

# Avanços e Desafios nos Modelos de Linguagem LLaMA e Video-LLaMA

Cristiano de Almeida Tomaz

Instituto Federal de Educação Ciência e Tecnologia de São Paulo]  
Campos do Jordão – SP – Brasil

Departamento de Análise e Desenvolvimento de Sistemas

cristiano.tomaz@aluno.ifsp.edu.br

**Abstract.** *This article explores the advances and challenges of the LLaMA and Video-LLaMA language models, emphasizing their contributions to artificial intelligence and practical applications in various fields. Developed by Meta AI, LLaMA is a text-based language model optimized for efficiency and accessibility, using public datasets to train models of up to 65 billion parameters, aiming to democratize the use of advanced AI in resource-limited contexts. Conversely, Video-LLaMA by Alibaba's DAMO Academy integrates visual and auditory data, offering ideal multimodal processing for complex real-time interactions, such as in security and medical assistance. While LLaMA stands out for its scalability and lightweight nature, Video-LLaMA broadens AI's interactive potential through audiovisual analysis. This study provides a comparative analysis, highlighting the strengths and limitations of each model, and suggests future applications in specific fields.*

**Resumo.** *Este artigo explora os avanços e desafios dos modelos de linguagem LLaMA e Video-LLaMA, destacando suas contribuições para a inteligência artificial e a aplicação prática em diversas áreas. O LLaMA, desenvolvido pela Meta AI, é um modelo de linguagem textual otimizado para eficiência e acessibilidade, usando dados públicos para treinar modelos de até 65 bilhões de parâmetros. Ele visa democratizar o uso de IA avançada em contextos com recursos computacionais limitados. Por outro lado, o Video-LLaMA, criado pela DAMO Academy da Alibaba, integra dados visuais e auditivos, oferecendo um processamento multimodal ideal para interações complexas e em tempo real, como em segurança e assistência médica. Enquanto o LLaMA se destaca pela leveza e escalabilidade, o Video-LLaMA expande as possibilidades de interação com IA através de análise audiovisual. Esse estudo fornece uma análise comparativa, destacando as vantagens e limitações de cada modelo e sugerindo aplicações futuras em áreas específicas.*

## 1. Introdução

Os modelos de linguagem baseados em inteligência artificial têm gerado grande impacto na sociedade nos últimos anos. Seus recursos de automatização de tarefas e geração de informação impressionam pela velocidade e precisão cada vez maiores.

Neste cenário tem destaque o LLaMA e o Video-LLaMA, modelos de linguagem de acesso gratuito.

O LLaMA tem sido projetado para fornecer um desempenho competitivo em tarefas de compreensão e geração de texto, utilizando uma arquitetura eficiente e acessível. E surge como uma alternativa eficiente e aberta, treinada exclusivamente em dados disponíveis publicamente. Projetado com uma arquitetura otimizada e dimensionado entre 7 bilhões e 65 bilhões de parâmetros, o LLaMA busca alcançar um equilíbrio entre alto desempenho e acessibilidade, permitindo sua execução em hardware menos robusto. E assim democratizando o acesso à pesquisa com modelos de linguagem avançados.

O modelo Video-LLaMA é projetado para processar simultaneamente dados visuais e auditivos, permitindo uma compreensão multimodal profunda de vídeos. Sua arquitetura de dois ramos, com processamento separado de quadros e áudio, oferece uma integração eficaz em um espaço de embedding comum. Isso permite que o modelo responda a perguntas contextuais de forma precisa, tornando-o ideal para aplicações interativas complexas, como assistência médica e segurança. Sua maior vantagem está na capacidade de lidar com dados multimodais, aprimorando a compreensão de cenários dinâmicos.

Neste trabalho vamos abordar dois artigos que são referência nos assuntos relacionados a estes modelos de linguagem, que oferecem arquiteturas multimodais capazes de compreender informações textuais, visuais e auditivas.

A análise realiza uma visão geral sobre o modelo LLaMA, um modelo de base eficiente desenvolvido pela Meta AI. E também sobre o Video-LLaMA, da Alibaba DAMO Academy, projetado para capturar dinâmicas audiovisuais em vídeos. Os dois artigos analisados discutem a otimização e eficiência no treinamento e na inferência, propondo soluções para desafios como escalabilidade, diversidade de dados e integração de múltiplas modalidades, consolidando as bases para futuros desenvolvimentos em LLMs.

