iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH
IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 Issue: VII Month of publication: July 2025

DOI: https://doi.org/10.22214/ijraset.2025.73464

The Conversational AI: The Prompt Formation and Response Matching in ChatGPT

Nivedita J[1], Manju Sadasivan[2]
School of Science and Computer Studies, CMR University, Bengaluru, India

Abstract: The introduction of the mind-blowing Large Language Models (LLMs) such as ChatGPT imposes the image of the brilliant future of smart chat practices in diverse circumstances. The issue of establishing good and representative prompts becomes increasingly burning as the models evolve to accept a broader range of variety of what they read as input, text, graphics, noise, audio, and video. The article suggests the quality, coherence, and depth of the answers that ChatGPT generates in a multimodal setting, specifically, the visual data interpretation using Tableau. This paper discusses the response of ChatGPT to some technical questions that people face in the real world regarding image recognition and manipulation of texts and emotions. The primary question, according to the middle section of the experiment, is whether the sufficiency of the model compliance is defined by the accuracy and clarity of the prompts set by the user, or the scenario of the issue at hand itself. Traces of experimental evidence indicate that the performance of ChatGPT depends on the variations of the ways the questions are formulated and on the realms of discussion. Further, the paper looks into the application of the visualization tools as a way of realizing the behavior of the models, assisting in the process of assessing it in a less wordy way through Tableau. Overall, this contribution helps to comprehend the possibilities and boundaries of LLMs in connection with the actual applications of multimodal systems and emphasizes the impact of the speed of design that helps to improve the results obtained through the assistance of AI.
Keywords: Prompt Engineering, ChatGPT, Multimodal AI, Conversation Agents, Large language models.

I. INTRODUCTION

Artificial Intelligence and its aspects have taken a very fast track, and opened the era of Conversational Artificial Intelligence, which has given the ability to communicate with the computer by using a natural language by listening or typing, or even chatting visually. Owing to such technological advancements, how human beings communicate with computers has been transformed totally, as it has become simplified, sensitive to emotions, and friendly to its users [1]. These tools can include such tools as COFFEE, which are adapting into adaptive systems so that an object can match the information to the emotional and cognitive state of the user, and personalization and the extent to which someone will be engaged [2]. Among the most topical phenomena in the field, one must mention the Human-in-the-Loop process (HITL): the human input is being added to the training of the AI regularly, with the thought to give the system some context and sense of moral resistance to the trajectory [3]. Prompt engineering process is the creation of input prompts explicitly, precisely, and purposefully, and it has since grabbed the leading position of controlling the quality of responses provided by the Large Language Model (LLM), such as ChatGPT [4]. These skills are being deployed more frequently in the emotionally tense areas of elderly care, mental health, and personal care techniques [5, 6]. Multimodal AI -interactivity- possibility to integrate text, image, and audio processing and generation- has increased the scope of context and time-sensitive interactive. This kind of system has now found its way in smart homes where it interprets behavior and messaging voices and intends to respond to them [7, 8]. The other advancement related to the LLM application was the formation of high-speed and speedy prompt-generation techniques, which have no form of retraining, e.g., Quick Turnaround Mechanisms (QTM) [9]. Through these advances, such as AudioGPT and LLaVA, cross-modal reasoning, situational awareness, or emotional recognition can be used in multimodal frameworks, and a move toward human-level comprehension is within reach [10, 11]. However, the same systems continue to experience scalability issues with computation, fixed training data, and slowness of response [12].

Multi-modal models can be powerful in generating coherent images and texts based on an extremely limited input and provide a step beyond imagination in AI, in the case of creative generative applications such as DALL-E and LISA [13, 14]. Nevertheless, technical progress notwithstanding, real-world implementations of technical progress have, in many instances, been hampered by a dearth of emotionally intelligent feedback mechanisms, by a lack of cultural customization by the absence of strong ethical frameworks [15].The development of new services creates wider movement towards the democratization of AI-based data analytics.

Put more simply, at the interface of the non-technical user with potentially massive, multidimensional data, LLM-integrated platforms have the potential to enable a state of natural language interaction that streamlines the interaction between raw analytics and intuitive decision-making [16]. The importance of tools of data visualization, such as Tableau, in such a situation comes in handy as it provides transparency into the workings of an LLM and helps to understand AI behaviors better through the visual representations[17].

This article finds out whether time-sensitive transparency and domain complexity affect the performance of ChatGPT on multimodal situations, and presents the results, using Tableau visualizations, to interpret and analyze them. It aims to get a picture of how multimodal LLMs can work on real-world, emotionally complex tasks and what visualization tools can help us manage their performance and interpretability.

Objectives and research questions are as follows:

All the existing gaps are discussed in the paper, and a set of 100 human-rated prompts is provided that have the potential to offer five alternative versions of the same AI that is also supposed to help with cross-modal performance, text generators, audio models, and models aimed at interacting with visual information. In the research under the dynamic visual analytics, Tableau, the following issues are going to be answered as follows:

1) Can the instruction on the type of prompt (e.g, a personal reflection, an educational guidance, or a tool review) matter in the performance of the model in the various LLMs and multimodal AI models?
2) What kind of measurement of conclusion (such as clarity, validity, ty, or accuracy) is most correlated with a higher placement by a human being?
3) What do visual patterns in a bar and box plot look like compared to the behavior of models and variability of performance? What can be stated concerning visual tools in a bar and box plot and a scatter plot?

