J. Electrical Systems 20-2 (2024): 2558-2567

[1]Manasa S A

[2]Dr. Pankaj Agarwal

[3]Goutam Pal

[4]P. Mahendra

[5]Shankar Rao Pendyala

[6]Sachin Marge

Towards Seamless Air Travel: Developing a Next-Generation Flight Booking Assistant

Abstract: - In the dynamic landscape of flight booking, passengers seek an optimal and streamlined experience encompassing factors such as flight duration, onboard services, and pricing. This paper introduces developing a Next-Generation Flight Booking Assistant, leveraging Meta's LLAMA 2, a transformer-based auto-regressive causal language model. Our approach not only provides a comprehensive solution to the traditional challenges of flight booking but introduces several novel features. We fine-tune the LLAMA 2 model using the Low-Rank Adaptation (LoRA) technique, enabling efficient handling of user queries. Furthermore, we introduce a destination recommendation model based on travel history, employing the Retrieval-Augmented Generator (RAG) for data retrieval on booking and pricing information. The Next-Generation Flight Booking Assistant incorporates distinctive features, including real-time interpretation of social media sentiment to grasp user preferences, seam- less handling of multiple languages, and the provision of post-booking recommendations for top places to visit. These distinguishing aspects set our assistant apart by actively adapting to the evolving needs of users. To justify the significance of our solution, we present comprehensive performance statistics derived from user testing and feedback collection. The evaluation underscores the assistant's proficiency in understanding user preferences, navigating through available options, and efficiently completing the booking process. The results demonstrate not only the effectiveness but also the user satisfaction with the introduced features, solidifying the Next-Generation Flight Booking Assistant as a cutting-edge and user-centric solution in the realm of digital flight assistance.

Keywords: LLAMA 2, RAG, LoRA, LLM

I. INTRODUCTION

In the swiftly evolving landscape of technological advancements, digital assistants have emerged as transformative agents, reshaping the dynamics of human-computer interaction. As these intelligent virtual entities become increasingly integrated into our daily lives, their multifaceted roles extend beyond mere task automation, reaching into realms of personalized assistance, information retrieval, and seamless connectivity IpKin et al.[1]. In the contemporary travel landscape, most travelers book their flights online, where they input their details and swift through a long list of options to select the flight that aligns with their preferences. Therefore, recommending flights to users is becoming interesting and important as they offer a time- saving solution Jian et al.[3].

[1] Department of Data Science and Engineering, Birla Institute of Technology and Science, Pilani, India

*Corresponding author

Email: manasa.amar95@gmail.com

[2] Visiting Professor, Department of Data Science and Engineering, Birla Institute of Technology and Science, Pilani, Indi

Email: pankaj.a@wilp.bits-pilani.ac.in

[3] Department of Data Science and Engineering, Birla Institute of Technology and Science, Pilani, India

Email: pal.goutam88@gmail.com

[4] Department of Data Science and Engineering, Birla Institute of Technology and Science, Pilani, India

Email: mahipalukuri@gmail.com

[5] Department of Data Science and Engineering, Birla Institute of Technology and Science, Pilani, India

Email: shankarpendyala@gmail.com

[6] Department of Data Science and Engineering, Birla Institute of Technology and Science, Pilani, India

Email: marge.sachin@gmail.com

To understand how a digital assistant could help in this process, a survey was conducted across different groups, checking what people really need and want when booking flights on- line Hakseung et al.[9] Among the crucial factors for users, price and flight duration stand out as the most vital considerations. Imagine the simplicity if users could express their preferences through a chat, and a digital assistant could comprehensively understand and provide tailored recommendations based on these preferences Zelin et al. [4].

This streamlined process not only saves time but also ensures that users receive personalized flight options aligned with their priorities Cathy et al. [7]. In the digital age, people often check social media to decide on airlines. They read discussions and reviews to make choices. Digital assistants help users find the right reviews, giving airlines a chance to improve based on feedback mitigating potential risk Joyjit et al. [6]. Both travelers and airlines can gain, building a better connection between them through Digital assistants. Existing chat bots are often traditional and rule-based, struggling with handling multiple user queries and providing re- sponsors beyond predefined rules. Moreover, these chat bots face challenges in addressing users in various languages, limiting their effectiveness. Digital assistants, powered by large language models, offer a solution by seamlessly handling di- verse queries and accommodating multiple languages Alec et al.; Pei-Fang et al. [10, 11). Unlike rule-based systems, they can adapt and respond intelligently to a wide range of user inputs, enhancing the user experience and expanding the capabilities of conversational interfaces Junbo et al. [8].

