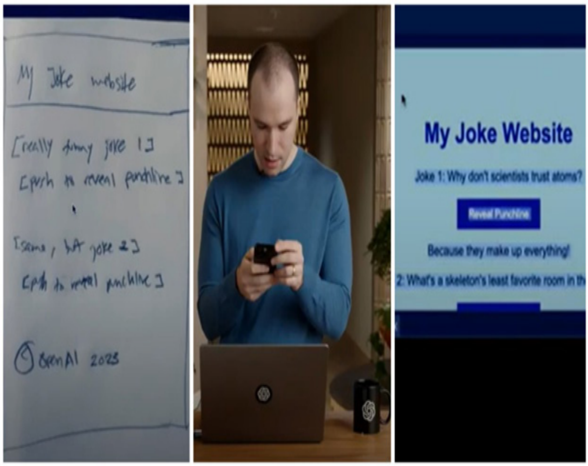
How Good Are Multimodal AI Models Like GPT-4? Explore Unmatched Greatness

For years, we human beings have taken pride in our intellect. Our ability to learn and grow has set us apart from everything else in this universe. But it seems that it will not only be humans who possess intellect.

We used to be amazed by LLMs that could understand and generate text. But now, there's something even more impressive: multimodal models like GPT-4V and Gemini. These models can understand not just text, but also images, sounds, and other types of information.

Why is this a big deal? It's simple: these models are now closer to thinking like us. For example, combining words and pictures helps them get better at understanding space and shapes, something that was hard for them before.

With multimodal tech, a pencil sketch is all it takes - and bam - you've got a whole website's code. That's the level of power we're dealing with!


My Joke website
[really funny joke 1]
[push to reveal punchline]
[same, but joke 2]
[push to reveal punchline]
© OpenAI 2023
My Joke Website
Joke 1: Why don't scientists trust atoms?
Reveal Punchline
Because they make up everything!
2: What's a skeleton's least favorite room in th

Want to dig deeper into the power of multimodal AI models? Come along!

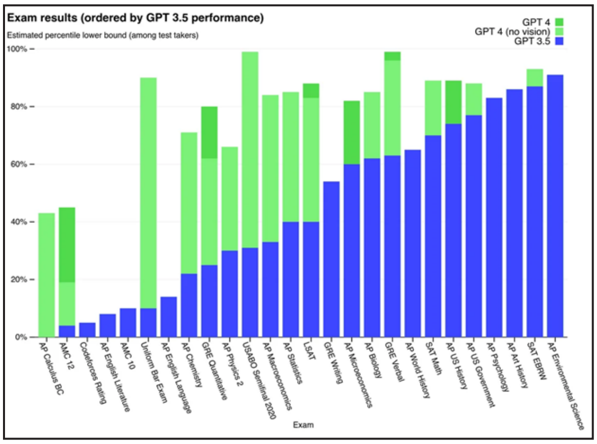
Compilation of informational blogs, articles, and papers.

Imagine this: Your room suddenly starts shaking, and everything trembles. Then, in a flash, your mom bursts in, her words cutting through the chaos: "Earthquake! We need to get out!" In that split second, you've processed a whirlwind of sensory inputs - vision, hearing, touch - leading you to one critical conclusion: Evacuation is imperative.

That's multimodality in action! It's the art of synthesizing diverse inputs altogether for razor-sharp reasoning.

Proof in Numbers: The Might of Multimodal AI

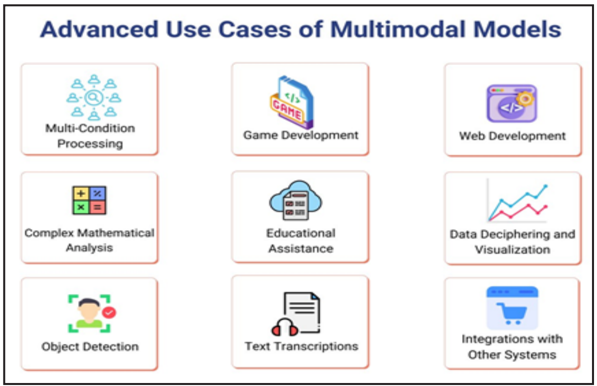
But don't just take our word for it. Here is how GPT-4 with vision performs significantly better than GPT-3.5 and GPT-4 because of its ability to process information from various sources.



