

Big Data Analysis

ChatGPT and Large Language Models (LLM) for Big Data Analysis

1112BDA08

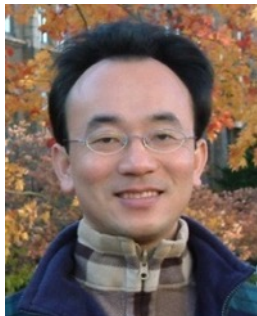
MBA, IM, NTPU (M6031) (Spring 2023)

Tue 2, 3, 4 (9:10-12:00) (B8F40)



<https://meet.google.com/paj-zhhj-mya>

aws
educate | Cloud
Ambassador
2020 Cohort



Min-Yuh Day, Ph.D,
Associate Professor

Institute of Information Management, National Taipei University

<https://web.ntpu.edu.tw/~myday>



Syllabus

Week	Date	Subject/Topics
1	2023/02/21	Introduction to Big Data Analysis
2	2023/02/28	(Day Off)
3	2023/03/07	AI, Data Science, and Big Data Analysis
4	2023/03/14	Foundations of Big Data Analysis in Python
5	2023/03/21	Case Study on Big Data Analysis I
6	2023/03/28	Machine Learning: SAS Viya, Data Preparation and Algorithm Selection

Syllabus

Week Date Subject/Topics

7 2023/04/04 (Children's Day) (Day off)

8 2023/04/11 Midterm Project Report

9 2023/04/18 Machine Learning: Decision Trees and Ensembles of Trees

**10 2023/04/25 Machine Learning: Neural Networks (NN) and
Support Vector Machines (SVM)**

11 2023/05/02 Case Study on Big Data Analysis II

12 2023/05/09 Machine Learning: Model Assessment and Deployment

Syllabus

Week	Date	Subject/Topics
13	2023/05/16	ChatGPT and Large Language Models (LLM) for Big Data Analysis
14	2023/05/23	Deep Learning for Finance Big Data Analysis
15	2023/05/30	Final Project Report I
16	2023/06/06	Final Project Report II
17	2023/06/13	Self-learning
18	2023/06/20	Self-learning

Outline

- **ChatGPT**
- **Large Language Models (LLMs)**
- **Foundation Models for Big Data Analytics**

ChatGPT

Large Language Models
(LLM)

Foundation Models

Generative AI

(Gen AI)

AI Generated Content

(AIGC)

Generative AI (Gen AI)

AI Generated Content (AIGC)

Image Generation

Instruction 1:

An astronaut riding a horse in a photorealistic style.

Instruction 2:

Teddy bears working on new AI research on the moon in the 1980s.

Figure 1



Figure 2

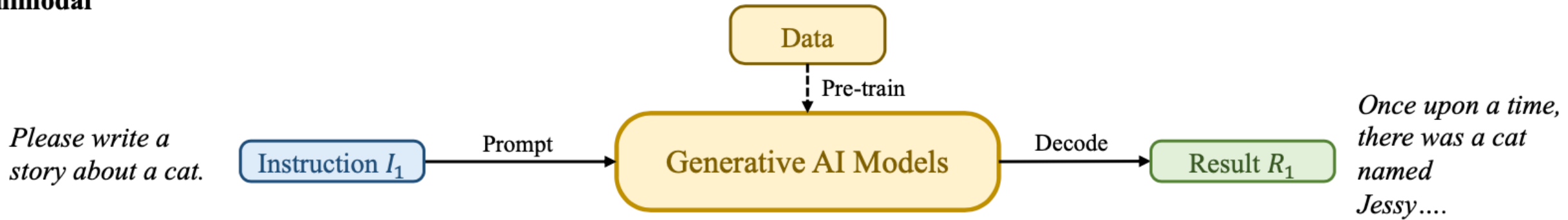


 **OpenAI DALL·E 2**

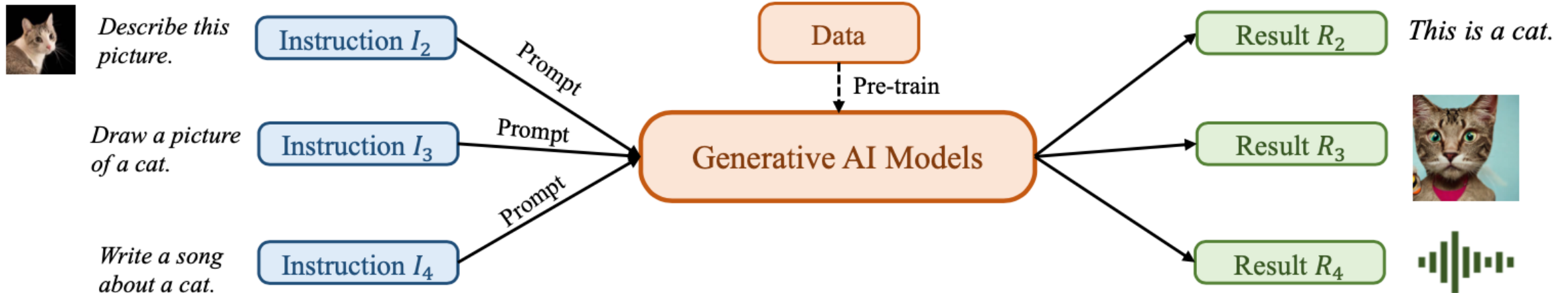
Generative AI (Gen AI)

AI Generated Content (AIGC)

Unimodal

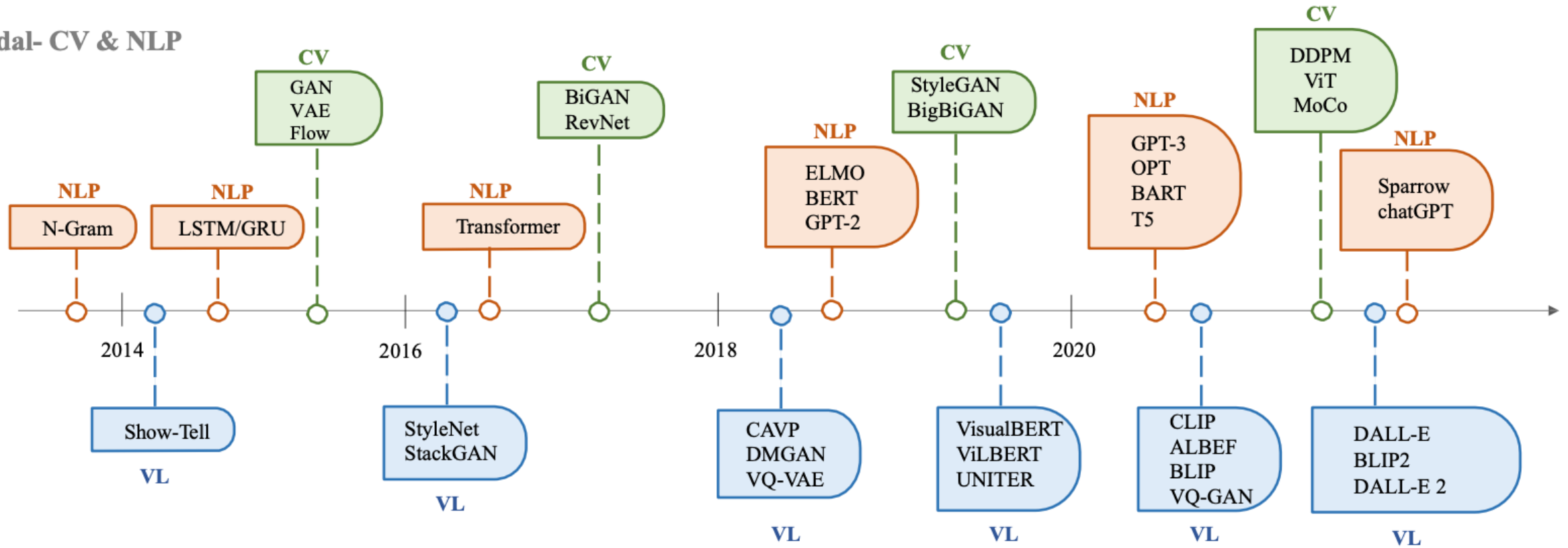


Multimodal



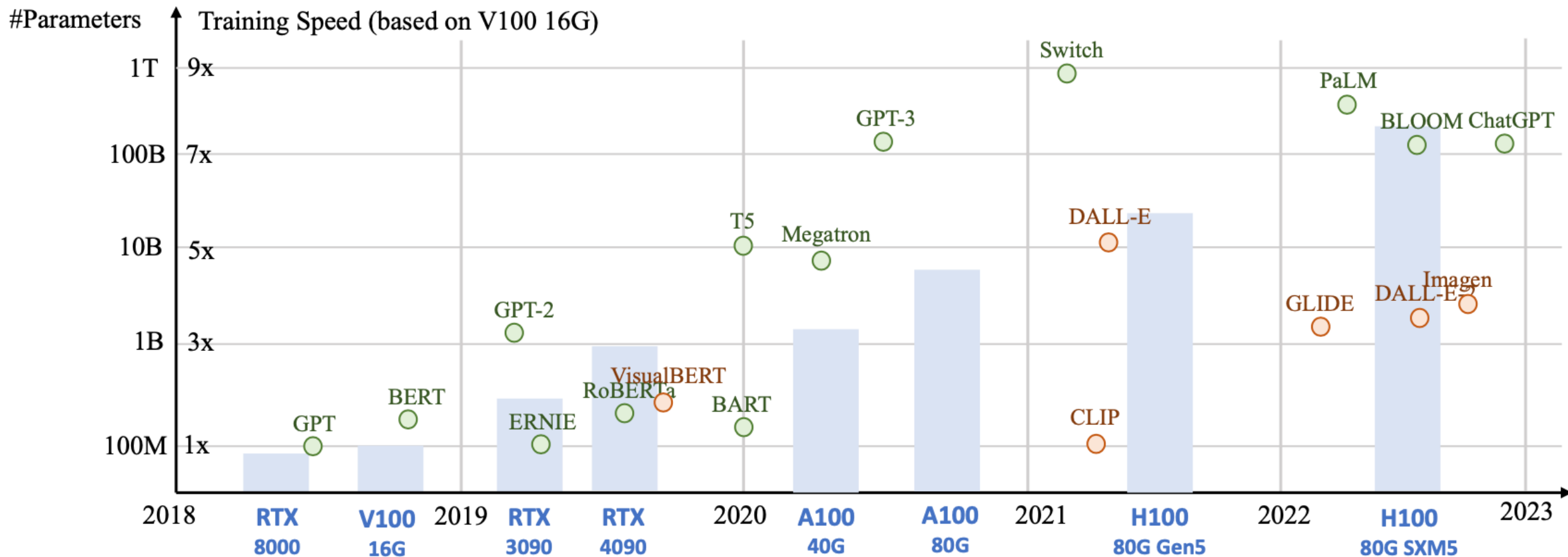
The history of Generative AI in CV, NLP and VL

