GPT-4o vs. Gemini 1.5 Pro vs. Claude 3 Opus: Multimodal AI Model Comparison

May 15, 2024 | 8 mins

ENCORD

< Back to Blogs

Contents

The multimodal AI war has been heating up. OpenAI and Google are leading the way with announcements like GPT-4o, which offers real-time multimodality, and Google's major updates to Gemini models. Oh, and let's not forget Anthropic's Claude 3 Opus, too.

These models are not just about understanding text; they can process images, video, and even code, opening up a world of possibilities for annotating data, creative expression, and real-world understanding.

But which model is right for you? And how can they help you perform critical tasks like labeling your images and videos? In this comprehensive guide, we'll learn about each model's capabilities, strengths, and weaknesses, comparing their performance across various benchmarks and real-world applications.

Let's get started.

Understanding Multimodal AI

Unlike traditional models that focus on a single data type, such as text or images, multimodal AI systems can process and integrate information from multiple modalities, including:

- **Text:** Written language, from documents to social media posts.
- **Images:** Photographs, drawings, medical scans, etc.

- **Audio:** Speech, music, sound effects.
- **Video:** A combination of visual and auditory information.

This ability to understand and reason across different types of data allows multimodal AI to tackle tasks previously beyond the reach of AI systems. For example, a multimodal AI model could analyze a video, understanding the visual content, spoken words, and background sounds to generate a comprehensive summary or answer questions about the video.

> ⓘ **Recommended Read:** Introduction to Multimodal Deep Learning.

GPT-4o: Open AI's Multimodal AI

OpenAI's GPT-4o is a natively multimodal AI that can understand and generate content across text, images, and audio inputs. The native multimodality in GPT-4o provides a more comprehensive and natural interaction between the user and the model.

GPT-4o is not just an incremental upgrade; it introduces several features that set it apart from **previous models** like GPT-4 and GPT-4 Turbo. Let's examine them.

> ⓘ **See Also 👀:** Exploring GPT-4 Vision: First Impressions.

GPT-4o: Benefits and New Features

GPT-4o, with the "o" for "omni," represents a groundbreaking shift towards more natural and seamless human-computer interactions. Unlike its predecessors, GPT-4o is designed to process and generate a combination of text, audio, and images for a more comprehensive understanding of user inputs.

1. High Intelligence: GPT-4o matches the performance of GPT-4 Turbo in text, reasoning, and coding intelligence but sets new benchmarks in

multilingual, audio, and vision capabilities.

2. Faster Response Times: With optimized architecture, GPT-4o provides quicker responses by generating tokens up to 2x faster than GPT-4 Turbo for more fluid real-time conversations. It can respond to audio inputs in as little as 232 milliseconds, with an average response time of 320 milliseconds.

> ⓘ Faster response times allow for more engaging, human-like interactions, ideal for chatbots, virtual assistants, and interactive applications.

3. Improved Multilingual Support: A new tokenizer allows GPT-4o to handle non-English languages better, expanding its global reach. For example, compared to previous models, it requires 4.4x fewer tokens for Gujarati, 3.5x fewer for Telugu, and 3.3x fewer for Tamil.

4. Larger Context Window: GPT-4o's context length is 128K tokens, equivalent to about 300 pages of text. This allows it to handle more complex tasks and maintain context over longer interactions. Its knowledge cut-off date is October 2023.

5. Enhanced Vision Capabilities: The model has improved vision capabilities, allowing it to better understand and interpret visual data.

6. Video Understanding: The model can process video inputs by converting them into frames, enabling it to understand visual sequences without audio.

7. More Affordable Pricing: GPT-4o matches the text and code capabilities of GPT-4 Turbo in English while significantly improving upon non-English language processing. It is also 50% cheaper than its predecessor in the API, making it more accessible to a wider range of users and developers.

8. API Enhancements: The GPT-4o API supports various new features, including real-time vision capabilities and improved translation abilities. Higher rate limits (5x GPT-4) make GPT-4o suitable for large-scale, high-traffic applications.

GPT-4o is currently available in preview access for select developers, with general availability planned in the coming months.

GPT-4o: Limitations

- **Transparency:** Limited information is available about the data used to train GPT-4o, the model's size, its compute requirements, and the techniques used to create it. This lack of transparency makes it difficult to fully assess the model's capabilities, biases, and potential impact. More openness from OpenAI would help build trust and accountability.
- **Audio Support:** While GPT-4o has made significant strides in multimodal capabilities, its API currently does not support audio input. This limitation restricts its use in applications that require audio processing, although OpenAI plans to introduce this feature to trusted testers soon.

ⓘ **Might Be Helpful:** GPT-4 Vision Alternatives.

Gemini 1.5 Pro and Gemini 1.5 Flash: Google's Multimodal AI Models

Gemini 1.5 Pro is Google's flagship multimodal model, providing advanced features for complex tasks and large-scale applications. It's designed to be versatile and capable of handling everything from generating creative content to analyzing intricate data sets.

