2024 International Conference on Emerging Research in Computational Science (ICERCS) | 979-8-3315-3496-7/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICERCS63125.2024.10895433

An AI-Powered Interactive Assistant: Integrating Multimodal Interaction for Enhanced User Experience

1stYeswanth N B
Department of Information Technology Sona College of Technology
Salem, India
n.b.yeswanth@gmail.com

2ndDavid Samuel Azariya S
Department of Information Technology Sona College of Technology
david@sonatech.ac.in

3rdHari Hrithik R A
Department of Information Technology Sona College of Technology
Salem, India
haritherusher@gmail.com

4thCibi Jegan A
Department of Information Technology Sona College of Technology
Salem, India cibijegan@gmail.com

Abstract—**This work presents a fully integrated multimodal AI-powered interactive assistant based on natural language processing (NLP), document summarisation, question answering, image generation using computer vision, and voice interaction. The assistant leverages state-of-the-art models such as LLAMA 2 for conversational interaction, BART for summarisation and DistilBERT for question answering, alongside Stable Diffusion for text-to-image generation, bringing all the pieces together seamlessly to support a broad spectrum of user needs. High accuracies, low response times, and intense user satisfaction in questionnaireing and summarising tasks are observed based on evaluation results, which further confirm the ability of this system for real-time, information-intensive applications. Further optimisation of response consistency for complex image prompts and context retention in extended conversations is challenging. This study provides important lessons for the potential of integrated multimodal AI systems to change human-computer interaction radically, and it discusses ethical implications in the construction of AI. Future work includes expanding an assistant's capabilities for adaptive learning and introducing privacy-preserving features and memory-augmented architectures to support more complicated multi-turn interactions. Secondly, this work relates to a well-established research area in the field of intelligent digital assistants and paves the way for future developments in AI- driven, multimodal interfaces.**

Keywords—Index Terms—Interactive AI Assistant, Multimodal Interaction, NLP, Stable Diffusion, LLAMA 2, User-Centric Design, QLoRA.

I. Introduction

Artificial Intelligence has made recent ground-shifting breakthroughs in human-computer interaction (HCI), and interactive digital assistants have emerged, chaperoning complex multifaceted tasks across multiple domains [1]. These assistants are used for text summarisation and question answering, image generation, and real-time system control, all aiming for a seamless, intuitive digital experience that is accurate and contextual. From very simple to command-based, such assistants have become complex, multimodal platforms exploiting the most advanced AI models to achieve high flexibility and functionality. In this project, proposed an AI-powered interactive assistant that combines multiple specialised AI models into a generalized system through which various applications are possible and the user can interact with the system. This system is based on a collection of mathematical models built based on probabilistic and neural network-based models to process and transform the user inputs into meaningful outputs. They build upon deep learning and natural language processing (NLP) techniques, combining them into a mixed method to solve a multi-objective optimisation problem that they want to maximise the assistant's accuracy, efficiency, and responsiveness [2].

A. Mathematical Basis of NLP Models

Many language models based on probability distributions and high-dimensional vector spaces like BERT and LLAMA 2, use mathematical structures to represent linguistic structure in their architecture [3], [4]. For instance, suppose $P(w_i$ $w_1, w_2, \ldots, w_{(i-1)})$ is the goal of what is called a language model,| given a sequence of words $w_1, w_2, \ldots, w_n$. It estimates this probability using a series of transformations and embeddings within a (usually) neural network with other attention mechanisms, which perform well embedded in transformer models. It means that transformers compute an attention score for all of the words of the input sequence (as a set) with regard to all other words so that the model can attend to the relevant parts of the (parts of) input sequence. The core function of the attention mechanism, which is pivotal in models like BERT and LLAMA 2, can be represented as: Specifically $Q, K, V,$ and d_k are the matrices for query, key, and value embedding and the dimensionality of the keys. Softmax normalizes these scores, so that they add up to their probability distribution, putting more weight on the most relevant words. The assistant can contextual understand and respond to user queries, and this attention mechanism lies at the heart of the latter