Os modelos LLaMA e Video-LLaMA, abordados nos artigos de Touvron et al. (2023) e Zhang et al. (2023), respectivamente, representam dois enfoques distintos, mas complementares, no uso de LLMs. O primeiro concentra-se em otimizar a eficiência e acessibilidade dos modelos de linguagem, enquanto o segundo explora a integração de dados visuais e auditivos, ampliando as capacidades dos LLMs em contextos multimodais. Esses avanços buscam não só aprimorar a precisão, mas também promover a democratização e a aplicabilidade prática desses modelos, abordando desafios técnicos e éticos na construção de IA de uso amplo.

## **2. Fundamentação Teórica**

O modelo LLaMA, como descrito por Touvron et al. (2023), foi projetado com o objetivo de maximizar o desempenho em diferentes orçamentos de inferência, atendendo à necessidade de modelos de IA acessíveis e escaláveis. Esse modelo foi treinado com uma abordagem que prioriza o uso de conjuntos de dados públicos e arquiteturas eficientes para reduzir os custos computacionais durante a inferência. Já o Video-LLaMA, como discutido por Zhang et al. (2023), foca na integração de multimodalidades, treinando LLMs para processar simultaneamente dados visuais e auditivos presentes em vídeos. Ao introduzir um método de treinamento multimodal de várias etapas, que inclui a aprendizagem de correspondências entre visão-linguagem e áudio-linguagem, o Video-LLaMA se destaca por sua capacidade de resposta em tempo real a estímulos audiovisuais, representando um avanço significativo para aplicações interativas.

## **3. Metodologia**

Este artigo foi realizado a partir da análise dos artigos citados, onde pudemos levantar tópicos importantes que pudessem ser relevantes nas questões estudadas. Também foi avaliado o cenário atual de evolução da Inteligência Artificial e seu uso em possíveis campos de estudo.

No desenvolvimento do LLaMA, Touvron et al. (2023) empregaram um extenso conjunto de dados de fontes públicas, totalizando 1,4 trilhões de tokens, para treinar modelos de diferentes tamanhos, variando de 7 a 65 bilhões de parâmetros. O processo de treinamento incluiu técnicas de otimização de arquitetura, como a normalização de camadas e o uso de embeddings rotativos, buscando reduzir o consumo de memória e otimizar a velocidade de processamento. Esse enfoque visa proporcionar um equilíbrio entre o desempenho e a viabilidade de uso em ambientes com limitações de recursos.

Para o Video-LLaMA, Zhang et al. (2023) propuseram uma arquitetura de dois ramos, onde dados visuais e auditivos são processados separadamente e depois integrados em um espaço de embedding comum. Utilizando uma combinação de datasets de vídeo e áudio-texto, o modelo é treinado para compreender conteúdos complexos de vídeos, incorporando o contexto temporal e espacial das cenas. Esse processo exige não apenas uma alta capacidade de processamento, mas também uma adaptação específica para que o modelo interprete corretamente as variações dinâmicas dos estímulos visuais e auditivos em sincronia.

#### **4. Resultados e Discussão**

Os testes realizados por Touvron et al. (2023) demonstram que o modelo LLaMA é capaz de alcançar um desempenho competitivo em uma ampla variedade de benchmarks de linguagem, superando o GPT-3 em muitos aspectos, mesmo com uma contagem de parâmetros significativamente menor. Esse feito se deve ao uso de dados públicos, como Common Crawl e C4, em um total de 1,4 trilhões de tokens, e a otimizações arquiteturais que visam melhorar a eficiência do treinamento e da inferência. O LLaMA utiliza camadas de normalização e embeddings rotacionais para reduzir o consumo de memória, o que contribui para uma execução mais rápida e acessível, permitindo que o modelo seja utilizado em máquinas com menos recursos computacionais. Além disso, a metodologia do LLaMA enfatiza a escalabilidade e o custo-benefício, sendo uma opção viável para pesquisadores e instituições que desejam explorar modelos de linguagem de alta qualidade sem a dependência de infraestrutura de ponta. Contudo, sua limitação está na exclusividade de dados textuais, o que restringe seu uso a aplicações que não demandam inputs multimodais, como áudio ou vídeo, o que pode ser considerado um limitador em alguns cenários mais interativos.