Its primary objective would be to hypothesize on the Multimodal Prompt Evaluation Framework that an individual ought to apply to evaluate the output of the artificial intelligence models that participated in the writing process, production of the speech, and pictures. This kind of thinking happens and is demonstrated when it comes to the application of the Tableau visual analytics package.

II. LITERATURE REVIEW

Supriyono et al. performed a study of the dynamics of Natural Language Processing (NLP), including the implications, challenges, and future trends, which were explicitly studied in the work. They have shown that when a modern transformer-based NLP system comes into play, high rates of incorrectness in certain tasks can be reached: 92.7 percent on the comprehension of the meaning of a question, 93.7 percent on obtaining intent, 89.3 percent on elaborating the context, and 99.1 percent on detecting the PII. They have even gone further to find out that when all the text is mined, then an extra 37 percent can be achieved of actionable information, which can be achieved based on structured information alone, and therefore result in a 47 percent contraction of customer service activities. This kind of research supports the possibilities of NLP in terms of customer experience conversion and the accuracy, stability, and reliability of conversational AI [1]. The scoping review made by Kusal et al., AI-based conversational agents gave an outline of how the deep learning based conversational agent has shifted, hence, evaluating the manner under which it has shifted towards becoming more deep learning based. The review listed the potential relevance of Natural Language Understanding NLU, Natural Language Generation NLG components, and anticipated the considerably longer problems like recognition of emotions, memory of context, and lack of consistent tools to measure the same. Transformer designs have become the new normal to build scale and high-performance conversational agents [2]. Lima et al. talked about the design and use of conversational affective social robots (SAR) in managing the elderly and those already affected by dementia. The article has come up with the conclusion that SARs would lead to better socialization of the individuals with dementia, besides minimizing neuro-psychiatric symptoms. EMOTIONAL lying, morals, and ethics were, however, cited as one of the major impediments to the implementation in the long run due to the failure of ASR. Some of the concepts proved by the scholars are that they have popularised user-friendly, ethically aware construction as part of the method of successful integration with healthcare [3]. Teye et al., in turn, examined the role of the ability to pick the emotions in conversational AI, referencing peculiarities of cultures and contexts. They report an 85 to 96 percent accuracy of emotion recognition in their model, with coupled speech and imagery data in a 3-layer CNN and a novel AFME algorithm. The paper talked about how culture matters in emotion recognition systems and some of the issues that should be brought to consideration, which include the difficulty of recognizing micro-expressions, where one should have accessories on the face or have light. [4]

Hassija et al. described what such Laurent Macron language models (LLM) or, in this particular case, the ChatGPT phenomenon is. They searched through architecture, training policies, and assessment indicators of GPT-3.5. The prejudice and inability to reason, weakness, and low computing load are counted among the limits that were mentioned by the authors. A part of the paper guides what needs to be improved on the LLMs, which includes, among others, increased generalization, multimodal inclusion, and ethical design [5]. The researchers thought that Lima et al. would see the health aspects of dementia-affected people being managed in the comfort of households with the help of agent conversationalists such as Amazon Alexa. The system made a comparison of the behavior change, the change in engagement through clinical path, by keeping the merging of the IoT sensors and the conversation account in check. This reduction of popularity in its use could be explained by the fact that the novelty aspect of doing this was gone; however, the participants continued to use Alexa because they found it very useful in their lives, and the same thing could be said when agreeing on the concept as a way of remaining healthy through conversational AI [6].