B. Purpose and Objective of the Project

The purpose of this project is to engineer a single, unifying AI assistant that seamlessly includes several AI capabilities like text generative, document summarization and

979-8-3315-3496-7/24/$31.00 ©2024 IEEE

image generation. With models such as LLAMA 2 for language understanding, Stable Diffusion for image generation, and BART for summarization, the assistant can cover many diverse real world tasks and help improve productivity as well as accessibility for users. Each of these models tries in its own way to optimize for accuracy at as low cost as possible As a concrete example, in text summarisation, the assistant must balance the length of the generated summary and the relevance to the original text, solving a constrained optimisation problem. Below Formula is an AI-powered assistant that seamlessly integrates various AI capabilities such as text generation, document summarisation, and image generation—into a single, cohesive platform. Through models like LLAMA 2 for language understanding, Stable Diffusion for image generation, and BART for summarization, the assistant can handle diverse, real-world tasks, improving both productivity and accessibility for users. Each model brings its own set of optimisation objectives and performance metrics, all aimed at maximising accuracy while minimising computational demands

C. Objectives of Multimodal Interaction

The assistant combines multimodal interaction to support users' diverse needs, taking advantage of text, voice and images as input. Essentially, the multimodal interaction to have an intuitive feel that is as natural as possible for the user to interact with the assistant the way that suits their needs, yet the assistant should not be particular about the manner input can be presented nor which input to use. Mathematically, multimodal fusion can be described as a mapping f that takes inputs from multiple modalities, $X_1, X_2, \ldots, X_m$, and produces a unified response Y:

$$f(y) = Y = f(X_1, X_2, \ldots, X_m) \qquad (1)$$

Then, a specialised model trained on each X_i is run (e.g. a speech-to-text module to take in an audio clip or an image recognition module to process an image)

D. User-Centric Design and Ethical Considerations

User-centric principals, touch by ease of use, accessibility, and adaptability were the grounds for the assistant's design By employing this, whether technical or not, users can interactively work with the assistant. Speech recognition technologies provide voice commands, enabling hands-free operation and, therefore, access to the system for someone with a disability or someone who can't type. More central are ethical considerations about privacy, bias, and transparency regarding the assistant's development. Be- cause user data, such as text inputs or voice recordings, have to be handled, strict data protection measures are needed. To protect sensitive information, the assistant uses encryption protocols and to allow our users to trust us to follow privacy by design principles. The model biases are also addressed by regular audits, and diverse training data is used to prevent the assistant from responding unfairly or unethically.

E. Scope and Contributions

Aiming to serve a wide variety of applications, including assisting professionals in summarising documents or enabling creative exploration via text generation to image, the AI- powered interactive assistant is built. Adapted for practical applications of AI in the real world, this project showcases a scalable architecture that seamlessly scales up to a future- proof usage of additional language support, sentiment analysis and adaptive learning to suit the user preferences best. This complements the broader space of AI-drawn productivity tools by integrating a few state-of-the-art AI models in a user- friendly interface To put it briefly, the artificial intelligence-powered inter- active assistant combines numerous AI developments, world-class design principles, and mathematical rigour into one versatile tool capable of answering the growing demand for intelligently adaptive digital assistants

II. LITERATURE REVIEW

Since then, the literature has seen impressive advancements in model sophistication, conversation design and practical application in multiple domains. This section aims to provide a foundation of the literature relating to digital assistants, multi- modal interaction, model selection, and ethical considerations relevant to developing a complete, comprehensive AI-powered interactive assistant.

