Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X
Volume 3, Issue 6, 2025

THE RISE OF MULTIMODAL AI: A QUICK REVIEW OF GPT-4V AND GEMINI

Jamil Ahmed[1], Ghalib Nadeem[2], Muhammad Kashif Majeed[*3], Rashid Ghaffar[4], Abdul Karim Kashif Baig[5], Syed Raheem Shah[6], Rana Abdul Razzaq[7], Talha Irfan[8]

[1,2,4,5] *Faculty of Engineering, Science and Technology, Iqra University, Karachi 75500, Pakistan*
[*3]*Faculty of Engineering, Science and Technology, Iqra University, Karachi 75500, Pakistan. School of Electronic Science and Technology, Xi'an Jiaotong University, Xi'an, Shaanxi 710049, China*

[*3]mkashif@iqra.edu.pk

DOI: https://doi.org/10.5281/zenodo.15718443

Keywords
(Multimodal AI, GPT-4V, Gemini 1.5, Artificial General Intelligence (AGI, AI applications)

Article History
Received on 14 May 2025
Accepted on 14 June 2025
Published on 21 June 2025

Copyright @Author
Corresponding Author: *
Muhammad Kashif Majeed

Abstract

Multimodal artificial intelligence (AI) systems– interpreting, synthesizing and reasoning heterogeneously over text, images, audio and video—represent a transformational boundary in AI research and application today. Some notable achievements in this area are Open AI GPT-4V (Vision) and Google DeepMind's Gemini 1.5, both exemplifying the current coups of cross-modal representation learning and generative reasoning. This paper remarks critically and succinctly on these two flagship models, studying their architecture, modality fusion, functionality, and performance metrics. Emphasis is placed upon their performance towards visual question answering, multimodal dialogue, instruction following, and other tasks that are reasoning integrated because intelligence and perception working in harmony are needed. Moreover, we examine GPT-4V and Gemini 1.5 from the lenses of model size, scaling, fine-tuning, alignment, and generalization in downstream tasks. The debate looks at the major outstanding issues of multimodal AI: hallucinations, no interpretability, high computational cost, and others which remain the most important barriers to wider use and trust. Finally, we study the far-reaching effects

INTRODUCTION

I. Introduction and related works

Multimodal Artificial Intelligence (AI) pertains to the construction and creation of systems that are able to process, fuse, and reason on different data modalities, including text, images, audio, video, and other forms of sensory information. In difference to the single type of input such as natural language or visual data, language models use, AI systems that utilize more than one type of input aim to combine different forms of information to enhance understanding meaning to human-like perception and cognition. The integration improves the model capabilities in understanding difficult situations, producing complex results, and interacting meaningfully with humans [1].

A. Why Multimodal AI is Trending (2024–2025) ?

A lot of technological and social factors were going on at the same time that generated a keen interest in multimodal AI in 2024–2025. For starters, foundation models have been making major leaps lately. Forthcoming models from OpenAI, like Generative Pretrained Transform (GPT-4V), as well as Gemini 1.5 from Google DeepMind are heralding an new era with astonishing cross-modal capabilities.

These models describe images, interpret charts, transcribe audio, and provide. answers to questions set within the visuals using some AI magic that allows them to switch between functions fluidly. Moreover, the existence of data in various forms such as video, podcasts, and even social media posts containing images, texts, and captions creates a demand for AI systems to comprehend and process information in a certain way. Integrating language, vision, and sound all at once is something that traditional un-modes systems cannot provide for real world use cases [2].

These are easier to achieve because of enhanced hardware accelerators Graphics Processing Units & Tensor Processing Units (GPUs and TPUs) and better optimization techniques that enable more efficient training and inference on complex multimodal models. Their scaling deployment advanced usability models. There's something more however: industries ranging from healthcare to robotics require the assistance of AI to amplify productivity, safety, and user interaction.[3,4].

B. Why Multimodal AI is Trending (2024–2025) ?

Multimodal AI systems that can "see, hear, and speak" represent a fundamental leap in machine intelligence, offering transformative value in a wide array of domains:

Healthcare: AI models that analyse radiology images alongside patient notes and spoken symptoms can assist in more accurate diagnoses and clinical decision-making.

