SCIENTIA PLENA VOL. 21, NUM. 01 2025
www.scientiaplena.org.br doi: 10.14808/sci.plena.2025.XXXXXX

Exploring Multimodal Models: How Text, Audio, and Image Combine in AI

Explorando Modelos Multimodais: Como Texto, Áudio e Imagem se Combinam em IA

E. S. Dias; J. E. M. Apóstolo; J. M. R. dos Santos; M. E. P. P. dos Santos; U. J. Cavalcante

Department of Computing, Federal University of Sergipe, 49100-000, São Cristóvão-Sergipe, Brazil

edgarsz@academico.ufs.br
joao.apostolo@dcomp.ufs.br
matheus.ribeiro@dcomp.ufs.br
eduardapossari@academico.ufs.br
ulisses.cavalcante@dcomp.ufs.br

This study investigates the evolution of Large Language Models (LLMs), such as OpenAI's ChatGPT and Google's Gemini, which combine text, image, and audio to generate more accurate and coherent outputs. The article focuses on how these models process and integrate information from different modalities to solve complex tasks, such as generating descriptions and understanding multimodal contexts. To this end, experiments will be conducted to evaluate textual coherence, multimodal accuracy, and response time. As a result, the findings contribute to identifying approaches that can enhance the efficiency of these models in interpreting and combining different types of data, such as text, image, and audio.
Keywords: Multimodal models, modality integration, LLM.

Este estudo investiga a evolução dos Modelos de Linguagem de Grande Escala (do inglês, Large Language Models ou LLM), como o ChatGPT da OpenAI e o Gemini da Google, que combinam texto, imagem e áudio para gerar saídas mais precisas e coerentes. O artigo foca em como esses modelos processam e integram informações de diferentes modalidades para resolver tarefas complexas, como a geração de descrições e a compreensão de contextos multimodais. Para isto, serão realizados experimentos para avaliar a coerência textual, a acurácia multimodal e o tempo de resposta. Como consequência, os resultados obtidos contribuem para a identificação de abordagens que possam aprimorar a eficiência dos modelos ao interpretar e combinar diferentes tipos de dados, como texto, imagem e áudio.
Palavras-chave: Modelos multimodais, integração de modalidades, LLM.

1. INTRODUCTION

Lately, Artificial Intelligence (AI) has made significant progress and is increasingly embedded in everyday human life, with Large Language Models standing out as one of the field's main innovations. These models were initially developed to handle only textual data. However, recent advances in LLMs have driven the development of systems capable of understanding and generating information in multiple formats, including text, image, and audio. These multimodal models act as integrative systems, capable of correlating different data modalities to produce more coherent, accurate, and contextually relevant responses. This capability results from training on large volumes of data composed of interconnected representations of these modalities, enabling the modeling of relationships between visual, auditory, and textual elements.

Among the most advanced multimodal models today are OpenAI's ChatGPT and Google's Gemini, both capable of handling different types of input (known as prompts), such as text, image, or audio, and generating increasingly accurate, informative, and contextually coherent

responses (known as completions). For example, in the ChatGPT-4 demo, the model was able to transform a handwritten sketch of a website into a functional website within minutes, showcasing its impressive ability to interpret and generate across multiple media formats. These capabilities open up a range of practical applications, from the creation of immersive virtual environments to the enhancement of accessibility for individuals with visual or hearing impairments, as well as potential contributions in the medical field, such as aiding in disease diagnosis.

This article investigates how large language models process and integrate information from different modalities to solve complex tasks, such as multimodal context understanding and description generation. To achieve this, experiments will be conducted involving pre-trained models—namely ChatGPT and Gemini—applied to three main scenarios. First, the models' ability to handle multimodal input will be assessed. In this scenario, an image and an audio clip of the same scene will be provided, with the task of generating a textual summary. Subsequently, modality transfer will be evaluated. In this case, only an image will be presented in the prompt, and the LLM will be expected to produce an audio description of the given scene. Finally, the models' ability to recognize context across multiple prompts will be measured. In this scenario, a sequence of images accompanied by audio descriptions will be presented, and the objective is to assess the models' aptitude for recognizing the narrative progression of the inputs.

