



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

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

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

Departamento de Computação, Universidade Federal de Sergipe - UFS, 49100-000, São Cristóvão-Sergipe, Brasil

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

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.

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.

1. INTRODUÇÃO

Ultimamente, a Inteligência Artificial (IA) têm avançado significativamente e está cada vez mais inserida no cotidiano da humanidade, com os grandes modelos de linguagem se destacando como uma das principais inovações da área. Tais modelos, inicialmente foram desenvolvidos para lidar apenas com dados textuais. Porém, os recentes avanços nos LLMs têm impulsionado o desenvolvimento de sistemas capazes de compreender e gerar informações em múltiplos formatos — incluindo texto, imagem e áudio — dando origem aos modelos multimodais, que integram diferentes modalidades de dados, assim como é descrito em YANG et al. (2024) [1]. Esses modelos multimodais atuam como sistemas integradores, capazes de correlacionar diferentes modalidades de dados para produzir respostas mais coerentes, precisas e contextualizadas. Tal capacidade é resultado de treinamentos realizados em grandes volumes de dados compostos por representações interligadas dessas modalidades, o que possibilita a modelagem das relações entre elementos visuais, sonoros e textuais.

Dentre os modelos multimodais mais avançados da atualidade, destacam-se o *ChatGPT* da empresa *OpenAI* 2023 [2], conforme demonstrado em um artigo no site da empresa, e o *Gemini* da companhia *Google* 2023 [3], como pode ser observado em um blog no website da empresa. Ambos são capazes de lidar com diferentes tipos de entrada (chamadas de *prompts*), sendo texto, imagem ou áudio e gerar respostas (conhecidas como *completions*) cada vez mais precisas, informativas e contextualmente coerentes. Por exemplo, foi demonstrado em uma *livestream* da *OpenAI* (2023) [4] que na *demo* do *ChatGPT-4*, o modelo conseguiu transformar uma imagem de um rascunho de site manuscrito em um site funcional em questão de minutos, evidenciando sua grande capacidade de interpretar e gerar diferentes formatos de mídia. Essas habilidades abrem uma gama de aplicações práticas, desde a criação de ambientes virtuais imersivos até o aprimoramento da acessibilidade para pessoas com deficiências visuais ou auditivas, além de potenciais contribuições na área médica, como no auxílio do diagnóstico de doenças.

Neste artigo, será investigado como os *large language models* processam e integram informações provenientes de diferentes modalidades para resolver tarefas complexas, como a compreensão de contexto multimodal e a geração de descrições. Para isso, serão realizados experimentos envolvendo modelos pré-treinados, neste caso, o *ChatGPT* e o *Gemini*, aplicados a três cenários principais. Primeiramente, será a capacidade dos modelos de lidarem com uma entrada multimodal. Neste cenário, uma imagem e um áudio de uma mesma cena serão fornecidos, solicitando que os modelos gerem um resumo textual. Subsequentemente, a transferência de modalidade será avaliada. Sob tal situação, apenas uma imagem será disponibilizada no *prompt* para o LLM e este deverá produzir um áudio descrevendo a cena fornecida. Por fim, será aferida a capacidade das arquiteturas de reconhecer o contexto de vários *prompts*. Nesta conjuntura, serão apresentadas uma sequência de imagens acompanhadas por descrições de áudio e o objetivo é investigar a aptidão dos modelos de reconhecer a progressão narrativa das entradas.

A análise dos resultados será conduzida com base em três métricas principais. A coerência textual, que avalia a consistência e precisão das descrições geradas; a acurácia multimodal, que compara o desempenho do modelo frente a diferentes combinações de entrada (apenas texto, texto e imagem, texto e áudio); e o tempo de resposta, que mede a eficiência no processamento das solicitações.

Destarte, este estudo busca compreender a dinâmica da transferência de conhecimento multimodal em modelos de linguagem e explorar suas implicações para aplicações futuras. Espera-se que, com o contínuo aprimoramento desses sistemas, novas formas de interação e processamento de informações sejam possibilitadas, promovendo avanços significativos na forma como seres humanos e máquinas se comunicam.

