

From VLMs to MM-LLMs

LLaVA, Video ChatGPT, ChatGPT-4V, Gemini 1.5

Vineeth N Balasubramanian

Department of Computer Science and Engineering
Indian Institute of Technology, Hyderabad



LLaVA¹: Large Language and Vision Assistant

- First successful image-text conversation model

¹Liu et al, “Visual Instruction Tuning”, NeurIPS 2023

LLaVA¹: Large Language and Vision Assistant

- First successful image-text conversation model
- The LLaVA model integrates visual and language models for instruction tuning

¹Liu et al, "Visual Instruction Tuning", NeurIPS 2023

LLaVA¹: Large Language and Vision Assistant

- First successful image-text conversation model
- The LLaVA model integrates visual and language models for instruction tuning
- Utilizes GPT-4 for generating multimodal instruction-following data

¹Liu et al, "Visual Instruction Tuning", NeurIPS 2023

LLaVA¹: Large Language and Vision Assistant

- First successful image-text conversation model
- The LLaVA model integrates visual and language models for instruction tuning
- Utilizes GPT-4 for generating multimodal instruction-following data
- Employs a two-stage instruction-tuning process:
 - Feature alignment pre-training with image-text pairs
 - End-to-end fine-tuning on multimodal data

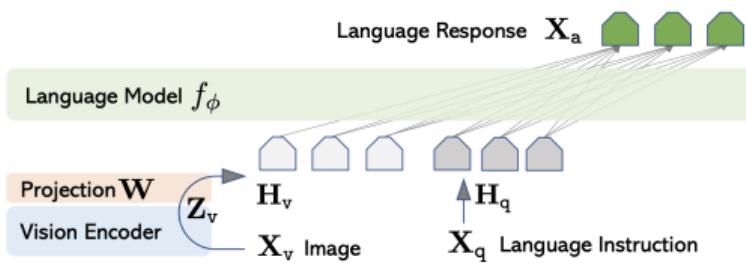
¹Liu et al, "Visual Instruction Tuning", NeurIPS 2023

LLaVA¹: Large Language and Vision Assistant

- First successful image-text conversation model
- The LLaVA model integrates visual and language models for instruction tuning
- Utilizes GPT-4 for generating multimodal instruction-following data
- Employs a two-stage instruction-tuning process:
 - Feature alignment pre-training with image-text pairs
 - End-to-end fine-tuning on multimodal data
- Aims to improve zero-shot capabilities on multimodal tasks

¹Liu et al, "Visual Instruction Tuning", NeurIPS 2023

LLaVA²: Architecture



- Goal is to leverage both the pre-trained LLM (Vicuna) and visual model capabilities
- For input images, pre-trained CLIP visual encoder ViT-L/14 is used to obtain visual features Z_v . A linear layer projects Z_v into the language embedding space, forming visual tokens $H_v = W \cdot Z_v$
- The projection scheme is simple and lightweight, allowing for quick iteration

²Liu et al, "Visual Instruction Tuning", NeurIPS 2023

LLaVA³: Training

- For each image X_v a multi-turn conversation is generated $(X_q^1, X_a^1 \dots X_q^T, X_a^T)$ where T is the total number of turns. Instruction X_{instruct}^t at t^{th} turn is

$$\mathbf{X}_{\text{instruct}}^t = \begin{cases} \text{Randomly choose } [\mathbf{X}_q^1, \mathbf{X}_v] \text{ or } [\mathbf{X}_v, \mathbf{X}_q^1], & \text{the first turn } t = 1 \\ \mathbf{X}_q^t, & \text{the remaining turns } t > 1 \end{cases}$$

- For a sequence of length L , the probability of the target answers X_a is computed by

$$p(\mathbf{X}_a | \mathbf{X}_v, \mathbf{X}_{\text{instruct}}) = \prod_{i=1}^L p_{\theta}(\mathbf{x}_i | \mathbf{X}_v, \mathbf{X}_{\text{instruct}, < i}, \mathbf{X}_{a, < i})$$

³Liu et al, "Visual Instruction Tuning", NeurIPS 2023

LLaVA: Two-stage training

Stage 1: Pre-training for Feature Alignment

- Filtered CC3M dataset to 595K image-text pairs for balance between concept coverage and training efficiency

LLaVA: Two-stage training

Stage 1: Pre-training for Feature Alignment

- Filtered CC3M dataset to 595K image-text pairs for balance between concept coverage and training efficiency
- Converted pairs to instruction-following data using a naive expansion method. Each sample is a single-turn conversation with an image, a randomly sampled question (language instruction), and the original caption as the ground-truth answer

LLaVA: Two-stage training

Stage 1: Pre-training for Feature Alignment

- Filtered CC3M dataset to 595K image-text pairs for balance between concept coverage and training efficiency
- Converted pairs to instruction-following data using a naive expansion method. Each sample is a single-turn conversation with an image, a randomly sampled question (language instruction), and the original caption as the ground-truth answer
- During training, with visual encoder and LLM weights frozen, only the projection matrix W is trained. This aligns image features H_v with pre-trained LLM word embedding, effectively training a visual tokenizer for the frozen LLM

LLaVA: Two-stage training

Stage 1: Pre-training for Feature Alignment

- Filtered CC3M dataset to 595K image-text pairs for balance between concept coverage and training efficiency
- Converted pairs to instruction-following data using a naive expansion method. Each sample is a single-turn conversation with an image, a randomly sampled question (language instruction), and the original caption as the ground-truth answer
- During training, with visual encoder and LLM weights frozen, only the projection matrix W is trained. This aligns image features H_v with pre-trained LLM word embedding, effectively training a visual tokenizer for the frozen LLM

Stage 2: Fine-tuning End-to-End

- Visual encoder weights are kept frozen while updating both pre-trained weights of the projection layer and LLM in LLaVA, with trainable parameters $\theta = \{W, \phi\}$.

LLAVA: GPT-assisted Visual Instruction Data Generation

- Use COCO images to generate three types of instruction-following data
 - Conversation
 - Detailed description
 - Complex reasoning

LLAVA: GPT-assisted Visual Instruction Data Generation

- Use COCO images to generate three types of instruction-following data
 - Conversation
 - Detailed description
 - Complex reasoning
- 158K unique language-image instruction-following samples
 - 58K in conversations
 - 23K in detailed description
 - 77k in complex reasoning

LLAVA: Instruction-following data

Context type 1: Captions

A group of people standing outside of a black vehicle with various luggage. Luggage surrounds a vehicle in an underground parking area. People try to fit all of their luggage in an SUV. The sport utility vehicle is parked in the public garage, being packed for a trip. Some people with luggage near a van that is transporting it.



Context type 2: Boxes

person: [0.681, 0.242, 0.774, 0.694], backpack: [0.384, 0.696, 0.485, 0.914], suitcase: ...<omitted>

Response type 1: conversation

Question: What type of vehicle is featured in the image?

