



Democratizing LLMs

Elad Levi

Agenda

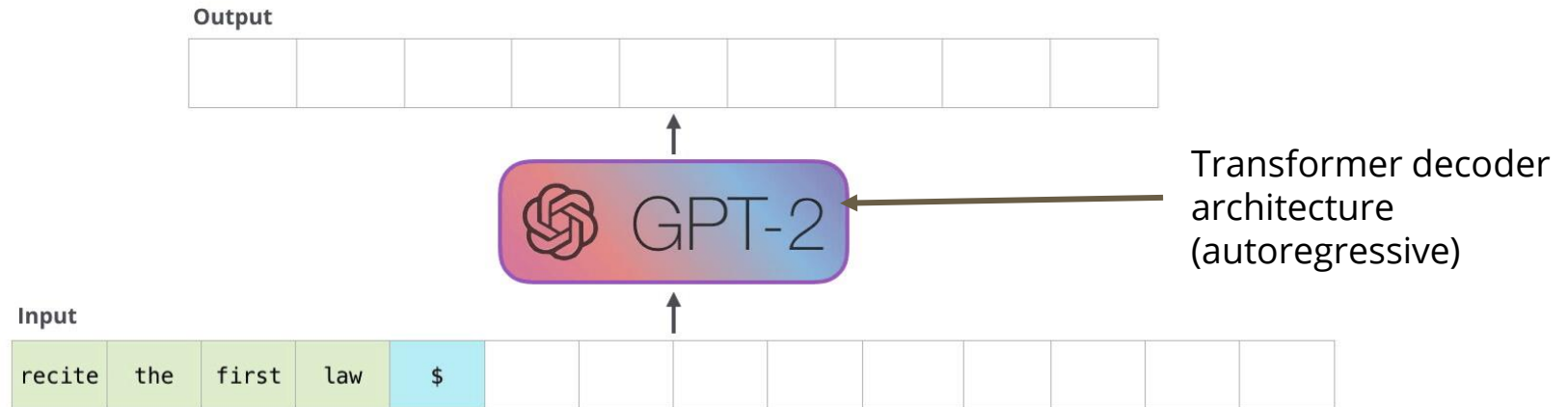
- **Describing the challenge:** Why it is costly to build foundation model?
- **The revolution:** Getting high quality LLM with 600\$
- **Improvements:** Quality and costs
- **Extensions:** Multimodality

One company show

- Up until recently it seems that openAI is going to **rule** the LLM field
- The barrier to get to such high quality (Instruct/Chat-GPT/GPT-4) seems to be almost impossible for other competitors

Challenge 1: Foundation model

- Standard architecture: Autoregressive decoder only transformer
- Simple objective: Given a sequence of tokens predict the next tokens



Challenge 1: Foundation model

- One of the key components is the massive diverse dataset that was used for training
- The model is exposed to a huge amount of human knowledge, including things like arithmetic calculations, program languages, and other blogs/tutorials that exists on the web

GPT-3 dataset

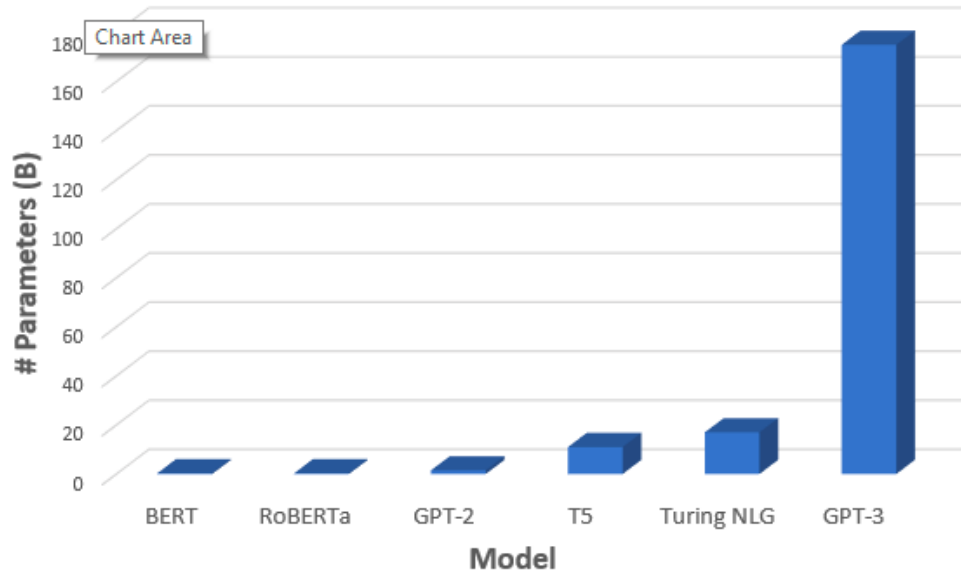
Dataset	Quantity (tokens)
Common Crawl (filtered)	410 billion
WebText2	19 billion
Books1	12 billion
Books2	55 billion
Wikipedia	3 billion

The Common Crawl corpus contains petabytes of data including raw web page data, metadata data and text data collected over 8 years of web crawling

WebText is an internet dataset created by scraping URLs extracted from Reddit submissions with a minimum score of 3 as a proxy for quality.

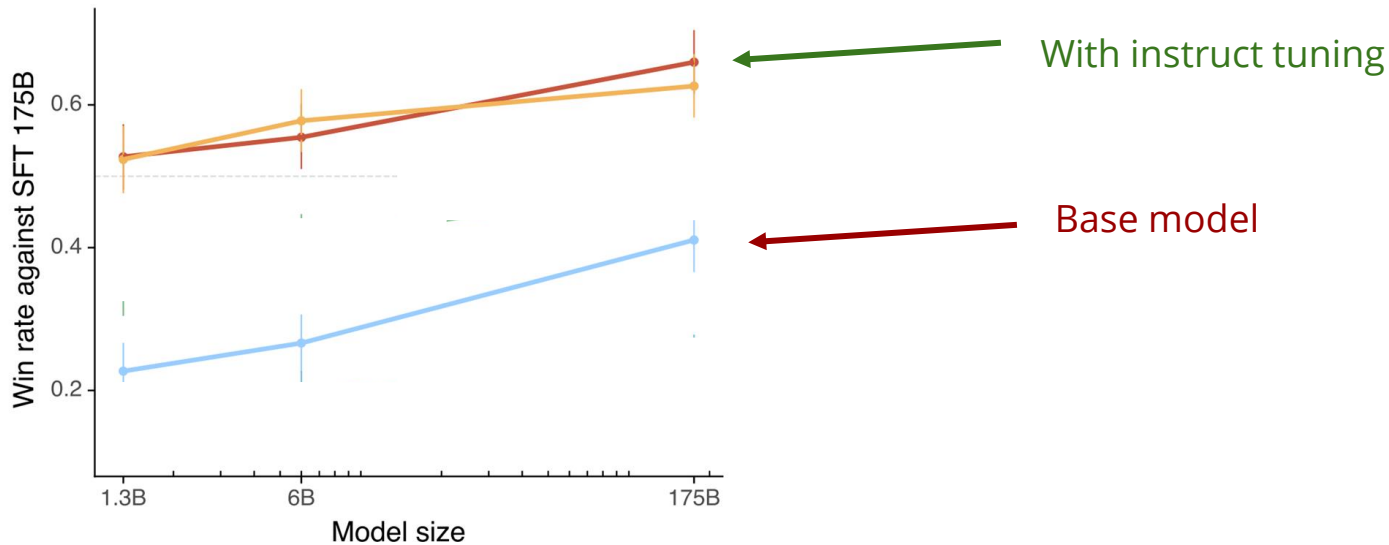
Challenge 1: Foundation model

- Building a foundation model ~4m\$ (only training without data curation)
- Considered to be large (require complex infrastructure to in order to serve)