Gemini 1.5 Flash, on the other hand, prioritizes speed and efficiency, making it ideal for scenarios where real-time responses or high throughput are crucial.

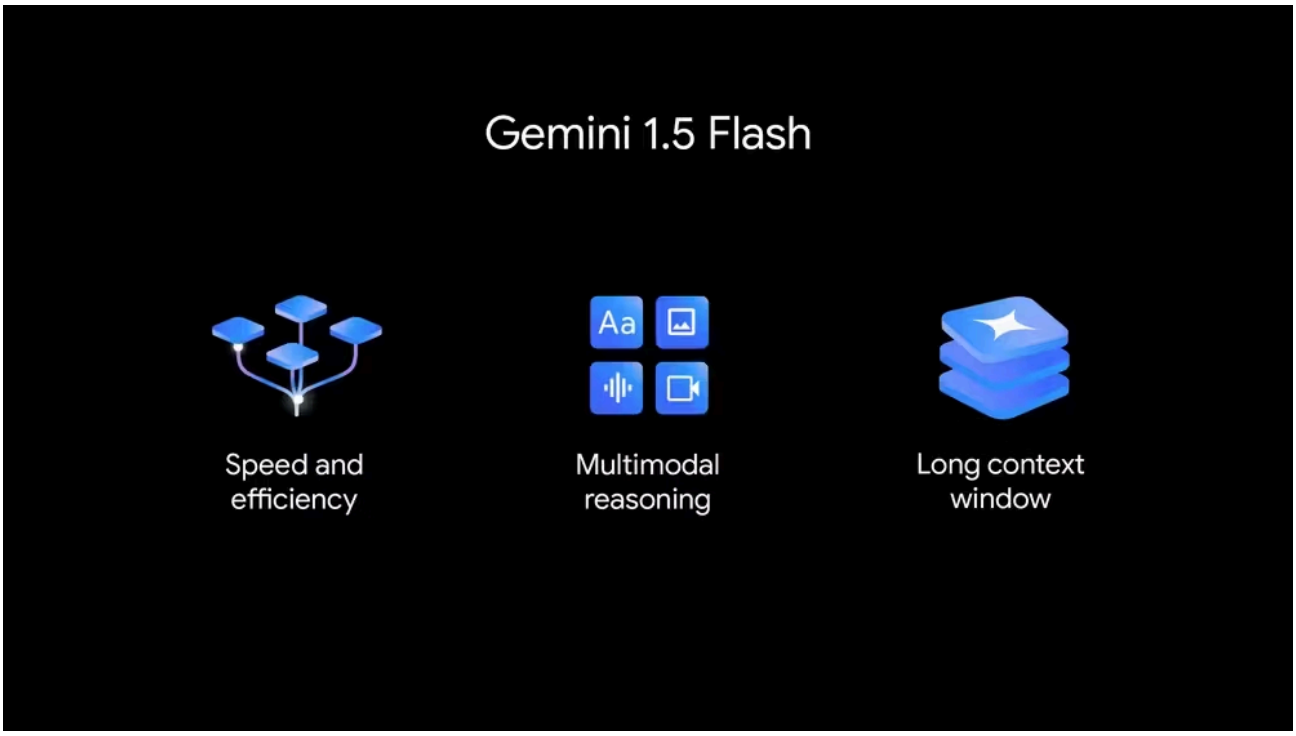
These models can process and generate content across text, images, audio, and video with minimal response latency, enabling more sophisticated and context-aware applications.

> ⓘ **See Also 👀: Gemini 1.5: Google's Generative AI Model with Mixture of Experts Architecture**.

Gemini 1.5 Pro: Benefits and New Features

At **Google I/O 2024**, several new features and updates for Gemini 1.5 Pro and the Gemini family of models were announced:

- **Gemini 1.5 Flash:** This Gemini model is optimized for narrower or high-frequency tasks where the speed of the model's response time matters the most. It is designed for fast and cost-efficient serving at scale, with multimodal reasoning and a similar context size to Gemini 1.5 Pro. It's great for real-time applications like chatbots and on-demand content generation.
- **Natively Multimodal with Long Context:** Both 1.5 Pro and 1.5 Flash come with our 1 million token context window and allow you to interleave text, images, audio, and video as inputs. There is a **waitlist in Google AI Studio** to access 1.5 Pro with a 2 million token context window.
- **Pricing and Context Caching:** Gemini 1.5 Flash costs $0.35 per 1 million tokens, and **context caching** will be available in June 2024 to save even more. This way, you only have to send parts of your prompt, including large files, to the model once, making the long context even more useful and affordable.
- **Gemini Nano:** Is expanding beyond text-only inputs to include images as well. Starting with Pixel, **applications using Gemini Nano** with Multimodality will be able to understand the world the way people do—not just through text but also through sight, sound and spoken language.
- **Project Astra:** The team also introduced Project Astra, which builds on Gemini models. It's a prototype AI agent that can process information faster by continuously encoding video frames, combining the video and speech input into a timeline of events, and caching this information for efficient recall.



The new Gemini 1.5 Flash model is optimized for speed and efficiency.