Llama2 is a powerful and open-source large language model (LLM) developed by Meta

AI Hugo et al. [2]. Llama2 represents a significant advancement in LLM technology, offering a powerful and cost-effective solution for various applications. Its open-source nature further expands its potential impact on the future of AI development. It is offered with seven billion (7B), 13 billion (13B) or 70 billion parameters (70B). The Llama 2 model is flexible, allowing fine- tuning to meet specific requirements. This adaptability enables the creation of a customized digital assistant tailored to our needs. Leveraging Llama 2 as a foundation, we can craft an efficient and personalized digital assistant. This paper discusses the design and model of a digital assistant designed for flight booking. The assistant not only offers personalized recommendations but also assists users in finding the most suitable flights based on their preferred time, destinations, and budget. Additionally, it incorporates social media sentiment analysis to help users better understand the experiences and preferences shared by others. This paper introduces a collaborative filtering method leveraging user travel history to generate personalized flight recommendations. To enhance the performance of LLAMA 2 without the need for extensive retraining, the Low-Rank Adaptation (LoRA) technique is employed E Edward J et al. [14]. Additionally, Retrieval-Augmented Generator (RAG) techniques are incorporated to efficiently retrieve information related to booking, pricing, and seating.

II. Methods

Large Language Models (LLMs), such as ChatGPT, LLAMA, PALM, FALCON, Have paved the way for the emergence of digital assistants and bots. Llama 2, an open-source LLM developed by Meta, trained on a substantial parameter range from 7 billion to 70 billion, stands out as a capable option for serving as a digital assistant. Llama 2 allows fine- tuning which can potentially enhance user interactions, provide intelligent responses, and

offer a versatile and powerful solution for various conversational applications. The beauty of fine-tuning lies in its ability to customize Large Language Models (LLMs) Llama 2 ac- cording to specific domain requirements, eliminating the need for extensive retraining. This not only saves costs but also conserves valuable computing power. Fine-tuning allows these models to be tailored to specific contexts, making them more efficient and practical for a wide range of applications, while still benefiting from the preexisting knowledge and capabilities of the base model Long et al. [13]. LLama2 stands out due to its innovative features, with a key differentiator being the Grouped Query Attention (GQA) mechanism, enhancing scalability and efficiency. It utilizes a standard transformer architecture with a unique autoregressive model, employing pre-normalization using RMSNorm to streamline computation. The model incorporates SWIGLU activation, a fusion of Swish and GLU, deviating from the conventional ReLU. Further, LLama2 adopts rotary positional embedding and introduces the concept of grouped-query attention, setting it apart with a comprehensive set of techniques aimed at improving performance and functionality which can be utilized for dialogue conversations in digital assistants or chat bots.

The process employed to achieve the full capabilities of the digital assistant is outlined below, with the primary algorithmic structure illustrated in Fig. 1

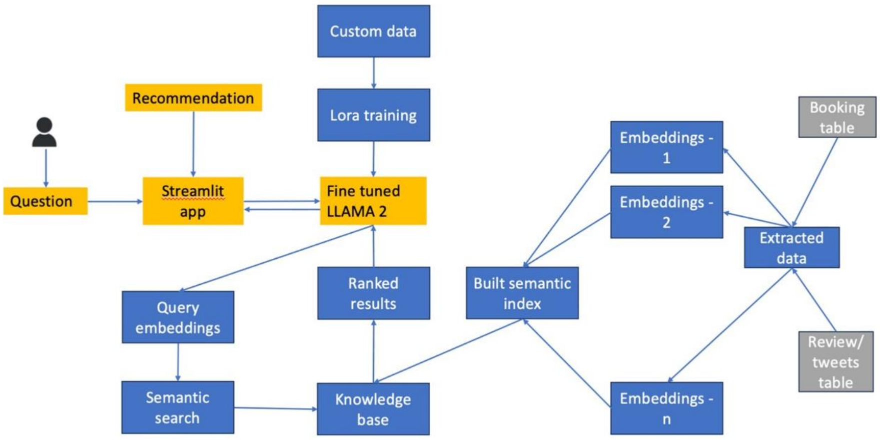

Figure 1: The Fine-tuned LLAMA 2 based mode

A. Exploratory Case Study

A comprehensive case study was undertaken, exploring how individuals interact with digital assistants during flight bookings. The study aimed to gain insights into user preferences, identify challenges, and gauge satisfaction levels. The findings from the case study helped to provide valuable input for enhancing digital assistants in the travel industry, ultimately contributing to an improved user experience and ease.