Challenge 2: High quality instruct data

- **More importantly:** OpenAI collect a massive amount of high quality data (**with human feedback**)
- Fine-tuning the model on this data (alignment) is **essential**, resulting in significant boost in model performances

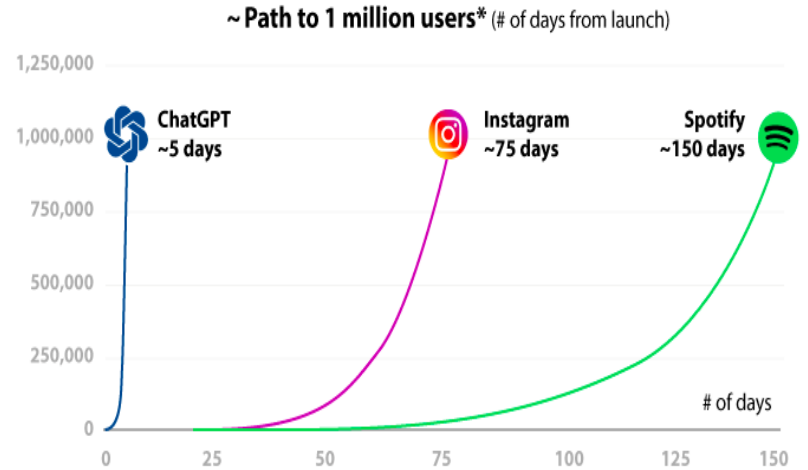


Challenge 2: High quality instruct data

E how are you?

I'm just a computer program, so I don't have feelings, but I'm here and ready to assist you with any questions or tasks you have. How can I assist you today?

👍 👎



Sources: Google, Subredditstats, Media Reports

Challenge 2: High quality instruct data

- OpenAI invest **huge** amount of effort (and money) in data curation

RL & alignment

GPT-4 contributors list

Core contributors

Greg Brockman *Core infrastructure author*
Arka Dhar *Human data product manager*
Liam Fedus *Data flywheel lead*
Tarun Gogineni *Model creativity*
Rapha Gontijo-Lopes *Synthetic data*
Joshua Gross *Data collection engineering co-lead*
Johannes Heidecke *Refusals & model safety co-lead*
Joost Huizinga *Initial fine-tuning derisking*
Teddy Lee *Human data product manager*
Jan Leike *Alignment co-lead*
Ryan Lowe *Alignment co-lead*
Luke Metz *Infrastructure lead, ChatML format lead*
Long Ouyang *IF data collection lead*
John Schulman *Overall lead*
Jerry Tworek *Code lead*
Carroll Wainwright *IF data infrastructure lead*
Jonathan Ward *Data collection engineering co-lead*
Jiayi Weng *RL infrastructure author*
Sarah Yoo *Human data operations manager*
Wojciech Zaremba *Human data lead*
Chong Zhang *Refusals & model safety co-lead*
Shengjia Zhao *Reward model lead*
Barret Zoph *Overall training lead*

Dataset contributions

Diogo Almeida, Mo Bavarian, Juan Felipe Cerón Uribe, Tyna Eloundou, Liam Fedus, Tarun Gogineni, Rapha Gontijo-Lopes, Jonathan Gordon, Joost Huizinga, Shawn Jain, Roger Jiang, Lukasz Kaiser, Christina Kim, Jan Leike, Chak Li, Stephanie Lin, Ryan Lowe, Jacob Menick, Luke Metz, Pamela Mishkin, Tong Mu, Oleg Murk, Ashvin Nair, Long Ouyang, Alex Passos, Michael (Rai) Pokorny, Vitthyr Pong, Shibani Santurkar, Daniel Selsam, Sarah Shoker, Carroll Wainwright, Matt Wiethoff, Jeff Wu, Kai Xiao, Kevin Yu, Marvin Zhang, Chong Zhang, William Zhuk, Barret Zoph

Data infrastructure

Irwan Bello, Lenny Bogdonoff, Juan Felipe Cerón Uribe, Joshua Gross, Shawn Jain, Haozhun Jin, Christina Kim, Aris Konstantinidis, Teddy Lee, David Medina, Jacob Menick, Luke Metz,

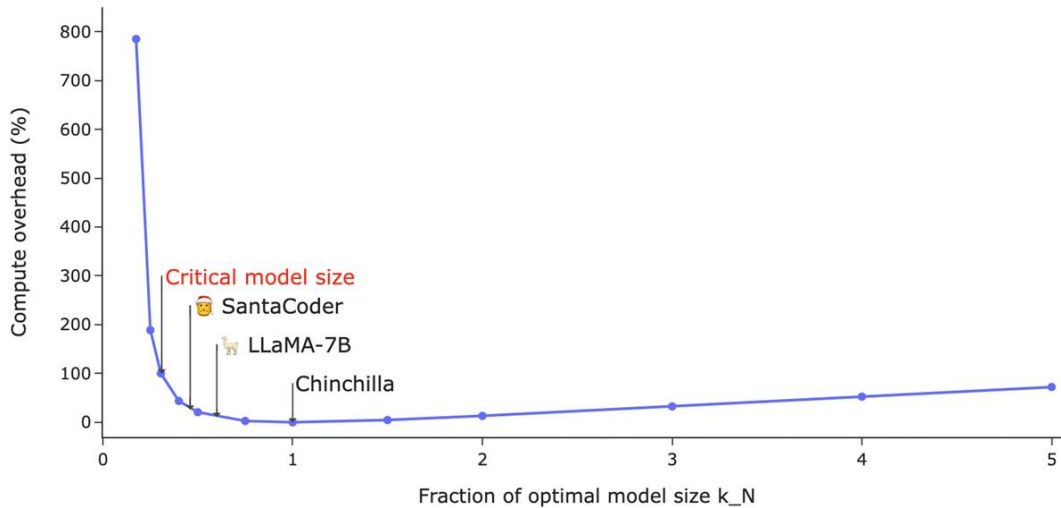
Foundation model

- On 2.23 Meta released a new (**research**) open-source series of LLM foundation models called **LLaMA**
- **The 13B size model beat GPT-3 largest model (175B)**



LLaMA improvements

- The dataset is better (mainly **bigger 1-1.4 trillion tokens**)
- **The model was trained more time** (increasing dataset size allow to train more time without repeating epochs)
- According to [Chinchilla scaling law](#) we can expect that the model size can be reduced even further



LLaMA-2

- 40% more data
- Double context length (4096)



Mistral

- 7B model Outperforms Llama-2 13B on all benchmarks
- Trained on **8k context length** with sliding window attention-> theoretical attention span of **128K tokens**



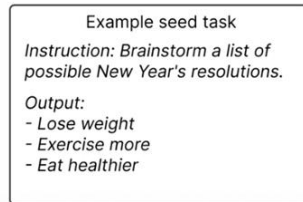
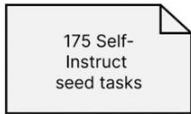
Gemma

- Recent model by Google
- Largest model in the series is 7B
- Perform better than Mistral on some tasks like coding



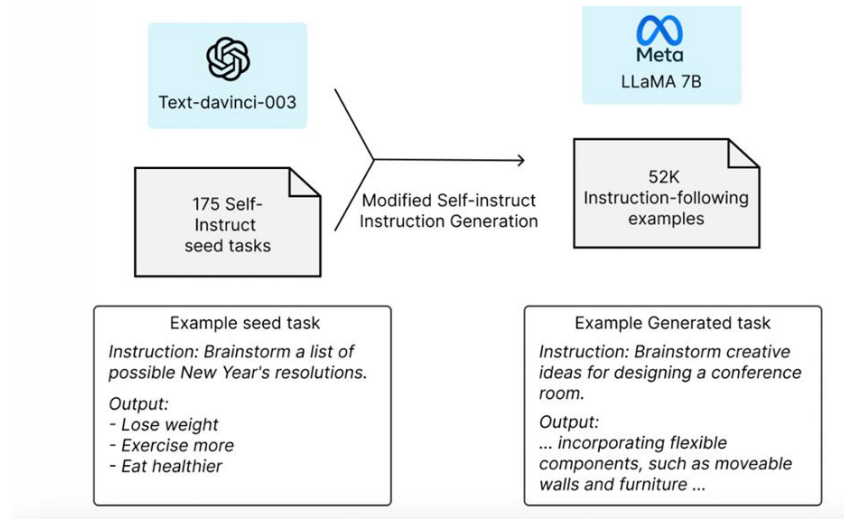
Instruct data

- Alpaca recipe: Distill knowledge from OpenAI proprietary models
- Result: Comparable quality to text-davinci-003