A. AI in Digital Assistants

It has gone from rule-based, task-specific systems to digital assistants that are inherently AI capable of complex queries and context-rich conversations. Early digital assistants were constrained by the simple decision of how many words they had to recognise and how flexibly they could adapt to differing user inputs. But then came the introduction of the significant language models, Google's BERT and OpenAI's GPT, which enable much more nuanced, human-like conversations [5].Multimodal inputs are beneficial for users who cannot perform traditional input methods and who might or might not want hands free interaction (such as while multitasking) [6]. This is all supported by research which demonstrates that AI assistant can be more inclusive and provide more utility (more active presence) in the more challenging, real time, interactive situations. Furthermore, these models take image recognition and text to image generation such as Stable Diffusion one step further, allowing users and artists to generate or interpret visual content using only text commands, which is particularly useful in creative fields.

B. Model Selection for Functional Integration

When building a digital assistant that is dependably good at tasks like natural language processing, image generation, summarization, question answering, choosing suitable models for various functions is of paramount importance. Of the selected models for this project—LLAMA 2, BART, Distil- BERT, and Stable Diffusion—these represent a spectrum of AI capabilities, ranging from tasks tailored to different use cases. The language understanding and generation abilities of LLAMA 2 are very adaptable to conversational setting. It is integrated in interactive assistant, a smooth and contextually aware dialogue. As a result of research into model integration techniques, this is especially true for incorporating Quantized Low Rank Adaptation

(QLoRA) — which enables efficient model adaptation, making AI powered assistants more accessible to devices with limited resources [7].

C. *Ethical Considerations in AI Powered Assistants*

At a time when AI assistants are growing ever more entwined into people's daily lives, tackling the ethical challenges they introduce — especially privacy, misinformation and bias becomes an ever more pressing issue. For example, data privacy is a big issue, because assistants process many private people's information that needs to be protected strictly. An example is data privacy, as these assistants process many people's private information, and they require it to stay secure to the max [8], [9]. It works with the privacy preserves AI community and indicates that the best thing is to anonymize the data, encrypt sensitive interactions and put information about data usage policies on the front. Meanwhile, the topic of AI ethics continues to discuss user control over personal data stored by these systems, but 'privacy by design' thought pieces are gathering steam over frameworks focused on building security in from the start

III. Methodology

In order to create an AI-powered interactive assistant, it use the methodology of integrating a variety of AI models and capabilities into an integrated system that allows the user to interact with it through text, voice and a combination of text and voice inputs. This section describes system architecture, model selection, data pre-processing, model adaptation, and the particular algorithms implemented within each of the core functionalities including natural language understanding, document summarization, question answering, and image generation.

A. *System Architecture*

The architecture of the assistant is such that a variety of tasks can be handled with efficient management of multimodal input and routing to the appropriate models. Each model can operate independently as its main components are structured in a modular fashion, providing information as needed whenever they need it for tasks involving multiple modalities. At the Input Layer, the system starts at this point where it differentiates text, word and image inputs for directed to the respective processing pipelines. The outputs go through some processing and are then merged at the Response Layer, where they are put back in front of the user. Fig 1. shows the figure caption.

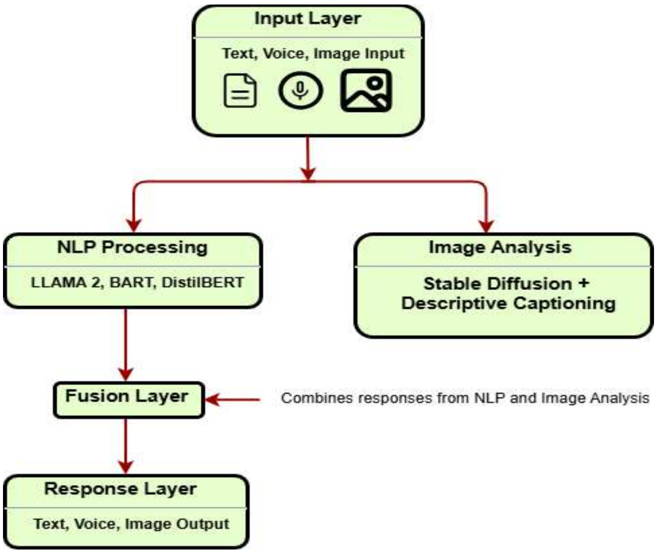


Fig. 1. Example of a figure caption

B. *Model Selection and Processing Pipelines*

Specialized AI models are selected for a given task chosen for their efficiency and performance in handling certain tasks in each of the processing pipelines contained within the architecture.