The case study revealed significant factors influencing people's flight bookings, with price, booking convenience, and flight duration emerging as major considerations observed in Fig. 2, Fig. 3 demonstrates that traditional chat bots were found to be less convenient for booking, highlighting the demand for more flexible and reliable digital assistants. Further- more, the study observed a preference among travelers for voice and type-based assistants in Fig. 4, indicating a strong inclination towards modern methods. The study also revealed a clear preference among individuals for personalized recommendations in the context of flight bookings which is observed in Fig. 5. These insights underscore the importance of adapting digital assistant technologies to align with user preferences and improve the efficacy of flight booking processes.

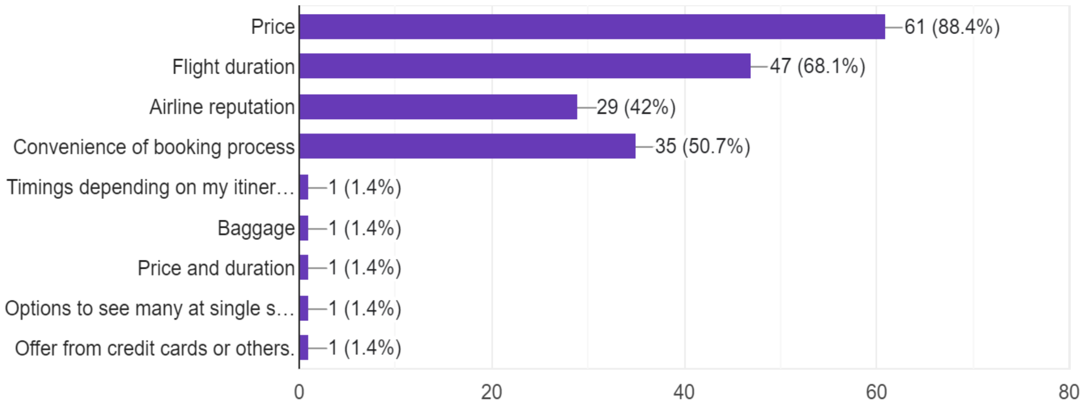

Figure 2: Factors for booking

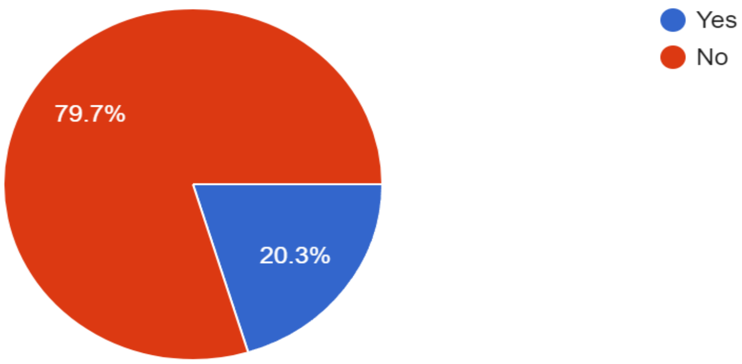

Figure 3: Traditional chatbots usage

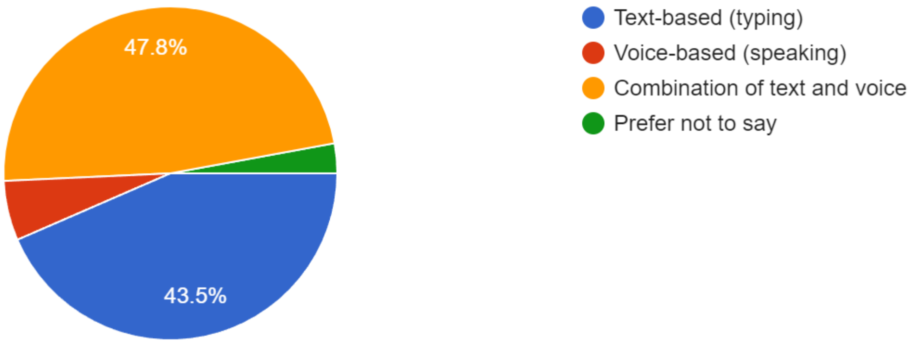

Figure 4: Interaction preference with Digital assistants

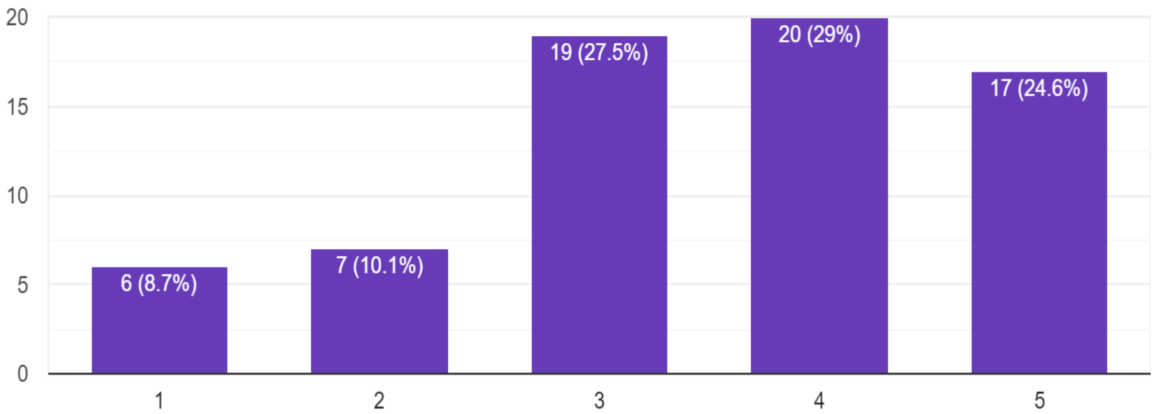

Figure 5: Recommendation Preferences

B. Data Preparation and Pre-processing

Historical data on flight bookings was gathered from SAP tables to analyze user travel patterns, followed by data cleaning, and preprocessing to eliminate duplicates and null values. Only Indian flight destinations are used for this dissertation and dataset includes 10 locations. Exploratory data analysis was performed on the data, and it is found that top origin and destination preference for passengers was Mumbai. Feature selection focuses on crucial elements such as Passenger ID, Age, date, origin, and destination, with a particular emphasis on Origin and Destination as key features for recommendations. Feature engineering was performed on the attributes Origin and Destination to combine and extract valuable information.