Answer: The image features a black sport utility vehicle (SUV) ...<omitted>

Response type 2: detailed description

The image is an underground parking area with a black sport utility vehicle (SUV) parked. There are three people in the scene, with one person standing closer to the left side of the vehicle, another person in the middle, and the third person on the right side. They are all working together to pack their luggage into the SUV for a trip. ...<omitted>

Response type 3: complex reasoning

Question: What challenges do these people face?

Answer: In the image, a group of people is standing outside a black SUV in a parking area, surrounded by various pieces of luggage, including suitcases and backpacks. They are facing the challenge of fitting all their luggage into the black SUV. There are multiple suitcases and backpacks to be packed, which suggests that the group has a significant amount of belongings ...<omitted>

- Top block shows the contexts such as captions and boxes used to prompt GPT
- Bottom block shows the three types of responses
- Note that the visual image is not used to prompt GPT, it is only shown here as a reference.

LLAVA: Multimodal Chatbot

	Conversation	Detail description	Complex reasoning	All
OpenFlamingo [5]	19.3 ± 0.5	19.0 ± 0.5	19.1 ± 0.7	19.1 ± 0.4
BLIP-2 [28]	54.6 ± 1.4	29.1 ± 1.2	32.9 ± 0.7	38.1 ± 1.0
LLaVA	57.3 ± 1.9	52.5 ± 6.3	81.7 ± 1.8	67.3 ± 2.0
LLaVA [†]	58.8 ± 0.6	49.2 ± 0.8	81.4 ± 0.3	66.7 ± 0.3

Figure 1: LLAVA-Bench (In-the-Wild)

- In order to evaluate generalizability to novel domains, a small pool of images containing 24 images with 60 questions in total is collected
- LLaVA achieves significantly better performance compared with BLIP-2 (+29%) and OpenFlamingo (+48%)

LLAVA: ScienceQA

Method	Subject			Context Modality			Grade		Average
	NAT	SOC	LAN	TXT	IMG	NO	G1-6	G7-12	
<i>Representative & SoTA methods with numbers reported in the literature</i>									
Human [34]	90.23	84.97	87.48	89.60	87.50	88.10	91.59	82.42	88.40
GPT-3.5 [34]	74.64	69.74	76.00	74.44	67.28	77.42	76.80	68.89	73.97
GPT-3.5 w/ CoT [34]	75.44	70.87	78.09	74.68	67.43	79.93	78.23	69.68	75.17
LLaMA-Adapter [59]	84.37	88.30	84.36	83.72	80.32	86.90	85.83	84.05	85.19
MM-CoT _{Base} [61]	87.52	77.17	85.82	87.88	82.90	86.83	84.65	85.37	84.91
MM-CoT _{Large} [61]	95.91	82.00	90.82	95.26	88.80	92.89	92.44	90.31	91.68
<i>Results with our own experiment runs</i>									
GPT-4 [†]	84.06	73.45	87.36	81.87	70.75	90.73	84.69	79.10	82.69
LLaVA	90.36	95.95	88.00	89.49	88.00	90.66	90.93	90.90	90.92
LLaVA+GPT-4 [†] (complement)	90.36	95.50	88.55	89.05	87.80	91.08	92.22	88.73	90.97
LLaVA+GPT-4 [†] (judge)	91.56	96.74	91.09	90.62	88.99	93.52	92.73	92.16	92.53

Science QA dataset: NAT = natural science, SOC = social science, LAN = language science, TXT = text context, IMG = image context, NO = no context, G1-6 = grades 1-6, G7-12 = grades 7-12

VideoChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models⁴

- Model capable of generating meaningful conversations about videos

⁴Maaz et al, "Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models", arXiv 2023

VideoChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models⁴

- Model capable of generating meaningful conversations about videos
- It merges LLMs with a pre-trained visual encoder for spatiotemporal video representations

⁴Maaz et al, "Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models", arXiv 2023

VideoChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models⁴

- Model capable of generating meaningful conversations about videos
- It merges LLMs with a pre-trained visual encoder for spatiotemporal video representations
- Introduces 100,000 high-quality video instruction pairs

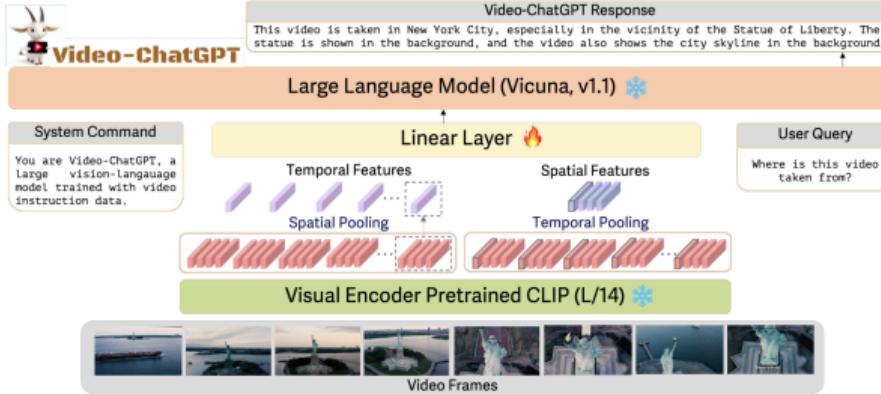
⁴Maaz et al, "Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models", arXiv 2023

VideoChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models⁴

- Model capable of generating meaningful conversations about videos
- It merges LLMs with a pre-trained visual encoder for spatiotemporal video representations
- Introduces 100,000 high-quality video instruction pairs
- First quantitative video conversation evaluation

⁴Maaz et al, "Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models", arXiv 2023

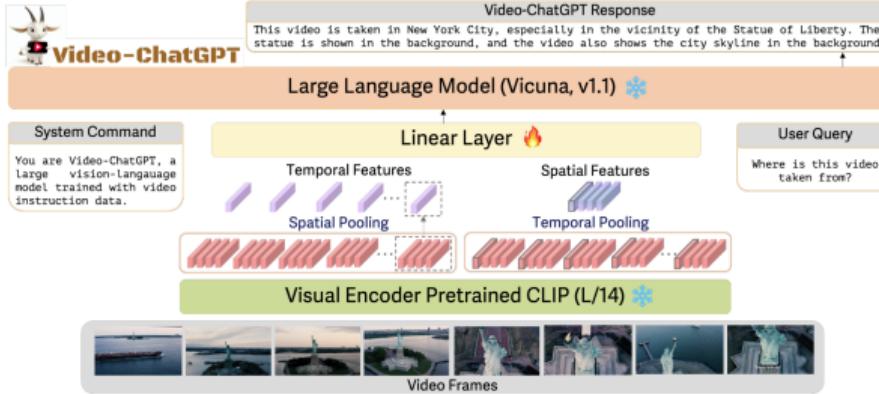
VideoChatGPT⁵: Architecture



- For a video sample $V_i \in \mathbb{R}^{T \times H \times W \times C}$ with T frames, the visual encoder outputs features $x_i \in \mathbb{R}^{T \times h \times w \times D}$