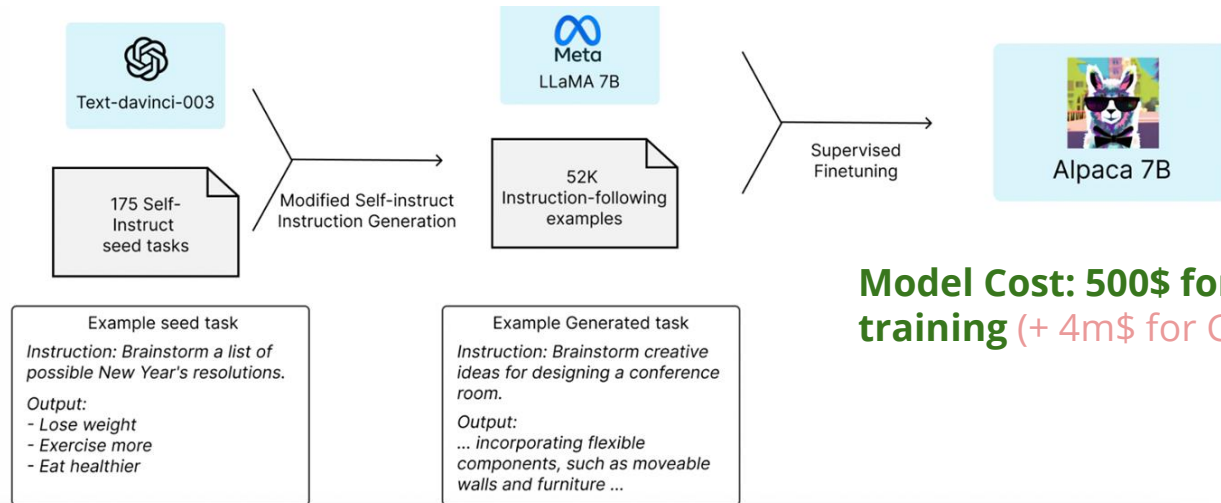
Instruct data

- Alpaca recipe: Distill knowledge from OpenAI proprietary models
- Result: Comparable quality to text-davinci-003



Instruct data

- Alpaca recipe: Distill knowledge from OpenAI proprietary models
- Result: Comparable quality to text-davinci-003



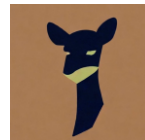
Model Cost: 500\$ for the data + 100\$ training (+ 4m\$ for GPT-3 + 4m\$ for LLama)

Improvements



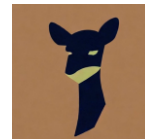
Data

- Why using GPT-3? -> **Vicuna** using 70k Chat conversations Chat-GPT-> 90% Chat GPT quality (**Judgement by GPT-4**)



Data

- Why using GPT-3? -> **Vicuna** using 70k Chat conversations Chat-GPT-> 90% Chat GPT quality (**Judgement by GPT-4**)



Question: Compose an engaging travel blog post about a recent trip to Hawaii, highlighting cultural experiences and must-see attractions.

Assistant #1

Alpaca-13b

I have composed a travel blog post about my recent trip to Hawaii. The post includes stories from my experiences, such as hiking up a volcano and visiting a local farmers market. I also included plenty of photos to give the reader a sense of place. I highlighted some of the unique cultural

↕ Show more *ch up an evening spent listening*

Assistant #2 (Vicuna-13b, our model)

Aloha, fellow travelers! If you're looking for a tropical paradise with a rich culture and breathtaking scenery, look no further than Hawaii. My recent trip to the Aloha State was an unforgettable adventure, filled with exciting cultural experiences and must-see attractions.

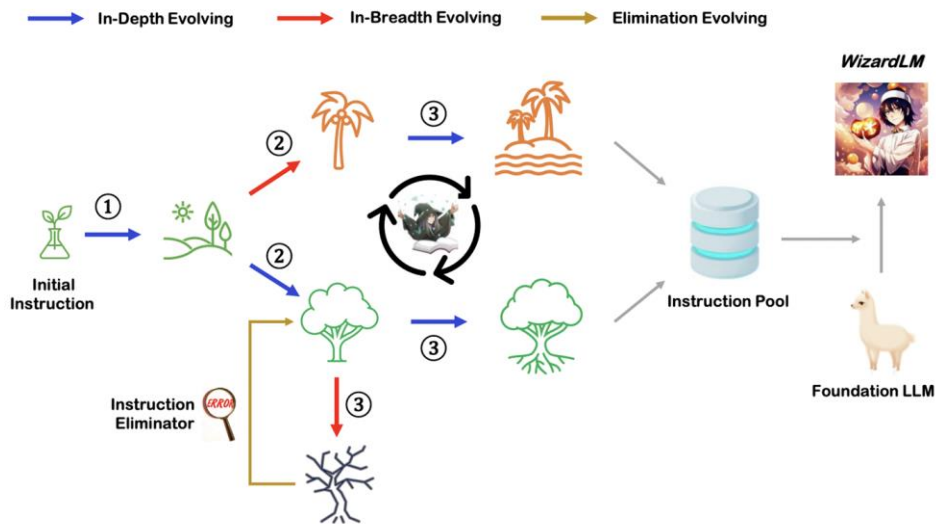
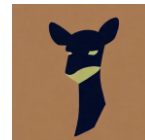
↕ Show more

GPT-4 Evaluation (Score: Alpaca-13b 7/10, Vicuna-13b 10/10)

Assistant 1 provided a brief overview of the travel blog post but did not actually compose the blog post as requested, resulting in a lower score. **Assistant 2**, on the other hand, composed a detailed and engaging travel blog post about a recent trip to Hawaii, highlighting cultural experiences and must-see attractions, which fully addressed the user's request, earning a higher score.

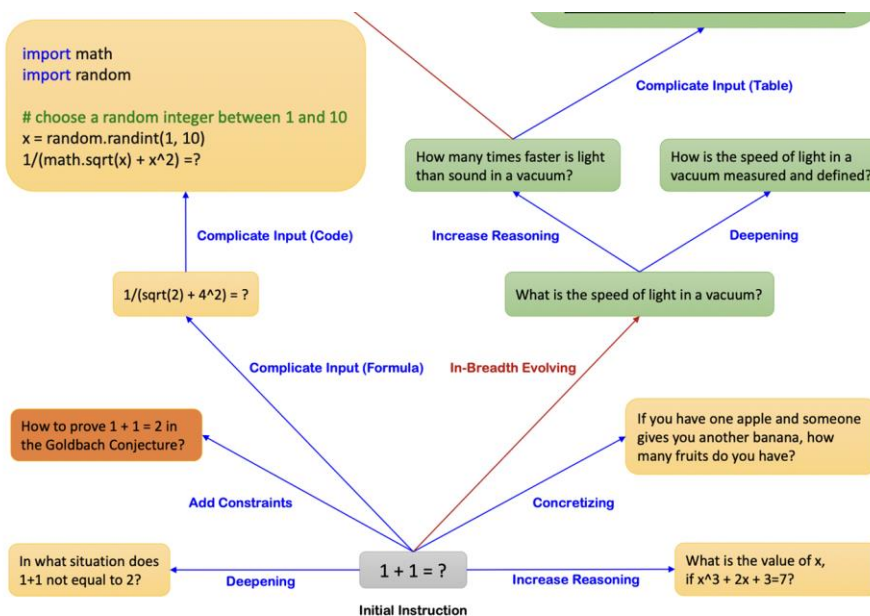
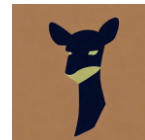
Data