2. METODOLOGIA

Mediante as necessidades impostas pelo âmbito comparativo do trabalho atual, alguns modelos de linguagem de grande escala pré-treinados, com a capacidade multimodal, podendo lidar com o processamento de imagens, textos e áudio, foram essenciais para a realização das experimentações que serão descritas posteriormente neste artigo. Nesse sentido, foram selecionados os seguintes modelos: *ChatGPT-4o*, *Gemini 1.5 Pro* e um modelo da ferramenta VOSK.

A arquitetura Transformer, que servirá como base para alguns dos modelos escolhidos, foi introduzida pelo artigo “*Attention Is All You Need*”, publicado pelo *Google* (2017) [5]. Inicialmente, a *Transformer* foi projetada para a realização de tarefas de tradução, mas foi capaz

de revolucionar o processamento de linguagem natural (PLN) ao trazer a possibilidade de reter contextos com entradas ainda maiores. Os elementos principais que a compõem são o codificador, responsável pelo processamento da entrada em uma representação intermediária, e o decodificador, encarregado de gerar a saída, que, por sua vez, podem ser repetidos n vezes. Uma camada responsável pelo processamento final do texto também faz parte desta composição, a *language modeling head*.

As três unidades citadas possuem alguns subcomponentes importantes que merecem uma breve menção. Para o codificador, temos o mecanismo de *multi-head attention*, permitindo que a relação entre as palavras de uma sentença possa ser ponderada considerando múltiplas perspectivas distintas, a *feed-forward*, uma camada que possibilita uma melhor compreensão do texto por meio do processamento das informações extraídas dos mecanismos de atenção, a normalização, mantendo as saídas em um formato compatível com a próxima camada, e as conexões residuais, garantindo que o contexto inicial não seja perdido durante o processamento.

No decodificador, temos o *masked multi-head attention*, que desempenha um papel semelhante ao *multi-head attention*, porém, com os *embeddings* — mapeamento das entradas para sequências numéricas que são capazes de representar diversas características, como a semântica da palavra, sua posição dentro do texto e o seu contexto — das palavras que serão geradas posteriormente mascarados, o *encoder-decoder attention*, responsável por identificar as partes mais relevantes da entrada original para gerar a próxima palavra, além dos três outros recursos também presentes no codificador.

Já na *language modeling head*, temos uma camada linear, encarregada de realizar o mapeamento entre a saída dos decodificadores e o vocabulário do modelo, e uma camada softmax, que, com base nas pontuações atribuídas pela linear, fornece as probabilidades das palavras serem a próxima na sequência.

O *GPT-4o* é um modelo de linguagem da *OpenAI* baseado em uma variação da arquitetura *Transformer* (*decoder-only*, ou seja, como o termo indica, utiliza apenas o decodificador da arquitetura original, sem o codificador) e treinado com uma grande quantidade de textos — extraídos de diversas fontes da internet, como sites, livros e artigos científicos —, dados visuais e sonoros. Como novidade com relação às versões anteriores, consegue lidar com entradas multimodais de forma nativa, podendo processar textos, imagens, áudios e vídeos.

Da linha de modelos *Gemini* do *Google*, a versão *Gemini 1.5 Pro* foi desenvolvida para ser um concorrente direto do *GPT-4* e, portanto, apresenta diversas semelhanças com o modelo descrito anteriormente, também sendo nativamente multimodal, possuindo uma alta capacidade de compreensão e geração de texto, imagens, áudios e vídeos, e que utiliza uma arquitetura baseada em *Transformer* (*decoder-only*).

Considerando a ferramenta de reconhecimento automático de fala VOSK da empresa Alphacephei, utilizamos especificamente um modelo pequeno para o idioma português, o “vosk-model-small-pt-0.3”, recomendado para a realização de tarefas limitadas em aplicações direcionadas para dispositivos móveis.

Para a realização dos testes — que serão detalhados adiante neste trabalho — e obtenção dos dados necessários para a análise do desempenho dos recursos mencionados anteriormente, elaboramos alguns algoritmos. Na Figura 1, é apresentado o algoritmo desenvolvido para avaliar a performance e precisão de modelos multimodais ao integrar informações de imagem e áudio.