Both models are in preview in **more than 200 countries** and territories and will be generally available in June 2024.

Gemini 1.5 Pro and Gemini 1.5 Flash: Limitations

- **Cost:** Access to Gemini 1.5 Pro, especially with the expanded context window, can be expensive for individual users or small organizations.
- **Access:** Both models are currently in limited preview, granting access to select developers and organizations.

> ⓘ **Recommended Webinar :Vision Language Models: How to Leverage Google Gemini in Your ML Data Pipelines.**

Claude 3 Opus: Anthropic's Multimodal AI

Claude 3 Opus is the most advanced model in **Anthropic's latest suite of AI models**, setting new benchmarks in various cognitive tasks. Opus offers the

highest performance and capabilities as part of the Claude 3 family, which also includes Sonnet and Haiku.

Claude 3 Opus: What's New?

One of the most significant advancements in Claude 3 Opus is its multimodal nature, enabling it to process and analyze text, images, charts, and diagrams. This feature opens up new possibilities for applications in fields like healthcare, engineering, and data analysis, where visual information plays a crucial role.

Opus also demonstrates improved performance in several key areas:

- Enhanced reasoning and problem-solving skills, outperforming GPT-4 and Gemini Ultra in benchmarks such as graduate-level expert reasoning (GPQA) and basic mathematics (GSM8K).
- Superior language understanding and generation, particularly in non-English languages like Spanish, Japanese, and French.
- Increased context window of up to 200,000 tokens, allowing for more comprehensive and contextually rich responses.

Claude 3 Opus: Benefits

The advanced capabilities of Claude 3 Opus offer several benefits for users and developers:

- Thanks to its enhanced reasoning and problem-solving abilities, it improved accuracy and efficiency in complex tasks.
- Expanded applicability across various domains, enabled by its multimodal processing and support for multiple languages.
- More natural and human-like interactions result from its increased context understanding and language fluency.

Claude 3 Opus: Limitations

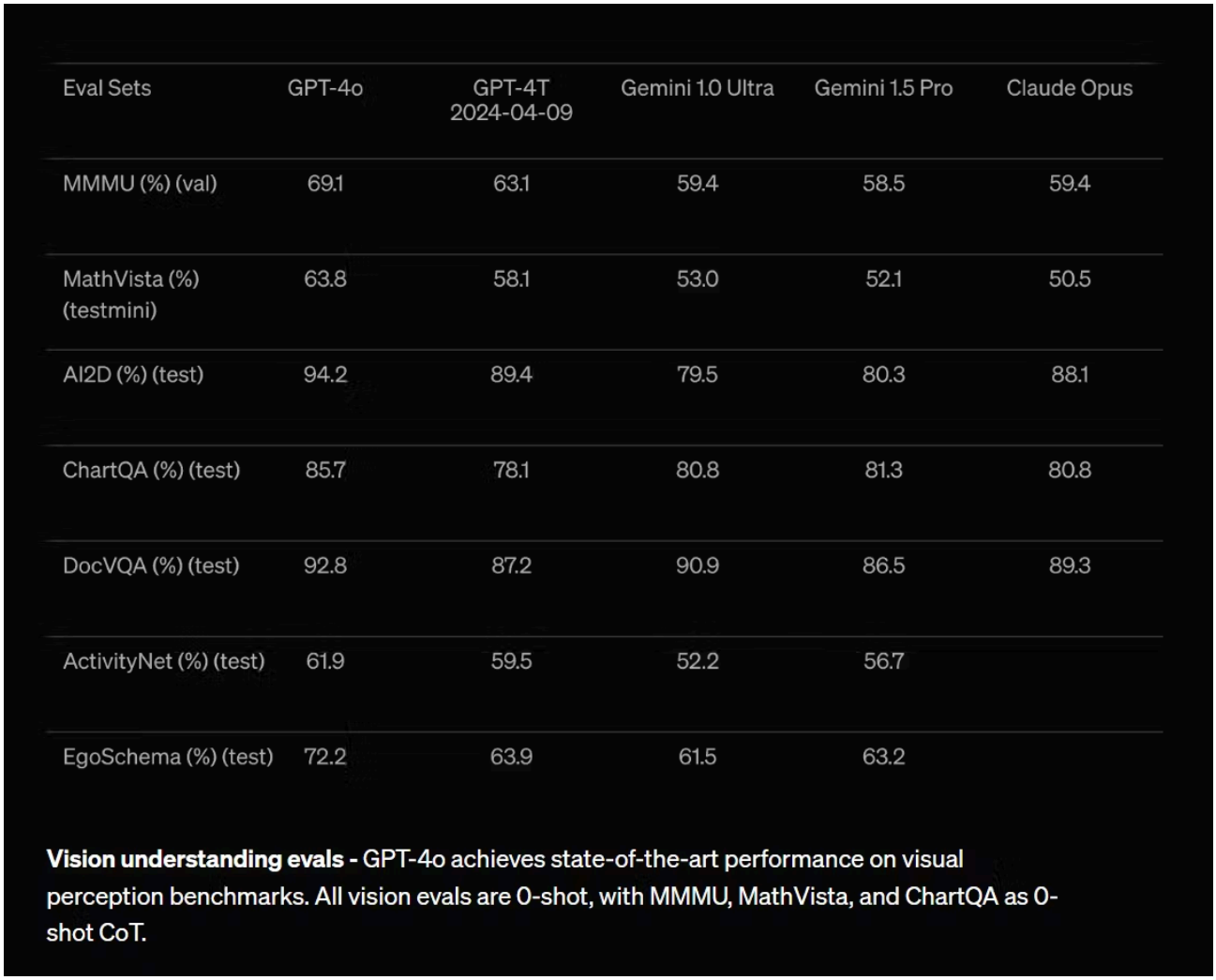
Despite its impressive performance, Claude 3 Opus has some limitations:

- Potential biases and inaccuracies, as the model may reflect biases present in its training data and occasionally generate incorrect information.
- Restricted image processing capabilities, as Opus cannot identify individuals in images and may struggle with low-quality visuals or tasks requiring spatial reasoning.
- Handling multimodal data, especially sensitive information, raises concerns about privacy and security. Ensuring compliance with relevant regulations and protecting user data remains a critical challenge.

Claude 3 Opus is also available through **Anthropic's API** and on **Amazon Bedrock**. However, it is in limited preview on platforms like Google Cloud's **Vertex AI**, which may limit its reach compared to other models.

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

The following table compares the performance of three multimodal AI models—GPT-4o, Gemini 1.5 Pro, and Claude 3 Opus—across various evaluation sets. The metrics are presented as percentages, indicating the accuracy or performance on each task.



Eval Sets	GPT-4o	GPT-4T 2024-04-09	Gemini 1.0 Ultra	Gemini 1.5 Pro	Claude Opus
MMMU (%) (val)	69.1	63.1	59.4	58.5	59.4
MathVista (%) (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	61.5	63.2	

Vision understanding evals - GPT-4o achieves state-of-the-art performance on visual perception benchmarks. All vision evals are 0-shot, with MMMU, MathVista, and ChartQA as 0-shot CoT.

GPT-4o model evaluations.

GPT-4o consistently outperforms the other models across most evaluation sets, showcasing its superior capabilities in understanding and generating content across multiple modalities.

- **MMMU (%)(val):** This metric represents the Multimodal Matching Accuracy. GPT-4o leads with 69.1%, followed by GPT-4T at 63.1%, and Gemini 1.5 Pro and Claude Opus are tied at 58.5%. This indicates that GPT-4o has robust multimodal capability and a strong grasp of reasoning.
- **MathVista (%)(testmini):** This metric measures mathematical reasoning and visual understanding accuracy. GPT-4o again has the highest score at 63.8%, while Claude Opus has the lowest at 50.5%.

- **AI2D (%)(test):** This benchmark evaluates performance on the **AI2D dataset** involving diagram understanding. GPT-4o tops the chart with 94.2%, and Claude Opus is at the bottom with 88.1%, which is still relatively high.
- **ChartQA (%)(test):** This metric measures the model's performance in answering questions based on charts. GPT-4o has the highest accuracy at 85.7%, with Gemini 1.5 Pro close behind at 81.3%, and Claude Opus matches the lower end of the spectrum at 80.8%.
- **DocVQA (%)(test):** This benchmark assesses the model's ability to answer questions based on document images. GPT-4o leads with 92.8%, and Claude Opus is again at the lower end with 89.3%.
- **ActivityNet (%)(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.
- **EgoSchema (%)(test):** This metric might evaluate the model's understanding of first-person perspectives or activities. GPT-4o scores 72.2%, Gemini 1.5 Pro is 63.2%, and Claude Opus is not listed.

From this data, we can infer that GPT-4o generally outperforms Gemini 1.5 Pro and Claude 3 Opus across the evaluated metrics. However, it's important to note that the differences in performance are not uniform across all tasks, and each model has its strengths and weaknesses.

The next section will teach you how to choose the right multimodal model for different tasks.

GPT-4o, Opus 3, Vs. Gemini 1.5 Pro: Choosing the RIght Multimodal Model

Data Annotation/Labelling Tasks with Encord's Custom Annotation Agents

We will use Encord's **Custom Annotation Agents (BETA)** to evaluate each model's capability to auto-classify an image as an annotation task.