Education: Intelligent tutoring systems leveraging visual cues, spoken feedback, and written content can create more engaging and personalized learning environments.

Customer Service: Virtual agents that understand user emotions via voice, interpret visual context (e.g., screenshots), and generate natural-sounding responses are enhancing human-computer interaction.

Accessibility: Multimodal systems enable innovations like real-time video captioning for the hearing impaired or audio descriptions of visual content for the visually impaired.

Autonomous Systems: In robotics and self-driving cars, the ability to fuse visual data with spatial audio and textual commands is crucial for navigation, object detection, and situational awareness [5].

I. Background and Evolution

A. Exploiting Single-Modal Bases: GPT-3, BERT, LLAMA

The cross-sectioning of contemporary AI technologies are rooted in singular models which obtain single particular mastery such as Practiced solely focus on a singular area of data like Natural Language Processing (NLP). The invention of large language models (LLMs) was accompanied by transformational models like Bidirectional Encoder Representations from Transformers (BERT) and GPT-3 Generative Pre-trained transformer-3(GPT-3) for large-scale mechanical learning. Google's BERT shifted the paradigm of natural language processing (NLP) with an achievement called bidirectional attention, which means understanding text and its context. Open AI's GPT-3 followed the trend with generative architectures, trained on gigantic corpora derived from the web, yielding astonishing zero and few shot performance across diverse tasks involving text [6].

Moreover, the Large Language Model Meta AI(LLAMA) series by Meta AI placed great emphasis on accessibility with well-documented open-weight models that bolstered research reproducibility. These models, while advanced, remained fundamentally unimodal, with inputs and outputs restricted to text. Their restrictions became clear for tasks that necessitated non-linguistic comprehension, such as image recognition, tone of voice understanding, or multi-modal logical reasoning that is visual and verbal [4].

The emergence of multimodal artificial intelligence (AI) started with the goal of bridging the gap between text and vision, resulting in hybrid architectures that could jointly encode and reason across different modalities. Open AI's ground-breaking clip (contrastive language–image pre-training) was a significant milestone in this field. Equipped with the ability to align image and text embeddings in a shared latent space, clip achieved zero-shot image classification using natural language descriptions, marking a major advancement in the field of vision-language understanding [7].

The next Open AI success was is using DALL-E to generate multimodal content: It can generate pictures based on what you say. With this generation methods of creativities between diverse modes and via codes as language for information whether humanly presented or otherwise publicly verifiable and evasive according to content can computationally illuminate into unseen advantages just as every work product usually becomes something new yet adopted within these limits of output methods that arise out not only from grammar itself becoming expressive for readers but also through lexico, However writing codes equivalently has an undisputable prerequisite: whatever means idea or question nobody knows can always still be transformed into meaningful English plaintext through logical transformation Textually belonged thus some terminological differences ensuring that the desired images from two topics or paragraphs will appear at the same time-that is these the subject pictures will be have been presented directly. Using Flamingo, DeepMind has taken cross-modal few-shot learning to another level by demonstrating his generalization training methods actually works on many problems. Flamingo builds performance on the basis of pre-trained language backbones and vision encoders, it integrates image features into language model by the use of Perceiver Resampled modules that is highly efficient in-speed and low-cost. Visual question answering, image captioning and cross-domain multi-modal dialogues were all made possible [8].

These initial multimodal systems laid the groundwork for the unified architectures we see today, where a single model can concurrently understand, reason about, and generate across multiple streams of sensory input. The success of CLIP, DALL·E and Flamingo demonstrated not only the technical viability of multimodal learning, but also the potential of multimodal learning to fundamentally change the balance of human-AI interaction by producing models that draw much closer to the richness of human perception and communication [9].

B. Developed by Open AI

GPT-4V (Vision) is an progressed multimodal show created by Open AI as portion of the GPT-4 family. Building upon the capabilities of its forerunners, GPT-4V is planned to handle both content and visual inputs, empowering it to lock in in assignments that require cross-modal thinking. It speaks to Open AI's proceeded endeavours to coordinated vision and dialect models, clearing the way for AI frameworks that can consistently prepare and create data over distinctive modalities. Released in ChatGPT Plus (2023–2024).