Source: OpenAI

Advanced Use Cases of Multimodal Models

Quite obviously, the use cases that multimodal AI models will bring are vast. Here are some important ones:

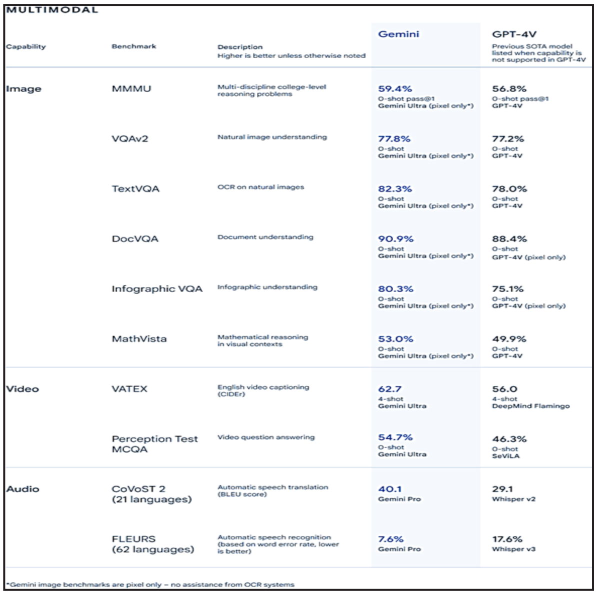

Advanced Use Cases of Multimodal Models
Multi-Condition Processing
Game Development
GAME
Web Development
Complex Mathematical Analysis
Educational Assistance
Data Deciphering and Visualization
Object Detection
Text Transcriptions
Integrations with Other Systems

Advanced Use-Cases of Multimodal Models

Read: Exploring GPT-4 Vision's Advanced Use-Cases

Latest Rival to GPT-4V - Cue Gemini

The long wait for Gemini has finally come to an end and we can see the excitement for obvious reasons. Google's most capable multimodal model has beaten the unbeatable OpenAI's GPT-4V in multimodality giving them a tough time for sure. Here's a comparison of Gemini with GPT-4V:



MULTIMODAL

Capability	Benchmark	Description Higher is better unless otherwise noted	Gemini	GPT-4V Previous SOTA model listed when capability is not supported in GPT-4V
Image	MMMU	Multi-discipline college-level reasoning problems	59.4% 0-shot pass@1 Gemini Ultra (pixel only*)	56.8% 0-shot pass@1 GPT-4V
	VQAv2	Natural image understanding	77.8% 0-shot Gemini Ultra (pixel only*)	77.2% 0-shot GPT-4V
	TextVQA	OCR on natural images	82.3% 0-shot Gemini Ultra (pixel only*)	78.0% 0-shot GPT-4V
	DocVQA	Document understanding	90.9% 0-shot Gemini Ultra (pixel only*)	88.4% 0-shot GPT-4V (pixel only)
	Infographic VQA	Infographic understanding	80.3% 0-shot Gemini Ultra (pixel only*)	75.1% 0-shot GPT-4V (pixel only)
	MathVista	Mathematical reasoning in visual contexts	53.0% 0-shot Gemini Ultra (pixel only*)	49.9% 0-shot GPT-4V
Video	VATEX	English video captioning (CIDEr)	62.7 4-shot Gemini Ultra	56.0 4-shot DeepMind Flamingo
	Perception Test MCQA	Video question answering	54.7% 0-shot Gemini Ultra	46.3% 0-shot SeViLA
Audio	CoVoST 2 (21 languages)	Automatic speech translation (BLEU score)	40.1 Gemini Pro	29.1 Whisper v2
	FLEURS (62 languages)	Automatic speech recognition (based on word error rate, lower is better)	7.6% Gemini Pro	17.6% Whisper v3

*Gemini image benchmarks are pixel only – no assistance from OCR systems

Read: What sets Gemini AI apart from GPT-4V?

Are Multimodal AI Models Taking Us Towards the Promised Neverland of Artificial General Intelligence (AGI)?

Well, yes! In the most recent paper by Microsoft Research, they talk about how GPT-4V has the sparks of AGI whereby they do have what sets humans apart i.e. common sense grounding which allows these models to not only reason but problem solve for novel situations, plan, and more.

Read: The Sparks of Artificial General Intelligence in GPT-4

Want to learn more about AI? Our blog is the go-to source for the latest tech news.

Live sessions and tutorial recommendations from experts.

The Paradox of Open-Sourcing AI

With AI becoming so powerful, we are surrounded by a paradoxical situation:

- LLMs should be open-sourced so that such a huge power is not in the hands of a few big tech companies.
- LLMs should not be open-sourced as such a powerful technology should be protected and regulated extensively.

Explore this important talk where experts in the field including Yann LeCun, Sebastien Bubeck, and Brian Greene explore artificial intelligence and the potential risks and benefits it poses to humanity. They also talk about the fact that big tech companies controlling AI is a bigger risk than AI itself taking over