Data for fine-tuning LLAMA 2 proved to be a crucial step in establishing a coherent and effective dialogue flow between users. Understanding the dialogue flow is crucial for the effective implementation of chatbots Michael et al.; Amelia et al. [5, 12). Around 500 records of dialogue flow conversation dataset were prepared in the format as shown in Fig. 6 to train the LLAM 2 model. Furthermore, it included German dialogues to ensure it comprehensively understands context and can adaptively handle multiple languages. These dialogues were sourced from user interactions and survey that was conducted survey which aimed at comprehending patterns and interactions. Additionally, exploration of various travel blogs and websites provided itinerary data, enriching the model's training dataset.

Lastly flight booking and pricing data was gathered from SAP tables to provide users a real-time view of the prices and availability. Dataset from Kaggle[15] for sentiment analysis that included mock tweets for flight booking and services was used for social media sentiment analysis.

```
<s>[INST] <<SYS>>
System prompt
<</SYS>>

User prompt [/INST] Model answer </s>
```

Figure 6: LLAMA 2 template for fine tuning

C. LLAMA 2 Model Fine tuning

In the rapidly evolving field of Generative AI, fine-tuning large language models like LLAMA 2 presents unique challenges due to the computational and memory demands of the workload. However, the newly enabled Low-Rank Adaptations (LoRA) technique present a powerful option for tuning state-of-the-art (SoTA) LLMs faster and at reduced costs. At its core, the theory behind LoRA revolves around matrix factorization and the principle of low-rank approximations. In linear algebra, any given matrix can be decomposed into several matrices of lower rank. In the context of neural networks, this decomposition can be viewed as breaking down dense, highly parameterized layers into simpler, compact structures without significant loss of information. By doing so, LoRA aims to capture a model’s most influential parameters while discarding the extraneous ones.