The results will be analyzed based on three main metrics: textual coherence, which assesses the consistency and accuracy of the generated descriptions—this metric is subjective and will be rated by the authors on a scale from 1 to 5; multimodal accuracy, which compares the model's performance across different input combinations (text only, text and image, text and audio); and response time, which measures the efficiency in processing the requests.

Thus, this study aims to understand the dynamics of multimodal knowledge transfer in language models and explore its implications for future applications. It is expected that, with the continued improvement of these systems, new forms of interaction and information processing will become possible, promoting significant advancements in the way humans and machines communicate.

2. METHODOLOGY

Given the requirements imposed by the comparative scope of the current work, some large-scale pre-trained language models with multimodal capabilities—able to handle image, text, and audio processing—were essential for conducting the experiments that will be described later in this paper. In this regard, the following models were selected: ChatGPT-4o, Gemini 1.5 Pro, and a model from the VOSK tool.

The Transformer architecture, which serves as the foundation for some of the selected models, was introduced in the paper "Attention Is All You Need", published by Google (2017) [5]. Initially, the Transformer was designed for translation tasks but revolutionized natural language processing (NLP) by enabling the retention of context with even larger inputs. Its main components are the encoder, responsible for processing the input into an intermediate representation, and the decoder, tasked with generating the output—both of which can be repeated n times. A layer responsible for the final text processing, the language modeling head, is also part of this architecture.

The three units mentioned have some important subcomponents that deserve brief mention. In the encoder, we have the multi-head attention mechanism, which allows the relationship between words in a sentence to be weighted considering multiple distinct perspectives; the feed-forward layer, which enables better text comprehension by processing information extracted from attention mechanisms; normalization, keeping outputs in a format compatible

with the next layer; and residual connections, ensuring the initial context is not lost during processing.

In the decoder, we have the masked multi-head attention, which plays a role similar to multi-head attention but with the embeddings—mapping inputs to numerical sequences capable of representing various features, such as word semantics, position within the text, and context—of words to be generated later being masked; the encoder-decoder attention, responsible for identifying the most relevant parts of the original input to generate the next word; and the three other features also present in the encoder.

As for the language modeling head, it consists of a linear layer, responsible for mapping the decoders' outputs to the model's vocabulary, and a softmax layer, which, based on the scores assigned by the linear layer, provides the probabilities of words being next in the sequence.

GPT-4o is an OpenAI language model based on a variation of the Transformer architecture (decoder-only, meaning it uses only the decoder of the original architecture, without the encoder) and trained on a vast amount of text—extracted from various internet sources, such as websites, books, and scientific papers—as well as visual and auditory data. As an innovation compared to previous versions, it can natively handle multimodal inputs, processing text, images, audio, and videos.

From Google's Gemini model line, the Gemini 1.5 Pro version was developed as a direct competitor to GPT-4 and thus shares many similarities with the previously described model. It is also natively multimodal, with high capabilities in understanding and generating text, images, audio, and videos, and uses a Transformer-based architecture (decoder-only).

Considering the VOSK automatic speech recognition tool by Alphacephei, we specifically used a small Portuguese-language model, "vosk-model-small-pt-0.3", recommended for limited tasks in applications designed for mobile devices.

To conduct the tests—which will be detailed later in this work—and obtain the necessary data for analyzing the performance of the aforementioned resources, we developed some algorithms. Figure 1 presents the algorithm designed to evaluate the performance and accuracy of multimodal models when integrating image and audio information.