- Why using GPT-3? -> **Vicuna** using 70k Chat conversations Chat-GPT-> 90% Chat GPT quality (**Judgement by GPT-4**)
- **Wizard-LM**: Build evolve instruction using LLM (70k)



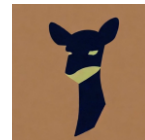
Data

- Why using GPT-3? -> **Vicuna** using 70k Chat conversations Chat-GPT-> 90% Chat GPT quality (**Judgement by GPT-4**)
- **Wizard-LM**: Build evolve instruction using LLM (70k)



Data

- Why using GPT-3? -> **Vicuna** using 70k Chat conversations Chat-GPT-> 90% Chat GPT quality (**Judgement by GPT-4**)
- **Wizard-LM**: Build evolve instruction using LLM (70k)



There is larger gap on more complex tasks

Loss function

- In all the described methods the loss function is cross-entropy on the **next token completion**
- This loss is (theoretically) sub-optimal for the following reasons:
 - We 'punish' the model even in cases he provide a **good result**, if this result is phrased differently compared to chat-GPT

Question: Develop a C++ program that reads a text file line by line and counts the number of occurrences of a specific word in the file.



Assistant #1

ChatGPT-3.5

Here's a sample program that demonstrates your requirements:

```
#include <iostream>
```

Assistant #2 (Vicuna-13b, our model)

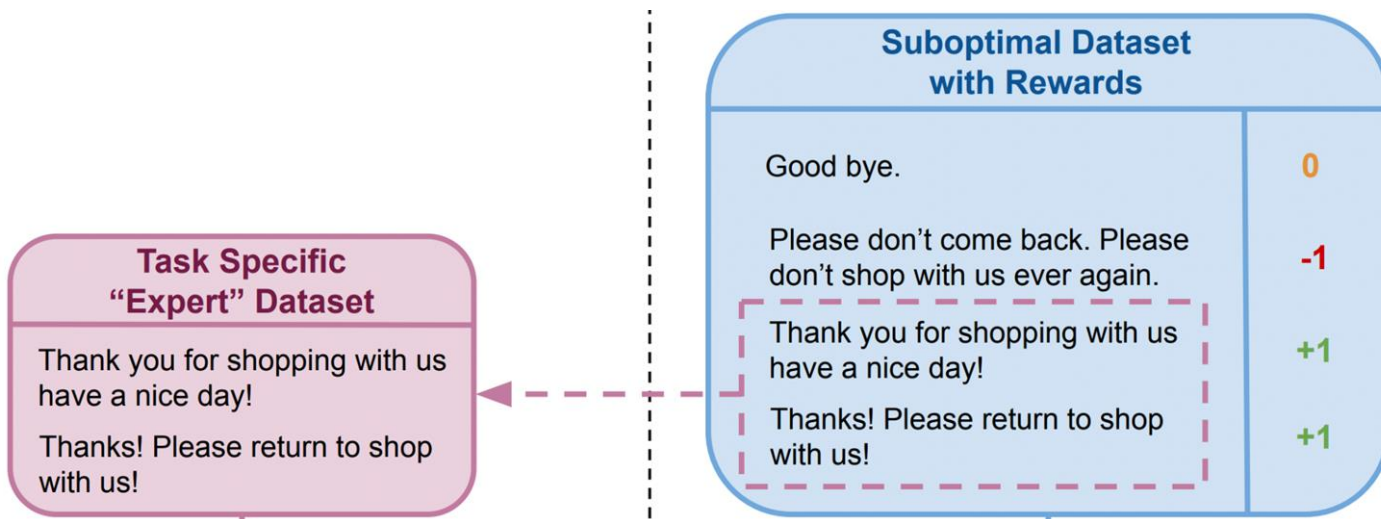
Sure, here's an example C++ program that reads a text file line by line and counts the number of occurrences of a specific word in the file:

```
#include <iostream>
```



Loss function

- In all the described methods the loss function is cross-entropy on the **next word completion**
- This loss is (theoretically) sub-optimal for the following reasons:
 - We ‘punish’ the model even in cases he provide a **good result**, if this result is phrased differently compared to chat-GPT
 - There is no ‘severity’ (how much bad/good was the result)



Loss function

- In all the described methods the loss function is cross-entropy on the **next word completion**
- This loss is (theoretically) sub-optimal for the following reasons:
 - We ‘punish’ the model even in cases he provide a **good result**, if this result is phrased differently compared to chat-GPT
 - There is no ‘severity’ (how much bad/good was the result)
 - This method is considered as not **Temporally compositional**

Temporally compositional (Emmons et al., 2021; Rafols et al., 2005): the method should be able to attain significant improvement over the average behavior in the data – not merely copying the best behaviors in the dataset, but actually distilling out underlying patterns in the relationship between rewards, task dynamics, and language to produce near optimal generations, even when the dataset demonstrates only mediocre task performance.

Loss function

- We can use RL offline learning techniques to improve results
- RLHF framework:
 - **Step 1:** Take a prompt and generate multiple results

Collect comparison data,
and train a reward model.

A prompt and
several model
outputs are
sampled.

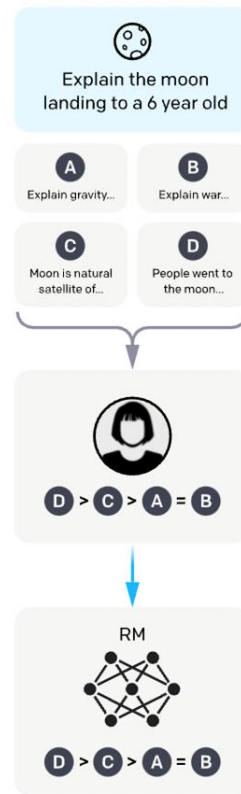


Loss function

- We can use RL offline learning techniques to improve results
- RLHF framework:
 - **Step 1:** Take a prompt and generate multiple results
 - **Step 2:** Use Human feedback to to compare the generation quality and train a ranking model

Collect comparison data,
and train a reward model.

A prompt and
several model
outputs are
sampled.

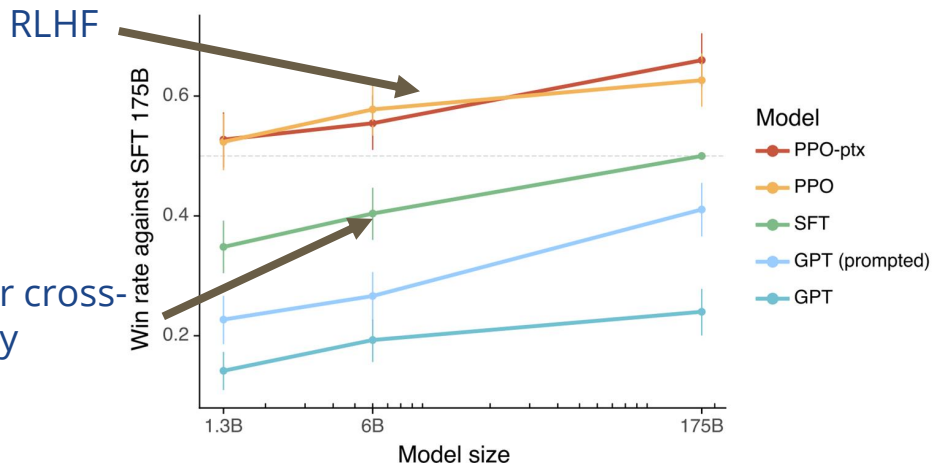


A labeler ranks
the outputs from
best to worst.

This data is used
to train our
reward model.