Algorithm 1 Processamento Multimodal com Avaliação de Acurácia

```

1: function ENTRADAMULTIMODAL(imagem, audio)
2:   transcricao  $\leftarrow$  TRANSCREVERAUDIO(audio)
3:   texto  $\leftarrow$  LIMITARTEXTO(transcricao)
4:   prompt  $\leftarrow$  “Aqui está uma descrição de áudio: ” + texto + “ Com base
   nesta descrição e na imagem fornecida, gere um resumo textual.”
5:   Iniciar cronômetro  $T_{ChatGPT}$ 
6:   respostaChatGPT  $\leftarrow$  GERARTEXTOOPENAI(imagem, prompt)
7:   Parar cronômetro  $T_{ChatGPT}$ 
8:   Iniciar cronômetro  $T_{Gemini}$ 
9:   respostaGemini  $\leftarrow$  GERARTEXTOGOOGLE(imagem, prompt)
10:  Parar cronômetro  $T_{Gemini}$ 
11:  acuraciaChatGPT  $\leftarrow$  CALCULARSIMILARIDADESEMANTICA(texto,
    respostaChatGPT)
12:  acuraciaGemini  $\leftarrow$  CALCULARSIMILARIDADESEMANTICA(texto,
    respostaGemini)
13:  Exibir respostas, tempos e acurácias
14: end function

```

Figura 1: Algoritmo para a avaliação de acurácia do processamento multimodal. Fonte: Elaborado pelos autores.

Com o objetivo de verificar a capacidade dos modelos em converter informações visuais em descrições textuais precisas, úteis para síntese de áudio, foi elaborado o algoritmo ilustrado pela Figura 2.

Algorithm 2 Transferência de Modalidade: Imagem para Áudio

```

1: function TRANSFERENCIAMODALIDADE(imagem, descricao_referencia)
2:   prompt  $\leftarrow$  “Descreva esta imagem minuciosamente para que esta possa
   ser convertida em formato de áudio”
3:   Iniciar cronômetro  $T_{ChatGPT}$ 
4:   descricaoChatGPT  $\leftarrow$  GERARTEXTOOPENAI(imagem, prompt)
5:   Parar cronômetro  $T_{ChatGPT}$ 
6:   acuraciaChatGPT  $\leftarrow$ 
    CALCULARSIMILARIDADESEMANTICA(descricao_referencia,
    descricaoChatGPT)
7:   Iniciar cronômetro  $T_{Gemini}$ 
8:   descricaoGemini  $\leftarrow$  GERARTEXTOGOOGLE(imagem, prompt)
9:   Parar cronômetro  $T_{Gemini}$ 
10:  acuraciaGemini  $\leftarrow$ 
    CALCULARSIMILARIDADESEMANTICA(descricao_referencia,
    descricaoGemini)
11:  Exibir respostas, tempos e acurácias
12: end function

```

Figura 2: Algoritmo para a avaliação de transferência de modalidade. Fonte: Elaborado pelos autores.

Outro elemento importante para a investigação é evidenciado pela Figura 3, que demonstra uma sequência de instruções para testar a capacidade do modelo de integrar e contextualizar informações multimodais (imagem + áudio) em uma narrativa coerente.

Algorithm 3 Avaliação do contexto multimodal de um modelo

```

1: procedure AVALIACAO_DO_CONTEXTO(pares_imagem_audio,
   avaliacao_referencia)
2:   Inicializar listas: textos_transcritos  $\leftarrow []$ , imagens_bytes  $\leftarrow []$ 
3:   for cada (imagem, audio) em pares_imagem_audio do
4:     texto  $\leftarrow$  TRANSCREVERAUDIO(audio)
5:     Adicionar (imagem, audio) em textos_transcritos
6:     Ler e adicionar bytes da imagem em imagens_bytes
7:     Exibir texto transcrito
8:   end for
9:   Inicializar lista mensagens com instrução de sistema
10:  for cada (imagem_bytes, texto) em
   (imagens_bytes, textos_transcritos) do
11:    Adicionar entrada multimodal à lista mensagens
12:  end for
13:  Adicionar pergunta final sobre progressão narrativa
14:  Enviar mensagens ao modelo
15:  Exibir resposta e tempo de resposta
16:  Calcular e exibir acurácia usando avaliacao_referencia
17: end procedure

```

Figura 3: Algoritmo para a avaliação do contexto multimodal de um modelo. Fonte: Elaborado pelos autores.