⁵Maaz et al, "Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models", arXiv 2023

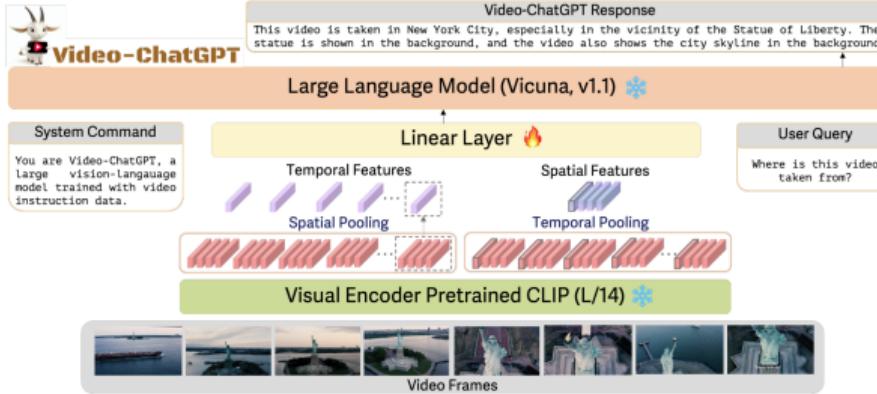
VideoChatGPT⁵: Architecture



- For a video sample $V_i \in \mathbb{R}^{T \times H \times W \times C}$ with T frames, the visual encoder outputs features $x_i \in \mathbb{R}^{T \times h \times w \times D}$
- Averaging frame-level embeddings across the temporal dimension gives a video-level temporal representation $t_i \in \mathbb{R}^{N \times D}$ where $N = h \times w$

⁵Maaz et al, "Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models", arXiv 2023

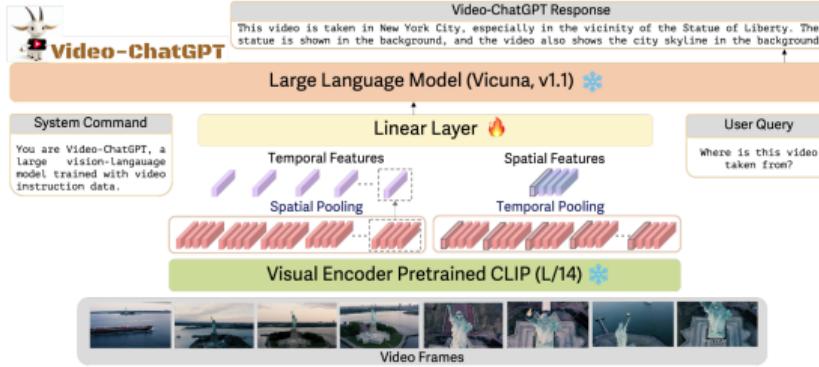
VideoChatGPT⁵: Architecture



- For a video sample $V_i \in \mathbb{R}^{T \times H \times W \times C}$ with T frames, the visual encoder outputs features $x_i \in \mathbb{R}^{T \times h \times w \times D}$
- Averaging frame-level embeddings across the temporal dimension gives a video-level temporal representation $t_i \in \mathbb{R}^{N \times D}$ where $N = h \times w$
- Averaging frame-level embeddings along the spatial dimension yields a video-level spatial representation $z_i \in \mathbb{R}^{T \times D}$

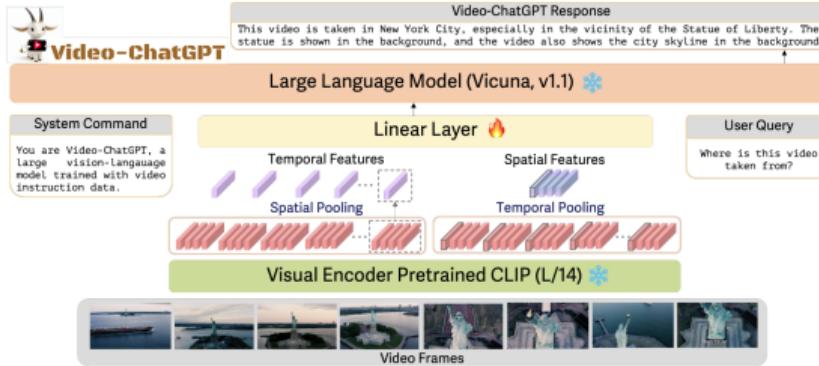
⁵Maaz et al, "Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models", arXiv 2023

VideoChatGPT: Architecture



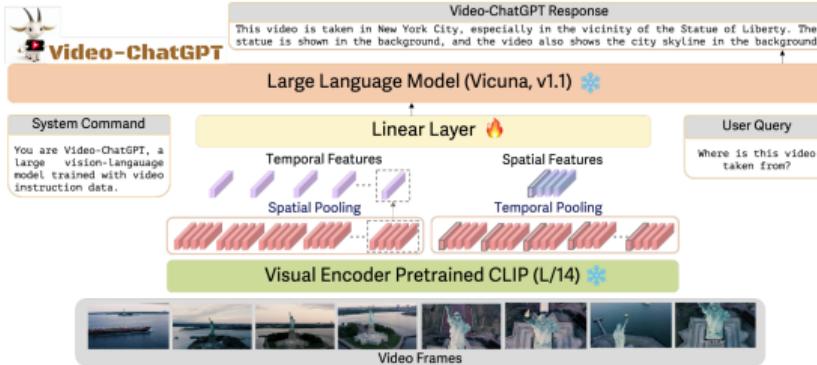
- The video-level features v_i are obtained by concatenating the temporal and spatial features: $v_i = [t_i \ z_i] \in \mathbb{R}^{(T+N) \times D}$

VideoChatGPT: Architecture



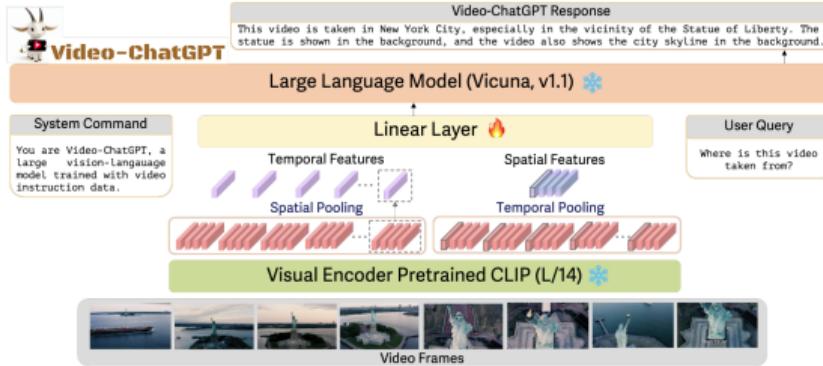
- The video-level features v_i are obtained by concatenating the temporal and spatial features: $v_i = [t_i \ z_i] \in \mathbb{R}^{(T+N) \times D}$
- A linear layer g projects v_i into the language decoder's embedding space, resulting in language embedding tokens Q_v :
$$Q_v = g(v_i) \in \mathbb{R}^{(T+N) \times K}$$

