than the other on-premises models, and closer to ChatGPT-3.5. Moreover, our model has the lowest WER score (9.7), indicating superior precision and quality in handling user inquiries compared to the other models.

6 Ethical Considerations

Ensuring that this produced technology serves the best interests of society and limiting possible harm requires us to give the highest priority to ethical issues when starting to develop and implement our chatbot project. Important ethical issues to consider at each stage of the project lifecycle include:

6.1 Data Privacy and Consent

Compliance to ethical data gathering norms is necessary to ensure the safeguarding of user confidentiality and privacy. Make obtaining users' informed consent a top priority before gathering any sensitive or personal data. Ensure there is unambiguous transparency regarding the utilization, retention, and dissemination of data. Furthermore, offer clients the capacity to control or choose not to utilize their data.

6.2 Bias Mitigation and Fairness

Ensure thorough scrutiny of biases in algorithms and data to ensure equitable treatment of all users. Implement measures to mitigate biases during data collection, data preprocessing, and model training. Regularly monitor the chatbot's responses for unintentional biases, particularly those related to gender, race, ethnicity, and other sensitive attributes.

6.3 Transparency and Accountability

Encourage openness by informing users and stakeholders about the chatbot's capabilities, constraints, and potential biases. Provide procedures for accountability, such as distinct responsibilities for handling mistakes, prejudices, or ethical transgressions. Offer channels for consumers to voice their concerns and seek redress if their interactions with chatbots have a negative effect.

6.4 User Well-being and Safety

Put user safety and well-being first by building the chatbot to deliver accurate, dependable, and useful information while preventing the spread of dangerous or deceptive content. Put precautions in place to stop users from acting dangerously or harmfully in response to chatbot suggestions. Give consumers going through difficult times or crisis situations access to interventions or support services.

6.5 Inclusivity and Accessibility

Make sure that everyone who uses the chatbot may access it, regardless of their language preference, skill level, or level of technology competence. Create interactions and user interfaces that are simple to use, intuitive, and compatible with assistive technology. Take into account the various requirements of user groups and work to ensure that chatbot services are accessible to all.

6.6 Continuous Monitoring and Improvement

Throughout the chatbot's career, pledge to continuously monitor and assess its effectiveness, impact, and ethical implications. Evaluate the chatbot on a regular basis for accuracy, efficacy, and user happiness. Iteratively improve the model and algorithms to fix any flaws or ethical issues that come up.

It can be ensured that the chatbot project complies with ethical standards, respects user rights and welfare, and improves the general welfare of society by taking these ethical issues into account. Transparency, fairness, privacy, participation, and responsible deployment are key components of a responsible technology deployment strategy that will optimize its benefits while mitigating the dangers.

7 Conclusion

The rapid progress of language models (LLMs) has completely transformed customer service, providing unparalleled possibilities to surpass the constraints of traditional online chatbots. Our study enhances the autonomy and adaptability of LLM-based solutions in providing customer assistance by using advanced AI algorithms and integrating data from various sources, such as social media and project documents. This contributes to the evolution of LLM-based solutions across different connectivity environments.

Our findings confirm that our offline chatbot system effectively tackles the difficulties of intermittent internet access, demonstrating better performance metrics than current benchmarks. The successful integration of Flash Attention technique and Probabilistic Inference into our chatbot model has significantly enhanced its performance, culminating in impressive results. These methods have notably improved the efficiency and accuracy of our CRM chatbot, showcasing their efficacy in advancing customer support systems. Our model showcases exceptional efficiency, accuracy, and responsiveness, as evidenced by its amazing 4.1-second inference time, a BLEU score of 90.2, and a WER score of 9.7. The results highlight the significant impact of our strategy in making customer service processes more efficient and improving user happiness.

The future of AI integration in customer care shows great potential for continued innovation and improvement. In light of our successful implementation, future research endeavors could focus on investigating sophisticated methodologies to augment the adaptability and scalability of offline chatbot systems. Furthermore, it is essential to make ongoing improvements to language models and broaden their functionalities in order to satisfy the changing requirements of the dynamic customer service sector. This will ultimately lead to the development of more robust and efficient customer support solutions. The chatbot system presented in this article is suitable for challenging scenarios that necessitate advanced data integration and offline capabilities to ensure uninterrupted assistance and optimal performance in situations when connections are variable [18].

References

[1] Mark Anthony Camilleri and Ciro Troise. “Live support by chatbots with artificial intelligence: A future research agenda”. In: *Service Business* 17.1 (2023), pp. 61–80.

[2] Dipanjan Saha et al. “Thinking like a machine: Alan Turing, computation and the praxeological foundations of AI”. In: *Science & Technology Studies* (2023).

[3] Robert Ciesla. “The Challenge of the Turing Test”. In: *The Book of Chatbots: From ELIZA to ChatGPT*. Springer, 2024, pp. 1–9.

[4] Bei Luo et al. “A critical review of state-of-the-art chatbot designs and applications”. In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 12.1 (2022), e1434.

[5] Eleni Adamopoulou and Lefteris Moussiades. "Chatbots: History, technology, and applications". In: *Machine Learning with applications* 2 (2020), p. 100006.
[6] Manik Rakhra et al. "E-commerce assistance with a smart chatbot using artificial intelligence". In: *2021 2nd International Conference on Intelligent Engineering and Management (ICIEM)*. IEEE. 2021, pp. 144–148.
[7] Guruswami Hiremath et al. "Chatbot for education system". In: *International Journal of Advance Research, Ideas and Innovations in Technology* 4.3 (2018), pp. 37–43.
[8] Rahim Sadigov et al. "Deep learning-based user experience evaluation in distance learning". In: *Cluster Computing* 27.1 (2024), pp. 443–455.
[9] Sezer Toprak et al. "LWE: An energy-efficient lightweight encryption algorithm for medical sensors and IoT devices". In: (2020).
[10] Akhan Akbulut et al. "Agent based pornography filtering system". In: *2012 International Symposium on Innovations in Intelligent Systems and Applications*. IEEE. 2012, pp. 1–5.
[11] Reenu Mohan. "The Chat bot revolution and the Indian HR Professionals". In: *International journal of information and computing science* 6.3 (2019), pp. 489–499.
[12] Soumi Majumder and Atreyee Mondal. "Are chatbots really useful for human resource management?" In: *International Journal of Speech Technology* 24.4 (2021), pp. 969–977.
[13] Josh Hanna et al. "HMM based classification of sports videos using color feature". In: *2012 6th IEEE International Conference Intelligent Systems*. IEEE. 2012, pp. 388–390.
[14] Alec Radford et al. "Robust speech recognition via large-scale weak supervision". In: *International Conference on Machine Learning*. PMLR. 2023, pp. 28492–28518.
[15] Hugo Touvron et al. "Llama 2: Open foundation and fine-tuned chat models". In: *arXiv preprint arXiv:2307.09288* (2023).
[16] Bing Wang, Shiyu Wang, and Qian Ouyang. "Probabilistic Inference Layer Integration in Mistral LLM for Accurate Information Retrieval". In: (2024).
[17] Tri Dao et al. "Flashattention: Fast and memory-efficient exact attention with io-awareness". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 16344–16359.
[18] Akhan Akbulut et al. "Wireless sensor networks for space and solar-system missions". In: *Proceedings of 5th International Conference on Recent Advances in Space Technologies-RAST2011*. IEEE. 2011, pp. 616–618.