Com relação às principais bibliotecas e frameworks que foram utilizados no desenvolvimento, podem ser mencionadas as seguintes tecnologias: a biblioteca gTTS (Google Text-to-Speech), possibilitando a conversão de texto em áudio por meio da interface do Google Tradutor; e o *PyTorch*, que oferece uma abordagem dinâmica na construção de redes neurais.

Quanto aos conjuntos de dados, como todos os modelos selecionados foram pré-treinados, nenhum conjunto de dados externo foi utilizado para a realização de um treinamento posterior. Além do mais, o ambiente de desenvolvimento e de realização de testes foi o *Google Colab*, utilizando uma máquina com a GPU NVIDIA T4 e 16 GB de memória RAM.

Além disso, um repositório no *GitHub*, uma plataforma de hospedagem de código-fonte e arquivos com controle de versão usando o *Git*, foi criado para dar suporte à este artigo. O link para acesso ao repositório é <https://github.com/JoaoEmanuel14/ufs-ia-trabalho-a2>, sendo que este possui o código-fonte desenvolvido para o trabalho, os recursos utilizados pelo código e informações importantes sobre todo o estudo.

3. EXPERIMENTOS

Os experimentos foram implementados no ambiente do *Google Colab*, a fim de obter maior controle sobre os *prompts* de entrada e sobre as *completions* das LLMs. Tal feito também foi necessário para avaliarmos com precisão as métricas de coerência textual, acurácia multimodal e tempo de resposta.

Em primeira instância, foram criadas funções auxiliares para processar as entradas multimodais para o *GPT-4o* e o *Gemini 1.5*. Em segunda instância, foram concebidas as funções

que testam, especificamente, a entrada multimodal, a transferência de modalidade e avaliação do contexto.

Antes da explicação das experimentações, é sabido que os modelos de linguagem de grande escala não recebem áudio e o analisam. Por isso, estes foram transcritos e passados textualmente para as arquiteturas. De maneira similar, na experiência da transferência de modalidade, a cena passada é transformada em texto e um áudio é gerado a partir dele.

Outrossim, as imagens utilizadas no terceiro ensaio foram geradas pelo *ChatGTP-4o* a partir de uma narrativa que, em suma, conta a história de um gato chamado Tobias que, ao andar por seu bairro encontra um guarda-chuva que, através de um portal, o transporta para outro mundo. A narrativa completa pode ser vista no repositório dedicado a este artigo. Os áudios utilizados neste mesmo experimento foram gerados via gTTS, tendo como base o nome do arquivo de cada imagem.

3.1. Funções Auxiliares

As funções auxiliares foram divididas em blocos de funções que tinham as mesmas funcionalidades para o *ChatGTP-4o* e *Gemini 1.5 Pro*, com o objetivo de facilitar a compreensão das ações e leitura do código. Dessa maneira, o primeiro bloco de funções foi criado para processar a entrada de uma imagem, juntamente de um texto. Tais funções serão utilizadas na primeira experimentação. O segundo bloco foi usado para converter texto em áudio, utilizando o TTS da *OpenAI* e o gTTS da *Google*. O terceiro e quarto blocos servem para converter o áudio “.mp3” para “.wav” para que o “vosk-model-small-pt-03” possa transcrever os áudios concebidos pelos modelos de linguagem baseados em IA. O quinto bloco de funções auxiliares serve para organizar a tupla de imagem e áudio, para que esta possa ser enviada para processamento na última experimentação. Enfim, o sexto bloco é caracterizado por representar as funções encarregadas de estimar a acurácia multimodal nos experimentos.

3.2 Métricas de Desempenho

Sobre a especificação de cada métrica, a coerência textual tem por principal característica ser uma avaliação subjetiva determinada pelos autores deste trabalho, podendo ser medida de 1 a 5. A acurácia multimodal será especificada através da função biblioteca *Sentence Transformer* (ou SBERT), com o uso do modelo multilíngue “paraphrase-multilingual-MiniLM-L12-v2”. Esta, usa a similaridade do texto com *embeddings* para os comparar. Assim, retorna como resultado um valor entre 0 e 1. O valor 1 indica que os textos são idênticos semanticamente e uma similaridade acima de 0.7 indica que existe uma alta semelhança entre os textos comparados. Finalmente, o tempo de resposta é estimado com o uso da biblioteca *time*.