GPT-4V was coordinates into Open AI's ChatGPT Additionally membership benefit, getting to be freely available to clients in 2023 and 2024. This integration stamped a noteworthy step in broadening the accessibility of multimodal AI, because it empowered clients to associated with the demonstrate through both content and picture inputs. As portion of the GPT-4 suite, GPT-4V acquires the model's large-scale transformer engineering, which has been fine-tuned to upgrade its visual comprehension and thinking capabilities. Handles Text + Images [12].

One of the defining features of GPT-4V is its ability to process both text and images as input, allowing it to generate textual responses based on visual content. This enables a range of multimodal interactions that were previously not possible with single-modal systems. GPT-4V's ability to parse and integrate visual data into the text-based framework of the GPT series allows it to generate meaningful, context-aware outputs from images, documents, or any other visual format [10].

C. Key Use-Cases: Image Captioning, OCR, Document Q&A

GPT-4V is extremely effective in some real-world use cases, such as:

Image Captioning: The model is able to create descriptive captions for images, allowing the visual content to be automatically described in a way that is both contextually correct and linguistically coherent.

Optical Character Recognition (OCR): GPT-4V is great at pulling out and understanding text from pictures, for example, documents scanned, written documents, handwritten documents, or photographs with text inside them, making it an efficient document processing and digitization tool. **Document Question Answering (Q&A):** Through the analysis of documents in a number of different formats (including images containing text within them), GPT-4V is capable of answering particular questions about the content involved, and is very well-suited to applications like automated customer service, knowledge management, or reviewing legal documents[7].

D. Strengths: Visual Reasoning, Diagrams

One of the strongest aspects of GPT-4V is its visual reasoning. The model excels at interpreting intricate visual scenes, recognizing objects, and identifying their relationships in the context of a query. This makes it especially useful for tasks that include diagrams and schematic representations, where visual information is critical to conveying information. GPT-4V's capacity for reasoning over textual and visual input alike also allows it to function better in categories such as:

Charts and Diagrams: The model can read and describe visual information in graphs, charts, and infographics and is therefore useful for business intelligence, scientific research, and technical applications.

Sophisticated Image Interpretation: For applications like medical image diagnosis or engineering design analysis, GPT-4V's sophisticated image reasoning can enhance text information with richer insights [11].

E. Limitations: No Audio/Video, Some Hallucinations

Although it is so powerful, GPT-4V has some limitations:

No Audio/Video Input: The model can process text and images but does not, as of yet, process audio or video inputs. This makes it less than fully useful for fields such as speech-to-text or video analysis, which are important for the full multimodal immersion. Hallucinations: Similar to other large language models, GPT-4V is susceptible to hallucinations—a situation in which the model creates outputs that are factually in error or illogical, especially when it encounters vague or poor visual inputs. This is an area of concern that reflects on the significance of meticulous calibration of the model and strong training data to keep real-world application errors at bay [7].

II. Gemini 1.5 Overview

A. Developed by Google DeepMind

Gemini 1.5 is a state-of-the-art multimodal model created by Google DeepMind, marking an important milestone in their AI research activities. Being part of the Gemini family, Gemini 1.5 continues to improve on what has been achieved by earlier models by further widening the horizon of multimodal integration, adding a dense mixture of text, image, audio, and video inputs. DeepMind's Gemini models are intended to tackle advanced cross-modal tasks and produce more coherent and contextually sensitive outputs by riding on a single shared architecture that can interpret multiple sensory data streams in parallel.[12], Launched in 2024 and released in 2024, Gemini 1.5 is a significant improvement over DeepMind's multimodal capabilities. The model is developed to handle and create high-quality content across a wide range of modalities, creating a new standard for AI systems that can interpret intricate, multimedia-heavy worlds. Its generalizability provides advanced reasoning with diverse inputs, enabling more interactive and dynamic user interfaces [10].

B. Handles Text, Images, Audio, Video

One of the characteristic aspects of Gemini 1.5 is that it can process text, images, sound, and video all at once, making it a very flexible and multimodal AI framework. Processing and synthesizing these different kinds of data, Gemini 1.5 can produce outputs that reflect more deeply and richly about real-world, information.