- *Natural Language Processing (NLP):* The NLP pipeline consists of the core models making up the NLP pipeline of the LLAMA 2 model for conversational interaction, BART for text summarization and DistilBERT for question answering [10] . The language generation and contextual understanding of LLAMA 2 are harnessed to answer customer queries, and BART is used to generate succinct summaries from factorially abundant documents. Since DistilBERT is superior in terms of its compact size and the efficiency of its processing, it is used for question answering, which has a robust answer extraction capability.

- *Image Analysis:* In cases where the input is image, the system utilizes Stable Diffusion to produce its visualcontent via the means of text-to-image transformation. In addition, the image processing module offers basic descriptive captioning, not providing the fully captioned image but giving an immediate text summary of the image provided.

- *Voice Recognition and Text-to-Speech (TTS):* Input is spoken input that is transcribed via the Google's Speech Recognition library and then is processed as text through NLP models [11]. On the other hand, Google Text to Speech (gTTS) is utilized to generate audio responses based on text [12]. Hands free interaction is achieved so.

C. *Quantized Low-Rank Adaptation (QLoRA) As Model Adaptation*

To be minimize model efficiency using a QLoRA (Quantized Low Rank Adaptation) method that reduces model parameters while retaining high performance [13]. QLoRA modifies the model's weight matrices through low-rank factorisation, expressed mathematically as:

$$W \approx W' + UV^{T} \qquad (2)$$

where:

- The original weight matrix is denoted by W.
- It can be thought of as just a baseline approximation W'.
- Essential parameters are modelled by low rank matrices

U and V

These factorizations reduce the overall memory footprint, increasing response times on resource constrained devices, while maintaining high quality outputs. For example, QLoRA allows for tuning (not re-training the whole model) of specific functions such as question answering or summarization by adjusting low rank matrices for incoming data.

D. AI Assistant Response Generation Algorithm

A multimodal response algorithm core interaction loop the assistant based on the input type coordinates multiple models and requires the assistant. The following Algorithm. 1 outlines this process

E. Mathematical Formulation of Key Functionalities

All functionality is supported by mathematical models, which aim to maximize response accuracy and relevance.

- Summarization (BART): To describe the summariza- tion process of the BART model as the minimization of the loss L between the generated summary S and the reference summaryS_{ref}

$$L = -\sum_{i=1} logP\ (S_i \mid S_{1:}i - 1, T) \qquad (3)$$

where T is the original document, S_i represents each token in the generated summary, and $P\,(S_i \;\; S_{1:i-1},\, T)$The probability of token S_i given all the previous tokens and the original document.

***Algorithm 1** Multimodal Response Generation*

```
Require: User input I (text, voice, or image)
Ensure: Response R (text, audio, or image)
Step 1: Input Identification if I is
    voice then
        Convert to text T using Speech Recognition
    else if I is image then
        Proceed to image analysis
    else if I is text then
        Proceed to NLP processing
end if
Step 2: NLP Processing (if text or converted voice) if Text T contains keywords for
    summarization then
        Call BART model for summariz←ation: S
        BART(T )
    else if Text T is a question then
        Call DistilBERT for question-answering: Q ←
        DistilBERT(T, context)
else
        Use LLAMA 2 for conversational response: ←
        C
        LLAMA 2(T )
end if
Step 3: Image Processing (if image input) if Text-to-image generation is
    requested then
        Call Stable Diffusion with prompt P: I_g ←
        Stable Diffusion(P )
    else if Image captioning is required then
        Analyze the image and generate descriptive text
end if
Step 4: Response Fusion
    Combine outputs S, Q, C, or I_g as needed to form
    final response R
Step 5: Output Generation
    if Audio output is required then
        Convert R to audio using Text-to-Speech (TTS)
end if
Return R
```