Elragal et al. have come up with a new learning chatbot, COFFEE, an artificial intelligence learning chatbot involved in helping learning disabled students. The system will incorporate the use of Dialogflow, Chat GPT, and learning analytics so that it provides a personal touch. When they tuned their model on the Conversational Question Answering (CoQA) set, they achieved around 98 and 96 percent accuracy on the training set and testing set, respectively. The work in point illustrates how conversational artificial intelligence can be used to create dynamic and versatile learning [7]. Ghosh and Deepa have created the QueryMintAI multimodal LLM that can take input in text, image, audio, video, and structured data. The significance of user privacy, the model could be optimized at the local level, and it could also perform well on various benchmarks such as Rouge and Perplexity. By having this integrated system, flexible and multimodal communication can be maintained with confidence in the safety of data [8]. Park et al. have considered the problem of rubric engineering in generative AI. They set up a Query Transformation Module (QTM) that queries into Preceding Phrases Query (PPQ), Cloze Query (CQ), and Purpose Explicit Query (PEQ). During the testing of Korean LLMs, an increase in the generation of answers in a natural and specific manner of responses by an average of 11.46 percent was seen, and it is possible to see how it can be assisted by QTM in making sure chatbot-like conversations do not fall into the trap of becoming more conversational and contextual [9]. In one of the previous surveys, the technical foundation and history of ChatGPT have been overviewed, its applications in healthcare, education, and research fields explored, and even the self-reflection quoted when chatting through ChatGPT was also attempted to be added. However, it never offered more radical criticism or longer-term visionary attitudes. Koubaa et al. conducted an overview of GPT models applied in the process of radiology to carry out image-based classification operations, image segmentation tasks, and further medical training. The analysis of ChatGPT in regards to the classic question-answering approaches over the knowledge graphs showed that ChatGPT was able to be more precise and natural in comparison to the classic approaches in most cases. It provided a workflow-based account of ChatGPT and its actions in the modern context, where it offered no objective judgment on whether gaps can be expected next time. The purpose of this up-to-date survey is to become the first critical survey to incorporate the technical novelty, correct taxonomy, coverage of the applications, assessments of the limitations, and prospects into future research requirements [10]. Alparslan was penning on the need for verification and accuracy of conversation in online analytics of businesses. The paper has mentioned that the viability of business KPIs and user goals, with the compatibility of the model, has been crucial, irrespective of having effective models. This work will fit into the thematic area of accountability and trust in the LLM deployments, but it is not your subordinate idea and is definitely within the thesis [11]. The other novel multimodal LLM proposed by Bai and Bai uses the global visual property of Contrastive Language-Image Pre-training (CLIP), Adapter-based fine-tuning, and an MLP projection module. They reported a new SoTA of 93.96% the Science QA dataset based on such a model. This paper demonstrated that the extensive use of comprehensive visual embeddings can drastically enhance the notion of multimodal reasoning [12]. In a subsequent publication, Bai and Bai came up with an improved architecture of projections that incorporated Resampler and MLP modules through CNN in enabling greater flexibility within token transformations. When they ran their model on Science QA (a model based on CLIP ViT-L/14 and Mistral-7B), they were able to achieve 94.27 percent. This is in comparison to other models, whose performance is worse relative to this model or costly in terms of training. Such developments pave the way to the future of more multifaceted and cost-effective MLLMs [13]. The issue has been resolved by Chen et al., who created a framework that led to Med3D Insight, which is composed of 2D MLLM and Plane-Slice-Aware Transformer (PSAT). They were GPT-4V CLIP-aligned annotated slices; they had increased Dice segments and classes than the 19 baselines, and increased the scores by 2-3 per cent. To the extent of diagnosis, the medical imaging provided by the invention becomes highly accurate [14]. Specifically, the article by Kang et al. addressed the problem regarding the possibility of improving the image editing space by complementing it with Chain-of-Thought (CoT) reasoning, filling in with ML models. They possessed a lightweight architecture that comprised computer-aided technology, where a small LLM is applied to finalize the CoT process, and the diffusion model instead generates the images.

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue VII July 2025- Available at www.ijraset.com

The method has provided high-quality and error-free revision and less computing expense, and presents a manageable resolution to make familiar and perform complicated prompts [15]. The problem of human-AI cooperation was also systematically reviewed by Puerta-Beldarrain et al., who also presented their uniform model of cooperation at five levels and the core principles of cooperation, i.e., trust, personalization, and communication. AI and human beings were the two areas where the authors saw potential in the application of data and moral decisions to be high. The above publication may be considered as the general rule of what to think and design about accountable and transparent systems involving humans and AI [16]. Rocchietti et al. compared the conversation query re-writing task that was carried out by ChatGPT (gpt-3.5-turbo) with smaller, fine-tuned, open-source LLMs, such as Llama-2 and Flan T5. Optimized models, e.g., Llama-2-13 B, can be used instead of ChatGPT, more effectively remembering the tasks and far less energy-consuming. As an illustration, using Flan T5-large, it required less energy up to 20 times. The observation is that the scalability of the conversational system entails small task-specific models [17].

III. METHODOLOGY

A. Types of modalities of prompts:

In the study that was conducted on a wide, multimodal prompt comprising 100 pre-selected tasks and displayed in multi-AI systems, the hand-picked tasks were experimented with:

- The natural language comprehension and production in perceptions of emotions, task generation, summary, and context-based reasoning.
- Generation and understanding of a new image: Image tasks, image-text, image language, visual question answering (VQA), and editing tasks.
- Sound signal: Identification of the speech, judgment of a tone of feeling, actual notes of the voice, and clarity of the sound.

Even the choice of the prompt itself was such that it was true to say that, to the extent that the prompt could have been in line with the work at the same period, particularly with the realms of personalized medicine, education, or assistive AI.

B. Model types

- Examples of Style Activities: Style in general, auto-summarization of a text, working with display summarization, and working with ChatGPT.
- Whisper (v2) Audio: The text is inclusive of the speech literature, and it will be used in.
- DALL-E: Image-to-text completion, Image perception, Image Text, Image description, and LLM Object Detection.
- AudioGPT: Multi-modal synthesis relative to an audio-text emotion detection.

The models were presented within the context of the strengths and the areas of weakness of the propositions of the modalities of the other honor of fidelity, and the cross-domain in which it had performed fairly well.

C. Task Verification

The outlined models were listed manually in a numerical score of rating (0 = poor, 10 = excellent) of person-in-the-loop or human-in-loop quantitative measurement in the multimetric design style as follows:

- Rightness: This is different in what rightness or suitableness of a description implies.
- Relevancy: The relevance of a response to an activity is discussed.
- Numeric: The plain old math, integrals, and calculus to assume.
- Syntactical persistence: Inventions /Creativity of surprise in the production of image/text.
- Emotion Recognition: A type of examination or perception of emotions.
- Audio Fidelity/Photo Fidelity: There is a use of recording and image photography.

Therefore, the outputs of any concrete model would be cross-linked, balanced to the personal rating, and would be replicated beside the Task Category, Model Type, Metrics, and would be carried out in a manner such that they prefill and balance the data structure with any particular Prompt ID.