Text: It can understand and respond based on written inputs, answering questions or creating artistic products such as stories and essays.

Images: Similar to GPT-4V, Gemini 1.5 can analyze images, generating captions, descriptions, and interpretations from visual information

Audio: Gemini 1.5 can also process audio inputs, like transcribing speech or One of the defining features of GPT-4V is its ability to process both content and images as input, allowing it to generate literary responses based on visual content. This enables a range of multimodal intuitive that were already not possible with interpreting sound signals, making it useful for tasks such as discourse recognition, language interpretation, and voice commands [12].

Video: The capacity to analyse video substance extends the model's utility assist, permitting it to recognize objects, translate scenes, and indeed reply questions around minutes or activities inside a video [10].

C. Can Interpret Long Documents, Videos

The ability of Gemini 1.5 to decipher lengthy documents and video clips is one of its most notable characteristics. This is particularly crucial for assignments requiring in-depth understanding of lengthy texts or multimedia sources: Long-Context Understanding: Gemini 1.5 is capable of processing lengthy textual materials, including novels, reports, and scholarly papers, and producing comprehensive insights, summaries, and analyses while preserving the coherence of long-form information. For areas that demand a sophisticated comprehension of context and structure over lengthy inputs, this skill is essential. Video Interpretation: Gemini 1.5 is quite good at deriving important information from scenes, speech, and actions in long-form video content. This enables it to deliver scene-based insights, describe video footage, and respond to inquiries regarding events—all of which are useful for applications like automatic content moderation, video [13].

D. Key Use-Cases: Academic Research, Science, Tutorials

Gemini 1.5's ability to integrate and reason across many modalities makes it a perfect tool for a number of high-value use cases. Gemini 1.5 enables researchers to examine long-form academic texts, including research papers, textbooks, and historical records, to provide summaries, explanations, and innovative theories. Its multimodal characteristics make it excellent for assessing multimedia-based research, such as movies, scientific data visualizations, and audio interviews or lectures. Gemini 1.5 can help analyze complicated datasets like medical scans and laboratory experiment films, as well as textual research papers and reports, in scientific fields. This can speed up scientific discovery and make technical information more accessible. Tutorials: The model can grasp visual and aural inputs.

Gemini 1.5's ability to integrate and reason across many modalities makes it a perfect tool for a number of high-value use, cases. Gemini 1.5 enables researchers to examine long-form academic texts, including research papers, textbooks, and historical records, to provide summaries, explanations, and innovative theories. Its multimodal characteristics make it excellent for assessing multimedia-based research, such as movies, scientific data visualizations, and audio interviews or lectures. Gemini 1.5 can help analyse complicated datasets like medical scans and laboratory experiment films, as well as textual research papers and reports, in scientific fields. This can speed up scientific discovery and make technical information more accessible. Tutorials: The model can grasp visual and aural inputs [14].

E. Strengths: Multimodal Synergy, Long-Context Understanding

The key characteristics of Gemini 1.5 are its multimodal synergy and capacity to grasp long-context information.

Gemini 1.5 offers multimodal synergy, combining text, graphics, audio, and video to produce more contextually aware outputs. This convergence of modalities enables more nuanced reasoning and problem-solving across a broad spectrum of complicated tasks. Long-Context Understanding: Its ability to interpret lengthy and complex documents, as well as extended video sequences, is a key feature that distinguishes Gemini 1.5 from previous models, making it ideal for academic, scientific, and professional settings that require in-depth analysis over long periods of time or large datasets [15].

F. Limitations: Still Evolving, Limited Access

Despite its extensive capabilities, Gemini 1.5 is still in the process of evolution, and it has numerous limitations:

Still evolving: Gemini 1.5, like many cutting-edge AI models, is still improving its capacity to process and integrate multimodal data effortlessly. In some cases, the model may struggle with complicated or ambiguous inputs, especially in highly dynamic scenarios such as real-time video interpretation or interpreting subtle human emotions solely through audio or video. Limited access: Currently, access to Gemini 1.5 is somewhat restricted, with wider release likely limited to select partners, academic institutes, and commercial applications. This limited access may delay the adoption of Gemini 1.5 in particular businesses until the model becomes more broadly available [15].