VideoChatGPT: Architecture



- The video-level features v_i are obtained by concatenating the temporal and spatial features: $v_i = [t_i \ z_i] \in \mathbb{R}^{(T+N) \times D}$
- A linear layer g projects v_i into the language decoder's embedding space, resulting in language embedding tokens Q_v :
$$Q_v = g(v_i) \in \mathbb{R}^{(T+N) \times K}$$
- Text queries are tokenized to dimensions $Q_t \in \mathbb{R}^{L \times K}$, where L is the text query length

VideoChatGPT: Architecture



- The video-level features v_i are obtained by concatenating the temporal and spatial features: $v_i = [t_i \ z_i] \in \mathbb{R}^{(T+N) \times D}$
- A linear layer g projects v_i into the language decoder's embedding space, resulting in language embedding tokens Q_v :
$$Q_v = g(v_i) \in \mathbb{R}^{(T+N) \times K}$$
- Text queries are tokenized to dimensions $Q_t \in \mathbb{R}^{L \times K}$, where L is the text query length
- The video embedding tokens Q_v are concatenated with the text tokens Q_t and input to the language decoder

VideoChatGPT: Video Instruction Tuning

- The pretrained LLM is fine-tuned on prediction tokens using its autoregressive objective and high-quality video-text pairs

VideoChatGPT: Video Instruction Tuning

- The pretrained LLM is fine-tuned on prediction tokens using its autoregressive objective and high-quality video-text pairs
- During fine-tuning, prompts are employed based on the template: “USER:
<Instruction><Vid – tokens> Assistant:”

VideoChatGPT: Video Instruction Tuning

- The pretrained LLM is fine-tuned on prediction tokens using its autoregressive objective and high-quality video-text pairs
- During fine-tuning, prompts are employed based on the template: “USER:
<Instruction><Vid – tokens> Assistant:”
- The *<Instruction>* in the prompt is a video-related question, and the model aims to predict the corresponding answer

VideoChatGPT: Video Instruction Tuning

- The pretrained LLM is fine-tuned on prediction tokens using its autoregressive objective and high-quality video-text pairs
- During fine-tuning, prompts are employed based on the template: “USER:
<Instruction><Vid – tokens> Assistant:”
- The *<Instruction>* in the prompt is a video-related question, and the model aims to predict the corresponding answer
- The model’s linear layer is adapted to align video features Q_v with the LLM’s word embeddings, enhancing response naturalness and reliability

VideoChatGPT: Visual Instruction Data Generation

- A data-focused approach uses human-assisted and semi-automatic annotation to create quality video instruction data for training Video-ChatGPT

VideoChatGPT: Visual Instruction Data Generation

- A data-focused approach uses human-assisted and semi-automatic annotation to create quality video instruction data for training Video-ChatGPT
- Expert analysis provides detailed data, while semi-automatic methods offer scalable, high-volume annotations with advanced vision-language models

VideoChatGPT: Visual Instruction Data Generation

- A data-focused approach uses human-assisted and semi-automatic annotation to create quality video instruction data for training Video-ChatGPT
- Expert analysis provides detailed data, while semi-automatic methods offer scalable, high-volume annotations with advanced vision-language models
- The dataset of 100,000 video-instruction pairs includes diverse content, covering a wide range of concepts for comprehensive model training

VideoChatGPT: Video Question-Answering Task

The screenshot displays two separate video recognition tasks within the VideoChatGPT interface.

Task 1: Playing Drums

User Query: **Describe the activity in the video.**

Model Response: In the video, a man is playing the drums while sitting in front of a microphone. He is wearing a black shirt and black pants.

Task 2: Horse Grooming

User Query: **Briefly describe the video.**

Model Response: The video shows a woman who is grooming a horse with a brush. She is also seen talking to the camera and explaining how to groom a horse.

User Query: **How many different horses are seen in the video?**

Model Response: There are two different horses seen in the video.

Examples from Video-ChatGPT's demonstration showcasing its performance in video action recognition tasks such as playing drums and grooming horse

VideoChatGPT: Evaluation

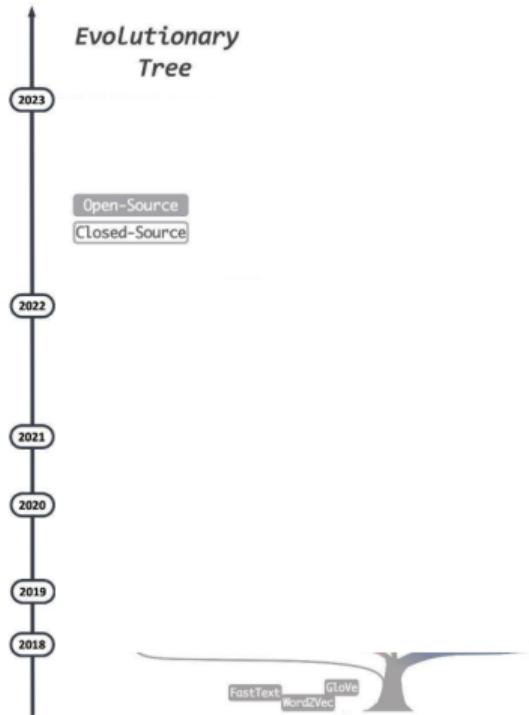
Evaluation Aspect	Video Chat	Video-ChatGPT
Correctness of Information	2.25	2.50
Detail Orientation	2.50	2.57
Contextual Understanding	2.54	2.69
Temporal Understanding	1.98	2.16
Consistency	1.84	2.20

Video-based Generative Performance Benchmarking

Model	MSVD-QA		MSRVTT-QA		TGIF-QA		Activity Net-QA	
	Accuracy	Score	Accuracy	Score	Accuracy	Score	Accuracy	Score
FrozenBiLM	32.2	–	16.8	–	41.0	–	24.7	–
Video Chat	56.3	2.8	45.0	2.5	34.4	2.3	26.5	2.2
Video-ChatGPT	64.9	3.3	49.3	2.8	51.4	3.0	35.2	2.7