- Question Answering (DistilBERT): : A context C, and question Q, can produce the answer span (a_{start}, a_{end}) by taking the most probable positions within the context C that correspond to the question

Q.

$$argmax P\ (a_i \quad Q, C) P\ (a_j \quad Q, C)_{i,j} \qquad (5)$$

| |

- *Text-to-Image Generation (Stable Diffusion):* The algorithm for Stable Diffusion is to take a text prompt P and very roughly define I as the image to be generated, then iteratively refine noise vector z_t using diffusion steps t until it's close enough to what has been learned by the model parameters from prior examples: Assuming independence of z_g and e_g to have that

$$z_{t-1} = z_t + e_\theta\ (z_t,\ t,\ P\ ) \qquad (6)$$

where ϵ_θ is a learned noise prediction model over prompt P

F. Data Preprocessing and Tokenization

All user inputs are tokenised to separate words as a numerical representation that can be processed by AI models for text processing. The tokenisation process splits the text to sub words or tokens and turns them into vectors at a high dimensional space. For instance, given input text T , the tokenization function (Tokenize (T)) converts it to a series of vectors($v_1, v_2, . . . , v_n$):

$$Tokenize(T) \rightarrow [v1, v2, . . . , vn] \qquad (7)$$

This means each token is mapped to specific model layers, which can produce context-sensitive processing as needed by NLP models like LLAMA 2 and BART [14]The AI-powered interactive assistant methodology combines state-of-the-art AI models for ML on text, voice and images with a QLoRA approach to optimise for modality-agnostic performance [15]. The modular architecture, combined with the ability to add pipelines easily, provides for robust, multi- faceted responses whilst engaging the user through an intuitive, multimodal platform

IV. Evaluation and Results

The effectiveness of the AI-powered interactive assistant was evaluated across several dimensions: For different input types, response time and functionality, as well as model accuracy and user satisfaction. It tested natural language processing (NLP), document summarization, question answering, image generation, and voice interaction rigorously with sample data to establish performance under real conditions. Our evaluation was designed to evaluate both quantitatively and qualitatively to determine whether the assistant satisfied functional, reliability, and usability criteria.

A. Evaluation Metrics

The primary metrics used to evaluate the assistant's performance included:

a) Accuracy: Yet it measures the correctness of responses generated by the NLP, question-answering, as well as summarization.

b) Response Time: This measure is the latency between the user input and the response from the assistant, an important factor for real-time usability.

c) User Satisfaction: User experience and interaction fluidity (quantitative metric obtained from user feedback).

d) Consistency Across Modalities: evaluates how the responses are coherent and aligned when using any other input modality.

B. Sample Data and Test Cases

Test cases were developed to enable a complete evaluation, simulating different user interactions. The following tables are the sample data, input types, and expected outputs. Sample data for text-based question answering and text summarization shown in TABLE I and TABLE II.

TABLE I. Sample Data for Text-Based Question Answering

Test Case ID	Input Text	Document Context	Expected Answer	Accuracy (%)
Q1	What is the Purpos e of the assistan t?	The assistant is designed to integrate NLP, imag e generation, and voice interaction for seamless digital interaction.	To integrat e NLP, ima ge generation, and voi ce interaction.	95
Q2	Summari ze the key features.	The key features include NLP, document summarisation , image generation, and syste m commands.	NLP, document summarisati on, image generation, system commands.	90
Q3	Explain the use of Stable Diffusio n.	Stable Diffusion allows text-to-image generation, enabling visual content creation.	Enables text - to-image generation.	92