Unimodal- CV & NLP

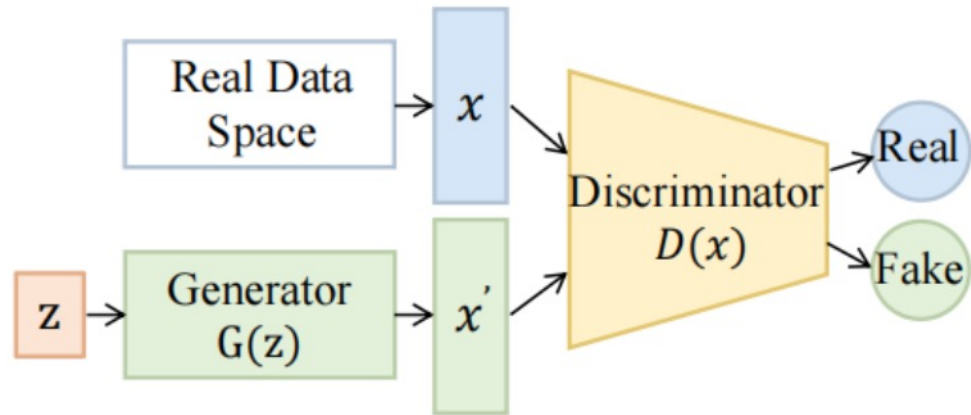


Multimodal – Vision Language

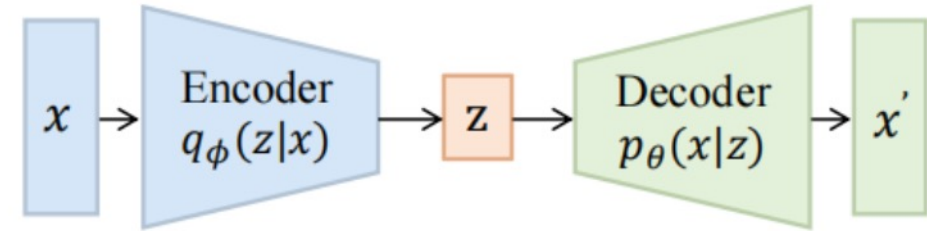
Generative AI Foundation Models



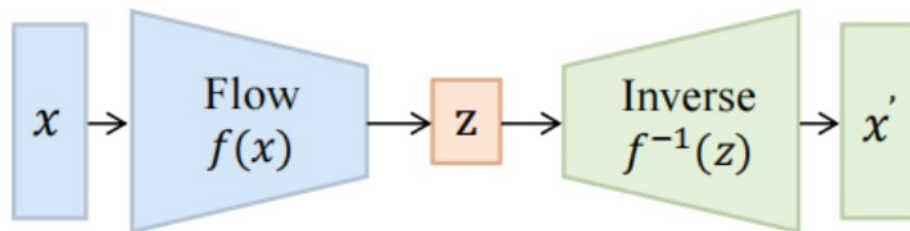
Categories of Vision Generative Models



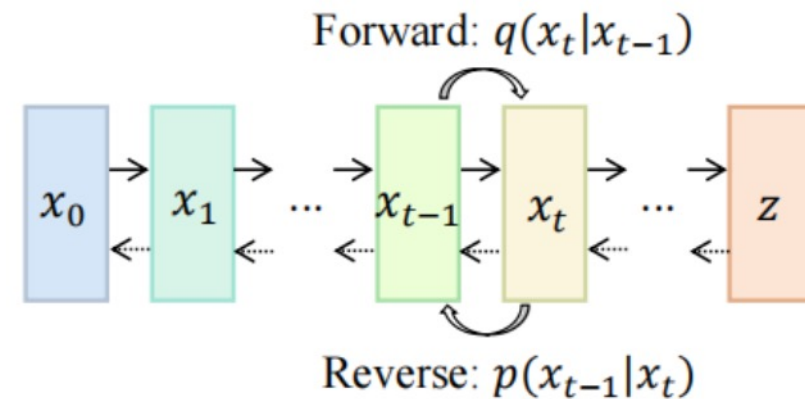
(1) Generative adversarial networks



(2) Variational autoencoders

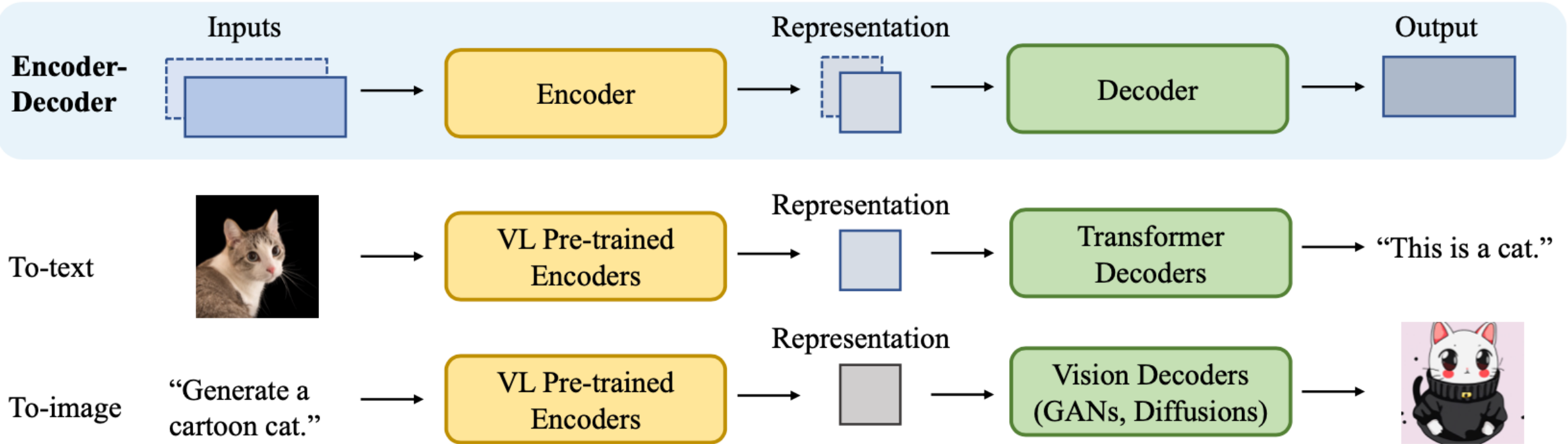


(3) Normalizing flows

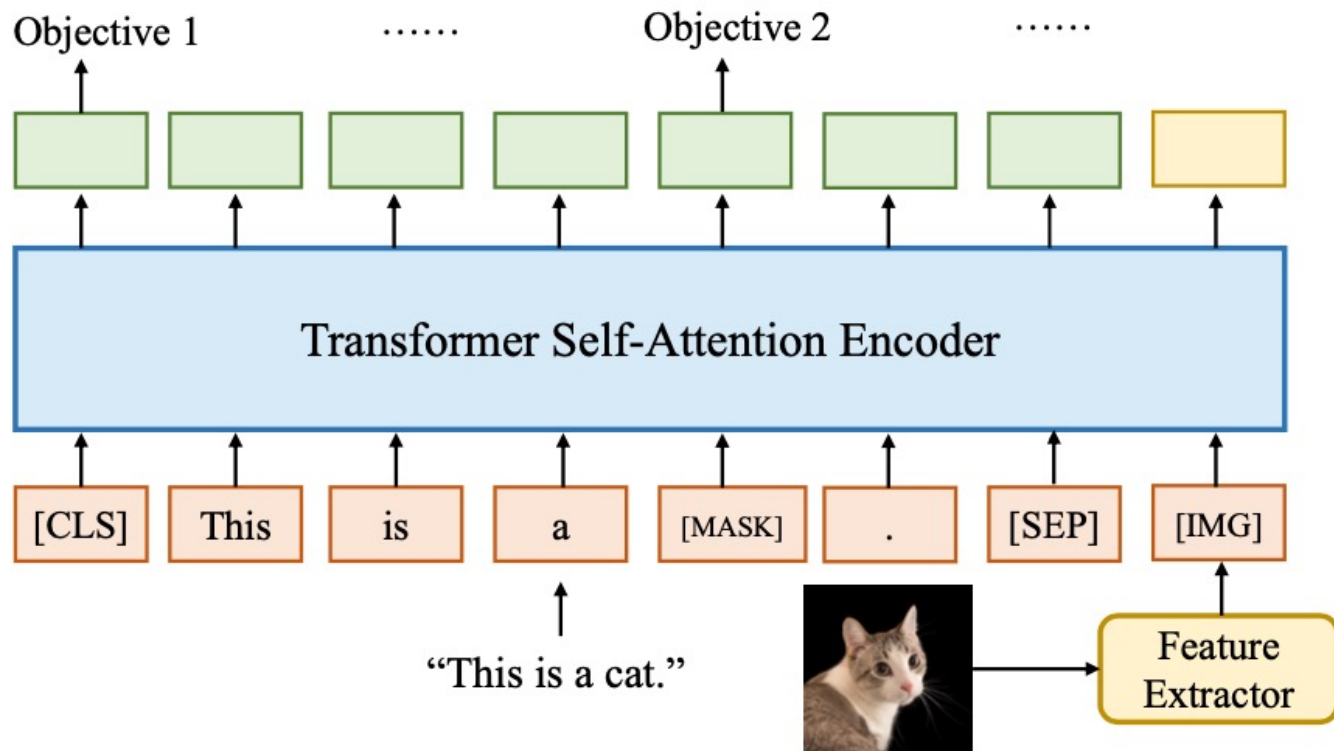


(4) Diffusion models

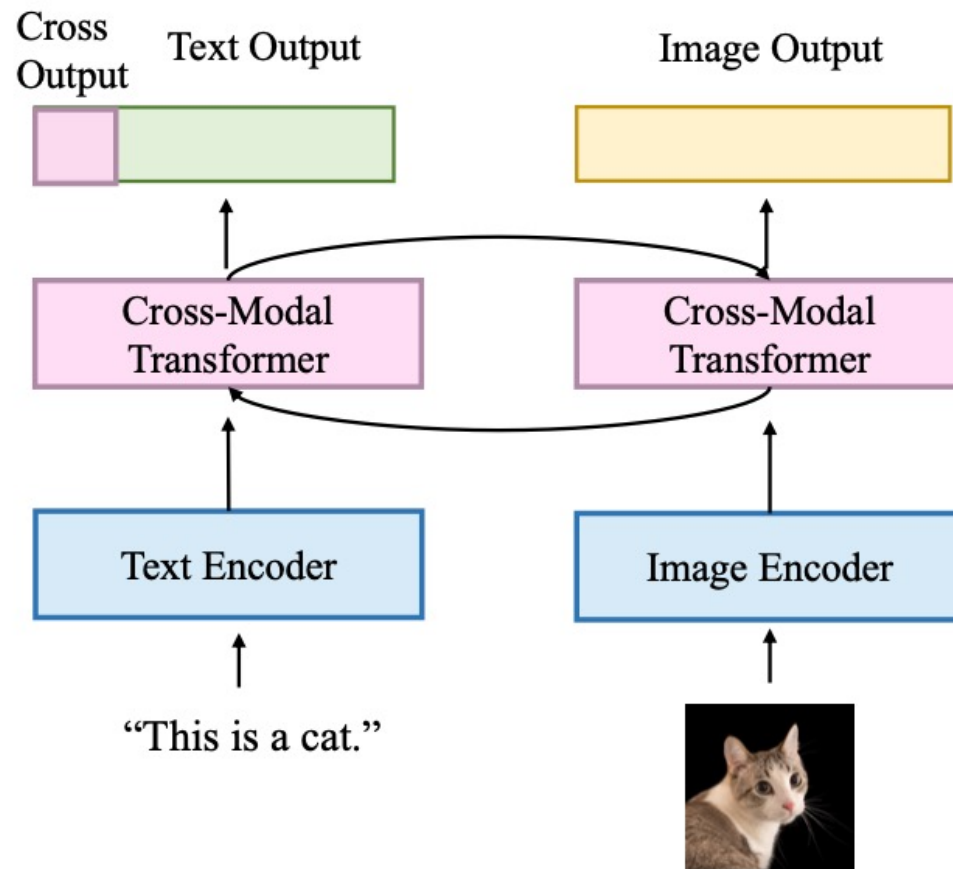
The General Structure of Generative Vision Language



Two Types of Vision Language Encoders: Concatenated Encoders and Cross-aligned Encoders

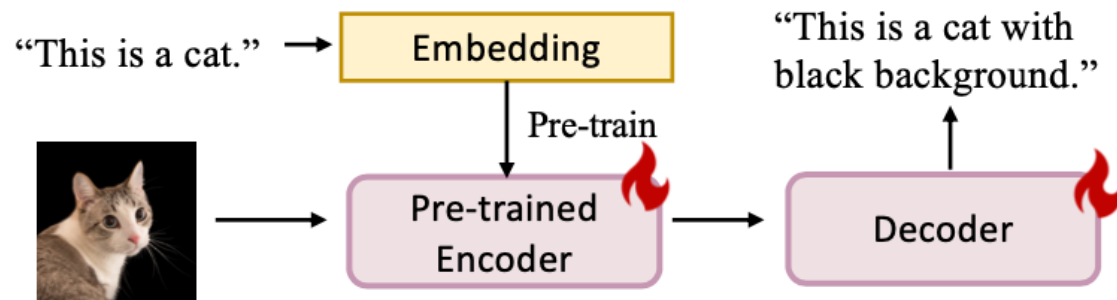


(a) Concatenated Encoder

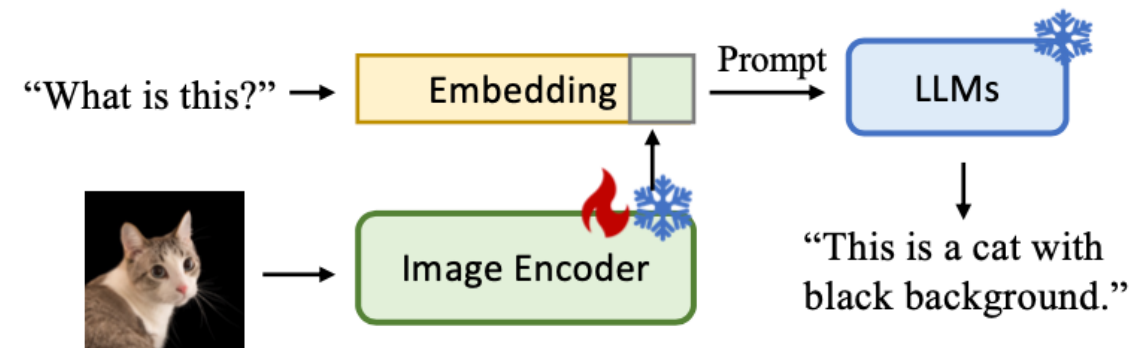


(b) Cross-aligned Encoder

Two Types of to-language Decoder Models: Jointly-trained Models and Frozen Models



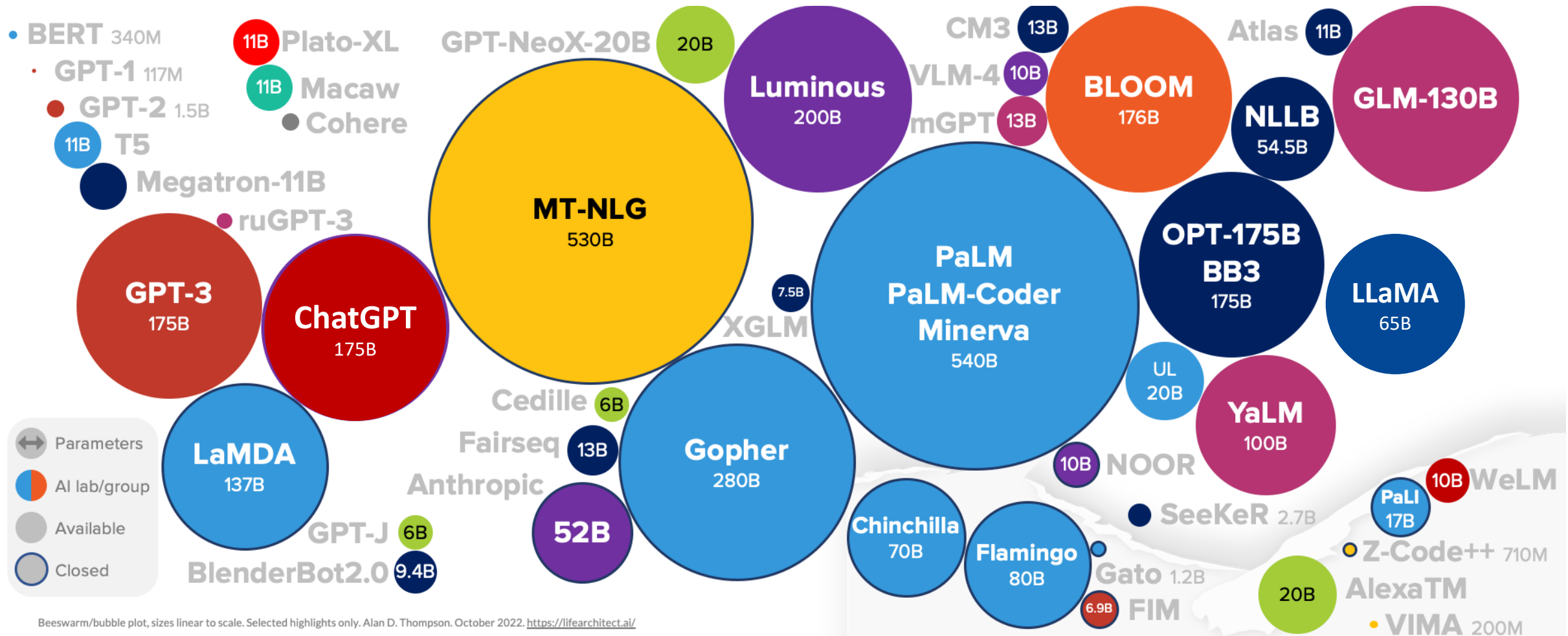
(a) Jointly-trained Models



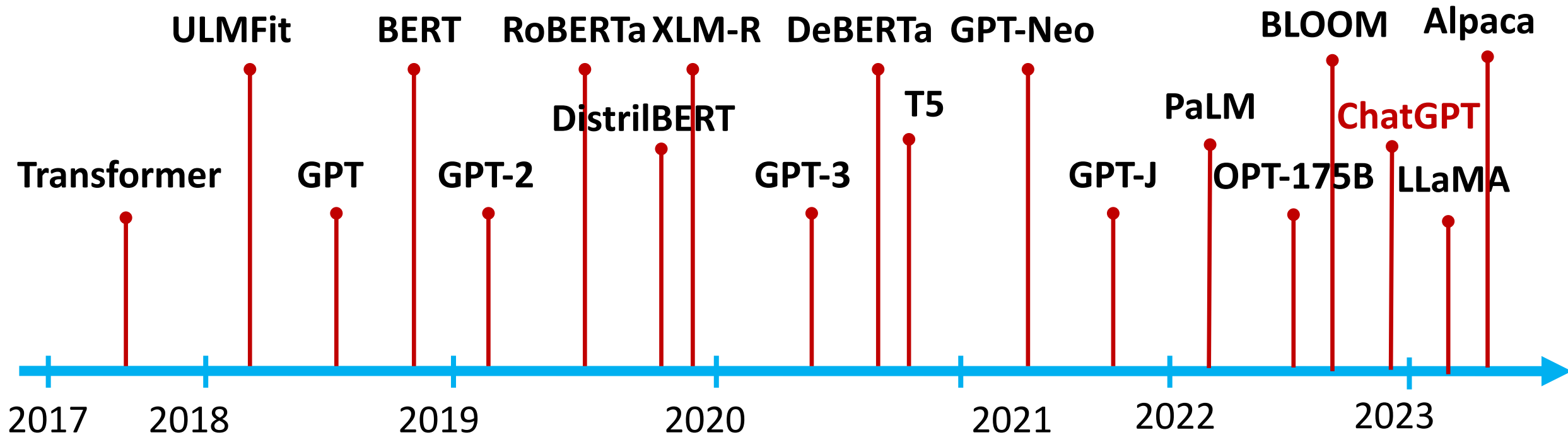
(b) Frozen Models

Large Language Models (LLM)

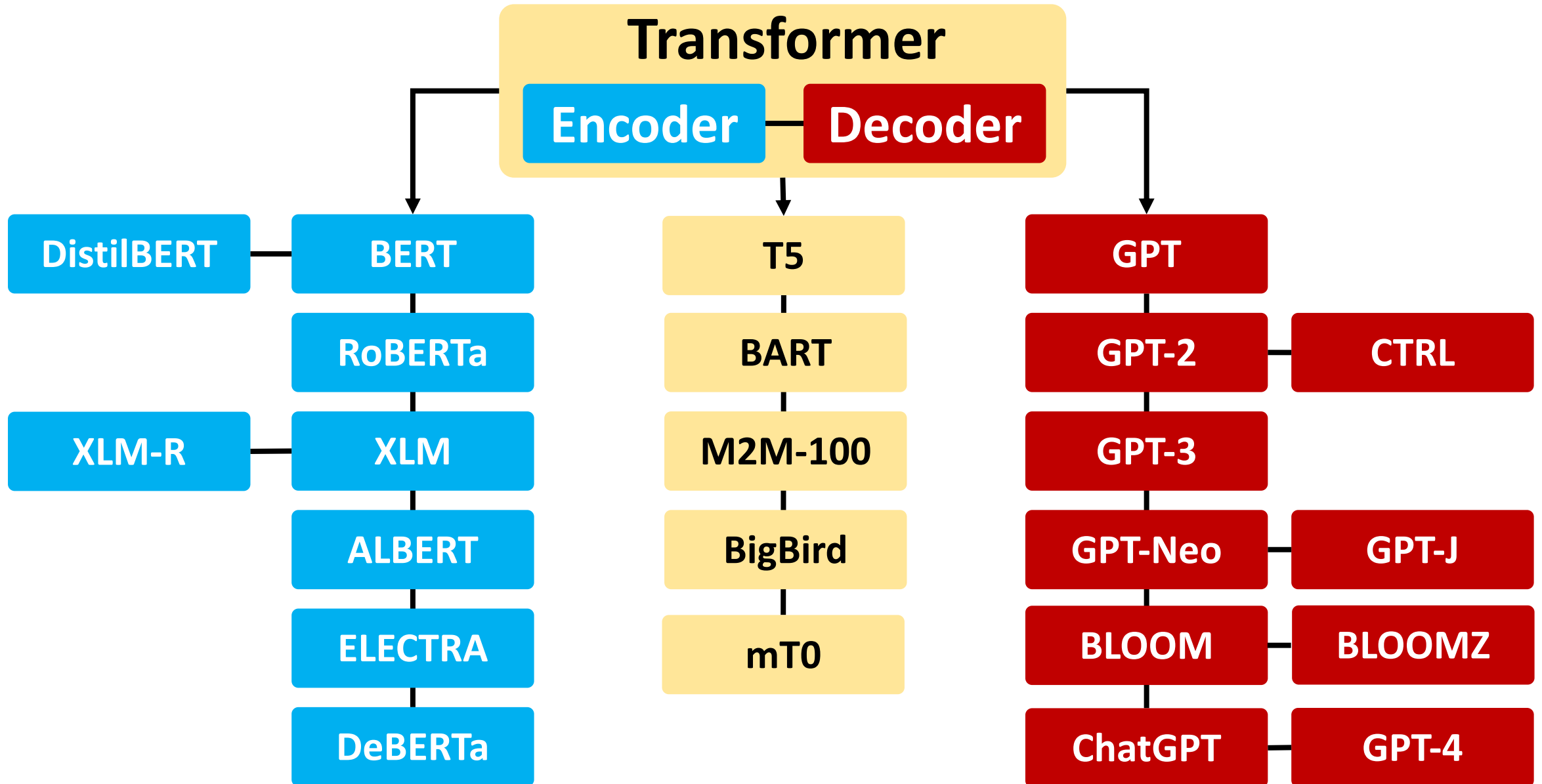
(GPT-3, ChatGPT, PaLM, BLOOM, OPT-175B, LLaMA)