Algorithm 1 Multimodal Processing with Accuracy Evaluation

function MULTIMODALINPUT($image$, $audio$)
  $transcription \leftarrow$ TRANSCRIBEAUDIO($audio$)
  $text \leftarrow$ LIMITTEXT($transcription$)
  $prompt \leftarrow$ “Here is an audio description: ” $+text +$ “ Based on this description and the provided image, generate a textual summary.”
  Start timer $T_{ChatGPT}$
  $response_{ChatGPT} \leftarrow$ GENERATETEXTOPENAI($image$, $prompt$)
  Stop timer $T_{ChatGPT}$
  Start timer T_{Gemini}
  $response_{Gemini} \leftarrow$ GENERATETEXTGOOGLE($image$, $prompt$)
  Stop timer T_{Gemini}
  $accuracy_{ChatGPT} \leftarrow$ CALCULATESEMANTICSIMILARITY($text$, $response_{ChatGPT}$)
  $accuracy_{Gemini} \leftarrow$ CALCULATESEMANTICSIMILARITY($text$, $response_{Gemini}$)
  Display responses, times, and accuracies
end function

Figure 1: Algorithm for evaluating multimodal processing accuracy. Source: Prepared by the authors.

To assess the models' ability to convert visual information into precise textual descriptions—useful for audio synthesis—the algorithm illustrated in Figure 2 was developed.

Algorithm 2 Modality Transfer: Image to Audio

```
function ModalityTransfer(image, reference_description)
    prompt ← "Describe this image thoroughly so it can be converted into audio format"
    Start timer T_ChatGPT
    description_ChatGPT ← GenerateTextOpenAI(image, prompt)
    Stop timer T_ChatGPT
    accuracy_ChatGPT ← CalculateSemanticSimilarity(reference_description, description_ChatGPT)
    Start timer T_Gemini
    description_Gemini ← GenerateTextGoogle(image, prompt)
    Stop timer T_Gemini
    accuracy_Gemini ← CalculateSemanticSimilarity(reference_description, description_Gemini)
    Display responses, times, and accuracies
end function
```

Figure 2: Algorithm for evaluating modality transfer. Source: Prepared by the authors.

Another key element of the investigation is illustrated in Figure 3, which outlines a sequence of instructions to test the model's ability to integrate and contextualize multimodal inputs (image + audio) into a coherent narrative.

Algorithm 3 Multimodal Context Evaluation of a Model

```
procedure CONTEXT_EVALUATION(image_audio_pairs, reference_evaluation)
    Initialize lists: transcribed_texts ← [], image_bytes ← []
    for each (image, audio) in image_audio_pairs do
        text ← TranscribeAudio(audio)
        Append (image, audio) to transcribed_texts
        Read and append image bytes to image_bytes
        Display transcribed text
    end for
    Initialize messages list with system instruction
    for each (image_bytes, text) in (image_bytes, transcribed_texts) do
        Add multimodal entry to messages list
    end for
    Add final question about narrative progression
    Send messages to the model
    Display response and response time
    Compute and display accuracy using reference_evaluation
end procedure
```

Figure 3: Algorithm for multimodal context evaluation of a model. Source: Prepared by the authors.

Regarding the main libraries and frameworks used in development, the following technologies can be mentioned: the gTTS (Google Text-to-Speech) library, enabling

text-to-audio conversion through the Google Translate interface; and PyTorch, which provides a dynamic approach for building neural networks.

Concerning datasets, since all selected models were pre-trained, no external datasets were used for additional training. Moreover, the development and testing environment was Google Colab, using a machine with an NVIDIA T4 GPU and 16 GB of RAM.

Additionally, a GitHub repository - a platform for hosting source code and version-controlled files using Git - was created to support this article. The repository access link is https://github.com/JoaoEmanuel14/ufs-ia-trabalho-a2, containing the source code developed for this work, resources used by the code, and important information about the entire study.

3. EXPERIMENTS

The experiments were implemented in the Google Colab environment to achieve greater control over input prompts and LLM completions. This approach was also necessary to accurately evaluate metrics for textual coherence, multimodal accuracy, and response time.