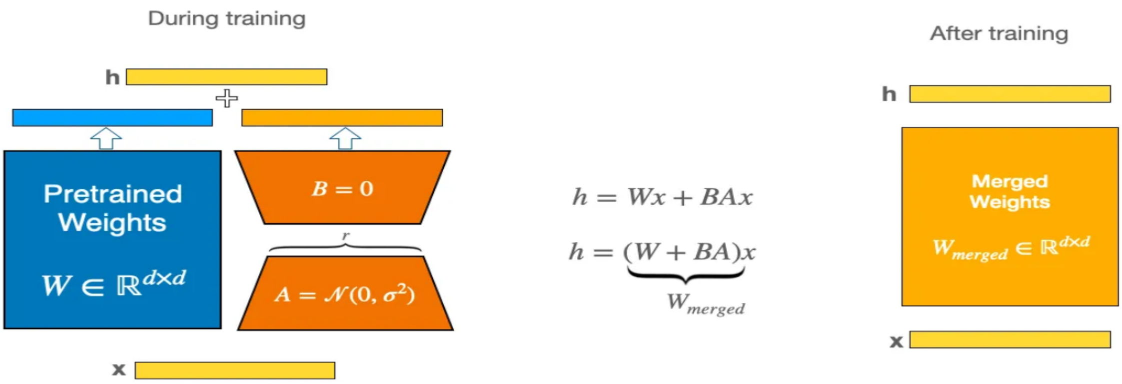

Figure 7: LoRA during and after training

High-dimensional data processed by these models often resides in lower-dimensional subspaces, allowing LoRA to create an efficient subspace for the neural network's parameters. By introducing task-specific parameters with constrained dimensionality through low-rank matrices, LoRA enables rapid adaptability to new tasks and computational efficiency, gaining new knowledge without retraining the entire parameter space.

Google Colab provides a 15GB Graphics Card with limited resources, sufficient to store LLAMA 2–7b's weights. Here the original LLAMA 2 model of Meta was accessed through hugging face which was then train and customize as per domain needs to assist passengers for flight bookings. Fine-tuning with Parameter-efficient fine-tuning (PEFT) techniques addresses RAM constraints by reducing VRAM usage, utilizing 4-bit precision with a rank of 64 and a scaling parameter of 16. The LLAMA 2 model is loaded directly in 4-bit precision using the NF4 type, trained for two epochs. Merging LoRA weights with the base model involves reloading the base model in FP16 precision and utilizing the PEFT library for integration. Despite the computational intensity and time-consuming nature of this fine-tuning process, it allows customization to meet the specific needs of assisting passengers in flight bookings.

The fine-tuned model is seamlessly implemented and accessed by loading it similar to any other Llama 2 model from the model hub, ensuring easy accessibility and integration into various applications. Utilizing Streamlit, a Python library for web application development, the front-end user interface for the flight assistant is constructed. Concurrently, Langchain, a conversational modeling tool, manages interactions in conversations involving the human user, the system, and the AI-powered assistant. The assistant is linked to an SAP table for real-time booking data. When users inquire about discounts or prices, retrieval techniques are employed to fetch results, which are then sent to the LLAMA 2 model for response generation, providing comprehensive and timely information to the user.

D. Recommendation of flight destinations

User-Based Collaborative Filtering was used for recommending origin-destination based on preferences of similar users. K-NN Model was used to find similar users, a common technique for collaborative filtering. Cold Start Handling to provide general recommendations for new users.

Recommendation Logic was to recommend origin - destinations that similar users have visited but the target user has not yet explored. User item matrix was created with the key features Passenger ID and new attribute which was a result of feature engineering origin and destination. Based on this matrix users are grouped together using KNN. Cosine similarity is used to find the similarity among the users. k-Nearest Neighbors (k-NN) is used here to find similar users, not to create clusters. New users are handled by providing general recommendations which included the top visited Origin-Destination pair. For existing users KNN is used to find similar booking patterns, KNN is used to locate the n+1 nearest neighbors and where-in the user himself is excluded. Here unique Origins-Destinations pairs visited by the similar users are identified and also ensure novelty in recommendations.

E. Real time Data Retrieval

Retrieval-Augmented Generation (RAG) is a technique that uses contextual information from external sources to improve a large language model's response.

In this research RAG is used to fetch answers from data tables. The first step is to load booking, pricing data from SAP tables and tweets data from Kaggle dataset to vector storage such as FAISS through an embedding model which converts table data to a vector form. Data from these tables are split and chunked which are then fed to Sentence transformer models from hugging face to generate embeddings. These embeddings are stored in FAISS vector store. Therefore, RAG technique here is used to serve and handle queries related to fetching real-time data about booking, pricing, seats check-in, discounts and understanding social media sentiment.

Intent is gathered from the user prompt which is queried on vector database. Top k results are retrieved from the database which is then sent to LLAMA 2 for generating the response, Fig. 8 demonstrates the flow of RAG techniques.