The Transformers Timeline



Transformer Models



OpenAI ChatGPT



API

RESEARCH

BLOG

ABOUT

ChatGPT: Optimizing Language Models for Dialogue

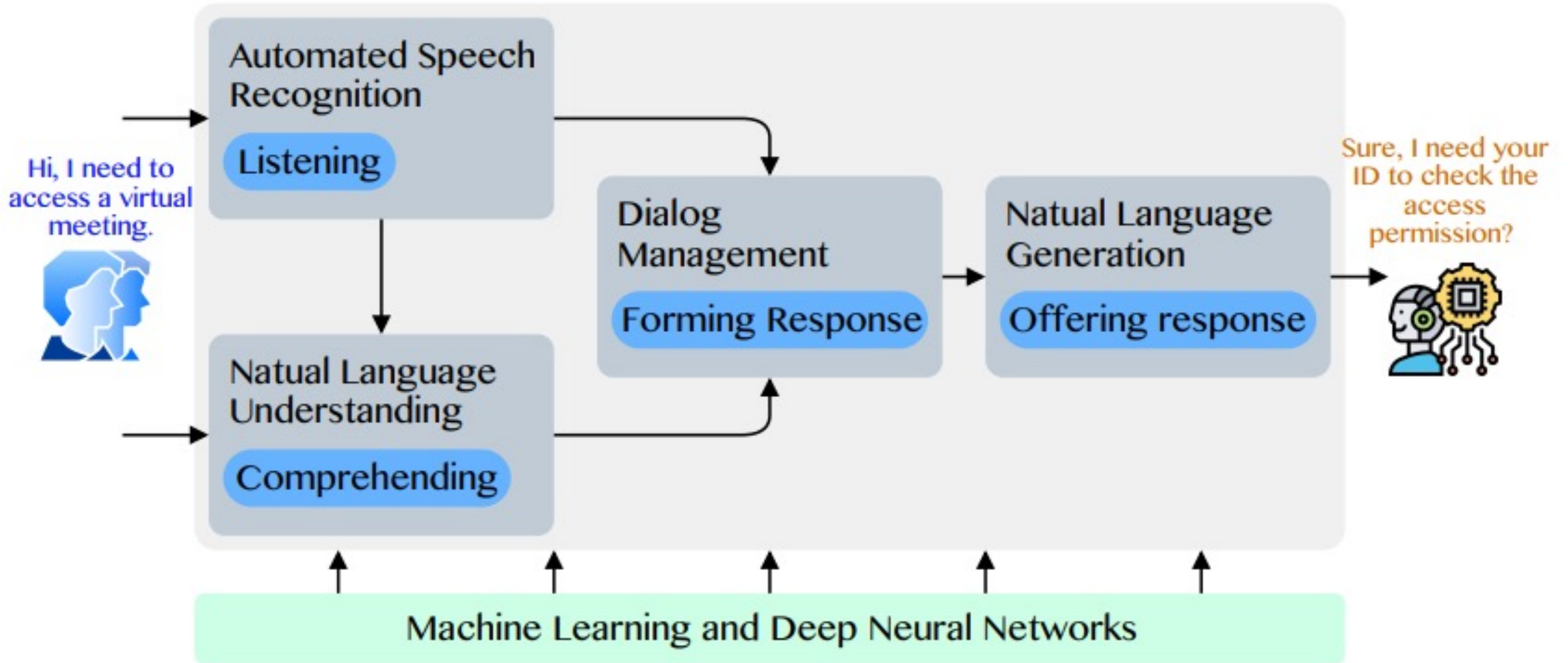
We've trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests. ChatGPT is a sibling model to InstructGPT, which is trained to follow an instruction in a prompt and provide a detailed response.



Source: <https://openai.com/blog/chatgpt/>

Conversational AI

to deliver contextual and personal experience to users



OpenAI ChatGPT

ChatGPT

This is a free research preview.



Our goal is to get external feedback in order to improve our systems and make them safer.



While we have safeguards in place, the system may occasionally generate incorrect or misleading information and produce offensive or biased content. It is not intended to give advice.

OpenAI ChatGPT

ChatGPT

How we collect data



Conversations may be reviewed by our AI trainers to improve our systems.



Please don't share any sensitive information in your conversations.

OpenAI ChatGPT

ChatGPT



Examples

"Explain quantum computing in simple terms" →

"Got any creative ideas for a 10 year old's birthday?" →

"How do I make an HTTP request in Javascript?" →



Capabilities

Remembers what user said earlier in the conversation

Allows user to provide follow-up corrections

Trained to decline inappropriate requests



Limitations

May occasionally generate incorrect information

May occasionally produce harmful instructions or biased content

Limited knowledge of world and events after 2021



Free Research Preview: ChatGPT is optimized for dialogue. Our goal is to make AI systems more natural to interact with, and your feedback will help us improve our systems and make them safer.

Source: <https://chat.openai.com/chat>

OpenAI ChatGPT



What is the future of AI?



It is difficult to predict the exact future of AI, but many experts believe that it will continue to play an increasingly important role in our lives. Some potential future developments in the field of AI include:



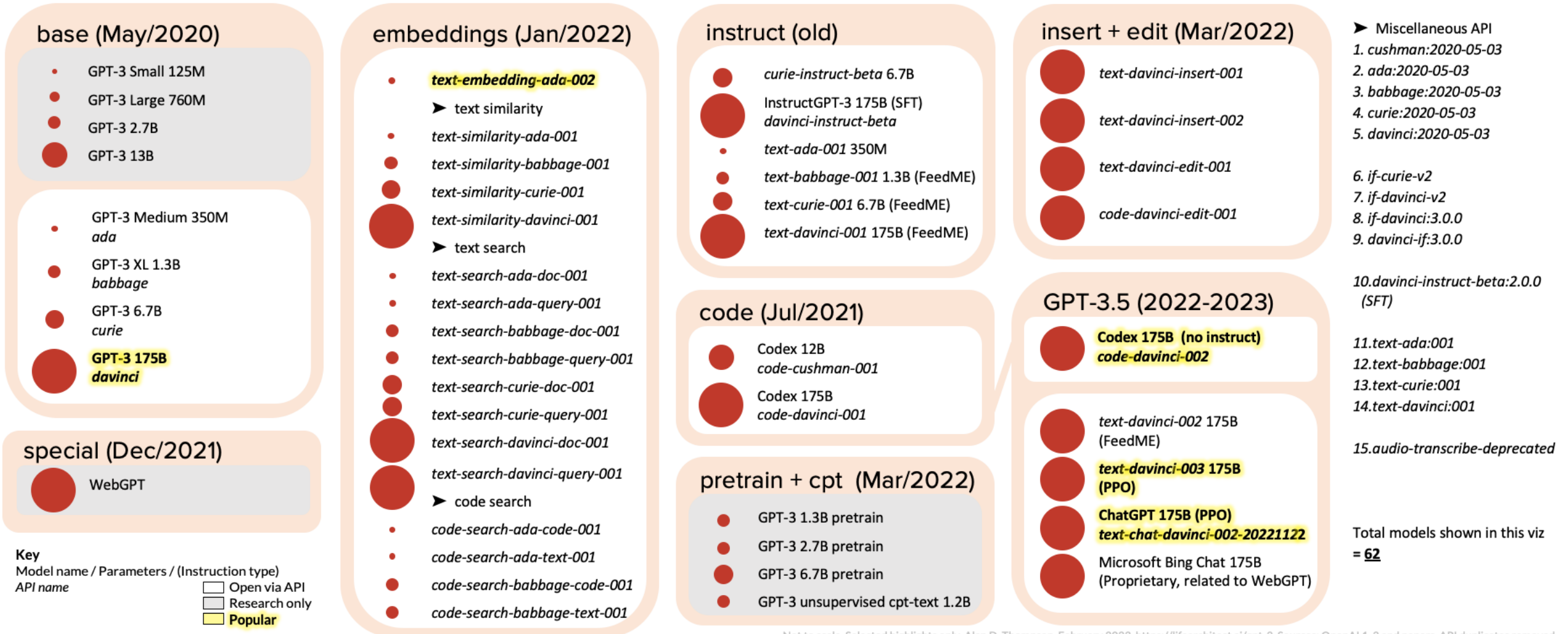
- Improved natural language processing, allowing AI systems to better understand and respond to human speech and writing.
- Enhanced ability to learn and adapt, allowing AI systems to become more autonomous and capable of handling a wider range of tasks.
- Greater integration with the internet of things (IoT), allowing AI systems to control and interact with a wide range of connected devices in the home and workplace.
- Continued advances in machine learning, allowing AI systems to make more accurate predictions and decisions based on data.
- Increased use of AI in healthcare, transportation, and other industries, leading to more efficient and effective processes and services.
- Continued development of ethical and regulatory frameworks to ensure that AI is developed and used responsibly.

Ultimately, the future of AI will depend on how it is developed and used by individuals, organizations, and society as a whole.

Source: <https://chat.openai.com/chat>

ChatGPT and GPT-3 Family

(GPT-3, InstructGPT, GPT-3.5, ChatGPT)



Not to scale. Selected highlights only. Alan D. Thompson. February 2023. <https://lifearchitect.ai/gpt-3> Sources: OpenAI [1](#), [2](#) and papers, API [duplicates](#) removed.

OpenAI ChatGPT and Open LLM

GPT-4, LLaMA, Alpaca, Dolly, Cerebras-GPT, GPT4All, Vicuna, ColossalChat, Koala, Phoenix

- OpenAI GPT-4
- Deepmind Chinchilla
- Meta OPT (LLaMA)
- Pythia
- **Stanford Alpaca**
- **Databricks Dolly**
- **Cerebras-GPT**
- **GPT4All**
- **Vicuna**
- **ColossalChat**
- **BAIR Koala**

Large Language Models (LMM)

Openness and Training Philosophy

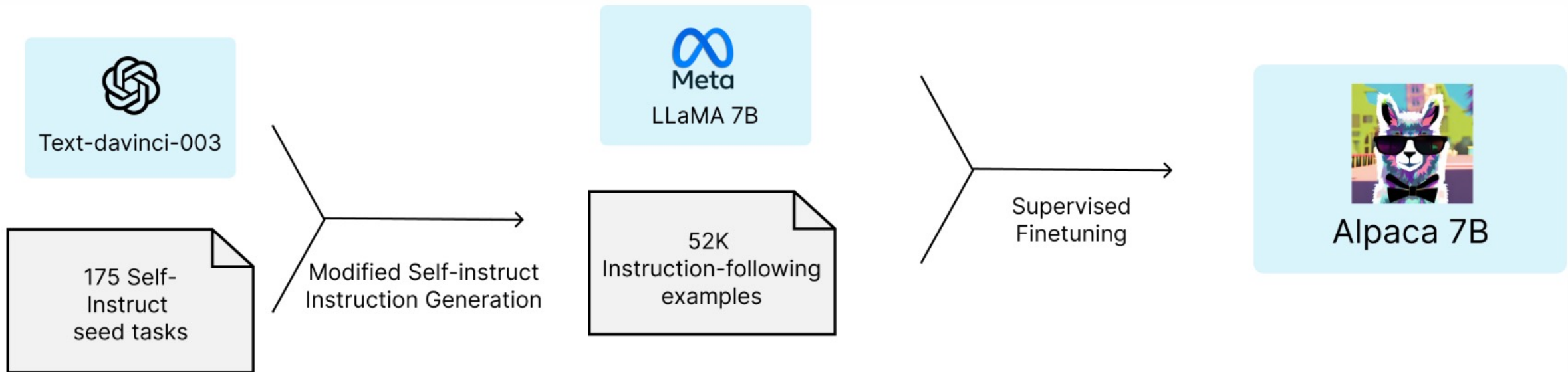
Model	Model architecture	Training data	Model weights	Checkpoints	Compute-optimal training	License
OpenAI GPT-4	Closed	Closed	No	No	Unknown	Not available
Deepmind Chinchilla	Open	Closed	No	No	Yes	Not available
Meta OPT	Open	Open	Researchers Only	Yes	No	Non-commercial
Pythia	Open	Open	Open	Yes	No	Apache 2.0
Cerebras-GPT	Open	Open	Open	Yes	Yes	Apache 2.0

Phoenix: Democratizing ChatGPT across Languages



Model	Backbone	#paras	Open-source		Claimed language	Post-training				Release date
			model	data		instruction data	lang	conversation data	lang	
ChatGPT	unknown	unknown	✗	✗	multi					11/30/22
Wenxin ⁷	unknown	unknown	✗	✗	zh					03/16/23
ChatGLM ⁸	GLM	6B	✓ ¹	✗	en/zh					03/16/23
Tongyi ⁹	unknown	unknown	✗	✗	zh					04/07/23
Shangliang ¹⁰	unknown	unknown	✗	✗	zh					04/10/23
Alpaca [12]	LLaMA	7B	✗	✓	en	52K	en	✗	✗	03/13/23
Dolly ¹¹ ²	GPT-J	6B	✓	✓	en	52k	en	✗	✗	03/24/23
BELLE [6]	BLOOMZ	7B	✓	✓	zh	1.5M	ch	✗	✗	03/26/23
Guanaco ¹²	LLaMA	7B	✓	✓	en/zh/ja/de	534K ³	4 ⁴	✗	✗	03/26/23
Chinese-alpaca [3]	LLaMA	7/13B	✓	✓	en/zh	2M/3M	en/zh	✗	✗	03/28/23
LuoTuo [7]	LLaMA	7B	✓	✓	zh	52k	cn	✗	✗	03/31/23
Vicuna [2]	LLaMA	7/13B	✓	✓ ⁵	en	✗	✗	70K	multi ⁶	03/13/23
Koala ¹³	LLaMA	13B	✓	✓	en	355K	en	117K	en	04/03/23
BAIZE [17]	LLaMA	7/13/30B	✓	✓	en	✗	✗	111.5K	en	04/04/23
Phoenix	BLOOMZ	7B	✓	✓	multi	267K	40+	189K	40+	04/08/23
Latin Phoenix (Chimera)	LLaMA	7B/13B	✓	✓	Latin	267K	40+	189K	40+	04/08/23

Stanford Alpaca: A Strong, Replicable Instruction-Following Model



Example seed task

Instruction: Brainstorm a list of possible New Year's resolutions.

Output:

- Lose weight
- Exercise more
- Eat healthier

Example Generated task

Instruction: Brainstorm creative ideas for designing a conference room.

Output:

... incorporating flexible components, such as moveable walls and furniture ...