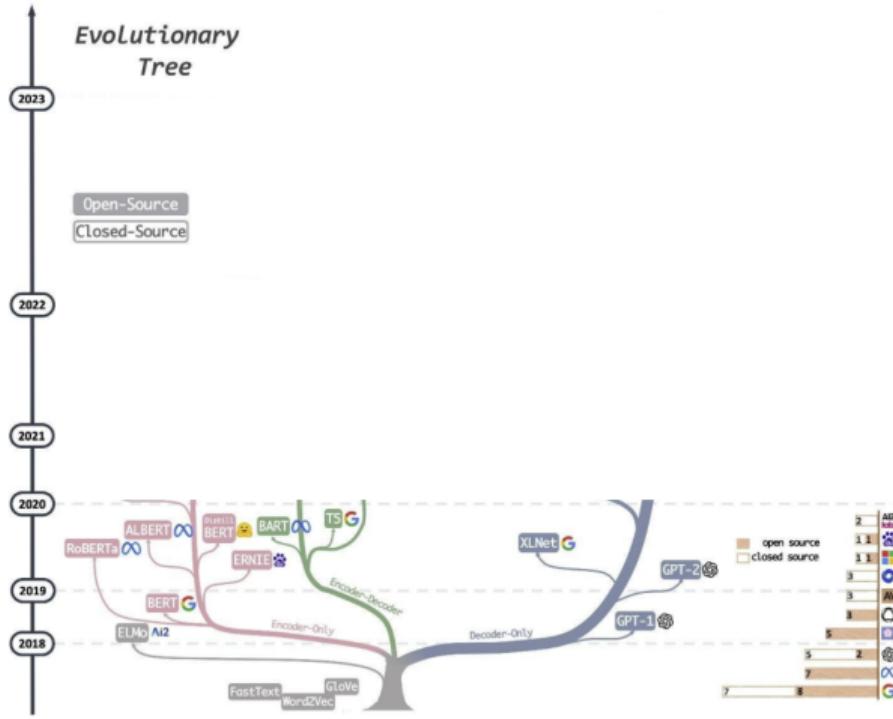
Zero-Shot Question-Answer Evaluation

A Concerning Shift towards Closed-Source Models⁶



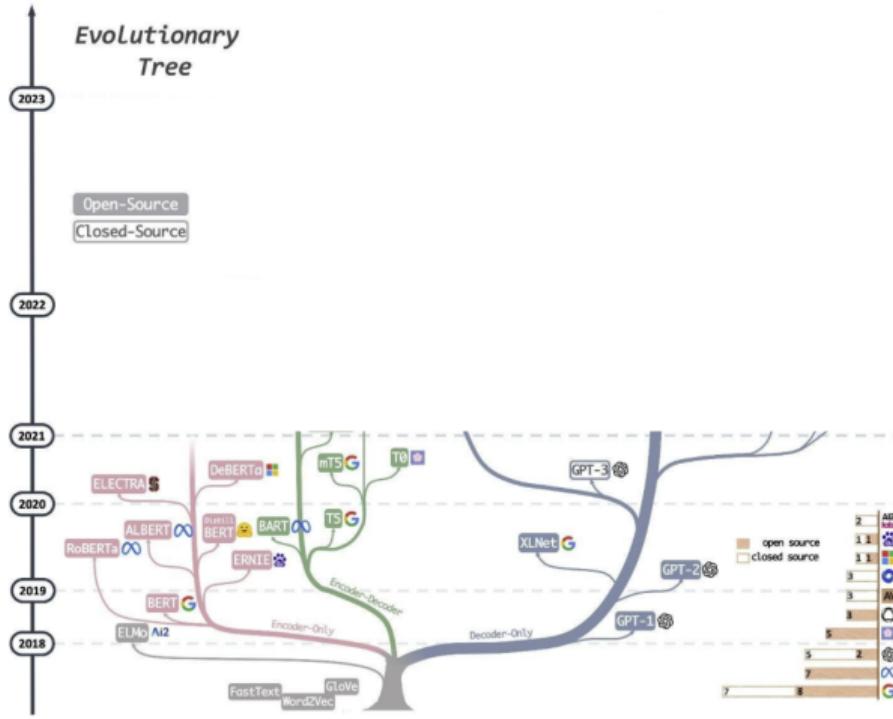
⁶ Source: Uspenskyi, Springsapps

A Concerning Shift towards Closed-Source Models⁶



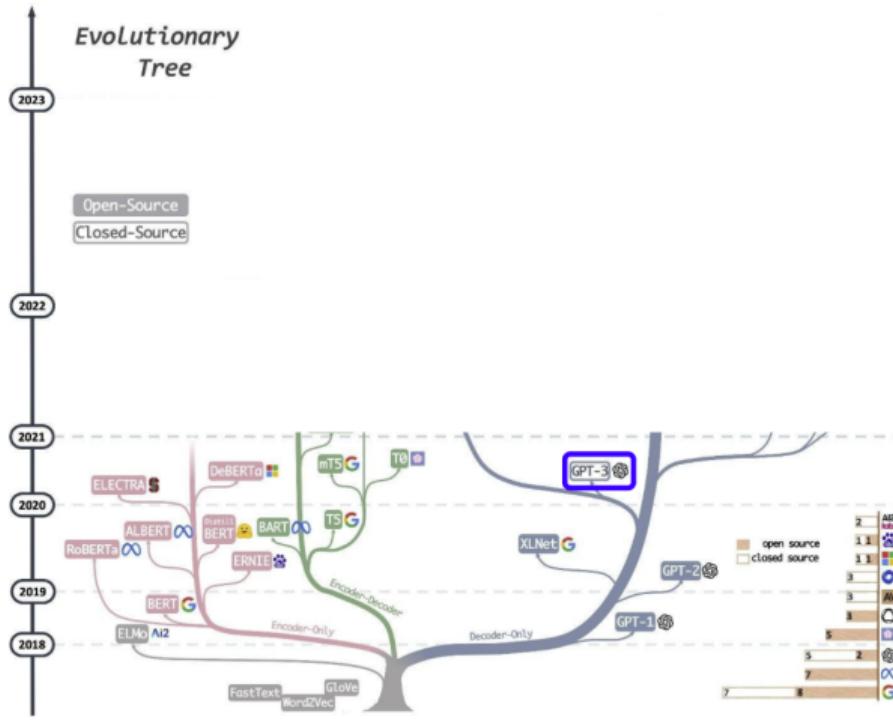
⁶ Source: Uspensky, Springsapps

A Concerning Shift towards Closed-Source Models⁶



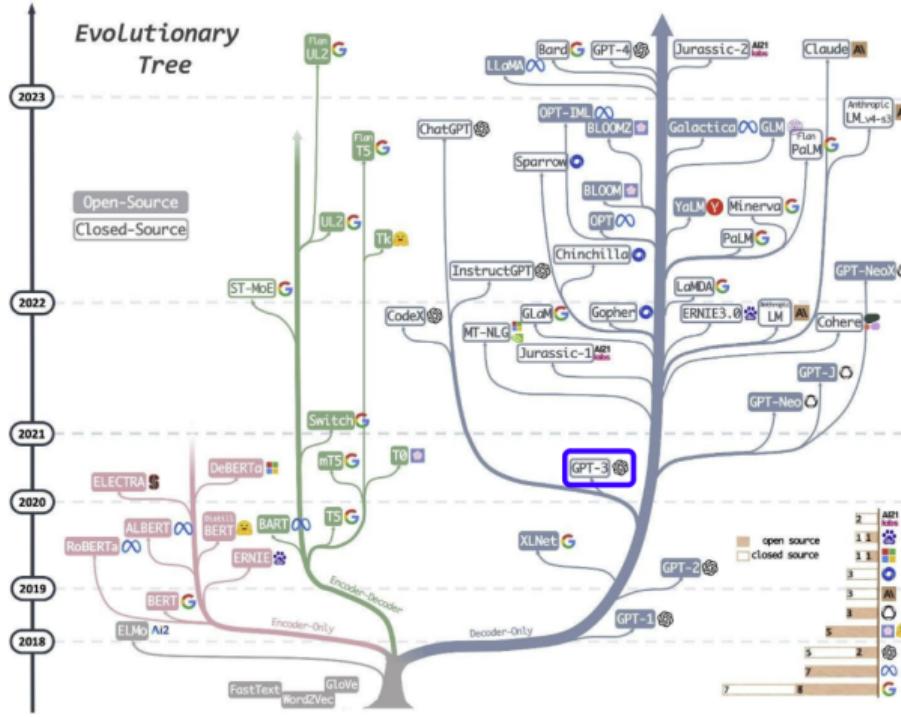
⁶ Source: Uspensky, Springsapps

A Concerning Shift towards Closed-Source Models⁶



⁶Source: Uspenskyi, Springsapps

A Concerning Shift towards Closed-Source Models⁶



⁶Source: Uspenskyi, Springsapps

Objectively Looking at the Closed-Source Trend

- **Pros:**

- **Confidentiality and Protection:** Safeguards proprietary techniques and intellectual property from competitors
- **Risk Control:** Reduces security risks and potential vulnerabilities by limiting access to detailed methodologies

Objectively Looking at the Closed-Source Trend

- **Pros:**

- **Confidentiality and Protection:** Safeguards proprietary techniques and intellectual property from competitors
- **Risk Control:** Reduces security risks and potential vulnerabilities by limiting access to detailed methodologies

- **Cons:**

- **Transparency and Trust Issues:** Lack of transparency may lead to skepticism and hinder collaboration
- **Reproducibility Challenges:** Independent verification becomes difficult without access to detailed training methods
- **Innovation Limitations:** Limits knowledge sharing and collaboration, stifling innovation within the field
- **Accessibility Barriers:** Creates barriers to entry for smaller teams, exacerbating disparities in access to technology and expertise

ChatGPT-4V⁷: Overview

- The release of GPT-4 without disclosing its training methodology establishes a significant precedent in the field

⁷Achiam et al, "GPT-4 Technical Report", OpenAI Report 2023

ChatGPT-4V⁷: Overview

- The release of GPT-4 without disclosing its training methodology establishes a significant precedent in the field
- The landscape of NLP can be delineated into distinct periods: before and after release of GPT-4

⁷Achiam et al, "GPT-4 Technical Report", OpenAI Report 2023

ChatGPT-4V⁷: Overview

- The release of GPT-4 without disclosing its training methodology establishes a significant precedent in the field
- The landscape of NLP can be delineated into distinct periods: before and after release of GPT-4
- Following GPT-4's lead, numerous companies refrained from divulging architectural details and methodology, including models like Gemini 1.5

⁷Achiam et al, "GPT-4 Technical Report", OpenAI Report 2023

ChatGPT-4V: Capabilities

- **Object Identification:** GPT-4 Vision accurately pinpoints and describes objects in images

ChatGPT-4V: Capabilities

- **Object Identification:** GPT-4 Vision accurately pinpoints and describes objects in images
- **Text Transcription:** Seamlessly extracts written or printed text from images

ChatGPT-4V: Capabilities