Figure 8: Retrieval Augmented Generation on Table data

III. Results

The proposed model contains fine-tuned LLAMA 2 model, recommendation model and Data retrieval model. Firstly, the LLAMA 2 model was fine-tuned on the gathered dataset using Supervised Fine-Tuning (SFT) parameters for two epochs. Encouragingly, a consistent decrease in loss was observed across the epochs, indicating positive progress in model refinement. Lower loss values indicate that the model's predictions are becoming closer to the actual values in the training set. We decided to stop the training at this stage since we noticed that Training loss is almost stable, Fig. 9 shows the details of training loss for two epochs, Fig. 10 indicates the graph where training loss is almost stable. We decided to stop at this point because we wanted to retain some parameters from the existing LLAMA 2 model as well. The evaluation process is crucial during fine-tuning, especially for Large Language Models (LLMs), where performance extends beyond mere accuracy to encompass the overall value of generated text. Conventional metrics like loss or validation scores lack insightful assessments, and metrics like perplexity and accuracy may fall short in capturing the entirety of model performance. Recognizing this challenge, task-specific evaluation matrix Recall-Oriented Understudy for Gisting Evaluation (ROUGE) was used. ROUGE is a metric used to assess the quality of generated text by measuring recall, precision, and F1-score based on shared n-grams (subsequences of words) between the generated content and reference texts. In the evaluation process, a series of questions were posed, and the responses produced by a finetuned LLAMA 2 model were compared with reference texts. The ROUGE score was then calculated by analyzing the overlap and similarity of n-grams, providing a quantitative measure of the model's performance in generating relevant and accurate content. This was computed for a set of questions and average ROUGE-L was found to be around 0.7 – 0.8. While ROUGE scores provide quantitative insights into the model's performance based on text overlap, human evaluation remains paramount in assessing contextual understanding. Comparing generated text against human judgments ensures analysis, considering factors beyond n-gram matching. This comprehensive approach captures the model's ability to grasp context, coherence, and relevance, providing invaluable insights into its real-world applicability and user-centric effectiveness. Therefore, integrating human evaluation alongside ROUGE scores ensures a holistic assessment, enhancing the reliability and validity of the evaluation process. During the process it was found that 90% of the times the context was right yielding promising results of fine-tuning.

$$\textit{ROUGE-L(summary, reference)} = (\Sigma\ L(LCS(summary\_i, reference))) / \Sigma\ |reference\_i|$$

To evaluate the recommendation model we used Hit Rate. The top-5 recommendation for users are provided. If one of the recommendations in a user's top-end recommendations is something they actually chose, then it’s considered a hit. Since the model managed to show the user something that they found interesting place to visit, and they chose it. So, to calculate we add up all the hits in the top-5 recommendations for every user and divide it by every user.

$$\textit{Hit Rate} = (hits / user)$$

A Hit Rate signifies that the model consistently aligns with user preferences, enhancing user satisfaction and engagement. This metric underscores the practical utility of the system, as it reflects the model's success in presenting appealing options within the initial recommendations.

Finally, the whole chat-model evaluation was done by collecting few common dialogue conversations that users usually use to converse with digital assistants while booking. It was noticed that the model was able to correctly understand and get the context of the conversations for 70% of the times sometimes the model would go into loop of booking process which could be improved by augment training data with diverse conversation samples, covering various scenarios and user intents. Additionally, we could incorporate user feedback mechanisms to iteratively refine the model's responses, ensuring continuous improvement and adaptability to user preferences and conversations.

We compared the Finetuned LLAMA 2 model with baseline LLAMA 2 model. The performance disparity between the baseline LLAMA 2 model and its finetuned counterpart was evident in dialogue conversations. The unmodified model exhibited an 80% context misunderstanding rate, often generating invalid responses or even code snippets. In contrast, the finetuned Llama 2 model showcased improved performance, demonstrating better contextual understanding, and consistently delivering more accurate and contextually relevant responses. This comparison highlights the significant enhancement achieved through finetuning, underscoring its effectiveness in refining the model's ability to comprehend and respond appropriately in dialogue scenarios.