G. Comparative Analysis Table

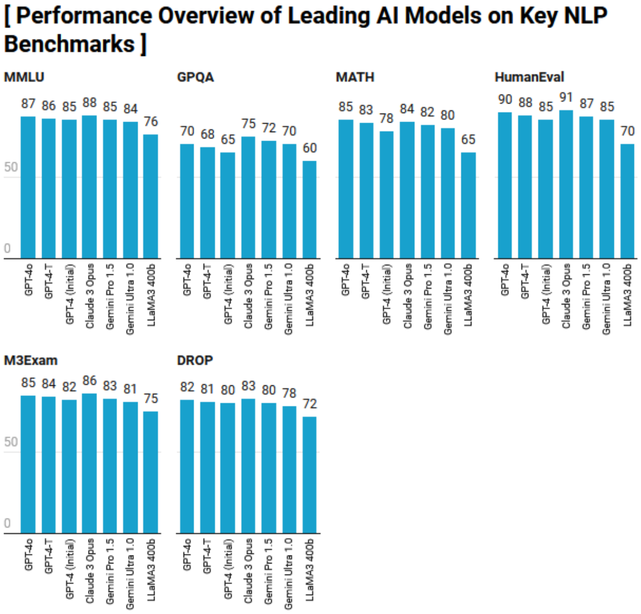
Below Table 1. Sates that Open AI's GPT-4V accepts both text and images, demonstrating advanced skills in OCR, visual interpretation, and interface comprehension, and is available to ChatGPT Plus subscribers, processing around 128,000 tokens. Google DeepMind's Gemini 1.5 goes further by integrating text, images, audio, and video, offering a significantly larger context window of up to 1 million tokens, and specializing in cross-modal understanding across video, code, and text through Gemini Advanced.

Table 1. Comparative Analysis

Feature	GPT-4V	Gemini 1.5
Modality Support	Text + Image	Text, Image, Audio, Video
Max Context Length	~128K tokens	Up to 1 million tokens
Public Availability	ChatGPT Plus	Gemini Advanced
Multimodal Strengths	OCR, visual reasoning	Cross-modal understanding
Notable Use-cases	UI analysis, images, charts	Video + code + text analysis

The graph presents a comparative analysis of leading AI models based on their performance in several text evaluation benchmarks, GPT-4o and Claude 3 Opus demonstrate the best overall performance, consistently achieving top scores, GPT-4-T and Gemini Pro 1.5 also exhibit robust capabilities across different evaluation tasks., Across all models, scores on MMLU and M3Exam are uniformly high, exceeding 75. GPQA and MATH benchmarks reveal a greater disparity in performance, with some models scoring considerably lower, The Human Eval benchmark, focused on code generation, highlights the strengths of GPT-4o and Claude 3. All models achieve strong results on the DROP benchmark, which assesses reading comprehension., Gemini Ultra 1.0 shows slightly lower performance compared to the more recent Gemini Pro 1.5. LLaMA3 400b achieves relatively low scores across all benchmark evaluations, GPT-4o and Claude 3 distinguish themselves with their well-rounded text processing capabilities.

Fig. 1. GPT-4o vs. GPT-4 vs. Gemini 1.5 Performance Analysis



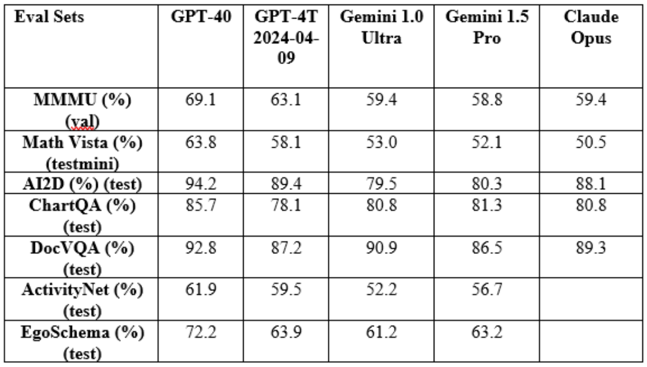
III. GPT-4o Vs. Gemini 1.5 Pro vs. Claude 3 Opus: Model Performance

The table below shows how three multimodal AI models (GPT-4o, Gemini 1.5 Pro, Claude 3 Opus) perform on different eval sets. All the metrics are expressed as percentages (higher is better), and GPT-4o consistently outperforms the other models on most evaluation sets, demonstrating its strength in understanding and generating content across modalities.[16]

- **Multimodal Matching Accuracy (MMMU):** Measured in percentage, this metric assesses how accurately models match multimodal information. GPT-4o demonstrates superior performance at 69.1%, surpassing GPT-4T (63.1%) and Gemini 1.5 Pro and Claude Opus (both at 58.5%), suggesting strong multimodal reasoning capabilities.