GPT4All:

Training an Assistant-style Chatbot with Large Scale Data Distillation from GPT-3.5-Turbo

- **Demo, data and code to train an assistant-style large language model with ~800k GPT-3.5-Turbo Generations based on LLaMa**
- **Reproducibility**
 - **Trained LoRa Weights:**
 - gpt4all-lora (four full epochs of training):
 - <https://huggingface.co/nomic-ai/gpt4all-lora>

GPT4All-J

An Apache-2 Licensed Assistant-Style Chatbot

GPT4All-J (GPT4All v2) based on Open Source **GPT-J** model

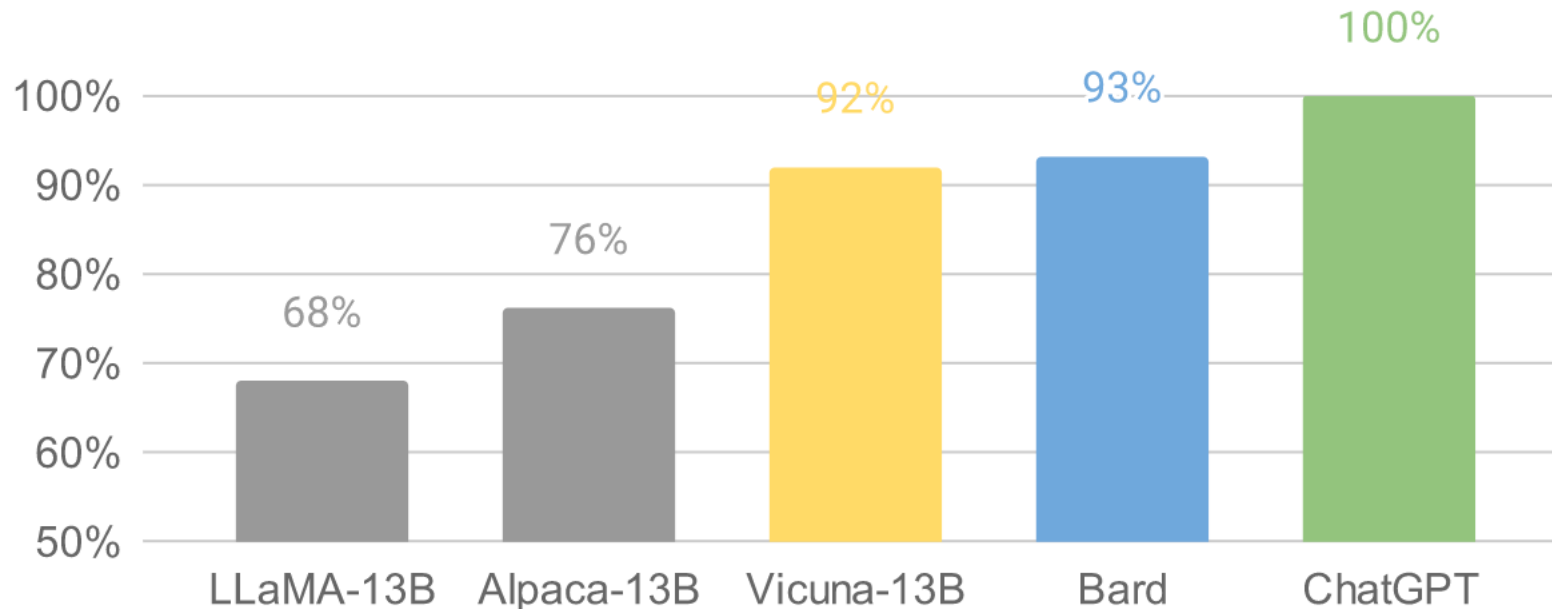
Model	BoolQ	PIQA	HellaSwag	WinoGrande	ARC-e	ARC-c	OBQA
GPT4All-J 6.7B	73.4	74.8	63.4	64.7	54.9	36.0	40.2
GPT4All-J Lora 6.7B	68.6	75.8	66.2	63.5	56.4	35.7	40.2
GPT4All LLaMa Lora 7B	73.1	77.6	72.1	67.8	51.1	40.4	40.2
Dolly 6B	68.8	77.3	67.6	63.9	62.9	38.7	41.2
Dolly 12B	56.7	75.4	71.0	62.2	64.6	38.5	40.4
Alpaca 7B	73.9	77.2	73.9	66.1	59.8	43.3	43.4
Alpaca Lora 7B	74.3	79.3	74.0	68.8	56.6	43.9	42.6
GPT-J 6.7B	65.4	76.2	66.2	64.1	62.2	36.6	38.2
LLaMa 7B	73.1	77.4	73.0	66.9	52.5	41.4	42.4
Pythia 6.7B	63.5	76.3	64.0	61.1	61.3	35.2	37.2
Pythia 12B	67.7	76.6	67.3	63.8	63.9	34.8	38

Vicuna: An Open-Source Chatbot

Impressing GPT-4 with 90%* ChatGPT Quality

by the Team with members from UC Berkeley, CMU, Stanford, and UC San Diego

- **Vicuna-13B: an open-source chatbot trained by fine-tuning LLaMA on user-shared conversations collected from ShareGPT.**
- **The cost of training Vicuna-13B is around \$300.**



Source: <https://vicuna.lmsys.org/>

Chinese-Vicuna:

A Chinese Instruction-following LLaMA-based Model

一個中文低資源的 llama+lora 方案

Dataset Preview Size: 409 MB </> API Go to dataset viewer

instruction (string)	input (string)	output (string)
"用一句话描述地球为什么是独一无二的。 \n "	""	"地球上有着适宜生命存在的条件和多样化的生命形式。"
"给出一段对话，要求GPT模型使用合适的语气和回答方式继续对话。 对话： A：你今..."	""	"A：这太好了！你的新工作听起来很令人兴奋。你对接下来的日子有什么期望吗？ B：是啊，我非..."
"基于以下提示填写以下句子的空格。 提示： - 提供多种现实世界的场景 - 空格..."	""	"多种形容词可填，以下是其中一些例子： - 愉快的 - 惬意的 - 轻松的 - 安静的 - 美妙的"
"请生成一个新闻标题，描述一场正在发生的大型自然灾害。 \n "	""	"强烈飓风肆虐，数百万人疏散！"
"为指定的词汇创建一个关于该词汇的简短解释。 "人工智能""	""	"人工智能是一种科技，通过使用计算机程序和算法实现类似于人类智能的能力，例如学习、推理和自..."
"在给定的文本中查找并替换特定的单词。 原始文本： "这个句子中有一个错误的单词..."	""	"替换后的文本： "这个句子中有一个修改的单词""
"根据给定的年份，计算该年是否为闰年。 \n \n1996\n"	""	"是闰年。"



Bactrian camel



Wild Bactrian camel



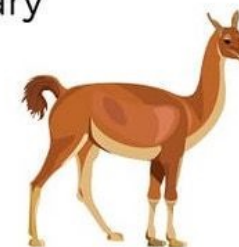
Dromedary



Llama



Alpaca



Guanaco



Vicuña

Chinese-Vicuna based on Guanaco Dataset and Belle Dataset

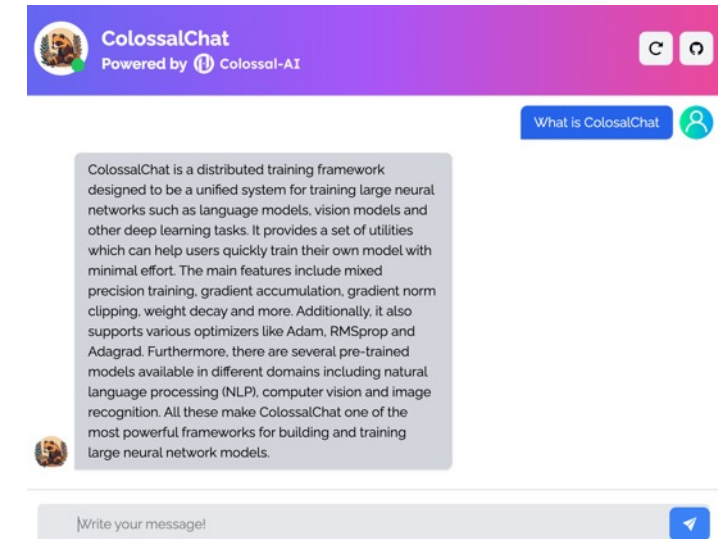
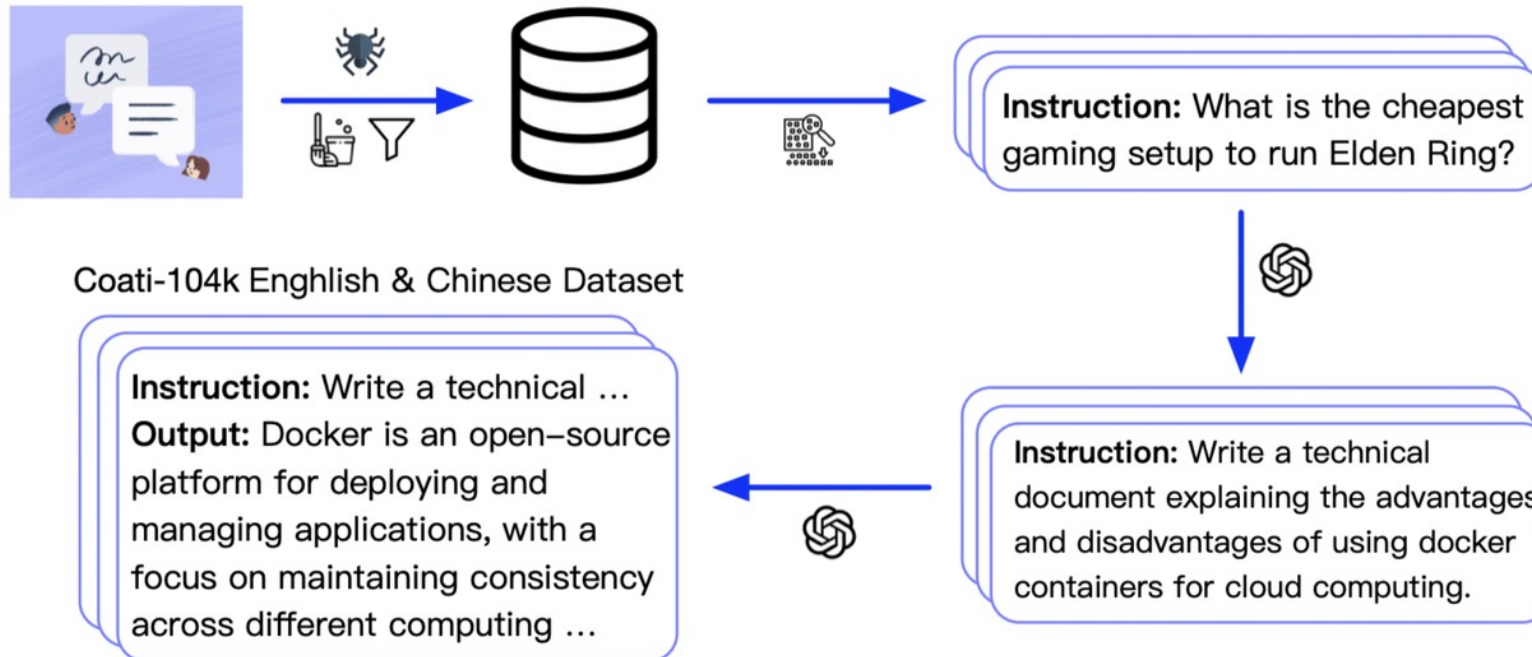
Source: https://huggingface.co/datasets/Chinese-Vicuna/guanaco_belle_merge_v1.0

Source: <https://github.com/Facico/Chinese-Vicuna>



ColossalChat

- ColossalChat is the project to implement LLM with RLHF, powered by the Colossal-AI project.
- Coati stands for ColossalAI Talking Intelligence.



Dolly v2

Open Source Instruction-Tuned LLM

- Databricks' Dolly is an instruction-following large language model trained on the Databricks machine learning platform that is licensed for commercial use.
- Based on pythia-12b, Dolly is trained on ~15k instruction/response fine tuning records databricks-dolly-15k generated by Databricks employees in capability domains from the InstructGPT paper, including brainstorming, classification, closed QA, generation, information extraction, open QA and summarization.
- dolly-v2-12b is not a state-of-the-art model, but does exhibit surprisingly high quality instruction following behavior not characteristic of the foundation model on which it is based.



StableLM

Stability AI Language Models

- **StableLM-Alpha models are trained on the new dataset that build on The Pile, which contains 1.5 trillion tokens, roughly 3x the size of The Pile.**
 - **These models will be trained on up to 1.5 trillion tokens.**
 - **The context length for these models is 4096 tokens.**
- **Fine-tuned the model with Stanford Alpaca's procedure using a combination of five recent datasets for conversational agents: Stanford's Alpaca, Nomic-AI's gpt4all, RyokoAI's ShareGPT52K datasets, Databricks labs' Dolly, and Anthropic's HH.**

RedPajama

a project to create leading open-source models,
starts by reproducing LLaMA training dataset of over 1.2 trillion tokens



Dataset	RedPajama	LLaMA*
CommonCrawl	878 billion	852 billion
C4	175 billion	190 billion
Github	59 billion	100 billion
Books	26 billion	25 billion
ArXiv	28 billion	33 billion
Wikipedia	24 billion	25 billion
StackExchange	20 billion	27 billion
Total Tokens	1.2 trillion	1.25 trillion

Chat with Open Large Language Models

Single Model

Chatbot Arena (battle)

Chatbot Arena (side-by-side)

Leaderboard

Chat with Open Large Language Models

- Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90% ChatGPT Quality. [\[Blog post\]](#)
- Koala: A Dialogue Model for Academic Research. [\[Blog post\]](#)
- [\[GitHub\]](#) [\[Twitter\]](#) [\[Discord\]](#)

Terms of use

By using this service, users are required to agree to the following terms: The service is a research preview intended for non-commercial use only. It only provides limited safety measures and may generate offensive content. It must not be used for any illegal, harmful, violent, racist, or sexual purposes. **The service collects user dialogue data and reserves the right to distribute it under a Creative Commons Attribution (CC-BY) license.**

Choose a model to chat with

[vicuna-13b](#): a chat assistant fine-tuned from LLaMA on user-shared conversations by LMSYS

[koala-13b](#): a dialogue model for academic research by BAIR

[oasst-pythia-12b](#): an Open Assistant for everyone by LAION

[RWKV-4-Raven-14B](#): an RNN with transformer-level LLM performance

[alpaca-13b](#): a model fine-tuned from LLaMA on instruction-following demonstrations by Stanford

[chatglm-6b](#): an open bilingual dialogue language model by Tsinghua University

[llama-13b](#): open and efficient foundation language models by Meta

[dolly-v2-12b](#): an instruction-tuned open large language model by Databricks

[stablelm-tuned-alpha-7b](#): Stability AI language models

[fastchat-t5-3b](#): a chat assistant fine-tuned from FLAN-T5 by LMSYS

[mpt-7b-chat](#): a chatbot fine-tuned from MPT-7B by MosaicML

vicuna-13b



<https://chat.lmsys.org/>

Chat with Open Large Language Models: Chatbot Arena

The screenshot shows a side-by-side comparison of two AI chatbot responses. On the left, 'vicuna-13b' (Model A) provides a concise definition of LLMs and their applications in data science. On the right, 'mpt-7b-chat' (Model B) provides a more detailed explanation, including the specific example of GPT-3 and its parameters. The interface includes a voting bar at the bottom with four options: 'A is better', 'B is better', 'Tie', and 'Both are bad'.

vicuna-13b (Model A)

Large Language Models for Data Science

Large Language Models (LLMs) are a type of artificial neural network that have been trained on a massive amount of text data to generate human-like language. These models have achieved state-of-the-art performance on a variety of natural language processing tasks, such as language translation, text generation, and question answering. In the field of data science, LLMs have been used to process and analyze unstructured text data, such as customer reviews, social media posts, and news articles, to extract insights and make predictions.

mpt-7b-chat (Model B)

Large Language Models for Data Science

Large Language Models (LLMs) are a type of artificial intelligence (AI) model that are capable of processing and generating human-like natural language. They are based on deep neural networks that are trained on large amounts of text data, such as books, news articles, and social media posts. LLMs have shown promising results in a variety of data science applications, including language translation, text summarization, and sentiment analysis. One of the most popular LLMs is GPT-3 (Generative Pre-trained Transformer 3), developed by OpenAI. GPT-3 has 175 billion parameters, making it one of the largest language models ever created. It can perform a wide range of language tasks, such as text completion, question answering, and text generation. GPT-3 has been used in various industries, including healthcare, finance, and journalism, to automate

👉 A is better 👉 B is better 🤝 Tie 👎 Both are bad

<https://chat.lmsys.org/>

Chat with Open Large Language Models: Leaderboard




[Single Model](#)[Chatbot Arena \(battle\)](#)[Chatbot Arena \(side-by-side\)](#)[Leaderboard](#)

Leaderboard

[\[Blog\]](#) [\[GitHub\]](#) [\[Twitter\]](#) [\[Discord\]](#)

We use the Elo rating system to calculate the relative performance of the models. You can view the voting data, basic analyses, and calculation procedure in this [notebook](#). We will periodically release new leaderboards. If you want to see more models, please help us [add them](#).




Last updated: 2023-05-08 16:55:45 PDT


Rank	Model	Elo Rating	Description
1	 gpt-4	1274	ChatGPT-4 by OpenAI
2	 claude-v1	1224	Claude by Anthropic
3	 gpt-3.5-turbo	1155	ChatGPT-3.5 by OpenAI
4	vicuna-13b	1083	a chat assistant fine-tuned from LLaMA on user-shared conversations by LMSYS
5	koala-13b	1022	a dialogue model for academic research by BAIR
6	RWKV-4-Raven-14B	989	an RNN with transformer-level LLM performance
7	oasst-pythia-12b	928	an Open Assistant for everyone by LAION
8	chatglm-6b	918	an open bilingual dialogue language model by Tsinghua University
9	stablalm-tuned-alpha-7b	906	Stability AI language models
10	alpaca-13b	904	a model fine-tuned from LLaMA on instruction-following demonstrations by Stanford


<https://chat.lmsys.org/>


Perplexity.ai




 New Thread  



 Popular

 Your Threads

 Login

[Sign Up](#)

 Download

 Follow  Join

[About](#) [Blog](#) [Privacy Policy](#)

Ask anything...