- **Object Identification:** GPT-4 Vision accurately pinpoints and describes objects in images
- **Text Transcription:** Seamlessly extracts written or printed text from images
- **Data Deciphering:** Extracts essential data from graphical content, translating it into text insights

ChatGPT-4V: Capabilities

- **Object Identification:** GPT-4 Vision accurately pinpoints and describes objects in images
- **Text Transcription:** Seamlessly extracts written or printed text from images
- **Data Deciphering:** Extracts essential data from graphical content, translating it into text insights
- **Multi-condition Processing:** Analyzes images under varying conditions with unmatched prowess

ChatGPT-4V: Capabilities

- **Object Identification:** GPT-4 Vision accurately pinpoints and describes objects in images
- **Text Transcription:** Seamlessly extracts written or printed text from images
- **Data Deciphering:** Extracts essential data from graphical content, translating it into text insights
- **Multi-condition Processing:** Analyzes images under varying conditions with unmatched prowess
- **Educational Assistance:** Converts diagrams and visual aids into detailed textual explanations

ChatGPT-4V: Capabilities

- **Object Identification:** GPT-4 Vision accurately pinpoints and describes objects in images
- **Text Transcription:** Seamlessly extracts written or printed text from images
- **Data Deciphering:** Extracts essential data from graphical content, translating it into text insights
- **Multi-condition Processing:** Analyzes images under varying conditions with unmatched prowess
- **Educational Assistance:** Converts diagrams and visual aids into detailed textual explanations
- **Coding Enhancement:** Interprets visual code structures and flowcharts into coding languages

ChatGPT-4V: Capabilities

- **Object Identification:** GPT-4 Vision accurately pinpoints and describes objects in images
- **Text Transcription:** Seamlessly extracts written or printed text from images
- **Data Deciphering:** Extracts essential data from graphical content, translating it into text insights
- **Multi-condition Processing:** Analyzes images under varying conditions with unmatched prowess
- **Educational Assistance:** Converts diagrams and visual aids into detailed textual explanations
- **Coding Enhancement:** Interprets visual code structures and flowcharts into coding languages
- **Design Understanding:** Breaks down design elements, offering textual insights for refining concepts

ChatGPT-4V: Qualitative Examples



ChatGPT-4V: Qualitative Examples

User What is funny about this image? Describe it panel by panel.



Source: [hmmm \(Reddit\)](#)

GPT-4 The image shows a package for a "Lightning Cable" adapter with three panels.

Panel 1: A smartphone with a VGA connector (a large, blue, 15-pin connector typically used for computer monitors) plugged into its charging port.

Panel 2: The package for the "Lightning Cable" adapter with a picture of a VGA connector on it.