3.3 Entrada Multimodal

A entrada multimodal entrega para as LLMs uma imagem e um áudio descrevendo uma cena e pede que os modelos criem um resumo textual baseado nisso. A função recebe a cena e a gravação, a transforma do formato de arquivo “.mp3” para a formatação “.wav”, a transcreve, cria o *prompt* com essa transcrição de som e envia o *prompt* com a imagem para os modelos multimodais mais avançados da atualidade. Entre estes procedimentos, ocorre o cálculo do tempo de resposta de ambos os modelos e a estimativa da acurácia multimodal.

Para realizar o experimento, foi criado um áudio usando gTTS com a sentença “Uma praia bonita. Possui coqueiros e água cristalina.”, e o enviamos juntamente da cena de uma praia para a função.

3.4 Transferência de Modalidade

A transferência de modalidade entrega aos modelos de linguagem uma imagem e pede que eles criem uma descrição minuciosa da mesma. Desse jeito, a cena pode ser convertida em um áudio que a descreva. De maneira que, a representação é passada para o TTS da *OpenAI* para

gerar os registros de voz e salvá-los. A imagem passada para os *large language models* é a mesma da entrada multimodal, uma praia.

Assim, para que o experimento tivesse sucesso, o seguinte *prompt* “Descreva esta imagem minuciosamente para que esta possa ser convertida em formato de áudio” foi enviado para as arquiteturas. Além disso, o tempo de resposta e a acurácia multimodal de cada uma foi calculado e os áudios foram gerados a partir do gTTS.

3.5 Avaliação de Contexto

A avaliação de contexto foi dividida em uma função para cada LLM, já que a passagem de dados é feita de forma distinta do que nas funções anteriores, ademais, as funções são maiores. Assim, a fim de facilitar a compreensão, tanto para os programadores, quanto para examinadores da função, essa ação foi tomada. A função recebe tuplas de imagens e áudios, que descrevem estas imagens, e pede que os modelos identifiquem se existe alguma progressão narrativa lógica nas tuplas fornecidas.

Destarte, cada arquitetura de modelo de linguagem recebe tuplas de cenas e gravações, transcreve esses áudios para as LLMs, inicia o *prompt* pedindo para que as IAs de linguagem natural, assumindo o papel de especialistas de análise narrativa, avaliem se há uma progressão narrativa lógica entre as cenas e gravações de voz transcritos, anexa a sequência de imagens e textos transcritos e envia, finalizando com uma pergunta final declarando se existe ou não uma narrativa congruente no encadeamento de tuplas.

4. RESULTADOS E DISCUSSÃO

É indubitável que os experimentos feitos foram de suma importância pois, a partir deles, a análise dos resultados pode ser feita, servindo assim de inspiração para a discussão. Primeiramente, como resultado da experiência de entrada multimodal, a *large language model* da *OpenAI* gerou como *completion* o texto “A imagem retrata uma praia paradisíaca com areia clara e fina, cercada por coqueiros que criam uma sombra relaxante. A água é cristalina e de um azul vibrante, misturando-se com o céu claro e quase sem nuvens. Ao fundo, veem-se colinas cobertas por vegetação exuberante, contribuindo para um cenário tropical e sereno.”, que os autores responsáveis por este estudo o classificaram com a nota 5 de coerência textual. Além disso, este foi gerado em 3.16 segundos e possui uma acurácia de 0.838.

Adicionalmente, o sistema de geração de texto da Google resultou na *completion* “Uma serena praia tropical é retratada, com palmeiras inclinadas lançando sombras suaves na areia branca e primitiva. Águas cristalinas e turquesa lambem suavemente a costa, encontrando a exuberante vegetação verde de uma colina ao longe. O céu está pintado com um azul brilhante, pontilhado de nuvens fofas e brancas.”, o qual também foi categorizado com a nota 5 de coerência do texto. Esta resposta foi gerada em 4.70 segundos e possui uma acurácia de 0.754.

A partir destes resultados, é perceptível que ambas as arquiteturas desempenham de maneira excelente. Pois as duas apresentaram uma acurácia alta e bem similar. Ademais, elas obtiveram nota máxima em coerência textual, representando o contexto apresentado a elas de maneira bem descrita, concisa e coesa. Tendo tais fatos em vista, qualquer um dos dois grandes modelos de linguagem desempenha satisfatoriamente o processamento de entrada multimodal. Logo, é sabido que o fornecimento de uma entrada multimodal para as LLMs resulta em uma resposta satisfatória, completa e adequada.