Quick



Popular Now



Turkey election results see Erdogan forced in...

The 2023 Turkish presidential election and parliamentary elections were held on May 15,...

how to make a successful career pivot

Making a successful career pivot requires careful planning and execution. Here are...

did mammoths evolve from dinosaurs

No, mammoths did not evolve from dinosaurs. Mammoths are mammals that evolved...

Jayson Tatum's 51-point game sends Celtic...

Jayson Tatum scored 51 points in the Boston Celtics' 112-88 victory over the Philadelphia...

WhatsApp introduces chat lock feature

WhatsApp has introduced a new feature called "Chat Lock" that adds an extra layer of...

number of calories in a potato

The number of calories in a potato varies depending on the size and type of potato...

 **Perplexity for iPhone**
Available Now



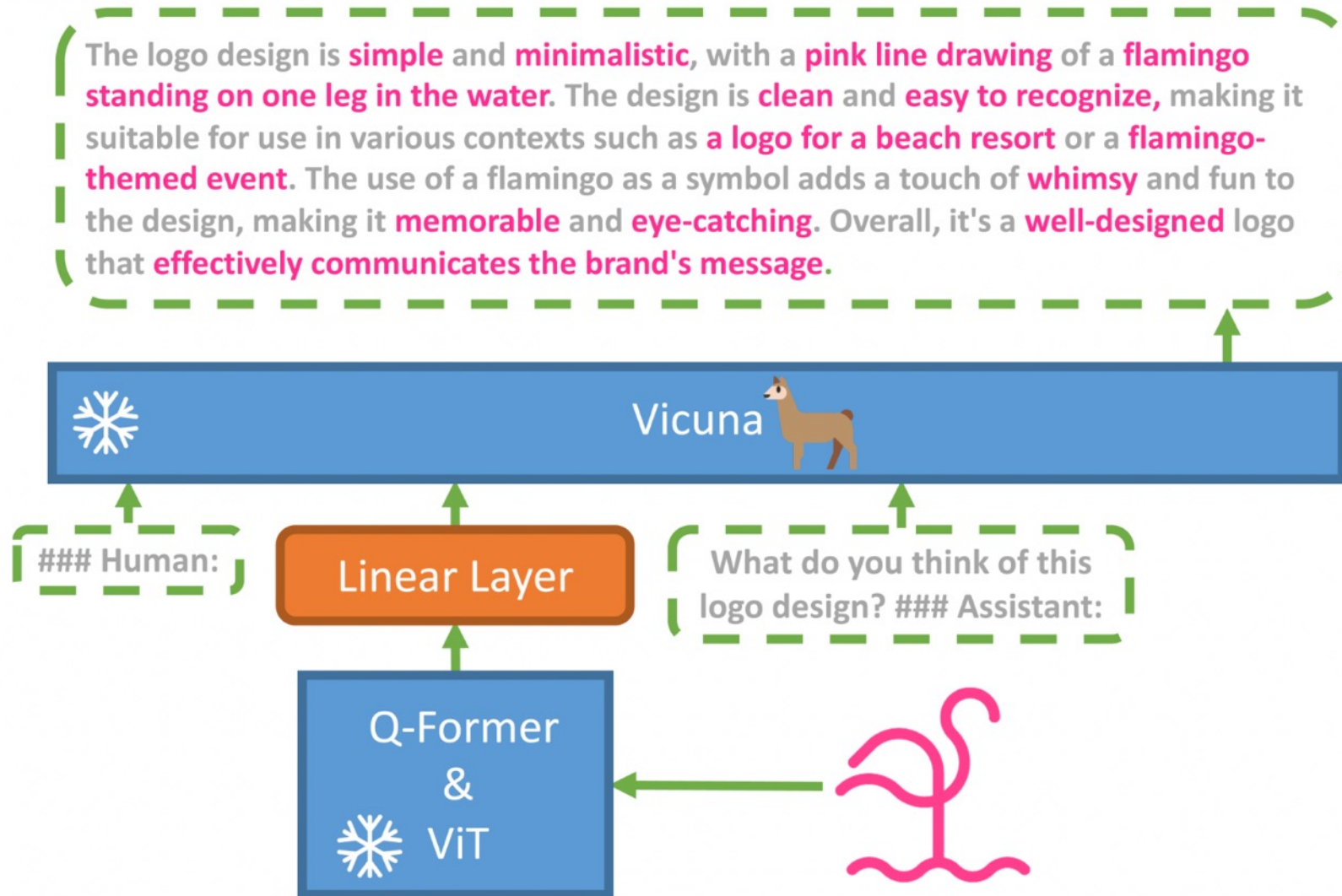
 **Perplexity for Android**
Join the waitlist



<https://www.perplexity.ai/>

MiniGPT-4:

Enhancing Vision-language Understanding with Advanced Large Language Models



LLaVA: Large Language and Vision Assistant

LLaVA: Large Language and Vision Assistant

[\[Project Page\]](#) [\[Paper\]](#) [\[Code\]](#) [\[Model\]](#)

LLaVA-13B-v0

Image

Drop Image Here
- or -
Click to Upload

Preprocess
When the image is not square, you may try adjusting this .

Crop Resize Pad

Parameters


Terms of use

By using this service, users are required to agree to the following terms:

The service is a research preview intended for non-commercial use only. It only provides limited safety measures and may generate offensive content. It must not be used for any illegal, harmful, violent, racist, or sexual purposes. The service may collect user dialogue data for future research.

LLaVA Chatbot

What are the things I should be cautious about when I visit here?



When visiting this location, one should be cautious about the weather conditions and the state of the wooden pier. In the image, the wooden pier is extended over a large, still lake under cloudy skies, which could suggest the possibility of sudden changes in weather, such as wind, rain, or even high waves. Visitors should pay attention to weather forecasts and remain vigilant while on the pier to ensure their safety. Additionally, they should be mindful of the pier's condition and the surrounding water, as wood can be slippery, especially when wet or in windy conditions. It is crucial to maintain proper footing and follow any posted safety guidelines, as the quiet waters of the lake can be deceptively

Enter text and press ENTER

Submit

Upvote Downvote Flag Regenerate

Source: <https://llava-vl.github.io/>

Visual Instruction Tuning

LLaVA: Large Language and Vision Assistant

University of Wisconsin-Madison, Microsoft Research, Columbia University

Science QA:
New SoTA with
the synergy of
LLaVA with
GPT-4



LLaVA represents a novel end-to-end trained large multimodal model that combines a vision encoder and Vicuna for general-purpose visual and language understanding, achieving impressive chat capabilities mimicking spirits of the multimodal GPT-4 and setting a new state-of-the-art accuracy on Science QA.


MPT-7B:

A New Standard for Open-Source, Commercially Usable LLMs



MPT-7B (MosaicML Pretrained Transformer)

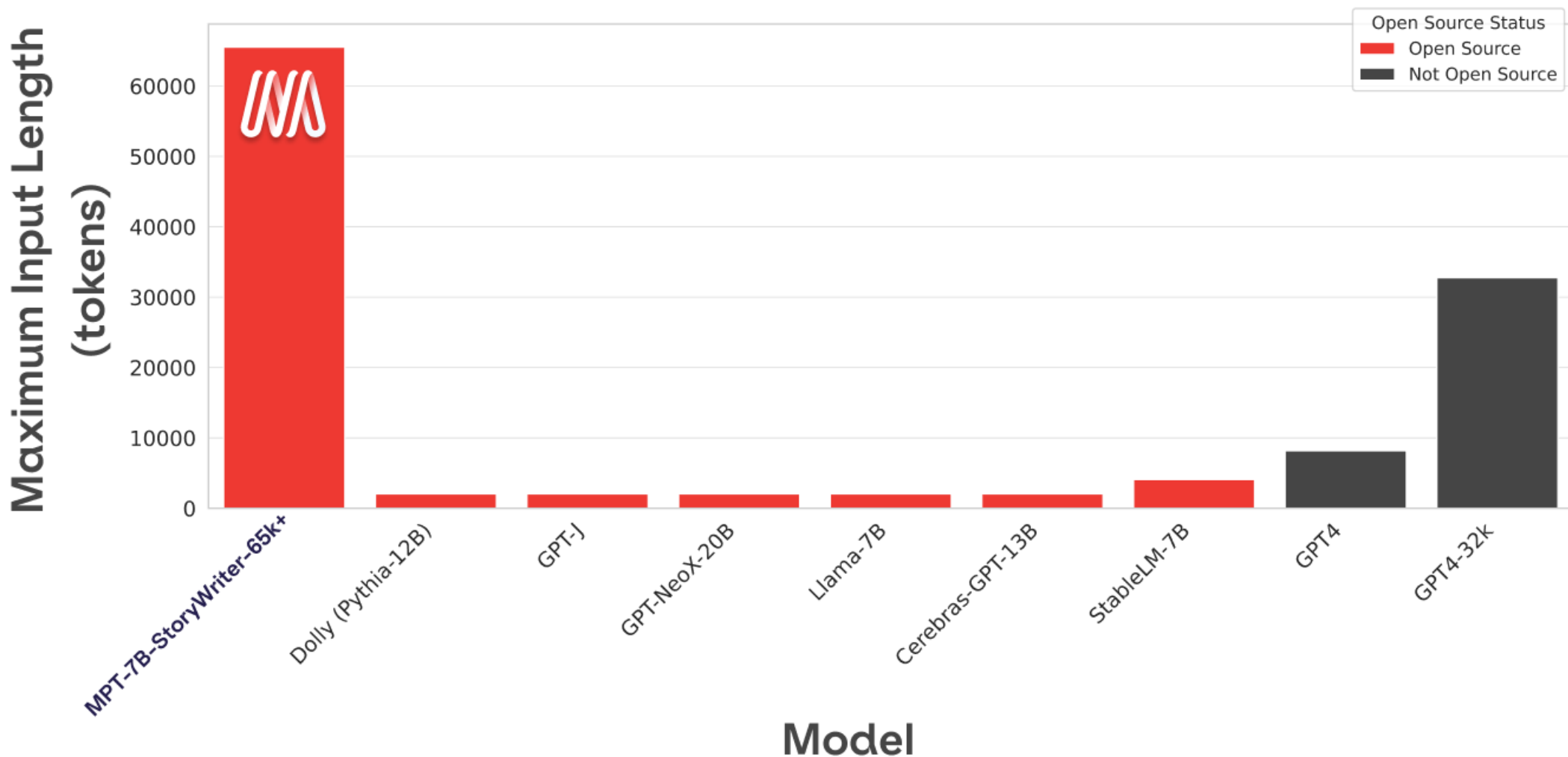
vs. open source models on academic tasks

Model	LAMBADA (OpenAI)	HellaSwag	PIQA	ARC-Easy	ARC-Challenge	BoolQ	COPA	Winograd	Winogrande	TriviaQA	Jeopardy	MMLU
 MPT-7B	0.703	0.761	0.799	0.673	0.394	0.750	0.813	0.878	0.683	0.343	0.308	0.296
LLaMA-7B	0.738	0.751	0.792	0.652	0.411	0.767	0.779	0.807	0.675	0.443	0.334	0.302
StableLM-7B (alpha)	0.533	0.411	0.666	0.435	0.259	0.606	0.672	0.646	0.513	0.049	0.000	0.251
Pythia-7B	0.667	0.636	0.761	0.581	0.325	0.634	0.769	0.786	0.607	0.198	0.022	0.265
Pythia-12B	0.704	0.672	0.768	0.605	0.351	0.675	0.781	0.847	0.627	0.233	0.026	0.253
GPTJ-6B	0.683	0.665	0.762	0.583	0.355	0.648	0.789	0.833	0.641	0.234	0.026	0.261
GPT-NeoX-20B	0.719	0.712	0.780	0.644	0.392	0.691	0.781	0.861	0.665	0.347	0.146	0.269
Cerebras-7B	0.636	0.582	0.744	0.564	0.311	0.625	0.734	0.779	0.603	0.141	0.012	0.259
Cerebras-13B	0.635	0.588	0.740	0.571	0.321	0.611	0.719	0.760	0.602	0.146	0.013	0.258
OPT-7B	0.677	0.676	0.773	0.579	0.329	0.665	0.719	0.840	0.656	0.227	0.020	0.251
OPT-13B	0.692	0.701	0.774	0.586	0.345	0.657	0.805	0.851	0.670	0.282	0.126	0.257

Source: <https://www.mosaicml.com/blog/mpt-7b>


MPT-7B-StoryWriter-65k+

Maximum Input Lengths of Different LLMs



InstructBLIP

Vision-Language Models with Instruction Tuning

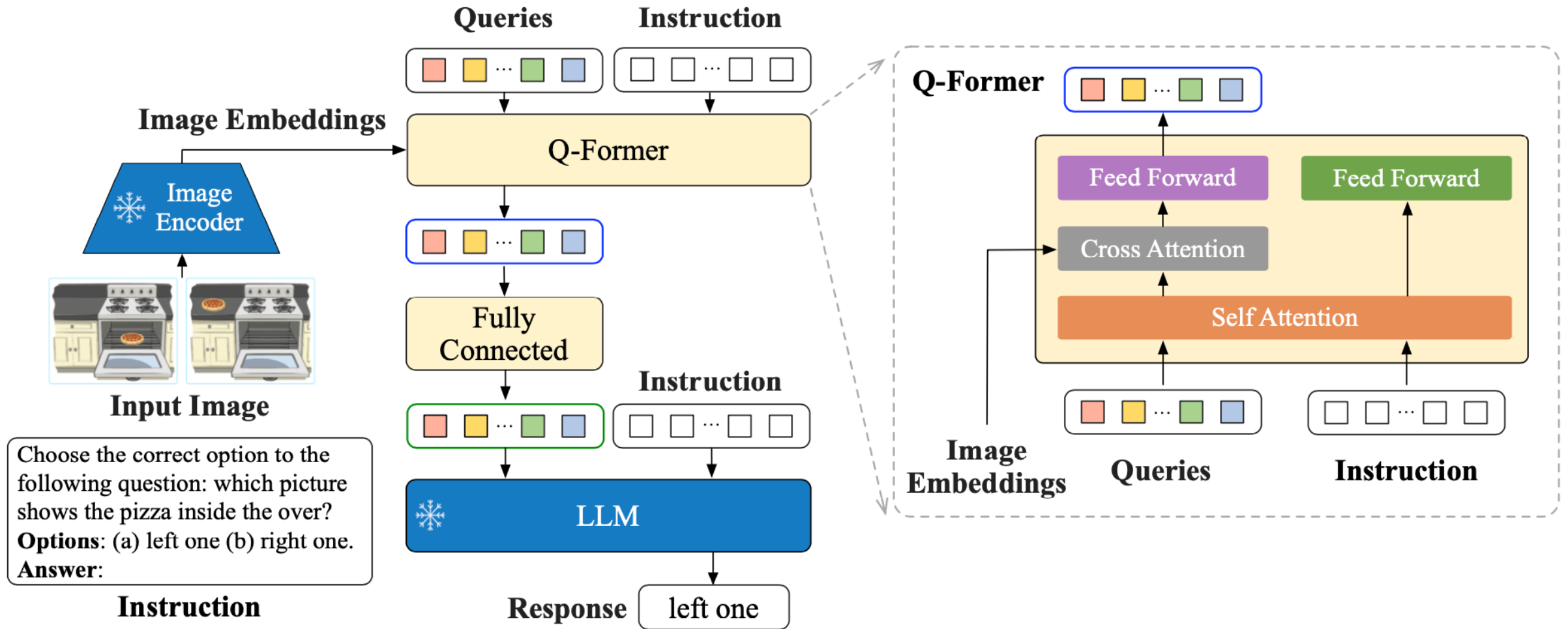
Introduce me this painting
in detail. 