Step	Training Loss
25	2.665300
50	1.603100
75	1.371300
100	1.216500
125	1.067800
150	1.095000
175	1.015500
200	1.035900

Figure 9: Training Loss for 2 epochs

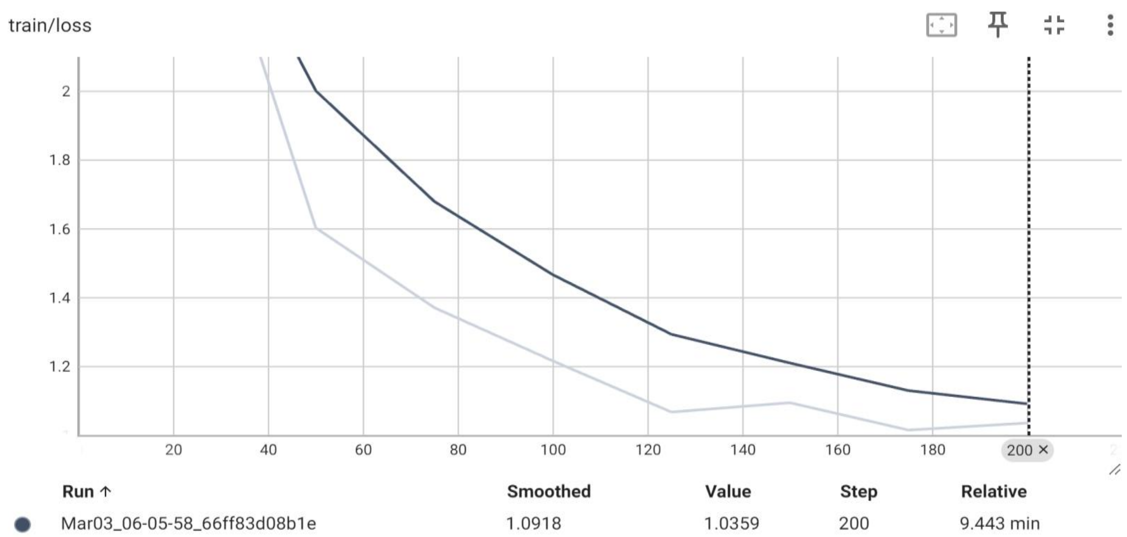

Figure 10: Graphical representation of training loss

IV. Discussion

Sometimes the model hallucinates which can be further handled by introducing feedback mechanism and retraining the model with more wide range of dataset. Future work could focus on implementing a robust feedback mechanism to systematically gather user corrections and preferences. This data-driven approach can enable continuous model improvement, fine-tuning responses based on real-world user interactions. Additionally, exploring advanced techniques like reinforcement learning and adversarial training may enhance the model's ability to handle ambiguous or unseen scenarios. Moreover, investigating interpretability and explain ability in the model's decision-making process would contribute to building an user -centric conversational AI systems. Future work could explore integrating voice chats, enhancing conversational AI with natural language processing in spoken interactions. Optimizing voice recognition, understanding accents, and enabling smooth transitions between text and voice inputs would contribute to a more versatile and user-friendly conversational experience

V. Conclusion

The survey findings suggest that existing chatbots are not extensively utilized, presenting an opportunity for digital assistants to establish a presence in the market. Notably, in Fig. 11 we can observe that around 70% of respondents express interest in exploring digital assistants for booking purposes underscores a significant demand and potential market need for such innovative solutions.

The proposed Llama Flight Booking Assistant based on LLAMA 2 transformer-based model, offers an innovative solution for simplifying flight bookings. Integrating a travel destination-based recommendation model, along with Retrieval-Augmented Generator (RAG) and fine-tuning via Low-Rank Adaptation (LoRA), enhances user personalization and efficiency also handles in addressing users with multiple languages.

Furthermore, these Flight booking assistants can be improved on user-friendliness by implementing a feedback mechanism. This not only allows users to provide input but also facilitates model retraining, enabling control over potential hallucinations and improving overall system accuracy and reliability.