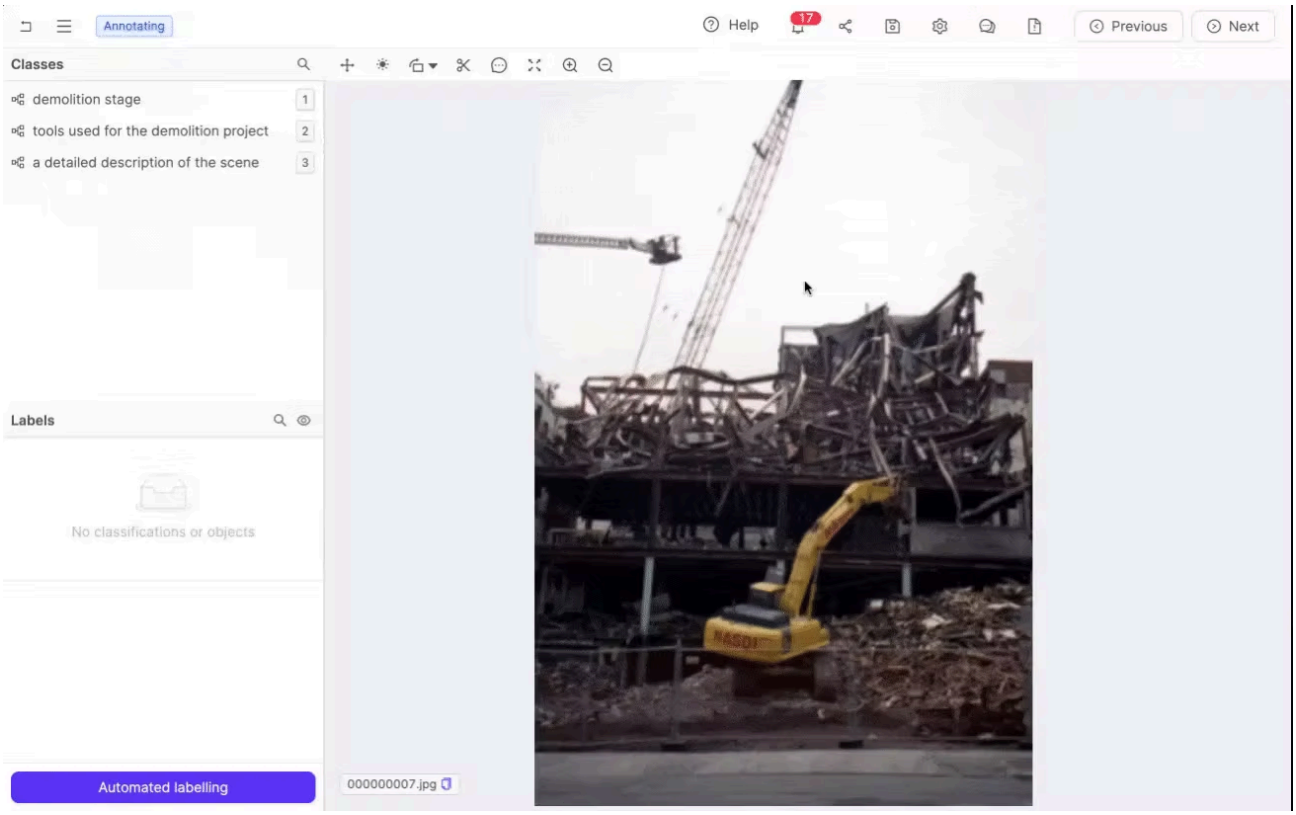
Agents within Encord enable you to integrate your API endpoints, such as your model or an LLM API from a provider, with Encord to automate your annotation processes. Agents can be called while annotating in the **Label Editor**.

> ⓘ Learn more in **the documentation**.

GPT-4o

With its multimodal capabilities, GPT-4o is well-suited for data annotation tasks, especially when annotating diverse datasets that include text, images, and videos.

Using Custom Agents, we show how GPT-4o can auto-classify a demolition site image. See the results below:



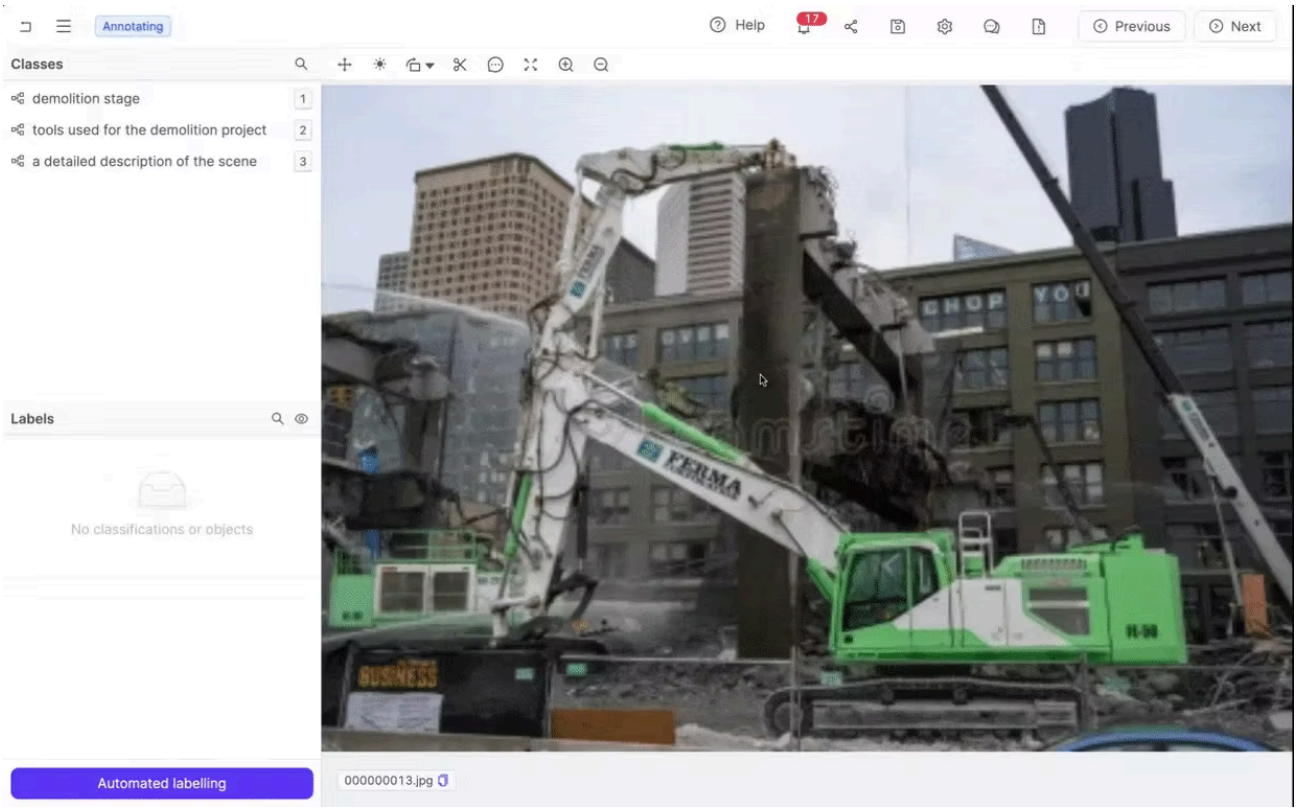
Encord AI Annotation Agents test with GPT-4o results.

The results are interesting! The GPT-4o endpoint we use for the annotation AI agent gives us a good head start with annotating the image with a few classes you can select from based on the scene and context.

Let's see how Gemini 1.5 Flash does on a similar image.

Gemini 1.5 Flash

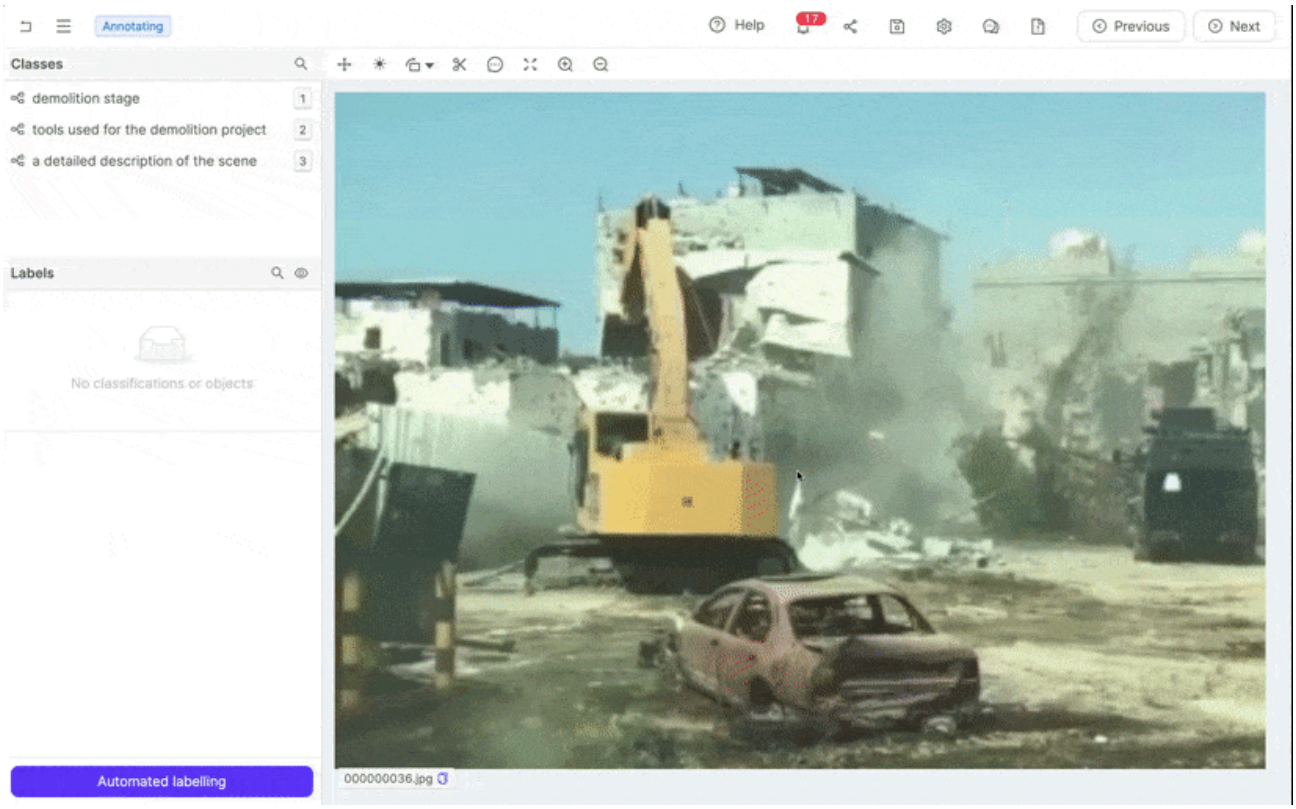
With Gemini 1.5 Flash, labeling is emphasized around speed, cost-effectiveness, and annotation quality. While it does not provide GPT-4o's level of annotation quality, it is quite fast, as you'll see in the demo below, and cheaper to run per 1 million tokens than GPT-4o.