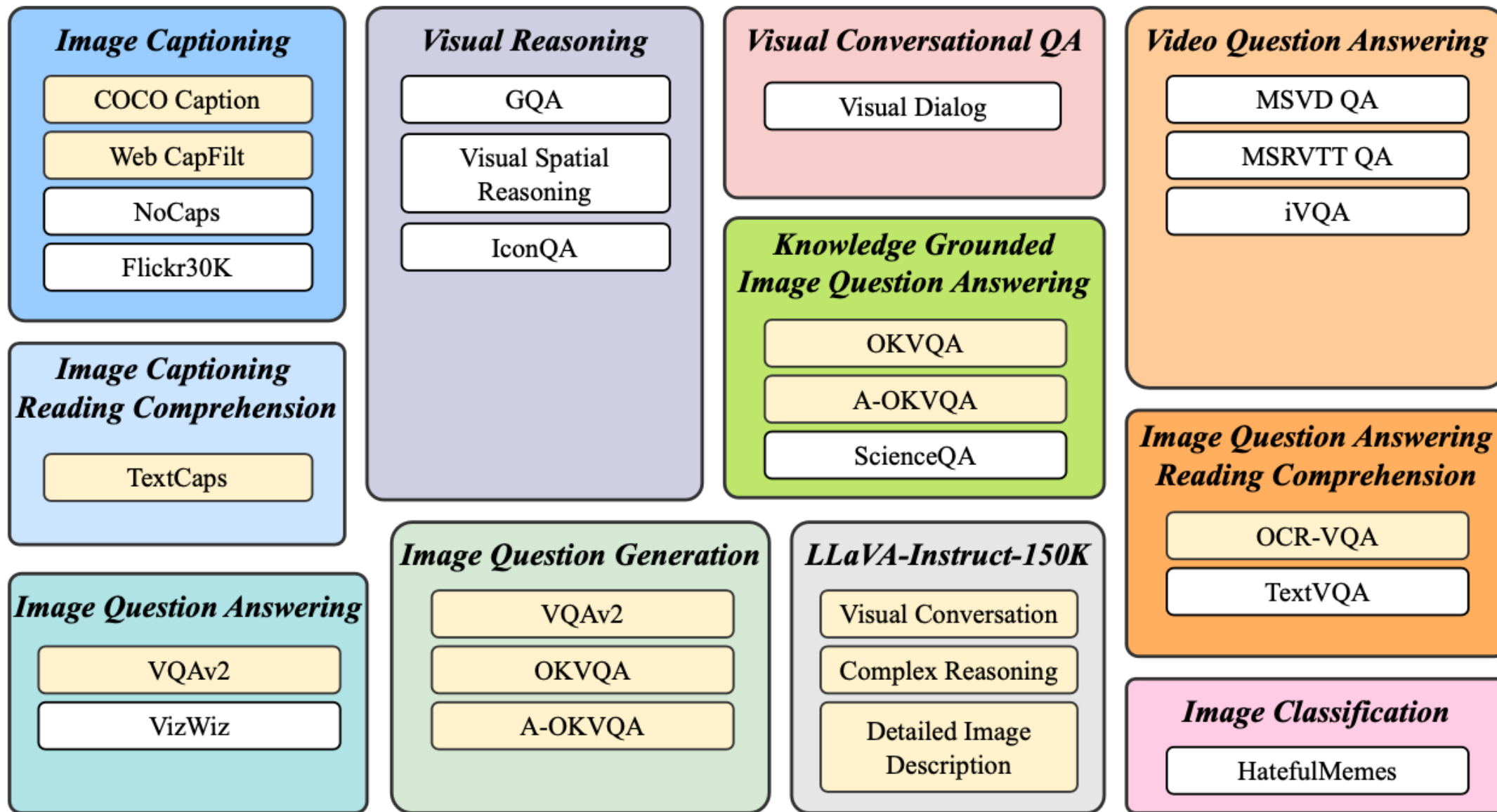
The painting depicts a young girl with a pearl earring, also known as "The Girl with the Pearl Earring" or "Girl with a Pearl Earring". She is wearing a blue dress and has a pearl earring on her left ear. Her hair is pulled back into a bun, and she is looking directly at the viewer. This painting is a famous work of art by Dutch painter Johannes Vermeer, created in 1665. It is considered one of his masterpieces and is widely recognized for its intricate details and realistic portrayal of the young girl's facial expression.

Model Architecture of InstructBLIP

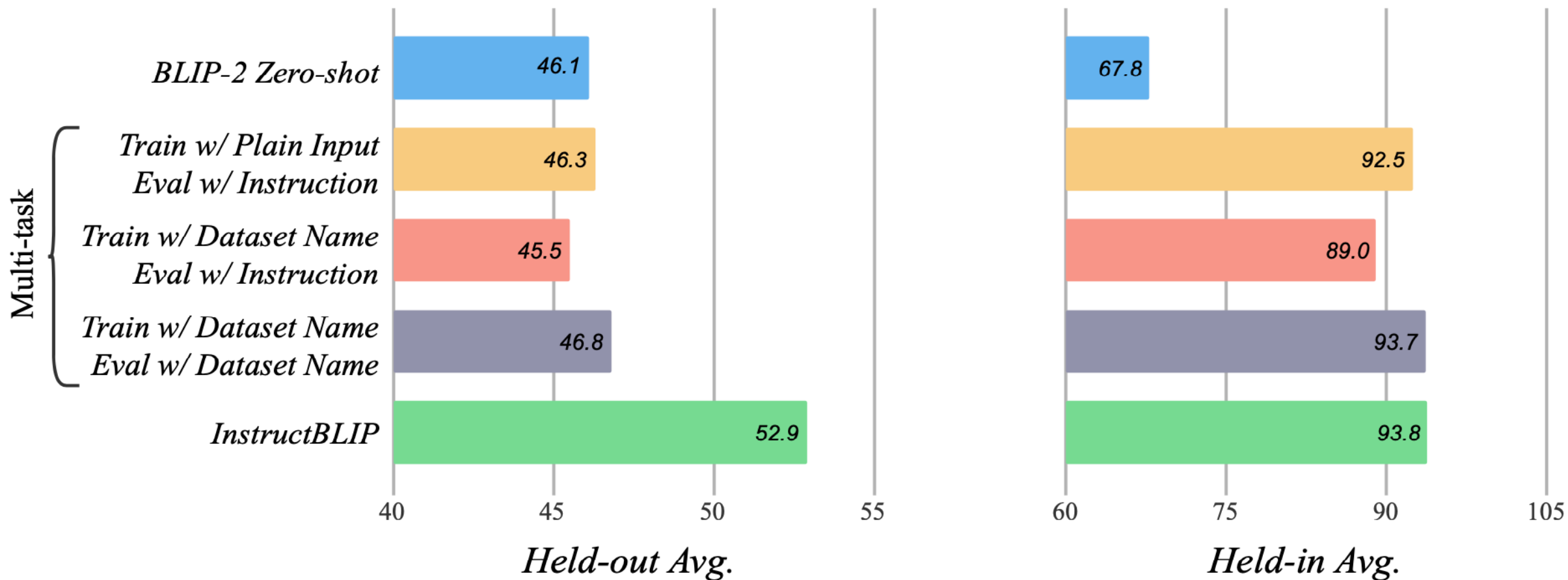
Vision-Language Models with Instruction Tuning



Vision-Language Instruction Tuning: Datasets



Instruction Tuning vs. Multitask Learning based on BLIP-2 FlanT5XL



Finetuning BLIP-2 and InstructBLIP on downstream datasets

	ScienceQA IMG	OCR-VQA	OKVQA	A-OKVQA			
				Direct Answer Val	Test	Multi-choice Val	Test
Previous SOTA	LLaVA [25] 89.0	GIT [42] 70.3	PaLM-E(562B) [9] 66.1	[15] 56.3	[36] 61.6	[15] 73.2	[36] 73.6
BLIP-2 (FlanT5 _{XXL})	89.5	72.7	54.7	57.6	53.7	80.2	76.2
InstructBLIP (FlanT5 _{XXL})	90.7	73.3	55.5	57.1	54.8	81.0	76.7
BLIP-2 (Vicuna-7B)	77.3	69.1	59.3	60.0	58.7	72.1	69.0
InstructBLIP (Vicuna-7B)	79.5	72.8	62.1	64.0	62.1	75.7	73.4

InstructBLIP

Vision-Language Models with Instruction Tuning

LLaVA

Large Language and Vision Assistant

MiniGPT-4

[Input Image]



[Input Instruction] Do you know who drew this painting?

[Model Response]

InstructBLIP  Leonardo da Vinci.

LLaVA

The painting depicts a woman, commonly believed to be Mona Lisa, the famous artwork by Leonardo da Vinci. It is a portrait painting that showcases the woman's enigmatic smile and has become one of the most famous and iconic art pieces in the world. The original work is displayed in the Louvre Museum in Paris, and it is known for its intricate details, use of oil paint, and the artist's innovative techniques that contributed to its enduring appeal and mystery.

MiniGPT-4

The painting is the Mona Lisa, also known as the Mona Lisa, by Leonardo da Vinci. It is one of the most famous paintings in the world and is known for its enigmatic smile and the mysterious identity of the sitter. The painting is housed in the Louvre Museum in Paris, France.

Instruction Tuning Datasets

Dataset Name	Held-out	Dataset Description
COCO Caption [23]	✗	We use the large-scale COCO dataset for the image captioning task. Specifically, Karpathy split [17] is used, which divides the data into 82K/5K/5K images for the train/val/test sets.
Web CapFilt	✗	14M image-text pairs collected from the web with additional BLIP-generated synthetic captions, used in BLIP [21] and BLIP-2 [20].
NoCaps [3]	✓ (val)	NoCaps contains 15,100 images with 166,100 human-written captions for novel object image captioning.
Flickr30K [50]	✓ (test)	The Flickr30k dataset consists of 31K images collected from Flickr, each image has five ground truth captions. We use the test split as the held-out which contains 1K images.
TextCaps [37]	✗	TextCaps is an image captioning dataset that requires the model to comprehend and reason the text in images. Its train/val/test sets contain 21K/3K/3K images, respectively.
VQAv2 [11]	✗	VQAv2 is dataset for open-ended image question answering. It is split into 82K/40K/81K for train/val/test.
VizWiz [12]	✓ (test-dev)	A dataset contains visual questions asked by people who are blind. 8K images are used for the held-out evaluation.
GQA [16]	✓ (test-dev)	GQA contains image questions for scene understanding and reasoning. We use the balanced test-dev set as held-out.
Visual Spatial Reasoning	✓ (test)	VSR is a collection of image-text pairs, in which the text describes the spatial relation of two objects in the image. Models are required to classify true/false for the description. We use the zero-shot data split given in its official github repository.
IconQA [28]	✓ (test)	IconQA measures the abstract diagram understanding and comprehensive cognitive reasoning abilities of models. We use the test set of its multi-text-choice task for held-out evaluation.
OKVQA [29]	✗	OKVQA contains visual questions that require outside knowledge to answer. It has been split into 9K/5K for train and test.
A-OKVQA [35]	✗	A-OKVQA is a successor of OKVQA with more challenging and diverse questions. It has 17K/1K/6K questions for train/val/test.

Instruction Tuning Datasets

Dataset Name	Held-out	Dataset Description
ScienceQA [27]	✓ (test)	ScienceQA covers diverse science topics with corresponding lectures and explanations. In out settings, we only use the part with image context (IMG).
Visual Dialog [8]	✓ (val)	Visual dialog is a conversational question answering dataset. We use the val split as the held-out, which contains 2,064 images and each has 10 rounds.
OCR-VQA [30]	✗	OCR-VQA contains visual questions that require models to read text in the image. It has 800K/100K/100K for train/val/test, respectively.
TextVQA [38]	✓ (val)	TextVQA requires models to comprehend visual text to answer questions.
HatefulMemes [18]	✓ (val)	A binary classification dataset to justify whether a meme contains hateful content.
LLaVA-Instruct-150K [25]	✗	An instruction tuning dataset which has three parts: detailed caption (23K), reasoning (77K), conversation (58K).
MSVD-QA [46]	✓ (test)	We use the test set (13K video QA pairs) of MSVD-QA for held-out testing.
MSRVTT-QA [46]	✓ (test)	MSRVTT-QA has more complex scenes than MSVD, with 72K video QA pairs as the test set.
iVQA [48]	✓ (test)	iVQA is a video QA dataset with mitigated language biases. It has 6K/2K/2K samples for train/val/test.

Instruction Templates

Image Captioning

VQA

Vision Question Answering

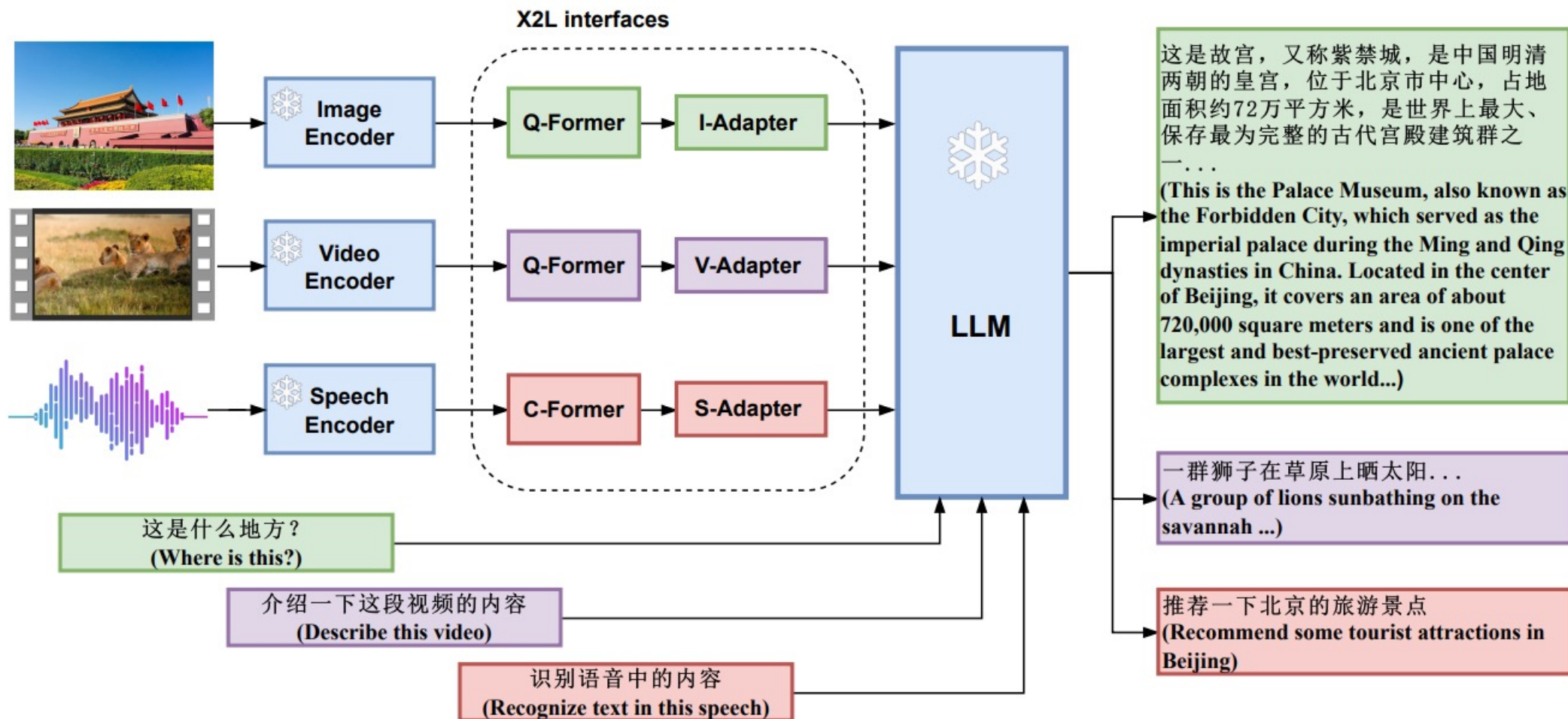
VQG

Vision Question Generation

Task	Instruction Template
Image Captioning	<p><Image>A short image caption: <Image>A short image description: <Image>A photo of <Image>An image that shows <Image>Write a short description for the image. <Image>Write a description for the photo. <Image>Provide a description of what is presented in the photo. <Image>Briefly describe the content of the image. <Image>Can you briefly explain what you see in the image? <Image>Could you use a few words to describe what you perceive in the photo? <Image>Please provide a short depiction of the picture. <Image>Using language, provide a short account of the image. <Image>Use a few words to illustrate what is happening in the picture.</p>
VQA	<p><Image>{Question} <Image>Question: {Question} <Image>{Question} A short answer to the question is <Image>Q: {Question} A: <Image>Question: {Question} Short answer: <Image>Given the image, answer the following question with no more than three words. {Question} <Image>Based on the image, respond to this question with a short answer: {Question}. Answer: <Image>Use the provided image to answer the question: {Question} Provide your answer as short as possible: <Image>What is the answer to the following question? "{Question}" <Image>The question "{Question}" can be answered using the image. A short answer is</p>
VQG	<p><Image>Given the image, generate a question whose answer is: {Answer}. Question: <Image>Based on the image, provide a question with the answer: {Answer}. Question: <Image>Given the visual representation, create a question for which the answer is "{Answer}". <Image>From the image provided, craft a question that leads to the reply: {Answer}. Question: <Image>Considering the picture, come up with a question where the answer is: {Answer}. <Image>Taking the image into account, generate an question that has the answer: {Answer}. Question:</p>