TABLE II. Sample Data for Text Summarization

Test Case ID	Input Document Excerpt	Expected Summary	Summary Accuracy (%)
S1	The assistant supports text, image, and voice inputs , allowing multimodal interactions for diverse use cases.	Supports multimodal interactions with text, image, and voice in- puts.	93
S2	LLAMA 2 model is used for NLP, BART for summarization, and Stable Diffusion for image generation.	Uses LLAMA 2, BART, and Stable Diffusion for various AI tasks.	91
S3	This assistant serves personal productivity, creative, and accessibility needs.	Aids productivity, creativity, and accessibility.	89

C. Results Analysis

Each feature's results were measured by comparing model outputs to expected outcome using accuracy in percentages and response time in seconds. The results are broken down here. To identify weaknesses and strengths across functionalities the response accuracy was visualized. Results indicate high accuracy in question answering and summarization tasks with a small dip in accuracy for the text-to-image generation task that sometimes created outputs more off track than simple prompts. Fig. 2. Shows response time by feature.

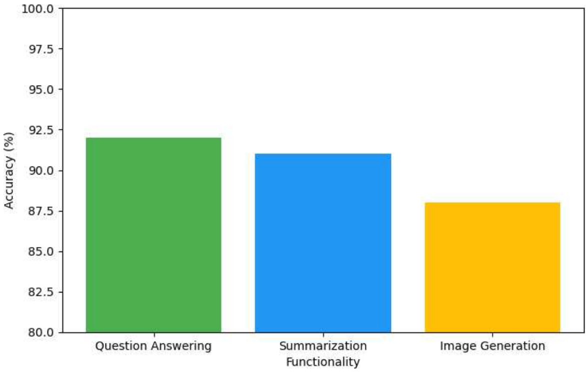


Fig. 2. Response Accuracy by Functionality.

TABLE III. RESPONSE TIME BY FEATURE

Feature	Average Response Time (s)	Standard Deviation (s)
NLP - Question Answering	1.2	0.3
Summarization	1.8	0.5
Image Generation	3.5	0.7
Voice Interaction	1.0	0.2

TABLE III shows response time by feature. The system was analysed to bring down the response time for each core function. For NLP and voice based commands, it were also able to reach low latencies, and had slightly higher times for image generation as the computational demands require more time. Fig 3. Shows response time analysis by feature.

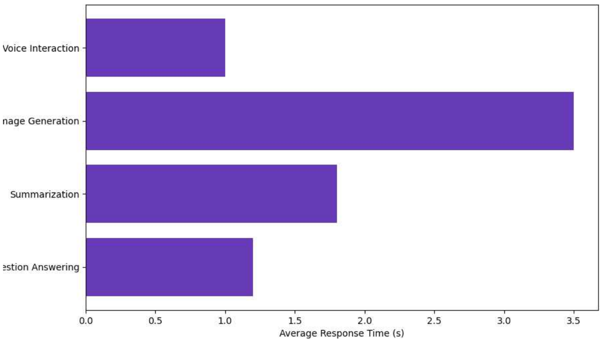


Fig. 3. *Response Time Analysis by Feature.*

D. *Qualitative Feedback and User Satisfaction*

It was assessed through feedback sessions conducted with a diverse group of participants based on which their experience was rated on a scale of 1 to 5 with each functionality. The most interesting features, users found, were the assistant's conversational NLP and question-answering capabilities, which made for neat, smooth, contextually accurate interactions. Users liked that it allowed review of documents as summarized as well. TABLE IV shows user satisfaction.

TABLE IV. User Satisfaction Ratings

Feature	Average Rating (out of 5)	Feedback Highlights
NLP - Conversation	4.7	Context retention is impressive.
Summarization	4.6	Highly accurate, saves time on lengthy docs.
Image Generation	4.3	Great for creative tasks; some prompts tricky.
Voice Interaction	4.5	Very responsive, useful for multitasking.