In the first stage, helper functions were created to process multimodal inputs for GPT-4o and Gemini 1.5. In the second stage, specific functions were developed to test multimodal input, modality transfer, and context evaluation.

Before explaining the experiments, it's important to note that large language models don't directly receive and analyze audio. Therefore, audio inputs were transcribed and passed to the architectures as text. Similarly, for the modality transfer experiment, visual scenes were converted to text, which was then used to generate audio.

Furthermore, the images used in the third experiment were generated by ChatGPT-4o based on a narrative about a cat named Tobias who, while walking through his neighborhood, finds an umbrella that transports him to another world through a portal. The complete narrative can be found in the repository dedicated to this article. The audio files used in this same experiment were generated via gTTS, using each image's filename as the base.

3.1 Helper Functions

The helper functions were organized into blocks with similar functionalities for both ChatGPT-4o and Gemini 1.5 Pro, aiming to facilitate understanding of the actions and code readability. The first function block was created to process image inputs along with text, which were used in the first experiment. The second block handled text-to-audio conversion using OpenAI's TTS and Google's gTTS. The third and fourth blocks converted ".mp3" audio to ".wav" format for transcription by the "vosk-model-small-pt-03" model. The fifth block organized image-audio tuples for processing in the final experiment. Finally, the sixth block contained functions responsible for estimating multimodal accuracy in the experiments.

3.2 Performance Metrics

Regarding the specific metrics, textual coherence was subjectively evaluated by the authors on a scale from 1 to 5. Multimodal accuracy was measured using the Sentence Transformer (SBERT) library with the multilingual "paraphrase-multilingual-MiniLM-L12-v2" model, which compares text similarity through embeddings and returns values between 0 and 1. A value of 1 indicates semantically identical texts, while scores above 0.7 suggest high similarity between compared texts. Finally, response time was measured using the time library.

3.3 Multimodal Input

The multimodal input provides LLMs with an image and an audio clip describing a scene, asking the models to generate a textual summary based on them. The function takes the scene and the recording, converts it from ".mp3" to ".wav" format, transcribes it, creates a prompt with the audio transcription, and sends both the prompt and the image to today's most advanced

multimodal models. Between these steps, the response time of both models is measured, and multimodal accuracy is estimated.

To conduct the experiment, an audio clip was generated using gTTS with the sentence: "A beautiful beach. It has palm trees and crystal-clear water." This audio was then sent along with a beach scene to the function.

3.4 Modality Transfer

Modality transfer provides language models with an image and asks them to generate a detailed description of it. This description is then converted into an audio clip. The representation is passed to OpenAI's TTS to generate and save the voice recordings. The image provided to the large language models is the same as in the multimodal input—a beach.

To ensure the experiment's success, the following prompt was sent to the architectures: "Describe this image in detail so that it can be converted into an audio format." Additionally, the response time and multimodal accuracy of each model were calculated, and the audio clips were generated using gTTS.

3.5 Context Evaluation

Context evaluation was divided into separate functions for each LLM, as data processing differs from previous functions, and the functions themselves are more extensive. This approach was taken to facilitate understanding for both programmers and reviewers. The functions receive tuples of images and their corresponding audio descriptions, then ask the models to determine whether there is a logical narrative progression across the provided tuples.

Thus, each language model architecture receives tuples of scenes and recordings, transcribes the audio for the LLMs, and initiates a prompt instructing the natural language AIs—acting as narrative analysis experts—to evaluate whether there is a logical narrative progression between the transcribed scenes and voice recordings. The sequence of images and transcribed texts is attached and sent, concluding with a final question asking whether a coherent narrative exists across the sequence of tuples.