D. System Design Process

The Tool has the following Working Mechanism of the System Design Process and flow:

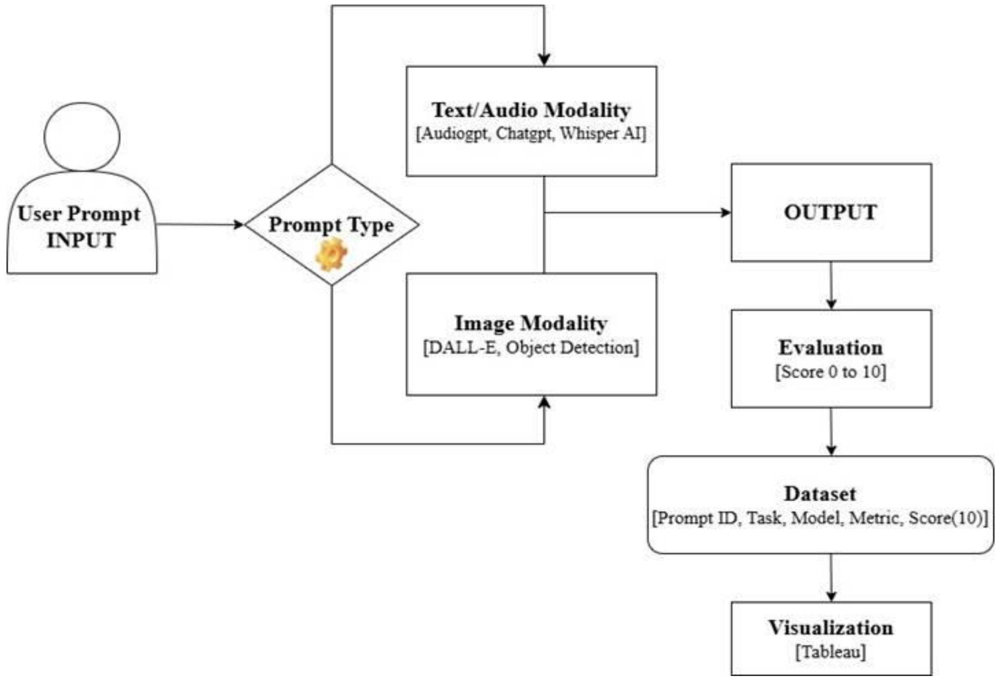


Fig 1. Mechanism of the System Design Work

Fig. 1 depicts the process by which the prompt is identified by a human-in-the-loop who manually evaluates on a scale of 0 - 10 then followed by the visualization.

1) User Prompt INPUT: This prompt is passed on to the user in the form of the search bar that may be connected to the text in the form of audio graphics that may be loaded in the prompt to retrieve the information, which is as per the job. The time LLM needs to accomplish the processing task is several seconds and the answer to the prompt can be corrected with the help of the tools such as: generate an image, search in the web, write or program, conduct deep research, and think longer, the material can be copied, further, the user may also access it using audio input and receive the response with the help of use voice mode.
2) Type of prompt: The following are the categorizations of the prompt type-
- Easy question and answer, text-based- upload an audible file or video and choose the audio input.
- Image-A follow-up to an image field, which is a linked picture file.
- Audio+ text- prompt/Query depending on the audio posted.
- The picture is a prompt/question about an image.

3) Text/ Audio Modality: This is the manner of recording the audio or the text prompt/question manually, and would be later manually captured in the dataset.
4) Image Modality: The image added/uploaded is handwritten, and it should also be repeated under the dataset.
5) OUTPUT: Answers given by ChatGPT are examined.
6) Evaluation: The process also measures and rates it on a scale of 1 to 10
7) Dataset (Record): The data will be entered into a record, and this record will be in the form as illustrated below.

Prompt ID	Task	Model	Metric	Score [10]

Fig. 2 Data entry format of data in the record

To make the recorded responses so accurate and consistent, all the administered responses were not just tabulated, but further, results and notations were achieved in the conventional and antique way of having the standardized tabular response form in handwritten form (see Fig. 2).

8) *Tableau:* Tableau has been used to arrive at such a visualization, and here, the same has been represented.

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue VII July 2025- Available at www.ijraset.com

E. ChatGPT Creation and Reactions Speed & Stability

ChatGPT does not perform analysis in a language; the probability rule, vectorization, and neural transformer are used. The text on input is also given in a smaller unit of length, and that is known as tokens, but Byte-Pair Encoding (BPE) is used in the majority of cases. ChatGPT works on the logic of changing words to mathematical formulas and according to likelihood and math, not a symbolicism, but a neural net, which is used by the back-end capacity. In this session, the reaction to the choice of selecting prompts shows coding of prompts, processing of response, generation of response, and assessment of the response.

1) Input Processing:

- The first one converts the input text into small units (e.g., subwords or words).
- All the tokens are assigned a whole dimension so that there is a numerical notion in the form of code, which is the semantics of the tokens.
- It is order sensitive, which is why the position of each of the tokens is mixed up with each vector.
- Attention process makes sure that such a complicated load of information is given to the transformer mode that can read contexts and movements of tokens.

2) Answer: Prediction of future token

- ChatGPT makes only one token prediction at a time; however, it carries out tasks into consideration of all tokens so far (both in the direction, the prompt and what it has already performed up to this point).
- It gives probabilities to other successive tokens at every step and selects the most probable one.
- The process is applied until it is a whole answer without incongruency in the model.
- ChatGPT does not make any rough computations either; it is more preferable to demonstrate what is to occur next by the patterns that it learned in the training process.