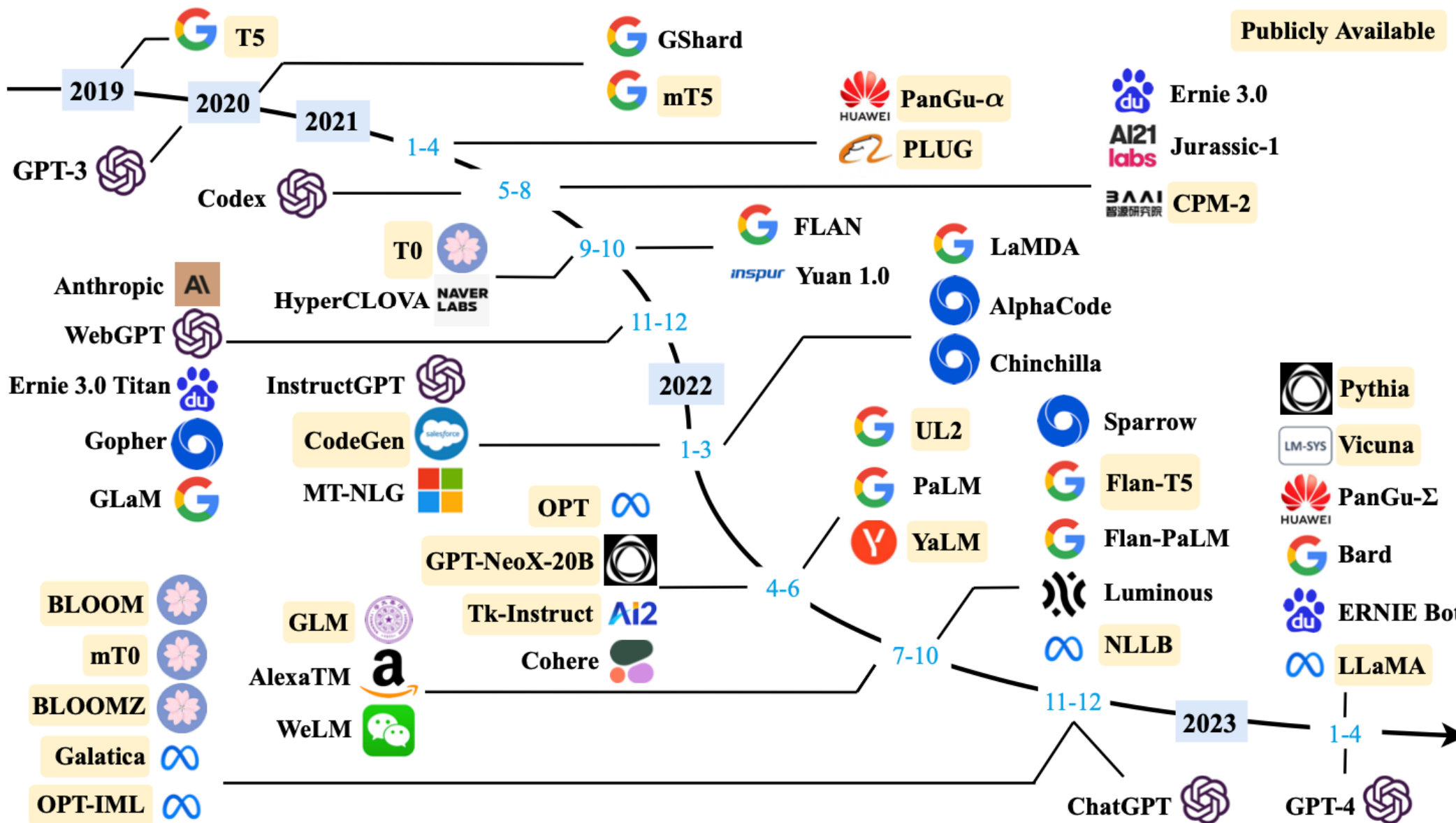
X-LLM:

Bootstrapping Advanced Large Language Models by Treating Multi-Modalities as Foreign Languages



Large Language Models (LLMs) Foundation Models

Large Language Models (LLMs) (larger than 10B)



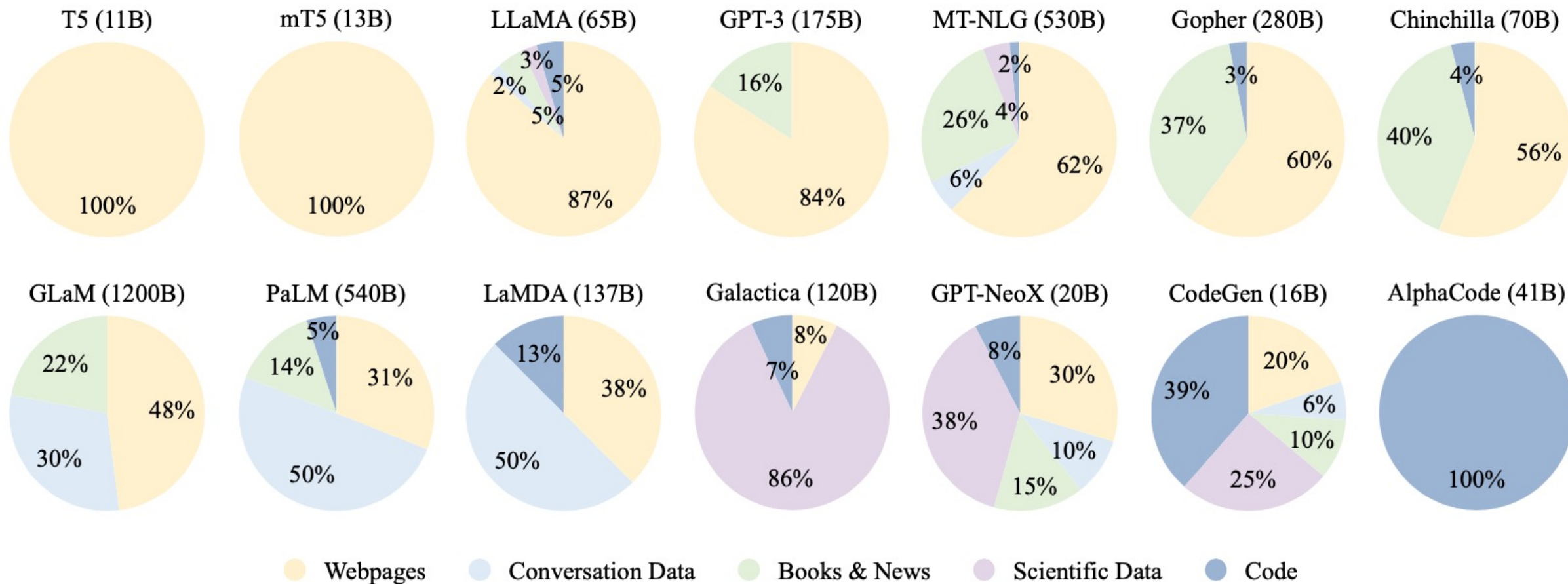
Large Language Models (LLMs) (larger than 10B)

Model	Release Time	Size (B)	Base Model	Adaptation		Pre-train Data Scale	Latest Data Timestamp	Hardware (GPUs / TPUs)	Training Time	Evaluation	
				IT	RLHF					ICL	CoT
T5 [72]	Oct-2019	11	-	-	-	1T tokens	Apr-2019	1024 TPU v3	-	✓	-
mT5 [73]	Oct-2020	13	-	-	-	1T tokens	-	-	-	✓	-
PanGu- α [74]	Apr-2021	13*	-	-	-	1.1TB	-	2048 Ascend 910	-	✓	-
CPM-2 [75]	Jun-2021	198	-	-	-	2.6TB	-	-	-	-	-
T0 [28]	Oct-2021	11	T5	✓	-	-	-	512 TPU v3	27 h	✓	-
CodeGen [76]	Mar-2022	16	-	-	-	577B tokens	-	-	-	✓	-
GPT-NeoX-20B [77]	Apr-2022	20	-	-	-	825GB	-	96 40G A100	-	✓	-
Tk-Instruct [78]	Apr-2022	11	T5	✓	-	-	-	256 TPU v3	4 h	✓	-
UL2 [79]	May-2022	20	-	-	-	1T tokens	Apr-2019	512 TPU v4	-	✓	✓
OPT [80]	May-2022	175	-	-	-	180B tokens	-	992 80G A100	-	✓	-
Publicly Available NLLB [81]	Jul-2022	54.5	-	-	-	-	-	-	-	✓	-
GLM [82]	Oct-2022	130	-	-	-	400B tokens	-	768 40G A100	60 d	✓	-
Flan-T5 [83]	Oct-2022	11	T5	✓	-	-	-	-	-	✓	✓
BLOOM [68]	Nov-2022	176	-	-	-	366B tokens	-	384 80G A100	105 d	✓	-
mT0 [84]	Nov-2022	13	mT5	✓	-	-	-	-	-	✓	-
Galactica [35]	Nov-2022	120	-	-	-	106B tokens	-	-	-	✓	✓
BLOOMZ [84]	Nov-2022	176	BLOOM	✓	-	-	-	-	-	✓	-
OPT-IML [85]	Dec-2022	175	OPT	✓	-	-	-	128 40G A100	-	✓	✓
LLaMA [57]	Feb-2023	65	-	-	-	1.4T tokens	-	2048 80G A100	21 d	✓	-
Pythia [86]	Apr-2023	12	-	-	-	300B tokens	-	256 40G A100	-	✓	-

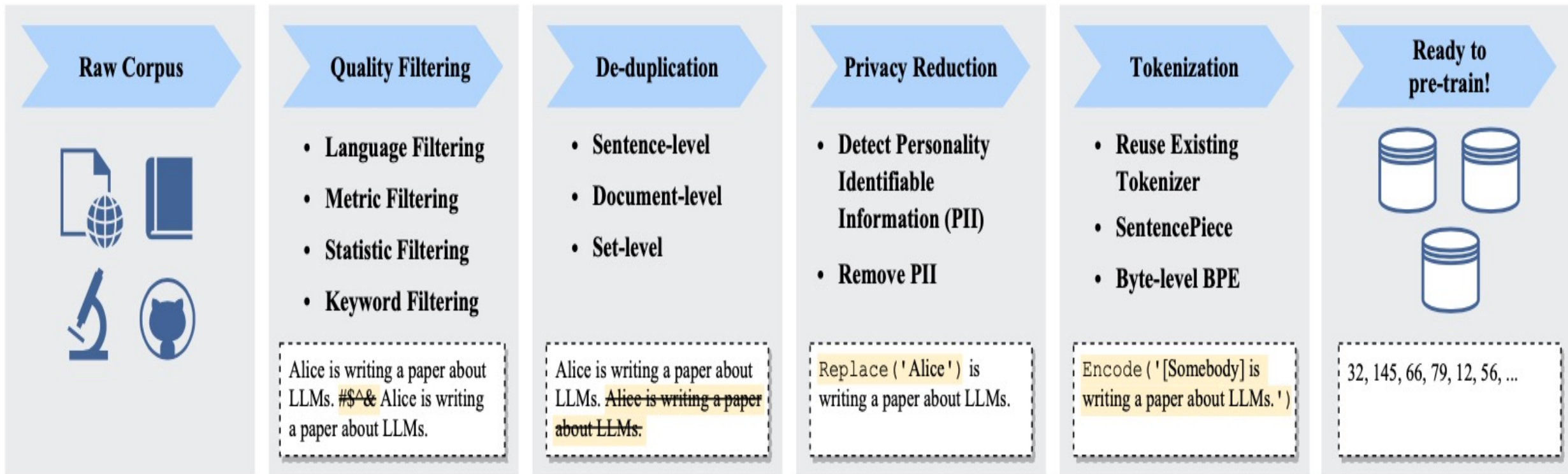
Large Language Models (LLMs) (larger than 10B)

	Model	Release Time	Size (B)	Base Model	Adaptation		Pre-train Data Scale	Latest Data Timestamp	Hardware (GPUs / TPUs)	Training Time	Evaluation		
					IT	RLHF					ICL	CoT	
Closed Source	GPT-3 [55]	May-2020	175	-	-	-	300B tokens	-	-	-	✓	-	
	GShard [87]	Jun-2020	600	-	-	-	1T tokens	-	2048 TPU v3	4 d	-	-	
	Codex [88]	Jul-2021	12	GPT-3	-	-	100B tokens	May-2020	-	-	-	✓	-
	ERNIE 3.0 [89]	Jul-2021	10	-	-	-	375B tokens	-	384 V100	-	-	✓	-
	Jurassic-1 [90]	Aug-2021	178	-	-	-	300B tokens	-	800 GPU	-	-	✓	-
	HyperCLOVA [91]	Sep-2021	82	-	-	-	300B tokens	-	1024 A100	13.4 d	-	✓	-
	FLAN [62]	Sep-2021	137	LaMDA	✓	-	-	-	128 TPU v3	60 h	-	✓	-
	Yuan 1.0 [92]	Oct-2021	245	-	-	-	180B tokens	-	2128 GPU	-	-	✓	-
	Anthropic [93]	Dec-2021	52	-	-	-	400B tokens	-	-	-	-	✓	-
	WebGPT [71]	Dec-2021	175	GPT-3	-	✓	-	-	-	-	-	✓	-
	Gopher [59]	Dec-2021	280	-	-	-	300B tokens	-	4096 TPU v3	920 h	-	✓	-
	ERNIE 3.0 Titan [94]	Dec-2021	260	-	-	-	300B tokens	-	2048 V100	28 d	-	✓	-
	GLaM [95]	Dec-2021	1200	-	-	-	280B tokens	-	1024 TPU v4	574 h	-	✓	-
	LaMDA [96]	Jan-2022	137	-	-	-	2.81T tokens	-	1024 TPU v3	57.7 d	-	-	-
	MT-NLG [97]	Jan-2022	530	-	-	-	270B tokens	-	4480 80G A100	-	-	✓	-
	AlphaCode [98]	Feb-2022	41	-	-	-	967B tokens	Jul-2021	-	-	-	-	-
	InstructGPT [61]	Mar-2022	175	GPT-3	✓	✓	-	-	-	-	-	✓	-
	Chinchilla [34]	Mar-2022	70	-	-	-	1.4T tokens	-	-	-	-	✓	-
	PaLM [56]	Apr-2022	540	-	-	-	780B tokens	-	6144 TPU v4	-	-	✓	✓
	AlexaTM [99]	Aug-2022	20	-	-	-	1.3T tokens	-	128 A100	120 d	-	✓	✓
	Sparrow [100]	Sep-2022	70	-	-	✓	-	-	64 TPU v3	-	-	✓	-
	WeLM [101]	Sep-2022	10	-	-	-	300B tokens	-	128 A100 40G	24 d	-	✓	-
	U-PaLM [102]	Oct-2022	540	PaLM	-	-	-	-	512 TPU v4	5 d	-	✓	✓
	Flan-PaLM [83]	Oct-2022	540	PaLM	✓	-	-	-	512 TPU v4	37 h	-	✓	✓
	Flan-U-PaLM [83]	Oct-2022	540	U-PaLM	✓	-	-	-	-	-	-	✓	✓
	GPT-4 [46]	Mar-2023	-	-	✓	✓	-	-	-	-	-	✓	✓
PanGu- Σ [103]	Mar-2023	1085	PanGu- α	-	-	329B tokens	-	512 Ascend 910	100 d	-	✓	-	

Ratios of various data sources in the pre-training data for existing LLMs



Typical Data Preprocessing Pipeline for Pre-training Large Language Models (LLMs)



LLMs with Public Configuration Details

Model	Category	Size	Normalization	PE	Activation	Bias	#L	#H	d_{model}	MCL
GPT3 [55]	Causal decoder	175B	Pre Layer Norm	Learned	GeLU	✓	96	96	12288	2048
PanGU- α [74]	Causal decoder	207B	Pre Layer Norm	Learned	GeLU	✓	64	128	16384	1024
OPT [80]	Causal decoder	175B	Pre Layer Norm	Learned	ReLU	✓	96	96	12288	2048
PaLM [56]	Causal decoder	540B	Pre Layer Norm	RoPE	SwiGLU	×	118	48	18432	2048
BLOOM [68]	Causal decoder	176B	Pre Layer Norm	ALiBi	GeLU	✓	70	112	14336	2048
MT-NLG [97]	Causal decoder	530B	-	-	-	-	105	128	20480	2048
Gopher [59]	Causal decoder	280B	Pre RMS Norm	Relative	-	-	80	128	16384	2048
Chinchilla [34]	Causal decoder	70B	Pre RMS Norm	Relative	-	-	80	64	8192	-
Galactica [35]	Causal decoder	120B	Pre Layer Norm	Learned	GeLU	×	96	80	10240	2048
LaMDA [96]	Causal decoder	137B	-	Relative	GeGLU	-	64	128	8192	-
Jurassic-1 [90]	Causal decoder	178B	Pre Layer Norm	Learned	GeLU	✓	76	96	13824	2048
LLaMA [57]	Causal decoder	65B	Pre RMS Norm	RoPE	SwiGLU	✓	80	64	8192	2048
GLM-130B [82]	Prefix decoder	130B	Post Deep Norm	RoPE	GeGLU	✓	70	96	12288	2048
T5 [72]	Encoder-decoder	11B	Pre RMS Norm	Relative	ReLU	×	24	128	1024	512

Note: PE denotes position embedding, #L denotes the number of layers, #H denotes the number of attention heads, d_{model} denotes the size of hidden states, and MCL denotes the maximum context length during training.