Panel 3: A close-up of the VGA connector with a small Lightning connector (used for charging iPhones and other Apple devices) at the end.

The humor in this image comes from the absurdity of plugging a large, outdated VGA connector into a small, modern smartphone charging port.

ChatGPT-4V: Qualitative Examples

User What is unusual about this image?



Source: [Barnorama](#)

ChatGPT-4V: Qualitative Examples

User What is unusual about this image?



Source: [Barnorama](#)

GPT-4 The unusual thing about this image is that a man is ironing clothes on an ironing board attached to the roof of a moving taxi.

Gemini 1.5⁸: Overview



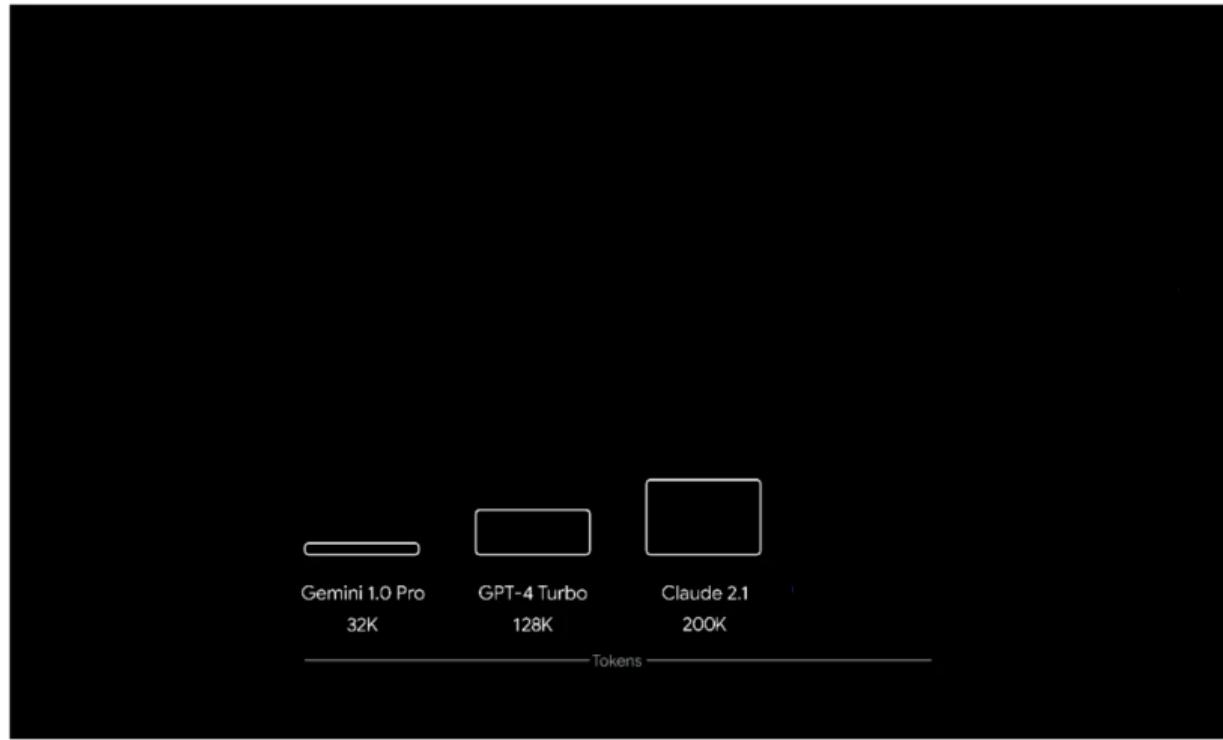
⁸Reid et al, "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context", arXiv 2024

Gemini 1.5⁸: Overview



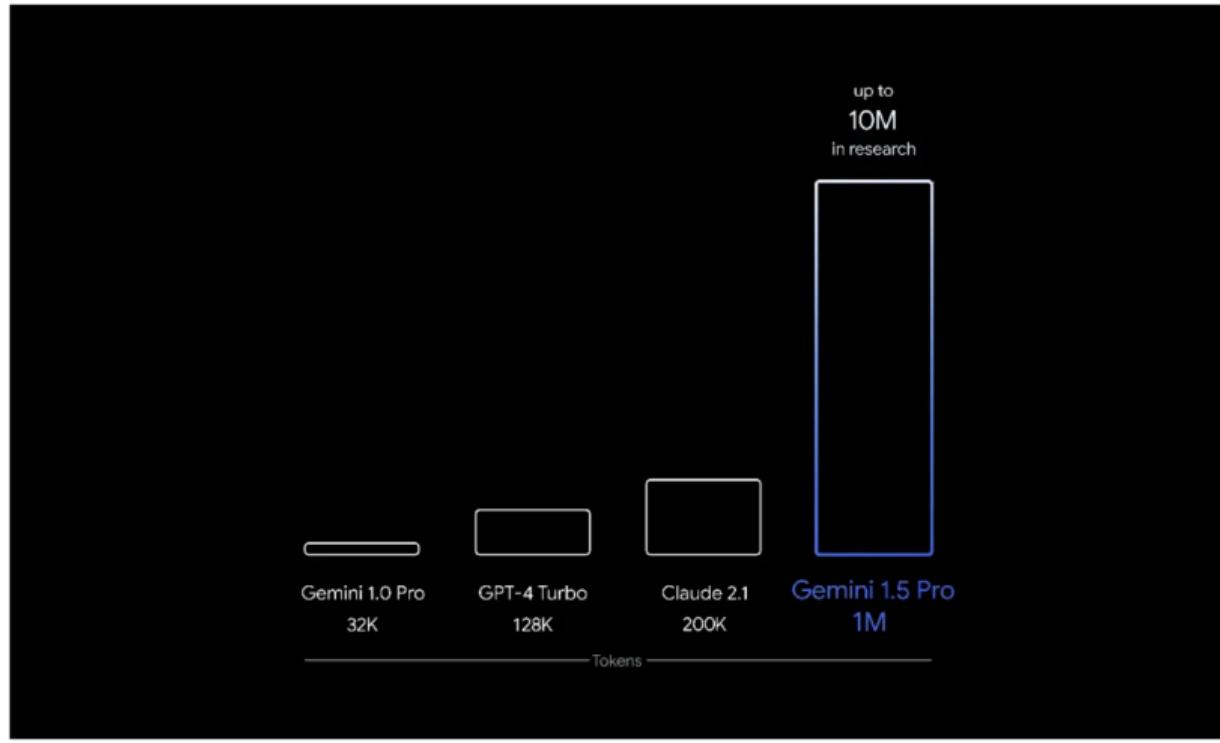
⁸Reid et al, "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context", arXiv 2024