Encord AI Annotation Agents test with Gemini 1.5 Flash results.

Claude 3

While Claude 3 Opus has a large context window and strong language capabilities, it is not as efficient as GPT-4o for annotation tasks requiring multimodal inputs.



Encord AI Annotation Agents test with Claude 3 Opus results.

It looks like we still need additional customization to get optimal results.

Text-Based Tasks

- **GPT-4o:** Inherits the exceptional text generation capabilities of GPT-4, making it ideal for tasks like summarization, translation, and creative writing.

- **Claude 3 Opus:** Boasts strong language understanding and generation skills, comparable to GPT-4o in many text-based scenarios.
- **Gemini 1.5 Pro:** While proficient in text processing, its primary strength lies in multimodal tasks rather than pure text generation.

Multimodal Understanding

- **GPT-4o:** Consistently demonstrates superior performance in understanding and generating content across multiple modalities, including images, audio, and videos.
- **Claude 3 Opus:** Shows promise in multimodal tasks but does not match GPT-4o's level of sophistication in visual and auditory comprehension.
- **Gemini 1.5 Pro:** Designed with multimodal capabilities in mind, it offers strong performance in tasks requiring understanding and integrating different data types.

Code Generation Capability

- **GPT-4o:** Excels at code generation and understanding, making it a valuable tool for developers and programmers.
- **Claude 3 Opus:** While capable of generating code, it might not be as specialized or efficient as GPT-4o in this domain.
- **Gemini 1.5 Pro:** Has some code generation capabilities, but it's not its primary focus compared to text and visual tasks.

Transparency

- **GPT-4o & Gemini 1.5 Pro:** Both models lack full transparency regarding their inner workings and training data, raising concerns about potential biases and interpretability.
- **Claude 3 Opus:** Anthropic emphasizes safety and transparency, providing more information about the model's architecture and training processes.

Accessibility

- **GPT-4o & Claude 3 Opus:** These are available through APIs and platforms, which offer relatively easy access for developers and users.
- **Gemini 1.5 Pro & Flash:** Currently in limited preview, access is currently restricted to select users and organizations.

Affordability

- **GPT-4o:** OpenAI offers various pricing tiers, making it accessible to different budgets. However, the most powerful versions can be expensive.
- **Claude 3 Opus:** Pricing details may vary depending on usage and specific requirements.
- **Gemini 1.5 Pro:** As a premium model, it is more expensive. Although Gemini 1.5 Flash is the cheapest of all the options, with a context window of up to 1 Million tokens and a price point of 0.53 USD

ⓘ A **Google Sheet**, kindly curated by **Médéric Hurier**, helps put the pricing comparison per context size in perspective.

Comparison Table

Feature	GPT-4o	Claude 3 Opus	Gemini 1.5 Pro
Data Annotation/Labeling	Strong	Moderate	Excellent
Text-Based Tasks	Excellent	Excellent	Good
Multimodal Understanding	Very Good	Good	Excellent
Code Generation	Excellent	Moderate	Very Good
Transparency	Limited	Relatively High	Limited
Accessibility	High	High	Limited (Preview)
Affordability	Varies (multiple tiers)	Varies (depends on usage)	Potentially Expensive, but Flash is the least expensive per token

Comparison Table - GPT-4o vs Gemini 1.5 vs Claude 3 Opus | Encord

This table provides a high-level overview of how each model performs across various criteria. When deciding, consider your application's specific needs and each model's strengths.

GPT-4o, Gemini 1.5 Pro, Opus 3: Key Takeaways

Throughout this article, we have focused on understanding how GPT-4o, Gemini 1.5 Pro and Flash, and Claude 3 Opus compare across benchmarks and use cases. Our goal is to help you choose the right model for your task.

Here are some key takeaways:

GPT-4o

- GPT-4o is OpenAI's latest multimodal AI model, capable of processing text, images, audio, and video inputs and generating corresponding outputs in real-time.

- It matches GPT-4 Turbo's performance on text and code while being significantly faster (2x) and more cost-effective (50% cheaper).
- GPT-4o demonstrates improved multilingual capabilities, requiring fewer tokens for non-English languages like Gujarati, Telugu, and Tamil.
- The model is free to all ChatGPT users, with paid subscribers getting higher usage limits.
- It is great for real-time interaction and harmonized speech synthesis, which makes its responses more human-like.