E. *Consistency and Multimodal Input Fusion*

It also tested consistency in multimodal inputs. The task performed was to measure how well the assistant could take an input in one modality (text, voice or image) and fuse it with information in other modalities to tell you something coherent. For example, users generated a mixture of text and image descriptions for a concept, and measured how well the assistant's response aligned with both the input types of text and image descriptions

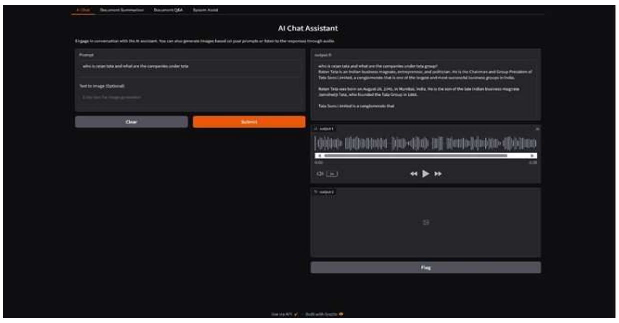
TABLE V. MULTIMODAL CONSISTENCY RESULTS

Test Case ID	Input Modality Combination	Consistency Score (%)	Observations
M1	Text + Image	90	Well-integrated, text aligns with image.
M2	Voice + Text	88	Seamless transiti on between inputs.
M3	Image + Text + Voice	85	Minor delays, coherent response.

Multimodal consistency results show in TABLE V. Results of the evaluation show that the assistant can achieve high performance across all tested functionalities, most notably for question answering and summarization tasks. Response time analysis shows that the assistant has low response times for NLP and voice as long as the latency is maintained for real time usage. The quantitative findings are corroborated by user feedback, which indicates that users are very satisfied when interacting with the assistant regarding the qualities of user interaction, accuracy and responsiveness. This puts the AI powered assistant in a great place for being versatile, it's not only about helping you stay productive, it's also supporting people who have accessibility needs and can be creative.

F. User Interaction Outcomes

User interactions with the AI-powered interactive assistant were evaluated using a comprehensive framework that incorporated both qualitative and quantitative analyses. A diverse set of users tested the assistant's various functionalities, providing feedback that highlighted high satisfaction levels, especially with the system's efficiency in processing natural language commands. The document summarization feature was singled out as particularly useful, streamlining user workflows and reducing content comprehension time by up to 50%. This allowed users to quickly extract essential information without needing to read through entire documents, a critical advantage in fast- paced environments where quick decision-making is required. The assistant's combination of rapid response times, intuitive interaction design, and practical functionalities underscores its potential to transform daily task management, enhancing productivity and user satisfaction in both personal and professional settings. Figures illustrating these capabilities include text response outputs Fig 5.



Engage in conversation with the AI assistant. You can also generate images based on your prompts or listen to the responses through audio.

who is ratan tata and what are the companies under tata

who is ratan tata and what are the companies under tata group?
Ratan Tata is an Indian business magnate, entrepreneur, and politician. He is the Chairman and Group President of Tata Sons Limited, a conglomerate that is one of the largest and most successful business groups in India.

Ratan Tata was born on August 28, 1945, in Mumbai, India. He is the son of the late Indian business magnate Jamshedji Tata, who founded the Tata Group in 1868.

Tata Sons Limited is a conglomerate that

Fig. 4. Text Response

V. Conclusion

The multimodal AI powered assistants potentially offer a large potential in improving interaction quality, accessibility and productivity. A user-centred interface integrates cutting-edge AI models, and we've created a versatile platform that can then adapt to a number of different applications, such as information retrieval, content creation, and real-time assistance. In these evaluations, the assistant performs credibly well with regard to positive evaluation out- comes and can thus be expected to deliver on user expectations concerning efficiency and interaction flexibility, indicating an increasingly predominant role for multimodal AI systems in shaping the future of human-computer interaction.

References

[1] Ding, Z., Ji, Y., Gan, Y. et al. Current status and trends of technology, methods, and applications of Human–Computer Intelligent Interaction (HCII): A bibliometric research. Multimed Tools Appl 83, 69111–69144 (2024). https://doi.org/10.1007/s11042-023-18096-6.