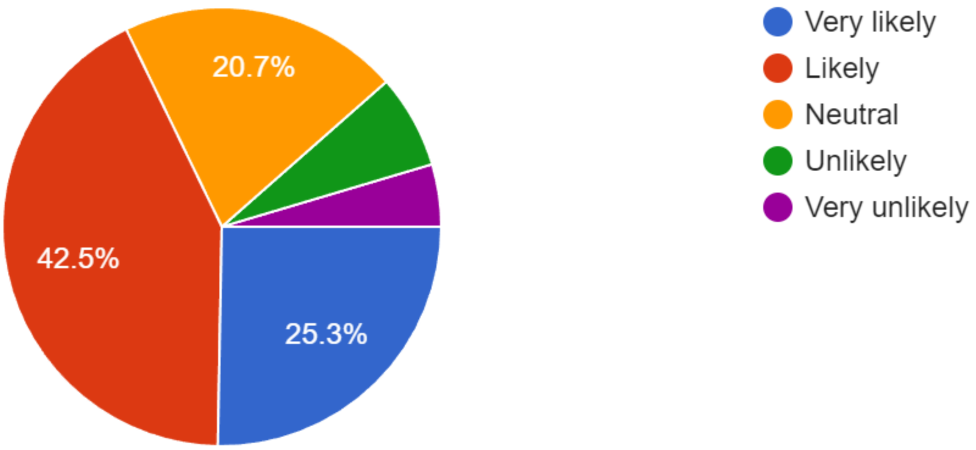

Figure 11: Explore Digital Assistants for booking

REFERENCES

[1] IpKin Anthony Wong, Qi Lilith Lian, Danni Sun, "Autonomous travel decision-making", Journal of Hospitality and Tourism Management, vol. 56 :253-263, 2023.

[2] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, Guillaume Lample, "LLaMA: Open and Efficient Foundation Language Models", Meta, arXiv pp. 2302.13971v1, 2023.

[3] Jian Cao; Yuchang Xu; Huisi Ou; Yudong Tan Quanwu Xiao, "PFS: A Personalized Flight Recommendation Service via Cross-Domain Triadic Factorization", IEEE International Conference on Web Services (ICWS), pp. 10.1109/ICWS.2018.00039, 2018.

[4] Zelin Liu, Jian Cao, Yudong Tan, Quanwu Xiao & Mukesh Prasad, "Planning Above the API Clouds Before Flying Above the Clouds: A Real-Time Personalized Air Travel Planning Approach", International Journal of Parallel Programming, vol. 48:137–156 2019.

[5] Michael Free, Andrew Langworthy, Mary Dimitropoulaki and Simon Thompson, "Towards Goal-Oriented Agents for Evolving Problems Observed via Conversation", Artificial Intelligence XL, vol. 14381:142-155, 2024.

[6] Joyjit Chatterjee, Nina Dethlefs, "This new conversational AI model can be your friend, philosopher, and guide ... and even your worst enemy", Patterns, vol. 4(4): 100731, 2023.

[7] Cathy H.C. Hsu, Guoxiong Tan, Bela Stantic, "A fine-tuned tourism-specific generative AI concept", Annals of Tourism Research, vol. 104: 103723, 2024.

[8] Junbo Zhang, Qi Chen, Jiandong Lu, Xiaolei Wang, Luning Liu, Yuqiang Feng, "Emotional expression by artificial intelligence chatbots to improve customer satisfaction: Underlying mechanism and boundary conditions", Tourism Management, vol. 100: 104835, 2024.

[9] Hakseung Shin, Juhyun Kang, "Bridging the gap of bibliometric analysis: The evolution, current state, and future directions of tourism research using ChatGPT", Journal of Hospitality and Tourism Management, vol. 57: 40-47, 2023

[10] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, "Improving Language Understanding by Generative Pre-Training", OpenAI in computer science and linguistics, 2018.

[11] Pei-Fang Hsu, Tuan (Kellan) Nguyen, Chen-Ya Wang & Pei-Ju Huang, "Chatbot commerce—How contextual factors affect Chatbot effectiveness", Electronic Markets, vol. 33(14), 2023

[12] Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, Lucy Campbell-Gillingham, Jonathan Uesato, Po-Sen Huang, Ramona Comanescu, Fan Yang, Abigail See, Sumanth Dathathri, Rory Greig, Charlie Chen, Doug Fritz, Jaume Sanchez Elias, Richard Green, Soňa Mokrá, Nicholas Fernando, Boxi Wu, Rachel Foley, Susannah Young, Iason Gabriel, William Isaac, John Mellor, Demis Hassabis, Koray Kavukcuoglu, Lisa Anne Hendricks, Geoffrey Irving, "Improving alignment of dialogue agents via targeted human judgements", Machine Learning; Computation and Language, arXiv pp. 2209.14375v1, 2022

[13] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, Ryan Lowe, OpenAI "Training language models to follow instructions with human feedback", Neural Information Processing, 2022

[14] E Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, "LoRA: Low-Rank Adaptation of Large Language Models", Microsoft, ICLR, arXiv pp. 2106.09685v2, 2022

[15] Kaggle.com, Airline Reviews dataset, Available at: https://www.kaggle.com/datasets/juhibhojani/airline-reviews