Gemini 1.5 Pro and Flash

- Gemini 1.5 Pro showcases enhanced performance in translation, coding, reasoning, and other tasks compared to previous versions.
- It is integrated with Google's suite of apps, potentially offering additional utility for users already within the Google ecosystem.
- Gemini 1.5 Pro's performance in multimodal tasks is strong but does not consistently outperform GPT-4o across all benchmarks.

Claude 3 Opus

- Claude 3 Opus has strong results in benchmarks related to math and reasoning, document visual Q&A, science diagrams, and chart Q&A.
- It offers a larger context window of 200k tokens, which can be particularly beneficial for tasks requiring a deep understanding of context.
- Despite its strengths, Claude 3 Opus has shown some limitations in tasks such as object detection and answering questions about images accurately.

In summary, GPT-4o appears to be the most versatile and capable across various tasks, with Gemini 1.5 Pro being a strong contender, especially within the Google ecosystem.

Claude 3 Opus offers cost efficiency and a large context window, making it an attractive option for specific applications, particularly those requiring deep context understanding. Each model has strengths and weaknesses,

and the choice between them should be guided by the task's specific needs and requirements.

WRITTEN BY

Stephen Oladele

View more posts →

Frequently asked questions

Is Gemini 1.5 Pro better than ChatGPT 4?

Is ChatGPT better than Gemini?

Is Gemini Advanced better than GPT-4?

How much is Gemini 1.5 Pro vs GPT-4o?

How do GPT-4o, GPT-4, and Gemini 1.5 compare performance metrics like accuracy and speed?

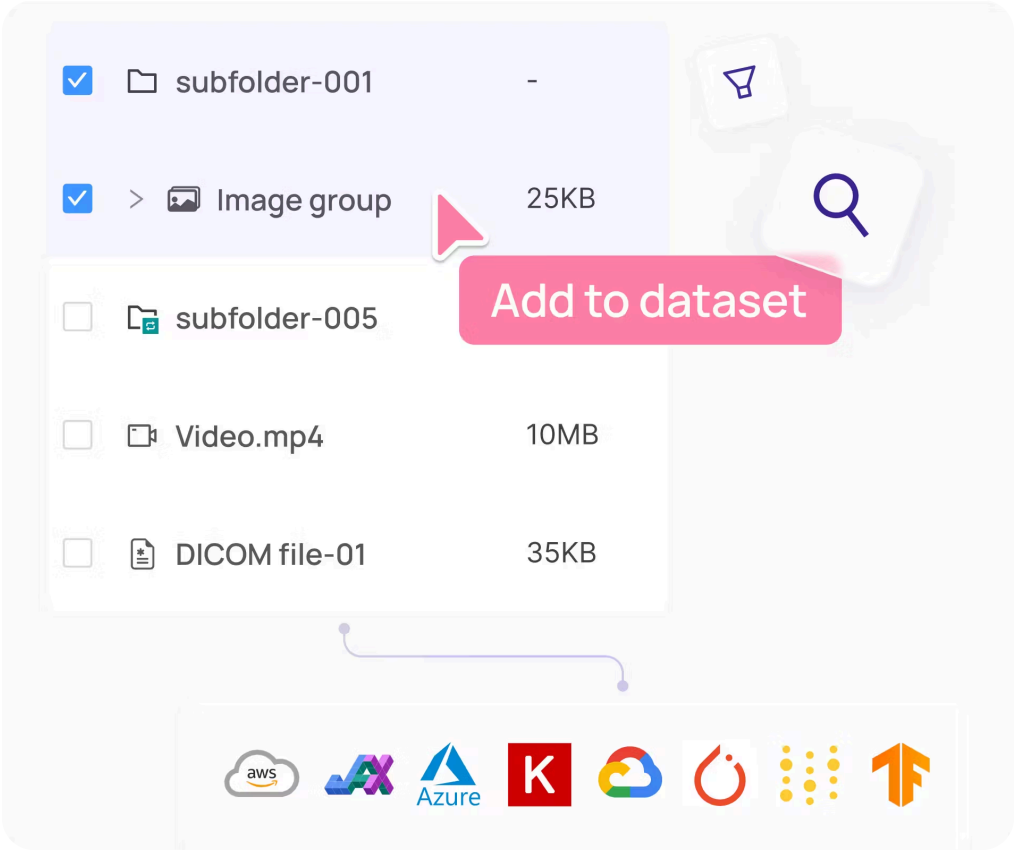
Are there specific industries or use cases where one model outperforms the others?

What are the key architectural or training differences between GPT-4o, GPT-4, and Gemini 1.5 that contribute to their varying performance levels?

PREVIOUS BLOG
How to Use GPT-4o for Model Development

NEXT BLOG
Meta Imagine AI Just got an Impressive GIF...

Explore our products



Index

Manage & curate your data

Understand and manage your visual data, prioritize data for labeling, and initiate active learning pipelines.

Explore Index →

Software To Help You Turn Your Data Into AI

Subscribe

Get occasional product updates and tutorials to your inbox.

Your work email