3) Output: Match the replies to

- In building an association between the answer and the prompt, the measures of similarity are applied to the answer produced and the expected answer or the correct answer.
- The comparison can be executed with verification of objective cases with a text match.
- Validation of factual cases by the text match.
- A comparison of the semantics of the meanings, an act of comparison, though not an act at all, and of a word-for-word kind.

Just to go through a few language models to decide on the correct or pertinent one. The answer to a prompt hinges a lot on how the prompt is written. The right prompt also informs the right direction the model is supposed to take to come up with proper and real answers. Clarity of goals and settings, and this gives a sample or a criterion of what the response should be like, are some of the strategies in good prompts. The refining of the products is updated and repetitive depending on the outputs. ChatGPT responds to its commands and generates answers due to more complex mathematical concepts in the manner of vector representation, attentive mechanisms, and probabilistic modeling. By getting a feel of these concepts and adhering to good prompt engineering, the users will be able to reach a far better quality and applicability of the given response.

IV. EXPERIMENT SETUP

To implement this study, several tools and technologies were introduced into the working process. Both of them had a role to contribute to developing a structured, reliable, and analyzable process.

TABLE I. EXPERIMENTAL SETUP OVERVIEW

Technology	Purpose
The primary chat AI system was ChatGPT (GPT-4)	It was provided with formulated prompts, which involved audio, visual, and text input to determine its effectiveness in multimodal responses and comprehension.
Whisper (v2)	In operations that involve speech recognition, transcription accuracy as it pertains to audio prompts, Whisper (v2) has been put to use.

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue VII July 2025- Available at www.ijraset.com

DALL-E	DALL-E is applied to visual generation, in which prompts are textual, and image interpretation.
AudioGPT	AudioGPT was trained to perform emotion detection and audio-text multimodal generation when performing task responses over audio prompt.
Draw.io	Draw.io is a software that is used to draw the system's architecture and flowcharts (see Fig. 1).
Microsoft Excel	Microsoft Excel enabled the entry of instructions manually, recoding of a response, scoring, and categorizing of the information in a structured manual manner.
Tableau	With the help of Tableau, visualization of the dataset and generation of analytical charts based on performance comparison between models and selection of the prompt are suggested.

V. MATHEMATICAL MODEL

The proposed system can be described as follows in terms of mathematical analysis.
Let us suppose that the response score is $R \in [0,10]$, which is given as:

$$R = \alpha \cdot f(P) + \beta \cdot g(M) + \gamma \cdot h(C) + \delta \cdot m(T) + \sum w_i \cdot s_i$$

Where:

- $f(P)$ = Prompt effectiveness (how entrepreneurially clear a prompt is, accurate enough, and what type it is (in other words, is it instructional or Q&A, or reviewing).
- $g(M)$ = Multiplier of modality
  - text = 1
  - audio = 0.8
  - image = 0.9
  - multimodal = 1.2
- $h(C)$ = Complicatedness in the marking of the task, where marking can be, or 1.
- $m(T)$ = Model efficacy index(empirical).
- s_i = Mark in per metriche (es, Relevancy, Emotion detection, Syntactic coherence, etc.).
- w_i= Weights of each of the metrics, and they are added up to 1.

VI. RESULT

Various trends and patterns of performance were revealed during the experimental analysis in models and modalities tested. The most significant findings are presented below and justified with the help of Figures 3 to 9:

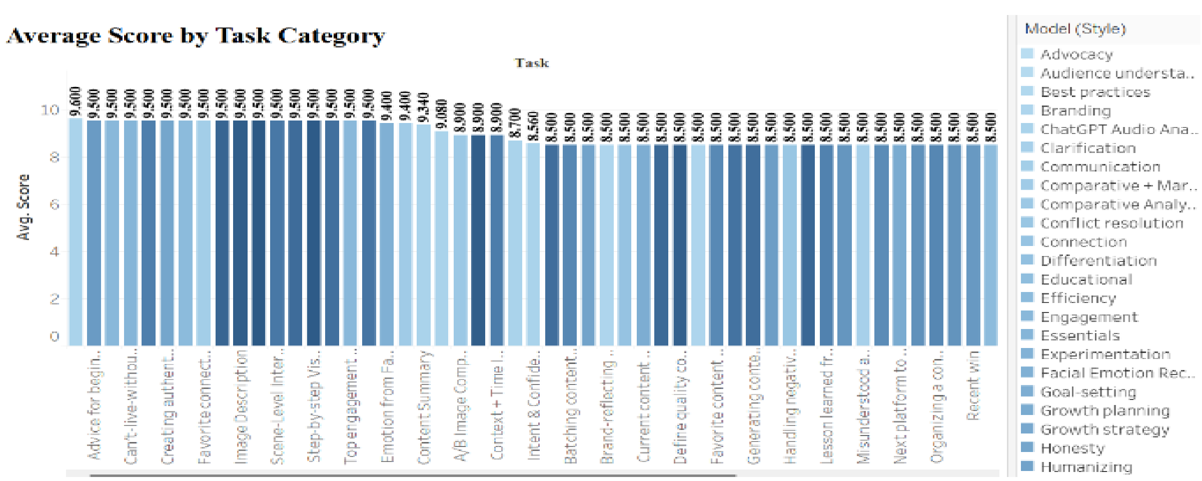


Fig. 3 In terms of the task completed, the average score

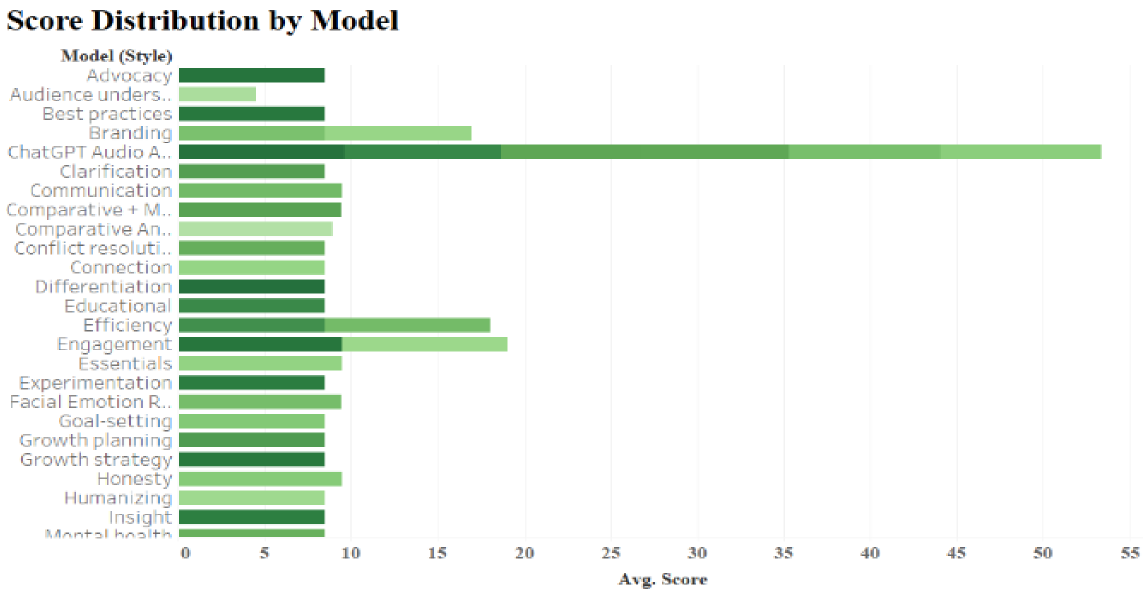


Fig. 4 Model Scoring

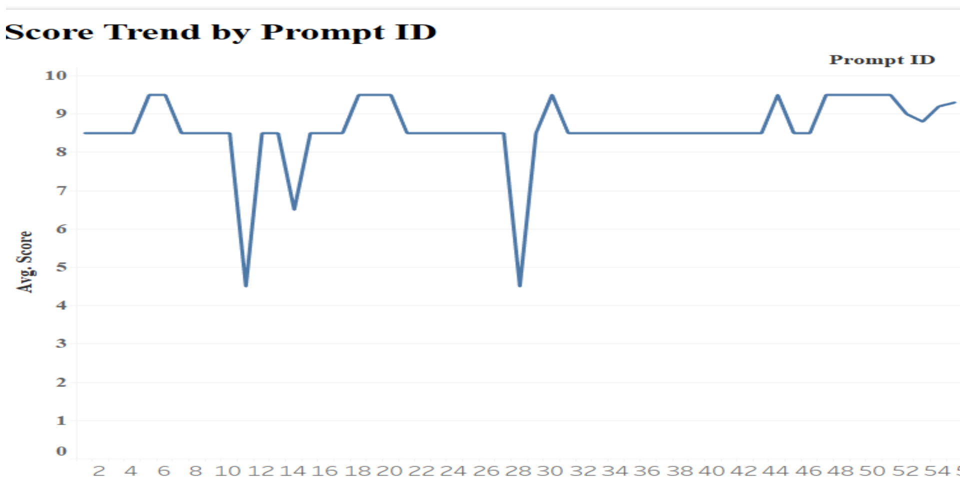


Fig. 5 Type of classification of Score based on Prompt ID

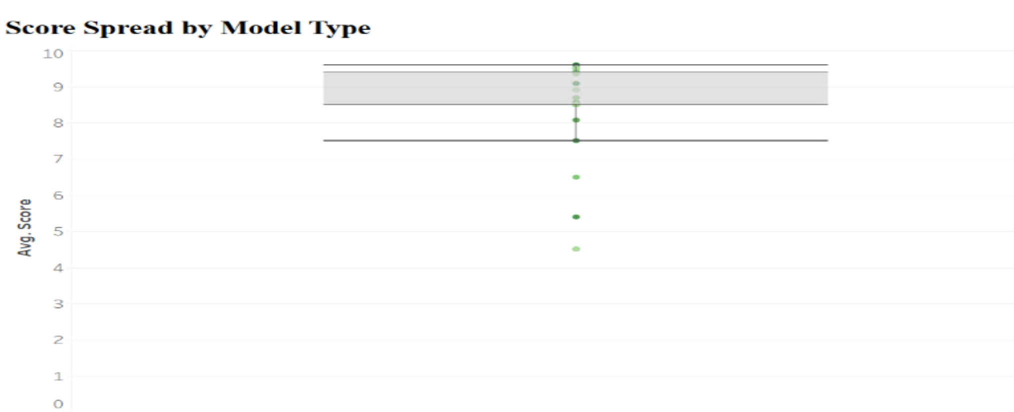


Fig. 6 The Dependency in the Type of Model was the Score Measured in the Form

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue VII July 2025- Available at www.ijraset.com