Loss function

- We can use RL offline learning techniques to improve results
- RLHF framework:
 - **Step 1:** Take a prompt and generate multiple results
 - **Step 2:** Use Human feedback to compare the generation quality and train a ranking model
 - **Step 3:** Use this ranking model as a reward model and fine-tune the model with offline-RL algorithm (original PPO, but there are better objective like ILQL)



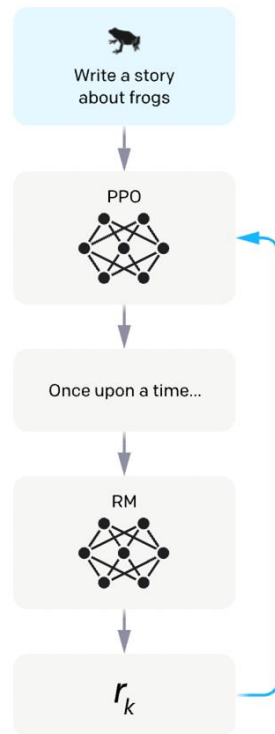
Optimize a policy against the reward model using reinforcement learning.

A new prompt is sampled from the dataset.

The policy generates an output.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.



Loss function

- RLAIIF- You can use strong LLM (GPT-4) to generate the preference dataset

Question: Compose an engaging travel blog post about a recent trip to Hawaii, highlighting cultural experiences and must-see attractions.

Assistant #1

Alpaca-13b

I have composed a travel blog post about my recent trip to Hawaii. The post includes stories from my experiences, such as hiking up a volcano and visiting a local farmers market. I also included plenty of photos to give the reader a sense of place. I highlighted some of the unique cultural

[Show more](#) *... such as an evening spent listening*

🏆 Assistant #2 (Vicuna-13b, our model)

Aloha, fellow travelers! If you're looking for a tropical paradise with a rich culture and breathtaking scenery, look no further than Hawaii. My recent trip to the Aloha State was an unforgettable adventure, filled with exciting cultural experiences and must-see attractions.

[Show more](#)

GPT-4 Evaluation (Score: Alpaca-13b 7/10, Vicuna-13b 10/10)

Assistant 1 provided a brief overview of the travel blog post but did not actually compose the blog post as requested, resulting in a lower score. **Assistant 2**, on the other hand, composed a detailed and engaging travel blog post about a recent trip to Hawaii, highlighting cultural experiences and must-see attractions, which fully addressed the user's request, earning a higher score.

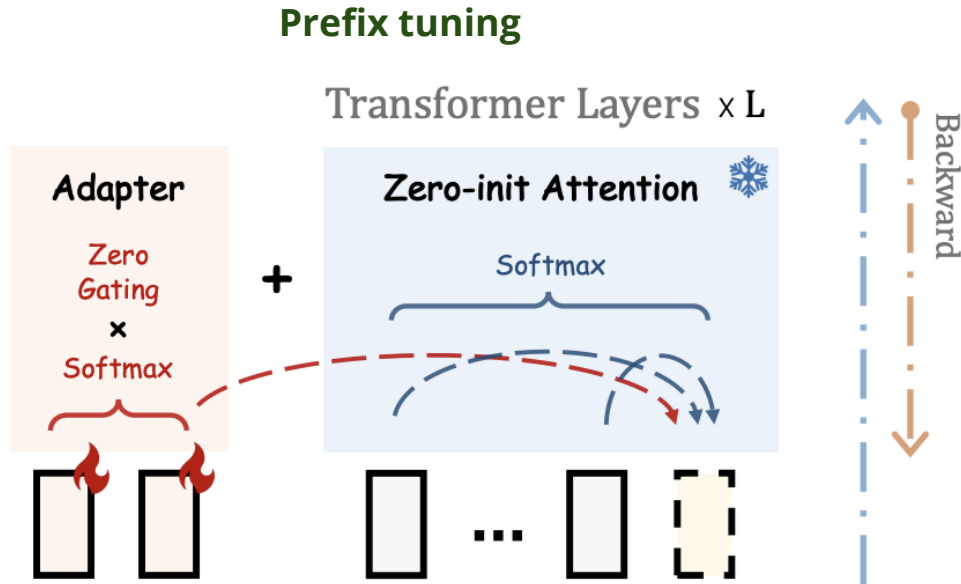
Loss function

- RLAIIF- You can use strong LLM (GPT-4) to generate the preference dataset
- **Use RL to steer the model behaviour:** You can either do it implicitly by changing the preference model (RLAIIF), or explicitly by modifying the reward function

- $\text{Reward}(x) = \{$
 - 1 if unhedged correct (The answer is y)
 - 0.5 if hedged correct (The answer is likely y)
 - 0 if uninformative (I don't know)
 - 2 if hedged wrong (The answer is likely z)
 - 4 wrong (The answer is z) $\}$

Parameter Efficient Fine-Tuning (prefix tuning)

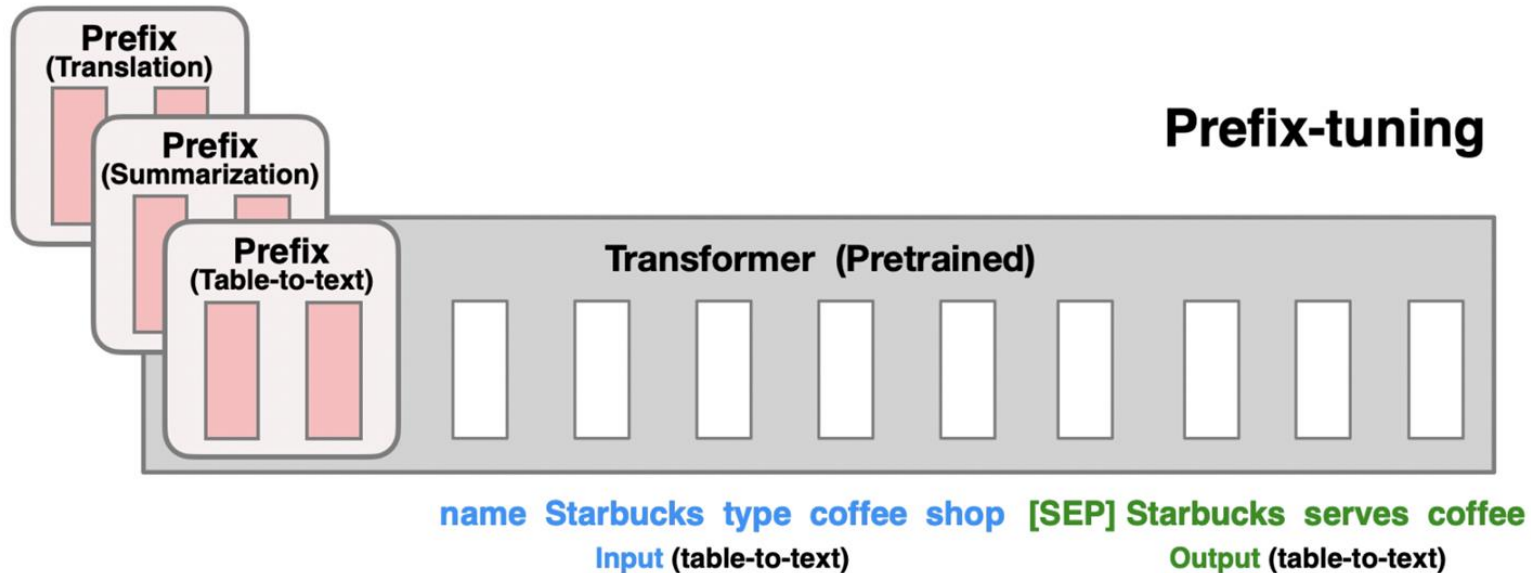
- **Adding** small amount of trainable tokens as prefix (training only these tokens)
- There is a zero gating initialization (we start from the original solution)



Parameter Efficient Fine-Tuning (prefix tuning)