Subsequentemente, tratando-se do experimento relacionado à transferência de modalidade, o resultado do *GPT-4o* para a descrição de áudio é a seguinte: “A imagem retrata uma praia tropical deslumbrante. No primeiro plano, há uma série de palmeiras esguias, cujas folhas verdes inclinam-se em direção à praia, criando sombras sobre a areia. A areia é clara e fina, estendendo-se em direção ao mar. À direita, o mar apresenta um degradê de tons azul-turquesa, com pequenas ondas espumando suavemente na margem. O céu está claro, com algumas nuvens brancas dispersas, e ao fundo, uma colina coberta por vegetação densa e tropical se ergue suavemente, completando o cenário paradisíaco.”, que foi classificada em nota 4 de coerência

textual. O tempo de resposta da mesma foi de 5.38 segundos e o modelo multilíngue “paraphrase-multilingual-MiniLM-L12-v2” julgou uma acurácia de 0.807.

Outrossim, a resposta do *Gemini* para a descrição é: “A imagem apresenta uma cena tropical idílica, semelhante a um paraíso. O primeiro plano é dominado por uma praia de areia branca e imaculada, que se estende para o quadro, convidando o espectador a dar um passeio. As sombras irregulares dos imponentes coqueiros projetam-se sobre a areia, acrescentando profundidade e textura. O canto inferior esquerdo é dominado por uma densa coleção de coqueiros, seus troncos inclinando-se graciosamente em direção ao oceano. Suas folhas verde-esmeralda vibrantes criam um contraste vibrante com a areia branca e o azul turquesa da água. O mar ocupa a parte central e direita da imagem, exibindo uma deslumbrante gama de tons azuis, do turquesa claro próximo à costa até um azul profundo mais adiante. Pequenas ondas suaves quebram suavemente na praia, criando uma linha branca de espuma na borda da água. Ao fundo, uma colina verdejante e exuberante eleva-se do mar, proporcionando um pano de fundo sereno à cena. Coberta por vegetação densa, a colina acrescenta um toque de mistério e magnificência selvagem. O céu acima é um azul brilhante vívido, salpicado de nuvens brancas e fofas que adicionam profundidade e textura ao cenário tropical. A luz geral é brilhante e ensolarada, reforçando a sensação de calor e tranquilidade da cena. Esta imagem evoca uma sensação de serenidade, beleza natural e o fascínio do paraíso tropical.”, que foi ranqueada em nota 5 de coerência do texto. O tempo de resposta para essa resolução foi de 6.10 segundos e foi julgada, por sua acurácia multimodal, em 0.794.

Tendo como base as respostas dadas pelas LLMs, é notável que, novamente, ambas desempenham a tarefa de maneira espetacular, contando com notas altas em acurácia e em coerência. Porém, elas divergem com relação ao tempo gasto para processar os dados e elaborar a *completion*. O *ChatGPT*, em sua versão 4o, foi capaz de gerar o resultado mais rapidamente do que o *Gemini 1.5 Pro*. Apesar disso, é notório que o modelo da *Google* foi capaz de cumprir a tarefa de maneira mais efetiva, gerando uma saída mais fiel ao que foi solicitado aos dois assistentes de IA baseado em linguagem. Dessa maneira, o modelo de linguagem de grande escala da *Google* demonstrou ser a melhor escolha para a tarefa de transferência de modalidade. Portanto, é visto que fornecer uma imagem à uma IA de linguagem natural e pedir por uma representação detalhada da mesma, em formato de voz, resulta em áudios que descrevem a cena de uma maneira ótima.

Finalmente, ao observar a experimentação da avaliação de contexto, é visto que tanto o *GPT-4o*, como o *Gemini 1.5 Pro*, retornaram que existe uma narrativa lógica nas tuplas passadas para eles, recebendo assim nota 5 em coerência textual. Os dois modelos de linguagem de grande escala conseguiram identificar todos os pontos da narrativa, oriundos das tuplas, corretamente. Porém, as arquiteturas divergem com relação ao tempo de resposta e de acurácia, apesar de que os resultados foram bem próximos. O sistema da *OpenAI* teve um tempo de resposta de 10.30 segundos e acurácia de 0.749, já o sistema da *Google* obteve os resultados em 9.39 segundos, com uma acurácia multimodal de 0.728.