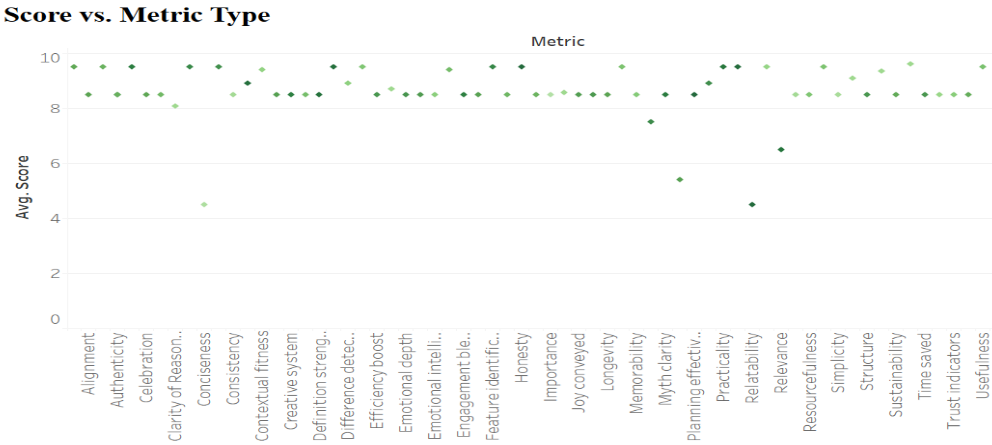


Fig. 7 Score v. Metric Type

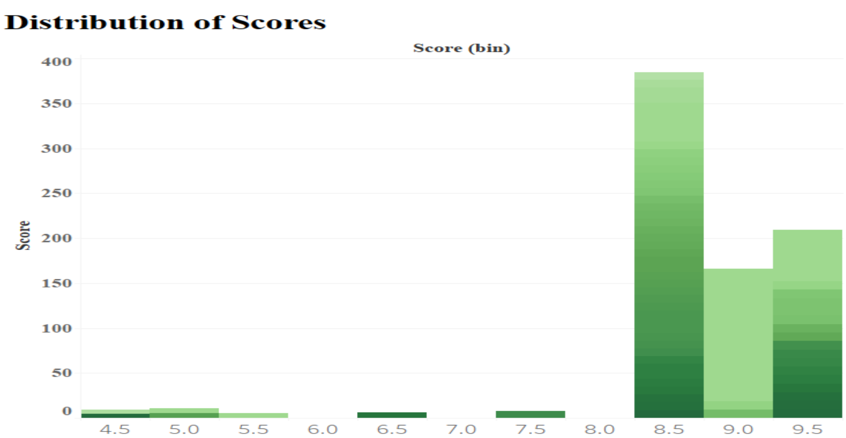
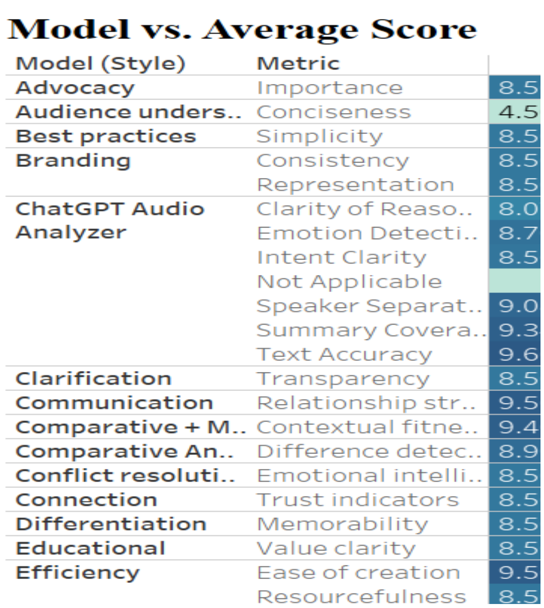


Fig. 8 Scores Distribution



Model vs. Average Score

Model (Style)	Metric	
Advocacy	Importance	8.5
Audience unders..	Conciseness	4.5
Best practices	Simplicity	8.5
Branding	Consistency	8.5
	Representation	8.5
ChatGPT Audio Analyzer	Clarity of Reaso..	8.0
	Emotion Detecti..	8.7
	Intent Clarity	8.5
	Not Applicable	
	Speaker Separat..	9.0
	Summary Covera..	9.3
	Text Accuracy	9.6
Clarification	Transparency	8.5
Communication	Relationship str..	9.5
Comparative + M..	Contextual fitne..	9.4
Comparative An..	Difference detec..	8.9
Conflict resoluti..	Emotional intelli..	8.5
Connection	Trust indicators	8.5
Differentiation	Memorability	8.5
Educational	Value clarity	8.5
Efficiency	Ease of creation	9.5
	Resourcefulness	8.5

Fig. 9 Model design Average Scoreboard

VII. CONCLUSION

This paper has compared five state-of-the-art multimodal AI systems against each other with 100 well-thought-out prompts across three domains: text, image, and audio, run through ChatGPT. The findings indicate that immediate engineering is vital to affect the quality of output and is especially important for complicated tasks, where the involvement of emotion recognition features or abstract visual analysis is required. It is worth mentioning that the performance of models differed dramatically, especially when evaluated against audio-based prompts, which indicated the issues with the stability of responses. Such variability was well translated into boxplots and line graphs in Tableau, and this underlines why modality alignment is critical in terms of producing a consistent and accurate output. The importance of strategic design of prompts and considerate modal selection is also underscored by such results. The efficacy and reliability of conversational AI systems could considerably improve by streamlining timely trends and outlooks in the real world.