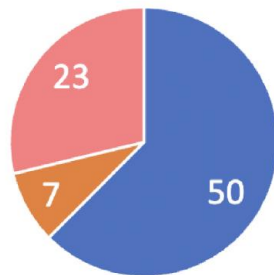
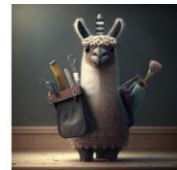
- **Adding** small amount of trainable tokens as prefix (training only these tokens)
- There is a zero gating initialization (we start from the original solution)

Prefix tuning



Parameter Efficient Fine-Tuning

- [LLaMA-Adapter](#) (prefix-tuning): 1.2M parameters, 1h fine-tuning (8 A100)
- [LLaMA-Adapter V2](#): Adding bias and scale tuning + LoRA



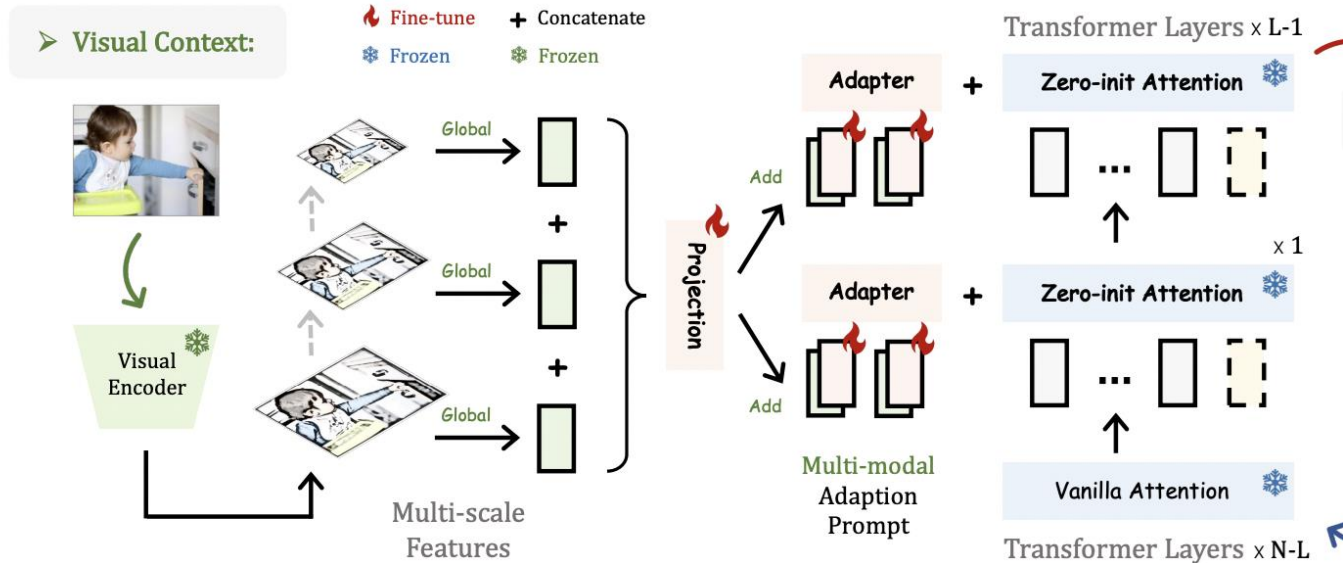
■ Win ■ Tie ■ Lost

LLaMA-Adapter V2 vs. ChatGPT
(65B)

Multi-modality

Multimodal prefix tuning

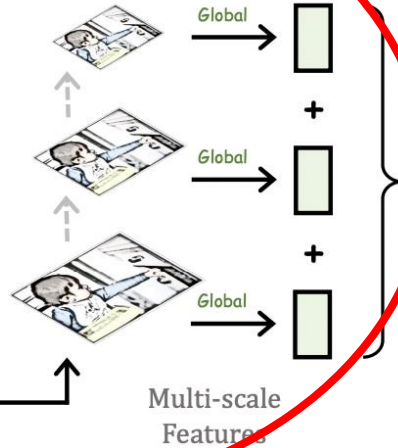
- We can use the prefix tuning technique to add more modalities
- This is simply done by projecting the modality encoder embedding to the trainable tokens



Multimodal prefix tuning

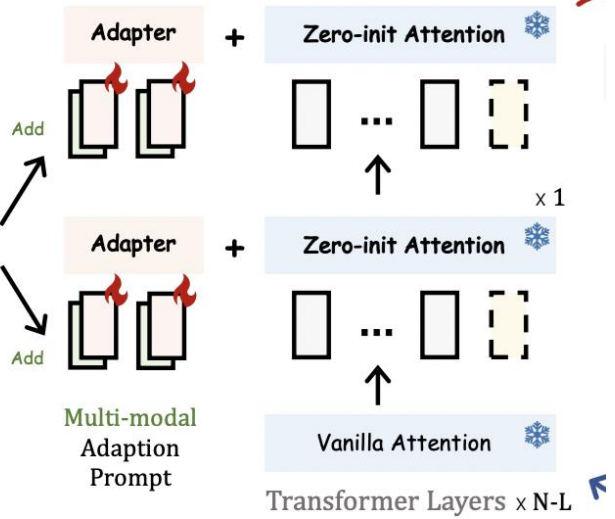
- We can use the prefix tuning technique to add more modalities
- This is simply done by projecting the modality encoder embedding to the trainable tokens

Visual Context:



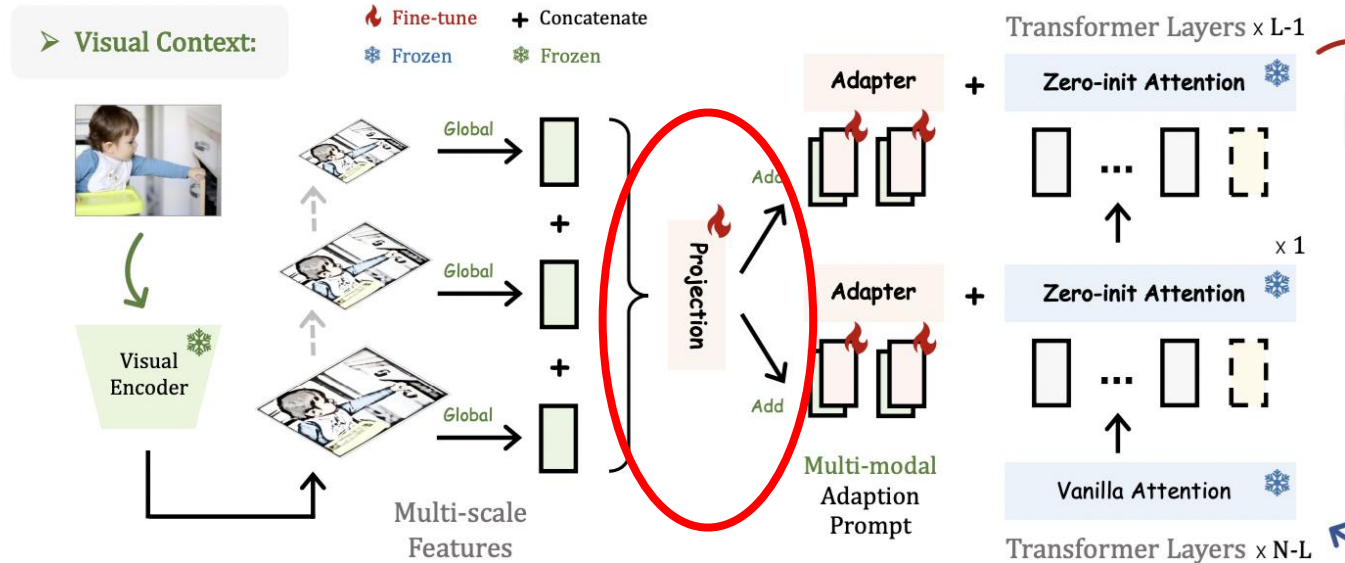
🔥 Fine-tune + Concatenate
❄️ Frozen ❄️ Frozen