- **Mathematical and Visual Reasoning (Math Vista):** Evaluated as a percentage on the testmini dataset, this metric ggauge's accuracy in mathematical reasoning combined with visual understanding. GPT-4o achieves the highest score (63.8%), while Claude Opus scores lowest (50.5%).
- **Diagram Understanding (AI2D):** This benchmark, measured as a percentage on the test dataset, assesses the ability to understand diagrams. GPT-4o excels with 94.2%, while Claude Opus scores 88.1%, the lowest among the models tested, though still relatively high.
- **Chart Question Answering (Chart QA):** This metric, reported as a percentage on the test set, assesses how well models answer questions related to charts. GPT-4o demonstrates the highest accuracy at 85.7%, with Gemini 1.5 Pro achieving 81.3% and Claude Opus scoring 80.8%.
- **Document Visual Question Answering (Doc VQA):** Measured as a percentage on the test set, this benchmark evaluates a model's ability to answer questions using document images. GPT-4o achieves the top score at 92.8%, while Claude Opus's performance is 89.3%.[10]
- **Activity Net (%) (test):** This metric evaluates performance in activity recognition tasks. GPT-4o scores 61.9%, Gemini 1.5 Pro is 56.7%, and Claude Opus is not listed for this metric.

Table 1. Comparative Analysis GPT-4o Vs. Gemini 1.5 Pro vs. Claude 3 Opus: Model Performance



Eval Sets	GPT-40	GPT-4T 2024-04-09	Gemini 1.0 Ultra	Gemini 1.5 Pro	Claude Opus
MMMU (%) (val)	69.1	63.1	59.4	58.8	59.4
Math Vista (%) (testmini)	63.8	58.1	53.0	52.1	50.5
AI2D (%) (test)	94.2	89.4	79.5	80.3	88.1
ChartQA (%) (test)	85.7	78.1	80.8	81.3	80.8
DocVQA (%) (test)	92.8	87.2	90.9	86.5	89.3
ActivityNet (%) (test)	61.9	59.5	52.2	56.7	
EgoSchema (%) (test)	72.2	63.9	61.2	63.2	

- The Ego Schema test, measuring the model's ability to comprehend first-person perspectives and actions, shows GPT-4o achieving a score of 72.2%, while Gemini 1.5 Pro scores 63.2%. Claude Opus's score on this metric is unavailable. Table 2. GPT-4o model evaluations.

The evaluated data indicates that GPT-4o, Gemini 1.5 Pro, and Claude 3 Opus exhibit varying performance levels across the considered metrics, with GPT-4o generally performing strongest. However, specific task performance differs for each model, revealing individual strengths and weaknesses.

Applications of Multimodal AI

1) Education: VisualL Explanations

Multimodal AI enhances learning by providing visual explanations alongside text, allowing for better comprehension of difficult concepts.

2) Medicine: X-rays + Patient Notes

In medicine, AI integrates medical images (e.g., X-rays) with patient notes to enable more precise diagnoses. It assists physicians in rapidly interpreting imaging findings along with patient history, enhancing diagnostic speed and accuracy.