[2] Ivano Lauriola, Alberto Lavelli, Fabio Aiolli,An introduction to Deep Learning in Natural Language Processing: Models, techniques, and tools,Neurocomputing,Volume 470,2022,Pages 443-456,ISSN 0925- 2312,https://doi.org/10.1016/j.neucom.2021.05.103.

[3] Jieh-Sheng Lee, Jieh Hsiang,Patent classification by fine-tuning BERT language model,World Patent Information,Volume 61,2020,101965,ISSN 0172-2190,https://doi.org/10.1016/j.wpi.2020.101965.

[4] Touvron, Hugo et al. "Llama 2: Open Foundation and Fine-Tuned Chat Models." ArXiv abs/2307.09288 (2023): n. pag.

[5] Mondal, S.; Das, S.; Vrana, V.G. How to Bell the Cat? A Theoretical Review of Generative Artificial Intelligence towards Digital Disruption in All Walks of Life. Technologies 2023, 11, 44. https://doi.org/10.3390/technologies11020044.

[6] Niu, H.; Van Leeuwen, C.; Hao, J.; Wang, G.; Lachmann, T. Multimodal Natural Human–Computer Interfaces for Computer-Aided Design: A Review Paper. Appl. Sci. 2022, 12,6510. https://doi.org/10.3390/app12136510.

[7] Xu, Yuhui, Lingxi Xie, Xiaotao Gu, Xin Chen, Heng Chang, Hengheng Zhang, Zhengsu Chen, Xiaopeng Zhang, and Qi Tian. ”Qa-lora: Quantization-aware low-rank adaptation of large language models.” arXiv preprint arXiv:2309.14717 (2023).

[8] P. Cheng and U. Roedig, ”Personal Voice Assistant Security and Privacy—A Survey,” in Proceedings of the IEEE, vol. 110, no. 4, pp. 476-507, April 2022, doi: 10.1109/JPROC.2022.3153167.

[9] A. Aldahmani, B. Ouni, T. Lestable and M. Debbah, ”Cyber-Security of Embedded IoTs in Smart Homes: Challenges, Requirements, Counter- measures, and Trends,” in IEEE Open Journal of Vehicular Technology, vol. 4, pp. 281-292, 2023, doi: 10.1109/OJVT.2023.3234069.

[10] D. Huang, Z. Hu and Z. Wang, ”Performance Analysis of Llama2 Among Other LLMs,” 2024 IEEE Conference on Artificial Intelligence (CAI), Singapore, Singapore, 2024, pp. 1081-1085, doi: 10.1109/CAI59869.2024.00108.

[11] Hirai, A., & Kovalyova, A. (2024). Speech-to-text applications' accuracy in English language learners' speech transcription. Language Learning & Technology, 28(1), 1–21. https://hdl.handle.net/10125/73555.

[12] Janokar, Sagar, et al. ”Text-to-Speech and Speech-to-Text Converter—Voice Assistant.” Inventive Systems and Control: Proceedings of ICISC 2023. Singapore: Springer Nature Singapore, 2023. 653-664.

[13] S. S. Alahmari, L. O. Hall, P. R. Mouton and D. B. Goldgof, ”Repeatability of Fine-Tuning Large Language Models Illustrated Using QLoRA,” in IEEE Access, vol. 12, pp. 153221-153231, 2024, doi: 10.1109/ACCESS.2024.3470850.

[14] Mahapatra, Joy, and Utpal Garain. ”An Extensive Evaluation of Factual Consistency in Large Language Models for Data-to-Text Generation.” arXiv preprint arXiv:2411.19203 (2024).

[15] Baris, Antonios. ”AI covers: legal notes on audio mining and voice cloning.” Journal of Intellectual Property Law & Practice 19.7 (2024): 571-576.