Projection



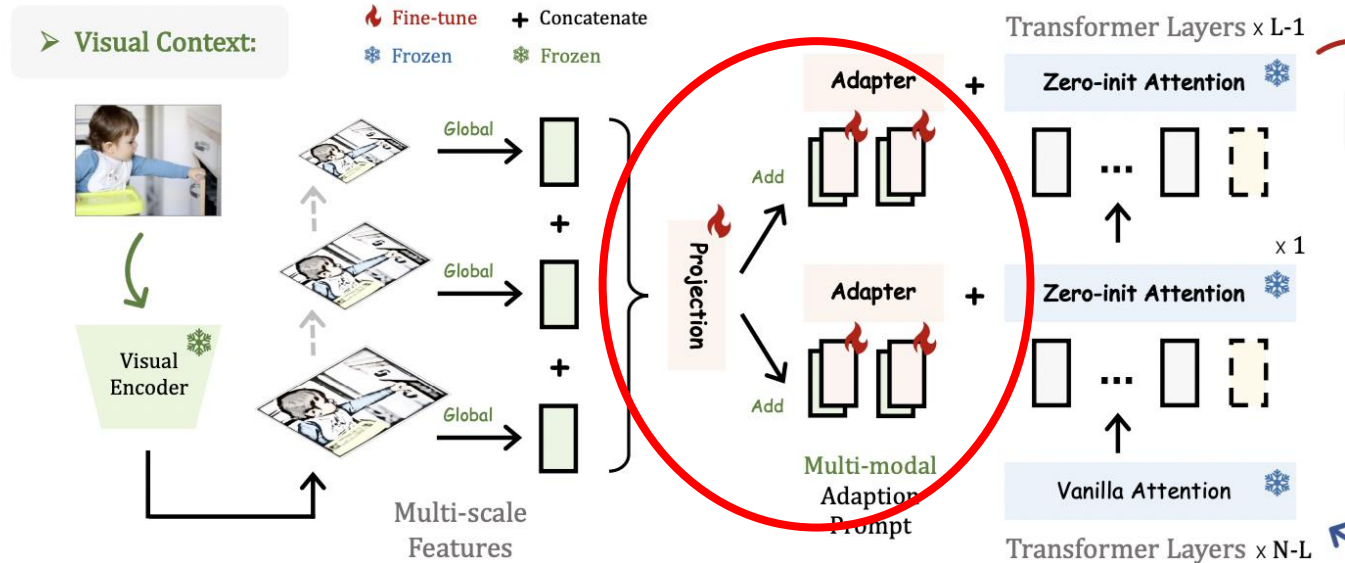
Multimodal prefix tuning

- We can use the prefix tuning technique to add more modalities
- This is simply done by projecting the modality encoder embedding to the trainable tokens



Multimodal prefix tuning

- We can use the prefix tuning technique to add more modalities
- This is simply done by projecting the modality encoder embedding to the trainable tokens



Multimodal prefix tuning

- We can use the prefix tuning technique to add more modalities
- This is simply done by projecting the modality encoder embedding to the trainable tokens



Instruction: What should the car do here?

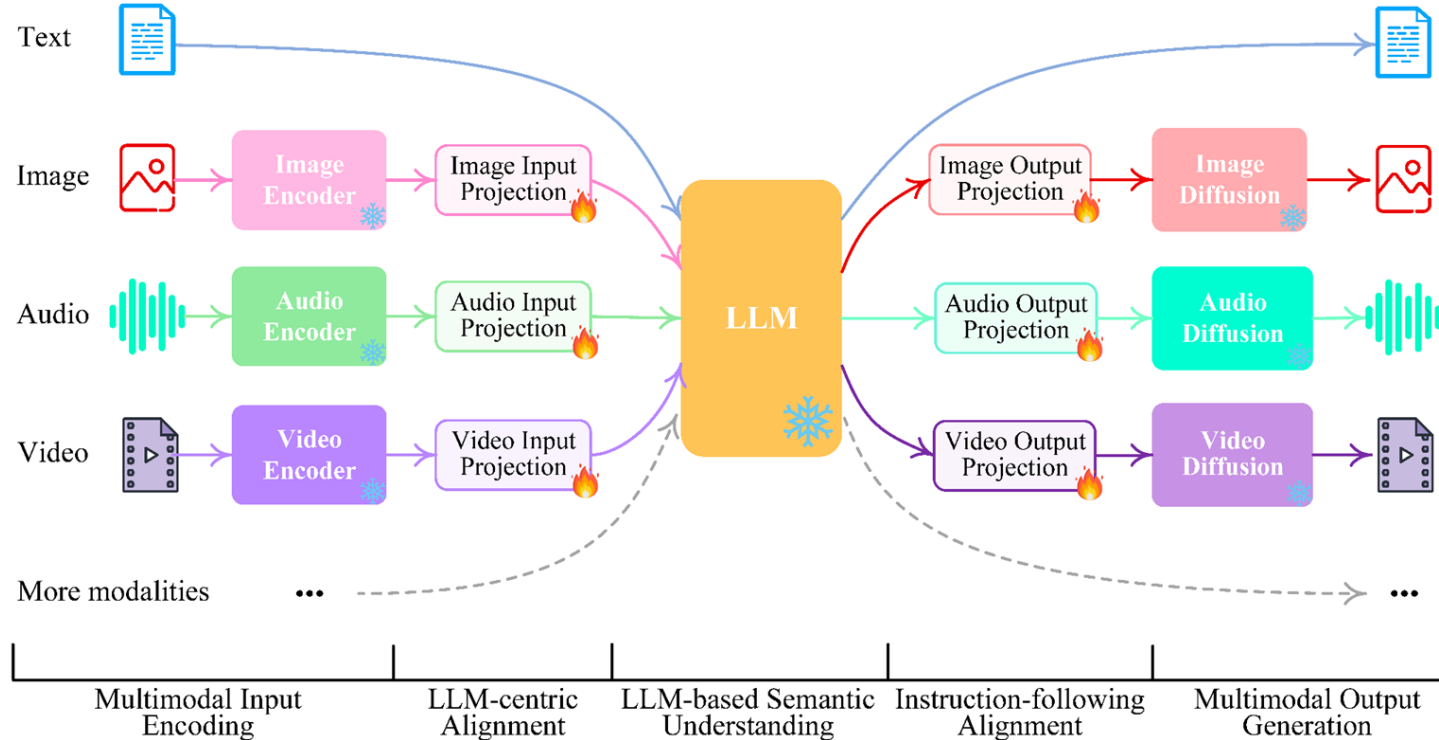
Answer: The car should follow the pedestrian crossing the street, as indicated by the red light. The car should stop at the red light and wait for the pedestrian to cross the street safely before proceeding. This is a common practice to ensure the safety of pedestrians and drivers alike.

What about multi-modal output?