4. RESULTS AND DISCUSSION

It is undeniable that the experiments carried out were of utmost importance, as they allowed for the analysis of the results, thus serving as inspiration for the discussion. Firstly, as a result of the multimodal input experiment, OpenAI's large language model generated the following completion: “The image depicts a paradisiacal beach with light, fine sand, surrounded by palm trees that create a relaxing shade. The water is crystal clear and vibrantly blue, blending with the clear, almost cloudless sky. In the background, hills covered in lush vegetation can be seen, contributing to a tropical and serene setting.” This output was rated a 5 for textual coherence by the authors responsible for this study. Additionally, it was generated in 3.16 seconds and has an accuracy score of 0.838.

Furthermore, Google’s text generation system produced the following completion: “A serene tropical beach is portrayed, with leaning palm trees casting soft shadows on the pristine white sand. Crystal-clear turquoise waters gently lap the shore, meeting the lush greenery of a distant hill. The sky is painted with a bright blue, dotted with fluffy white clouds.” This was also rated a 5 for textual coherence. This response was generated in 4.70 seconds and has an accuracy score of 0.754.

From these results, it is clear that both architectures perform excellently. Both presented high and very similar accuracy scores. Moreover, they achieved the highest possible rating in textual coherence, representing the context presented to them in a well-described, concise, and cohesive manner. Given these facts, either of the two large language models performs satisfactorily in processing multimodal input. Therefore, it is known that providing multimodal input to LLMs results in satisfactory, complete, and appropriate responses.

Subsequently, regarding the experiment related to modality transfer, GPT-4o's result for the audio description was: "The image depicts a stunning tropical beach. In the foreground, there are a series of slender palm trees, whose green leaves lean toward the beach, casting shadows over the sand. The sand is light and fine, stretching toward the sea. On the right, the sea displays a gradient of turquoise-blue tones, with small waves gently foaming at the shore. The sky is clear, with a few scattered white clouds, and in the background, a hill covered in dense tropical vegetation rises gently, completing the paradisiacal setting." This was rated a 4 in textual coherence. The response time was 5.38 seconds, and the multilingual model "paraphrase-multilingual-MiniLM-L12-v2" judged its accuracy as 0.807.

On the other hand, Gemini's response for the description was: "The image presents an idyllic tropical scene, akin to a paradise. The foreground is dominated by a beach of immaculate white sand stretching into the frame, inviting the viewer for a walk. The irregular shadows of imposing palm trees are cast upon the sand, adding depth and texture. The lower-left corner is filled with a dense collection of palm trees, their trunks gracefully leaning toward the ocean. Their vibrant emerald-green leaves contrast vividly with the white sand and turquoise-blue water. The sea occupies the central and right parts of the image, showcasing a stunning range of blue shades, from light turquoise near the shore to deep blue farther out. Small gentle waves softly break on the beach, creating a white line of foam along the water's edge. In the background, a lush green hill rises from the sea, providing a serene backdrop to the scene. Covered in dense vegetation, the hill adds a touch of mystery and wild magnificence. The sky above is a vivid bright blue, speckled with fluffy white clouds that add depth and texture to the tropical setting. The overall light is bright and sunny, reinforcing the feeling of warmth and tranquility in the scene. This image evokes a sense of serenity, natural beauty, and the allure of a tropical paradise." This was rated a 5 in textual coherence. The response time was 6.10 seconds, and its multimodal accuracy was judged to be 0.794.

Based on the responses given by the LLMs, it is notable that, once again, both perform the task spectacularly, with high scores in accuracy and coherence. However, they differ in the time taken to process the data and generate the completion. ChatGPT, in its 4o version, was able to generate the result faster than Gemini 1.5 Pro. Nevertheless, it is evident that Google's model fulfilled the task more effectively, generating an output that more faithfully aligned with what was requested from the two language-based AI assistants. Thus, Google's large-scale language model proved to be the best choice for the modality transfer task. Therefore, providing an image to a natural language AI and requesting a detailed representation in audio format results in outputs that describe the scene excellently.