REFERENCES

[1] Supriyono, Aji Prasetya Wibawa, Suyono, Fachrul Kurniawan, Advancements in natural language processing: Implications, challenges, and future directions, Telematics and Informatics Reports, Volume 16,2024,100173, ISSN 2772-5030,https://doi.org/10.1016/j.teler.2024.100173

[2] Kusal, S., Patil, S., Choudrie, J., Kotecha, K., mishra, S., &abraham, A. (2022). AI-Based Conversational Agents: A Scoping Review From Technologies to Future Directions: Conversational agents. IEEE Access, 10, 92337 - 92356. Article 22029817. https://doi.org/10.1109/ACCESS.2022.3201144

[3] Lima, Maria &Wairagkar, Maitreyee & Gupta, Manish & Baena, Ferdinando &Barnaghi, Payam & Sharp, David & Vaidyanathan, Ravi. (2021). Conversational Affective Social Robots for Ageing and Dementia Support. IEEE Transactions on Cognitive and Developmental Systems. PP. 1-1. DOI: 10.1109/TCDS.2021.3115228

[4] M. T. Teye, Y. M. Missah, E. Ahene, and T. Frimpong, "Evaluation of Conversational Agents: Understanding Culture, Context and Environment in Emotion Detection," in IEEE Access, vol. 10, pp. 24976-24984, 2022, doi: 10.1109/ACCESS.2022.3153787

[5] Hassija, Vikas & Chakrabarti, Arjab& Singh, Anushka &Chamola, Vinay & Sikdar, Biplab. (2023). Unleashing the Potential of Conversational AI: Amplifying Chat-GPT's Capabilities and Tackling Technical Hurdles. IEEE Access. PP. 1-1. DOI: 10.1109/ACCESS.2023.3339553

[6] M. R. Lima et al., "Discovering Behavioral Patterns Using Conversational Technology for In-Home Health and Well-Being Monitoring," in IEEE Internet of Things Journal, vol. 10, no. 21, pp. 18537-18552, 1 Nov.1, 2023, doi:10.1109/JIOT.2023.3290833

[7] A. Elragal, A. I. Awad, I. Andersson, and J. Nilsson, "A Conversational AI Bot for Efficient Learning: A Prototypical Design," in IEEE Access, vol. 12, pp. 154877-154887, 2024, doi:10.1109/ACCESS.2024.3476953

[8] [8] A. Ghosh and K. Deepa, "QueryMintAI: Multipurpose Multimodal Large Language Models for Personal Data," in IEEE Access, vol. 12, pp. 144631-144651, 2024, doi: 10.1109/ACCESS.2024.3468996

[9] D. Park, G.-t. An, C. Kamyod and C. G. Kim, "A Study on Performance Improvement of Prompt Engineering for Generative AI with a Large Language Model," in Journal of Web Engineering, vol. 22, no. 8, pp. 1187-1206, November 2023, doi: 10.13052/jwe1540-9589.2285

[10] A. Koubaa, W. Boulila, L. Ghouti, A. Alzahem, and S. Latif, "Exploring ChatGPT Capabilities and Limitations: A Survey," in IEEE Access, vol. 11, pp. 118698-118721, 2023, doi: 10.1109/ACCESS.2023.3326474

[11] A. Alparslan, "The Role of Accuracy and Validation Effectiveness in Conversational Business Analytics," in IEEE Access, vol. 13, pp. 29279-29291, 2025, doi: 10.1109/ACCESS.2025.3540975

[12] [12] Z. Bai and Y. Bai, "Exploring the Role of CLIP Global Visual Features in Multimodal Large Language Models," ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hyderabad, India, 2025, pp. 1-5, doi: 10.1109/ICASSP49660.2025.10889200.

[13] Z. Bai and Y. Bai, "Improving Multimodal Large Language Models through Combining Resampler and MLP Projections," ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hyderabad, India, 2025, pp. 1-5, doi: 10.1109/ICASSP49660.2025.10888340

[14] Q. Chen, X. Yao, H. Ye, and Y. Hong, "Enhancing 3D Medical Image Understanding with 2D Multimodal Large Language Models," ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hyderabad, India, 2025, pp. 1-5, doi: 10.1109/ICASSP49660.2025.10889731

[15] M. Kang, X. Zhang, F. Wei, S. Xu, and Y. Liu, "Enhancing Image Editing with Chain-of-Thought Reasoning and Multimodal Large Language Models," ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hyderabad, India, 2025, pp. 1-5, doi: 10.1109/ICASSP49660.2025.10890562.

[16] M. Puerta-Beldarrain, O. Gómez-Carmona, R. Sánchez-Corcuera, D. Casado-Mansilla, D. López-de-Ipiña and L. Chen, "A Multifaceted Vision of the Human-AI Collaboration: A Comprehensive Review," in IEEE Access, vol. 13, pp. 29375-29405, 2025, doi: 10.1109/ACCESS.2025.3536095

[17] G. Rocchietti, C. Rulli, F. Maria Nardini, C. Ioana Muntean, R. Perego, and O. Frieder, "ChatGPT Versus Modest Large Language Models: An Extensive Study on Benefits and Drawbacks for Conversational Search," in IEEE Access, vol. 13, pp. 15253-15271, 2025, doi:10.1109/ACCESS.2025.3529741