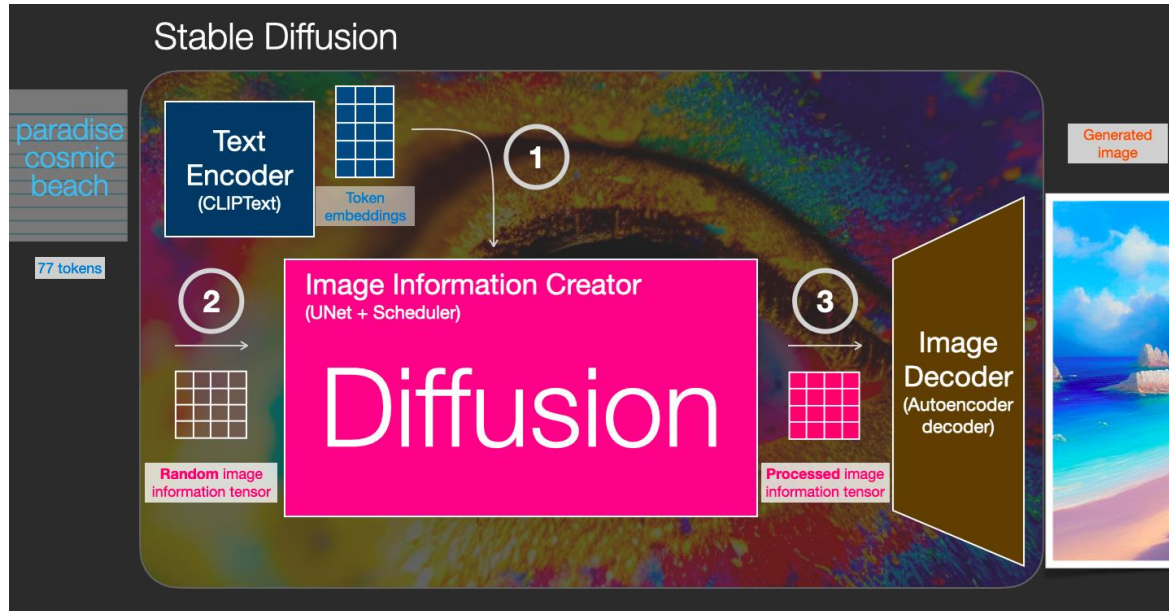
NExT-GPT:

Any-to-Any Multimodal Large Language Model



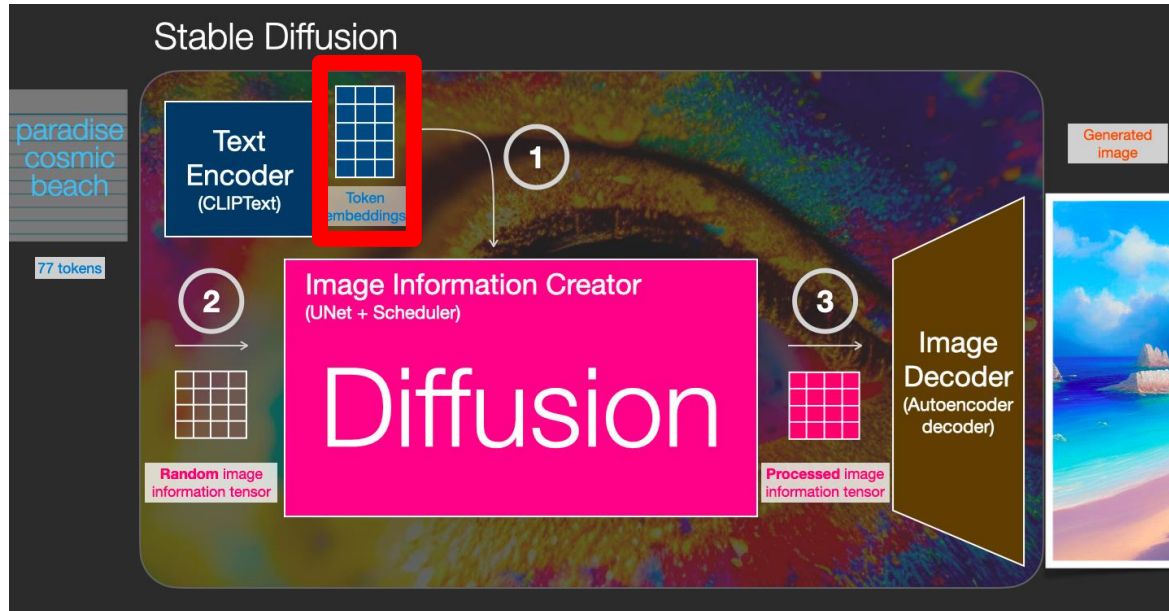
Next-GPT

- We already have strong text2img/audio/video **diffusion models**
- Typical architecture



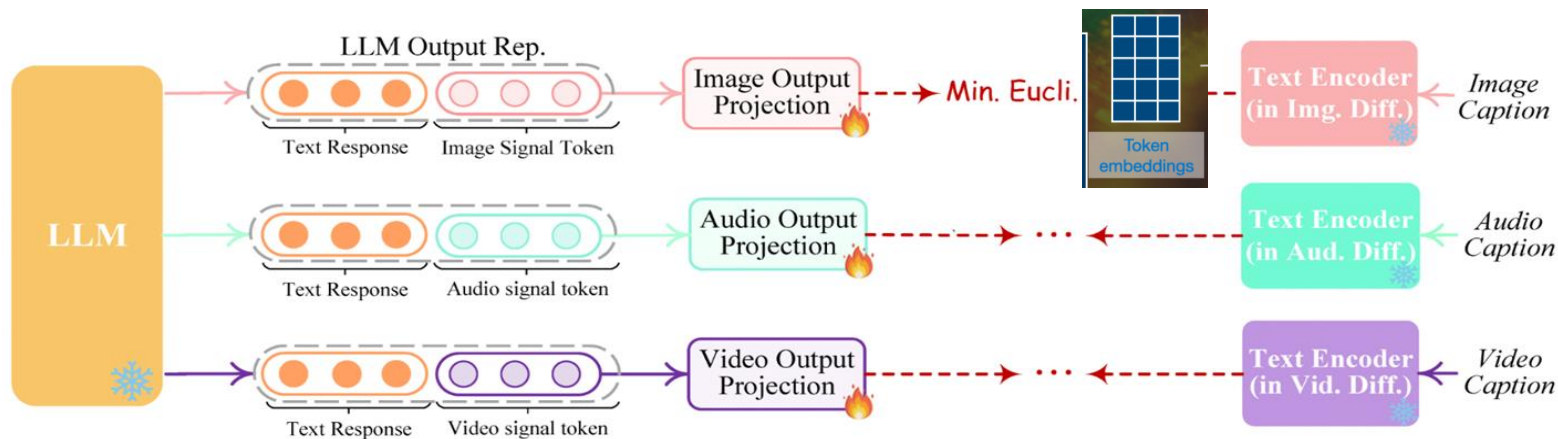
Next-GPT

- We already have strong text2img/audio/video **diffusion models**
- Typical architecture
- **We freeze the diffusion network** and train only a projection layer that align the image tokens with the **diffusion text encoder**



Next-GPT

- We already have strong text2img/audio/video **diffusion models**
- Typical architecture
- **We freeze the diffusion network** and train only a projection layer that align the image tokens with the **diffusion text encoder**



(b) Decoding-side Instruction-following Alignment

Next-GPT

Only 1% parameters are trainable

	Encoder		Input Projection		LLM		Output Projection		Diffusion	
	Name	Param	Name	Param	Name	Param	Name	Param	Name	Param
Text	—	—	—	—			—	—	—	—
Image					Vicuna [12]	7B❄️	Transformer	31M🔥	SD [68]	1.3B❄️
Audio	ImageBind [25]	1.2B❄️	Linear	4M🔥	(LoRA	33M🔥)	Transformer	31M🔥	AudioLDM [51]	975M❄️
Video							Transformer	32M🔥	Zeroscope [8]	1.8B❄️

How many cars are in this picture?



There are two cars shown in this picture.

Tell me what kind of car are they?

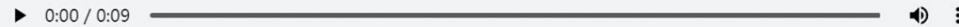


The two cars in the picture are SUVs.

Can you show me the sound of starting the engines of them.



Sure, here's the sound of the engine of the SUV:



Conclusion

- Today there are various advanced methods and tools that provide access to very powerful model (almost comparable to proprietary models)
- These models can be easily (and cheaply) customized according to the user specific domain and objective
- There are effective methods to extend the model capabilities, for example: adding more modalities, extending content window

Google "We Have No Moat, And Neither Does OpenAI"

Leaked Internal Google Document Claims Open Source AI Will Outcompete Google and OpenAI



DYLAN PATEL AND AFZAL AHMAD

MAY 4, 2023 · PAID