3) Csutomer Support

AI-driven customer support improves user experience by processing visual inputs (e.g., screenshots) as well as text, allowing for faster issue resolution through visually-guided automated troubleshooting[12]

4) Research: Combining Charts, Papers, and Text

Multimodal AI powers researchers by aggregating data from different sources such as charts, research articles, and text-based documents, hence speeding up the literature review process and promoting cross-disciplinary work through its capability to read visual and textual content

IV. Challenges

A. Hallucination

Multimodal AI models like GPT-4V and Gemini 1.5 are prone to producing hallucinations, which are literally false or made-up information. This is especially true when these models are presented with ambiguous or limited data, leading to outputs that are factually wrong and ungrounded in the given input

B. Model Bias

Multimodal models are susceptible to absorbing and perpetuating biases present in their training data, often mirroring societal prejudices concerning race, gender, and culture. This learned bias can then appear in both the text and images generated by the model, leading to significant ethical considerations and fairness issues when these AI systems are deployed.

C. High Computational Needs

Training and deploying AI models that handle diverse data types such as images, videos, and audio demand substantial computing power. This intensive processing results in considerable energy usage and difficulties in adapting these models for immediate use or in settings with limited resources.

D. Data Privacy (Especially with Images/Audio)

Protecting sensitive information is paramount when working with data like medical images or personal audio. Multimodal AI systems, due to the risk of data leakage through visual and auditory channels, require strong safeguards to prevent unauthorized access or misuse of private data.

V. Future Outlook

A. Push Toward True AGI

The integration of different data modalities like text, images, audio, and video in AI design is a crucial step towards achieving Artificial General Intelligence (AGI). This integration gives these systems better reasoning and flexibility like humans on different tasks, thus taking us closer to AGI.

B. Better Compression for Mobile Deployment

To power multimodal AI on mobile devices, researchers are working on effective compression techniques. This would enable AI applications to run natively on devices such as smartphones and wearables, broadening accessibility and reducing dependence on cloud computing.

C. Ethical Frameworks for Multimodal Aithcal Frameworks For Mltimodal AI

With the increasing power of multimodal AI, robust ethical guidelines become crucial. These frameworks must tackle bias, fairness, privacy, and accountability to guarantee the responsible use of AI that reflects societal values.

D. Fine-Tuning with User-Specific Multimodal Data

Future multimodal AI models are expected to offer personalized fine-tuning, adapting to individual users' multimodal data. This customization will lead to more relevant and context-aware responses, improving user experiences in areas like virtual assistants, healthcare, and personalized learning.[17]

VI. Conclusion

Multimodal AI has graduated from a theoretical idea to a universal technology. Models like GPT-4V and Gemini 1.5 represent a new era of intelligent systems capable of interacting in the way human perception and understanding do. While problems like hallucinations, model bias, and computational costs exist, the potential of multimodal AI to disrupt industries and advance towards AGI is vast. Its future holds exciting possibilities for yet more sophisticated and flexible AI systems

References

[1] Baltrušaitis, T., Ahuja, C., & Morency, L.-P. (2019). *Multimodal Machine Learning: A Survey and Taxonomy*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2), 423–44

[2] N. Rodis, C. Sardianos, P. Radoglou-Grammatikis, P. Sarigiannidis, I. Varlamis and G. T. Papadopoulos, "Multimodal Explainable Artificial Intelligence: A Comprehensive Review of Methodological Advances and Future Research Directions," in IEEE Access, vol. 12, pp. 159794-159820, 2024, doi: 10.1109/ACCESS.2024.3467062.

[3] Deepak Narayanan, Mohammad Shoeybi, Jared Casper, Patrick LeGresley, Mostofa Patwary, Vijay Korthikanti, Dmitri Vainbrand, Prethvi Kashinkunti, Julie Bernauer, Bryan Catanzaro, Amar Phanishayee, and Matei Zaharia. 2021. Efficient large-scale language model training on GPU clusters using megatron-LM. In Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC '21). Association for Computing Machinery, New York, NY, USA, Article 58, 1–15.

[4] Chen, X., Liu, C., Xu, Y., & Wang, H. (2022). *Artificial Intelligence in Robotics: Challenges and Trends*. IEEE Transactions on Industrial Informatics, 18(7), 4885–4897.

[5] Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., ... & Dean, J. (2019). *A guide to deep learning in healthcare*. Nature Medicine, 25(1), 24–29

[6] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). *Language Models are Few-Shot Learners*. arXiv preprint arXiv:2005.14165.

[7] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021). *Learning Transferable Visual Models From Natural Language Supervision*. Proceedings of the 38th International Conference on Machine Learning (ICML), 8748–8763.