Utilizando as informações encontradas como ponto de partida, é notório que os dois assistentes de IA desempenham a tarefa da avaliação de contexto de maneira satisfatória. À medida que, ambos identificaram uma progressão narrativa lógica a partir das tuplas, além de que eles descreveram todo o contexto da história de maneira perfeita. Logo, é evidente que utilizar LLMs para identificação de narrativas a partir de entradas multimodais é uma tarefa eficiente e que gera saídas satisfatórias.

5. CONCLUSÃO

Em síntese, com base nos experimentos realizados, é notável que ambos os assistentes de IA baseado em linguagem, tanto o *GPT-4o*, quanto o *Gemini 1.5 Pro*, conseguem alcançar resultados satisfatórios quando se trata de interpretação e geração de conteúdo ao trabalhar com entradas multimodais, demonstrando alta acurácia e coerência textual. Especificamente no experimento de entrada multimodal, pode-se perceber que, diante da obtenção de valores satisfatórios e bastante próximos, é garantido que os dois modelos são competentes em realizar a integração de imagens e áudios para a elaboração de resumos coesos.

Em contrapartida, ao trabalhar com a transferência de modalidade, houveram fatores divergentes que colocaram os modelos em patamares distintos. Dessa forma, o *GPT-4o*, apesar de necessitar de um tempo menor para cumprir a tarefa, obteve uma saída menos adequada com relação ao que foi requisitado, o que pôs o modelo da Google à frente, sendo mais eficiente para a execução da transposição entre modalidades distintas.

Para a avaliação de contexto foi demonstrado que ambos são capazes de identificar a existência de uma progressão narrativa ao receberem pares de imagens e áudios descritivos, apesar de uma diferença significativa entre as velocidades de resposta, o que demonstra a possibilidade dessa diferença se tornar ainda maior para um contexto com mais imagens e áudios. Destarte, o *GPT-4o* teve uma resposta mais satisfatória para o experimento.

Assim, visando planos futuros, seria viável o treinamento de uma LLM que receba texto, áudio e imagens em conjunto para que não fosse necessário o tratamento de conversão de áudio para texto e texto para áudio, e essa ação também poderia ser avaliada. Também poderíamos aprimorar as avaliações semânticas, criando nossas próprias funções para capturar nuances mais profundas da qualidade das respostas, abrindo a possibilidade para entradas e contextos mais complexos para testar melhor os limites e potencial das *large language models*.

6. REFERÊNCIAS BIBLIOGRÁFICAS

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. Disponível em: <https://www.mdpi.com/2076-3417/14/12/5068>. Acesso em: 2 abr. 2025.
2. OPENAI. GPT-4 Technical Report. OpenAI, 2023. Disponível em: <https://openai.com/research/gpt-4>. Acesso em: 2 abr. 2025.
3. PICHAR, Sundar. Introducing Gemini: our largest and most capable AI model. Blog oficial do Google, 6 dez. 2023. Disponível em: <https://blog.google/technology/ai/google-gemini-ai/>. Acesso em: 2 abr. 2025.
4. OPENAI. GPT-4 Developer Livestream. YouTube, 2023. Disponível em: <https://www.youtube.com/watch?v=outcGtbnMuQ>. Acesso em: 2 abr. 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. Disponível em: <https://arxiv.org/abs/1706.03762>. Acesso em: 08 abr. 2025.
6. Moveworks. Multimodal Language Models Explained: The next generation of LLMs. Youtube. 2 de Maio de 2023. Disponível em: <https://www.youtube.com/watch?v=-m4n3lsCtcA>. Acesso em: 2 de Abril de 2025.
7. AssemblyAI. How do Multimodal AI models work? Simple explanation. Youtube. 5 de Dezembro de 2023. Disponível em: <https://www.youtube.com/watch?v=WkoytLA3MoQ&t=203s>. Acesso em: 2 de Abril de 2025.
8. CAPITELLA, Donato. LLM Chronicles #6.3: Multi-Modal LLMs for Image, Sound and Video. Youtube. 1 de Julho de 2024. Disponível em: <https://www.youtube.com/watch?v=sGwL6RAsUc>. Acesso em: 2 de Abril de 2025.