Gemini 1.5⁸: Overview



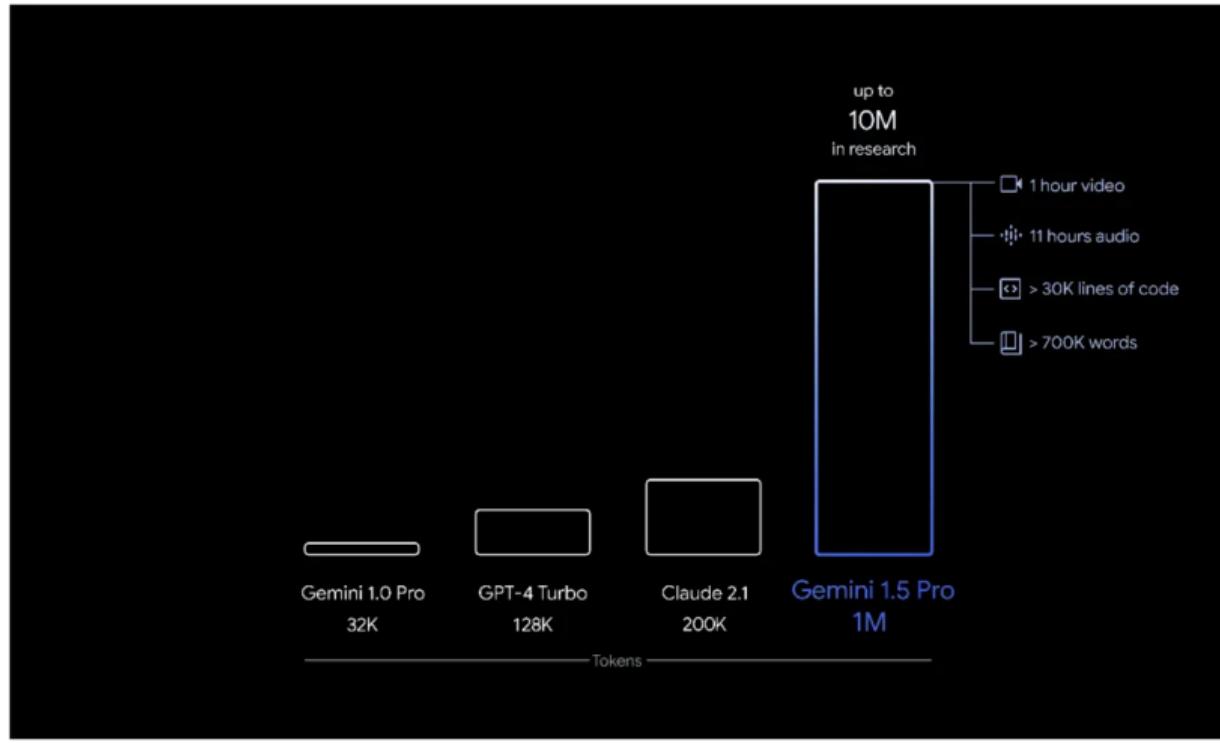
⁸Reid et al, "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context", arXiv 2024

Gemini 1.5⁸: Overview



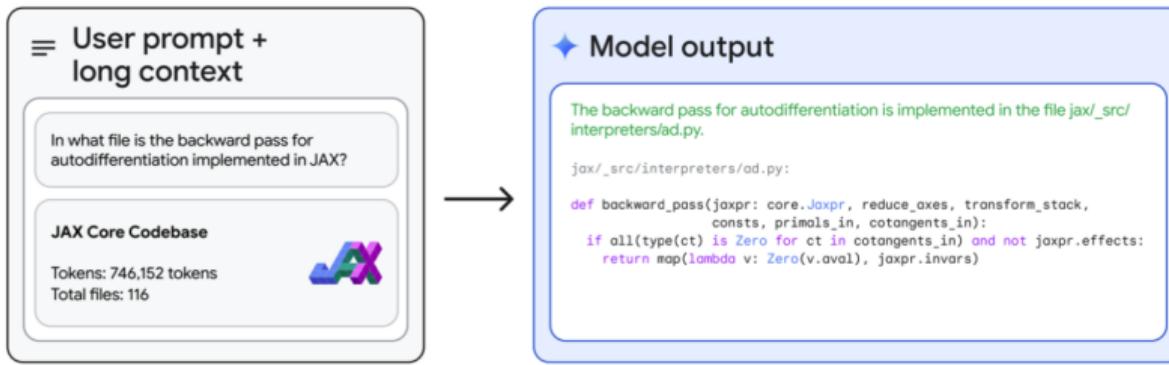
⁸Reid et al, "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context", arXiv 2024

Gemini 1.5⁸: Overview

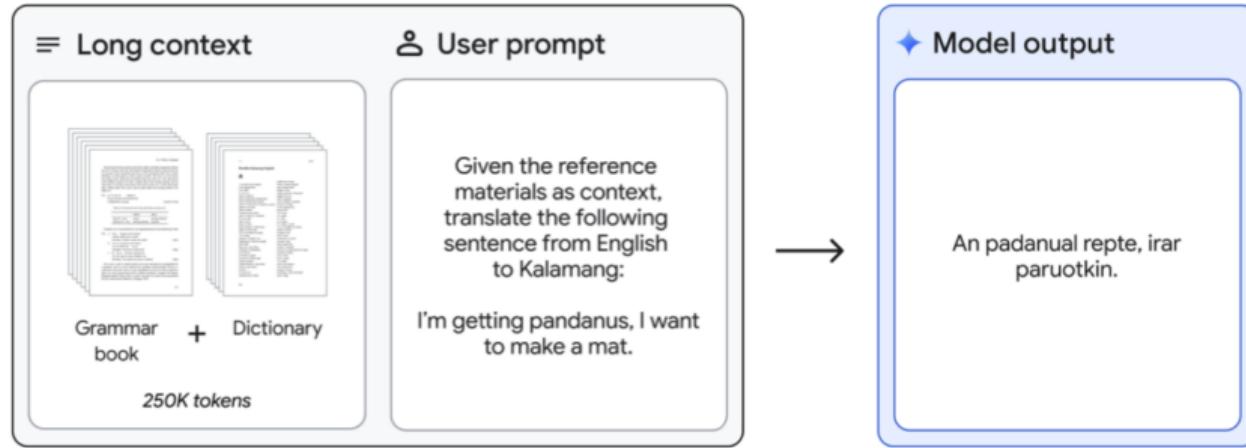


⁸Reid et al, "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context", arXiv 2024

Gemini 1.5: Qualitative Examples



Gemini 1.5: Qualitative Examples



References

- [1] Haotian Liu et al. "Visual Instruction Tuning". In: *Thirty-seventh Conference on Neural Information Processing Systems*. 2023.
- [2] Muhammad Maaz et al. "Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models". In: *arXiv:2306.05424* (2023).
- [3] OpenAI Josh Achiam et al. "GPT-4 Technical Report". In: 2023.
- [4] Machel Reid et al. "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context". In: 2024.