Finally, when observing the context evaluation experiment, it is seen that both GPT-4o and Gemini 1.5 Pro reported that there is a logical narrative within the tuples given to them, thus receiving a score of 5 in textual coherence. Both large language models were able to correctly identify all key points of the narrative originating from the tuples. However, the architectures differ in response time and accuracy, although the results were quite close. OpenAI's system had a response time of 10.30 seconds and an accuracy of 0.749, while Google's system returned results in 9.39 seconds, with a multimodal accuracy of 0.728.

Using the information gathered as a starting point, it is evident that both AI assistants perform the context evaluation task satisfactorily. Both identified a logical narrative progression from the tuples and described the story's entire context perfectly. Thus, it is clear that using LLMs for identifying narratives from multimodal inputs is an efficient task that produces satisfactory outputs.

5. CONCLUSION

In summary, based on the conducted experiments, it is evident that both AI language assistants - GPT-4o and Gemini 1.5 Pro - achieve satisfactory results when interpreting and generating content from multimodal inputs, demonstrating high accuracy and textual coherence. Specifically in the multimodal input experiment, the closely matched and satisfactory values confirm that both models are competent at integrating images and audio to produce cohesive summaries.

However, in modality transfer tasks, divergent factors placed the models on different performance levels. While GPT-4o completed tasks faster, its outputs were less aligned with requirements compared to Google's model, which proved more efficient at transposing information across different modalities.

The context evaluation revealed that both models can identify narrative progression when receiving image-audio description pairs, despite significant response time differences. This performance gap suggests the potential for even greater divergence in more complex scenarios with additional media inputs. Notably, GPT-4o delivered more satisfactory responses in this particular experiment.

For future development, training an LLM capable of processing text, audio, and images simultaneously would eliminate the need for audio-text conversion steps and could itself become an evaluation metric. Additionally, enhancing semantic evaluation methods by developing custom functions to capture deeper response quality nuances would enable testing with more complex inputs and contexts, better exploring the limits and potential of large language models.

6. BIBLIOGRAPHIC REFERENCES

1. YANG, Yujie et al. From Large Language Models to Large Multimodal Models. Applied Sciences, [S.l.], v. 14, n. 12, p. 5068, 2024. MDPI. DOI: 10.3390/app14125068. Available at: https://www.mdpi.com/2076-3417/14/12/5068. Accessed on: 2 Apr. 2025.
2. OPENAI. GPT-4 Technical Report. OpenAI, 2023. Available at: https://openai.com/research/gpt-4. Accessed on: 2 Apr. 2025.
3. PICHAR, Sundar. Introducing Gemini: our largest and most capable AI model. Google Official Blog, 6 Dec. 2023. Available at: https://blog.google/technology/ai/google-gemini-ai/. Accessed on: 2 Apr. 2025.
4. OPENAI. GPT-4 Developer Livestream. YouTube, 2023. Available at: https://www.youtube.com/watch?v=outcGtbnMuQ. Accessed on: 2 Apr. 2025.
5. VASWANI, Ashish; SHAZEER, Noam; PARMAR, Niki; USZKOREIT, Jakob; JONES, Llion; GOMEZ, Aidan N.; KAISER, Lukasz; POLOSUKHIN, Illia. Attention is all you need. arXiv preprint arXiv:1706.03762, 2017. Available at: https://arxiv.org/abs/1706.03762. Accessed on: 8 Apr. 2025.
6. Moveworks. Multimodal Language Models Explained: The next generation of LLMs. YouTube. 2 May 2023. Available at: https://www.youtube.com/watch?v=-m4n3lsCtcA. Accessed on: 2 Apr. 2025.
7. AssemblyAI. How do Multimodal AI models work? Simple explanation. YouTube. 5 December 2023. Available at: https://www.youtube.com/watch?v=WkoytlA3MoQ&t=203s. Accessed on: 2 Apr. 2025.
8. CAPITELLA, Donato. LLM Chronicles #6.3: Multi-Modal LLMs for Image, Sound and Video. YouTube. 1 July 2024. Available at: https://www.youtube.com/watch?v=_sGwL6RAsUc. Accessed on: 2 Apr. 2025.