No caso do Video-LLaMA, Zhang et al. (2023) os pesquisadores apresentam um modelo que não apenas processa informações textuais, mas também integra dados visuais e auditivos, posicionando-o como uma ferramenta inovadora no campo da inteligência artificial multimodal. O modelo foi testado em contextos complexos de compreensão de vídeo, onde ele mostrou capacidade de responder a perguntas específicas sobre conteúdo visual e auditivo, demonstrando uma compreensão contextual aprofundada. A arquitetura de dois ramos do Video-LLaMA é um diferencial, pois permite que o modelo processe quadros de vídeo e segmentos de áudio separadamente, alinhando-os em um espaço de embedding comum que facilita a integração dos dados multimodais. Esse método de treinamento, embora eficaz, exige grande poder computacional e uma quantidade significativa de dados de alta qualidade para otimizar a correspondência áudio-texto e visão-texto. Essa complexidade torna o Video-LLaMA menos acessível para aplicações que necessitam de um modelo leve e facilmente implantável. Ainda assim, seu potencial de aplicação é vasto, sendo especialmente útil para cenários que exigem análises simultâneas de áudio e vídeo, como em assistência médica, segurança e interfaces de realidade aumentada, onde uma compreensão holística de estímulos visuais e auditivos pode enriquecer a interação e a resposta do sistema.

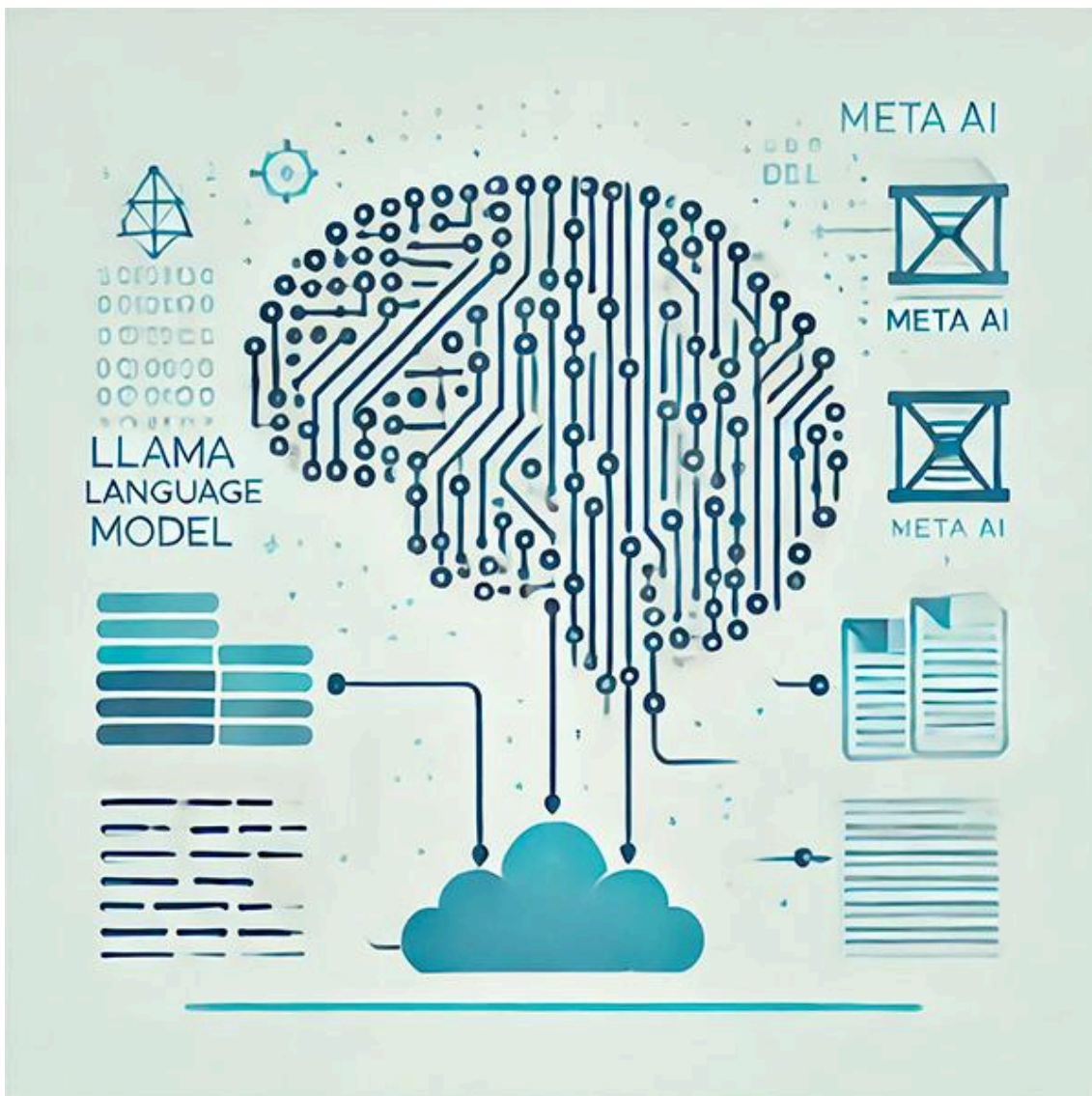


Figura 1: Representação esquemática do modelo LLaMA. A imagem ilustra um ícone de cérebro digital composto por nós interconectados, simbolizando o processamento eficiente de dados. Ao lado do cérebro, blocos empilhados de texto representam as diversas fontes de dados usadas no treinamento, incluindo ícones para texto, código e artigos de pesquisa. A imagem usa tons suaves de azul e verde, refletindo o foco em acessibilidade e eficiência do modelo (DALL·E 2024).



**Figura 2: Estrutura visual do modelo Video-LLaMA.** A imagem apresenta um ícone de player de vídeo estilizado com ondas sonoras e quadros de vídeo abstratos emergindo, simbolizando a capacidade de processamento multimodal do modelo (áudio e visual). Ao redor do player, nós conectados a um ícone de cérebro digital representam a integração da inteligência artificial. A paleta de cores em azul suave e detalhes em amarelo destaca o enfoque técnico e inovador do modelo para análise audiovisual em tempo real (DALL-E 2024)

## Conclusão

O LLaMA e o Video-LLaMA representam avanços distintos e complementares no campo dos modelos de linguagem, sendo que cada um possui vantagens e desvantagens em suas funcionalidades específicas. O LLaMA, conforme descrito por Touvron et al. (2023), é uma solução que prioriza eficiência e acessibilidade. Sua principal vantagem está na capacidade de oferecer um desempenho competitivo em benchmarks de linguagem, mesmo sendo substancialmente menor em tamanho quando comparado a modelos como GPT-3. Isso torna o LLaMA uma escolha ideal para aplicações que exigem processamento rápido e uso eficiente de recursos, como assistentes virtuais e chatbots em contextos educacionais e comerciais. Por outro lado, a limitação do LLaMA está no seu foco exclusivo em dados textuais, restringindo sua aplicação a interações baseadas unicamente em linguagem escrita.

Já o Video-LLaMA, de Zhang et al. (2023), se destaca por sua capacidade multimodal, permitindo que o modelo compreenda e responda a conteúdos de vídeo com componentes visuais e auditivos. A maior vantagem desse modelo está em sua arquitetura de dois ramos, que permite uma interpretação integrada e sincrônica dos dados visuais e auditivos, respondendo a estímulos contextuais com precisão. Essa funcionalidade é valiosa para aplicações interativas e de suporte em tempo real, como assistentes em ambientes médicos ou de suporte ao cliente, onde a compreensão completa de vídeos e sons pode melhorar a qualidade do atendimento. Já a desvantagem no uso do Video-LLaMA é a alta complexidade computacional necessária para seu treinamento e operação, além da dependência de conjuntos de dados de áudio-texto específicos, o que limita sua aplicabilidade em contextos onde esses dados não estão amplamente disponíveis.

Ambos os modelos têm potencial para transformar o uso de LLMs em áreas específicas. O LLaMA, pela sua acessibilidade, é particularmente adequado para instituições educacionais e pequenas empresas que desejam incorporar IA em suas operações sem altos custos de infraestrutura. Já o Video-LLaMA, com seu foco multimodal, apresenta oportunidades em áreas como segurança, saúde e entretenimento, onde a análise integrada de vídeo e áudio pode oferecer suporte a decisões complexas ou enriquecer experiências interativas. Esses modelos ilustram como o futuro dos LLMs pode se diversificar: enquanto o LLaMA sugere caminhos para uma IA mais acessível e eficiente, o Video-LLaMA aponta para um uso cada vez mais imersivo e multimodal da inteligência artificial.

## Referências

Touvron, H., et al. (2023). LLaMA: Open and Efficient Foundation Language Models. Meta AI.

Zhang, H., et al. (2023). Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding. DAMO Academy, Alibaba Group.