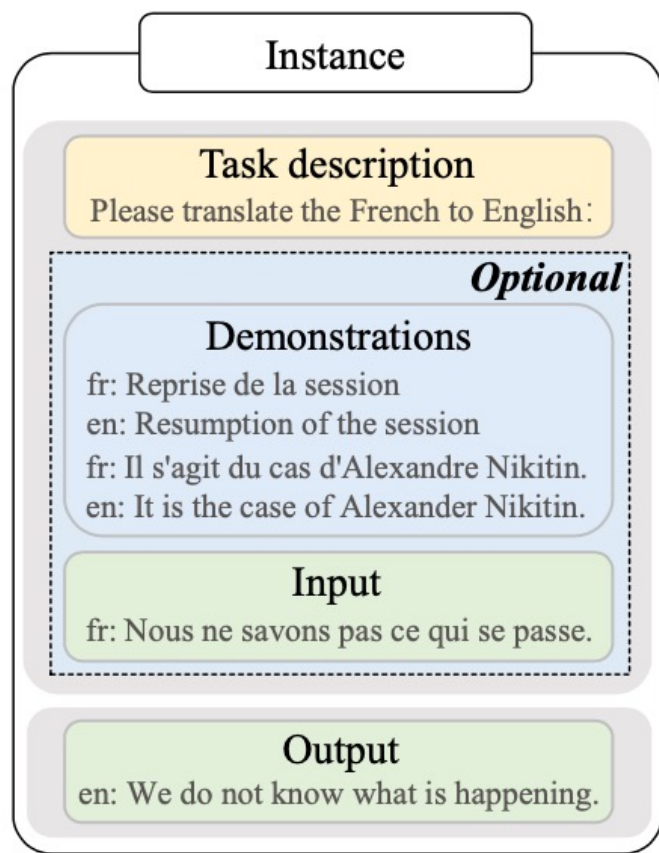
Detailed Optimization Settings of LLMs

Model	Batch Size (#tokens)	Learning Rate	Warmup	Decay Method	Optimizer	Precision Type	Weight Decay	Grad Clip	Dropout
GPT3 (175B)	32K→3.2M	6×10^{-5}	yes	cosine decay to 10%	Adam	FP16	0.1	1.0	-
PanGu- α (200B)	-	2×10^{-5}	-	-	Adam	-	0.1	-	-
OPT (175B)	2M	1.2×10^{-4}	yes	manual decay	AdamW	FP16	0.1	-	0.1
PaLM (540B)	1M→4M	1×10^{-2}	no	inverse square root	Adafactor	BF16	lr^2	1.0	0.1
BLOOM (176B)	4M	6×10^{-5}	yes	cosine decay to 10%	Adam	BF16	0.1	1.0	0.0
MT-NLG (530B)	64 K→3.75M	5×10^{-5}	yes	cosine decay to 10%	Adam	BF16	0.1	1.0	-
Gopher (280B)	3M→6M	4×10^{-5}	yes	cosine decay to 10%	Adam	BF16	-	1.0	-
Chinchilla (70B)	1.5M→3M	1×10^{-4}	yes	cosine decay to 10%	AdamW	BF16	-	-	-
Galactica (120B)	2M	7×10^{-6}	yes	linear decay to 10%	AdamW	-	0.1	1.0	0.1
LaMDA (137B)	256K	-	-	-	-	BF16	-	-	-
Jurassic-1 (178B)	32 K→3.2M	6×10^{-5}	yes	-	-	-	-	-	-
LLaMA (65B)	4M	1.5×10^{-4}	yes	cosine decay to 10%	AdamW	-	0.1	1.0	-
GLM (130B)	0.4M→8.25M	8×10^{-5}	yes	cosine decay to 10%	AdamW	FP16	0.1	1.0	0.1
T5 (11B)	64K	1×10^{-2}	no	inverse square root	AdaFactor	-	-	-	0.1
ERNIE 3.0 Titan (260B)	-	1×10^{-4}	-	-	Adam	FP16	0.1	1.0	-
PanGu- Σ (1.085T)	0.5M	2×10^{-5}	yes	-	Adam	FP16	-	-	-

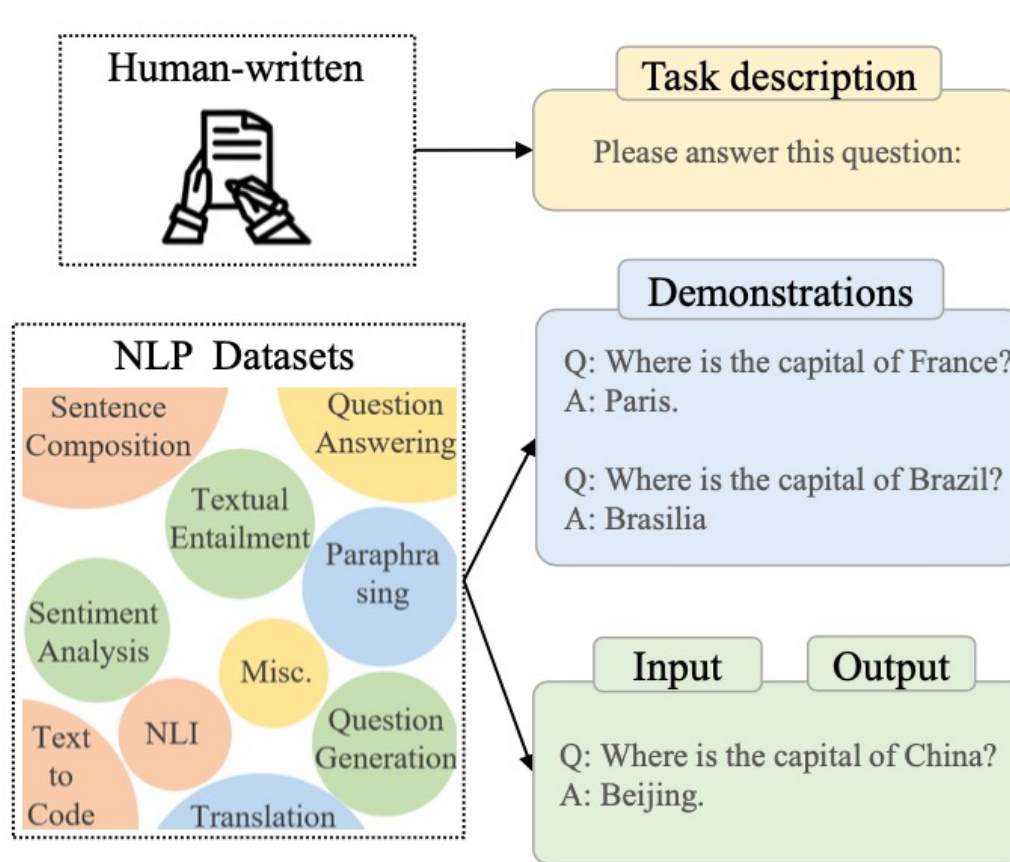
Available Task Collections for Instruction Tuning

Collections	Time	#Task types	#Tasks	#Examples
Nat. Inst. [193]	Apr-2021	6	61	193K
CrossFit [194]	Apr-2021	13	160	7.1M
FLAN [62]	Sep-2021	12	62	4.4M
P3 [195]	Oct-2021	13	267	12.1M
ExMix [196]	Nov-2021	11	107	18M
UnifiedSKG [197]	Jan-2022	6	21	812K
Super Nat. Inst. [78]	Apr-2022	76	1616	5M
MVPCorpus [198]	Jun-2022	11	77	41M
xP3 [84]	Nov-2022	17	85	81M
OIG ¹⁴	Mar-2023	-	-	43M

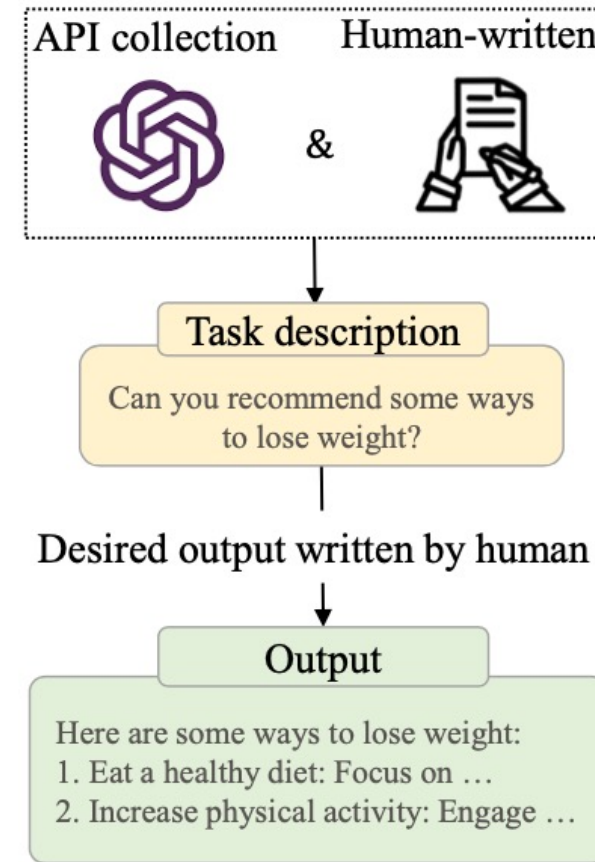
Instance Formatting and Two Different Methods for Constructing the Instruction-formatted Instances



(a) Instance format



(b) Formatting existing datasets



(c) Formatting human needs

In-context Learning (ICL) and Chain-of-thought (CoT) Prompting

In-Context Learning

Answer the following mathematical reasoning questions:

$N \times$

Q: If you have 12 candies and you give 4 candies to your friend, how many candies do you have left?

A: The answer is 8.

Q: If a rectangle has a length of 6 cm and a width of 3 cm, what is the perimeter of the rectangle?

A: The answer is 18 cm.

Q: Sam has 12 marbles. He gives $\frac{1}{4}$ of them to his sister. How many marbles does Sam have left?

A: The answer is 9.

Chain-of-Thought Prompting

Answer the following mathematical reasoning questions:

$N \times$

Q: If a rectangle has a length of 6 cm and a width of 3 cm, what is the perimeter of the rectangle?

A: For a rectangle, add up the length and width and double it. So, the perimeter of this rectangle is $(6 + 3) \times 2 = 18$ cm.


The answer is 18 cm.

Q: Sam has 12 marbles. He gives $\frac{1}{4}$ of them to his sister. How many marbles does Sam have left?

A: He gives $(\frac{1}{4}) \times 12 = 3$ marbles. So Sam is left with $12 - 3 = 9$ marbles. The answer is 9.

 : Task description

 : Demonstration

 : Chain-of-Thought

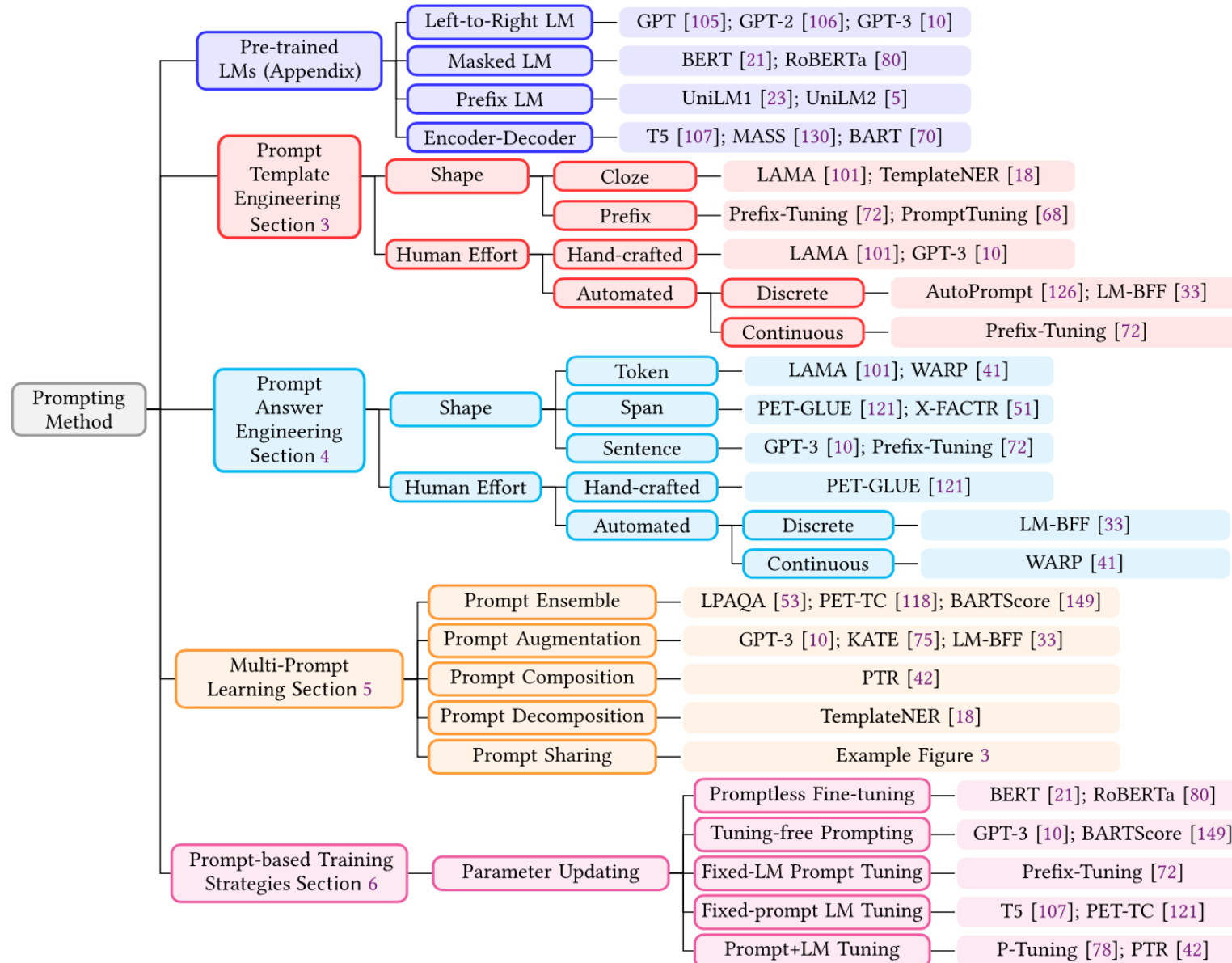
 : Query

Pre-train, Prompt, and Predict: Prompting Methods in Natural Language Processing (LLMs)

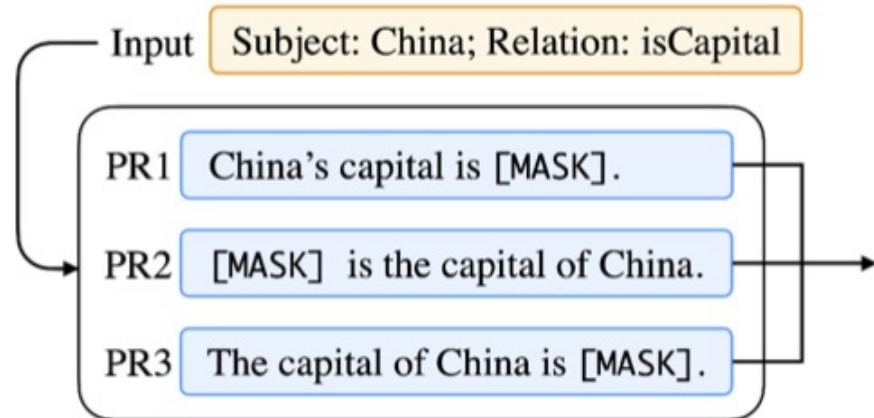
Four Paradigms in NLP

Paradigm	Engineering	Task Relation
a. Fully Supervised Learning (Non-Neural Network)	Feature (e.g. word identity, part-of-speech, sentence length)	
b. Fully Supervised Learning (Neural Network)	Architecture (e.g. convolutional, recurrent, self-attentional)	
c. Pre-train, Fine-tune	Objective (e.g. masked language modeling, next sentence prediction)	
d. Pre-train, Prompt, Predict	Prompt (e.g. cloze, prefix)	

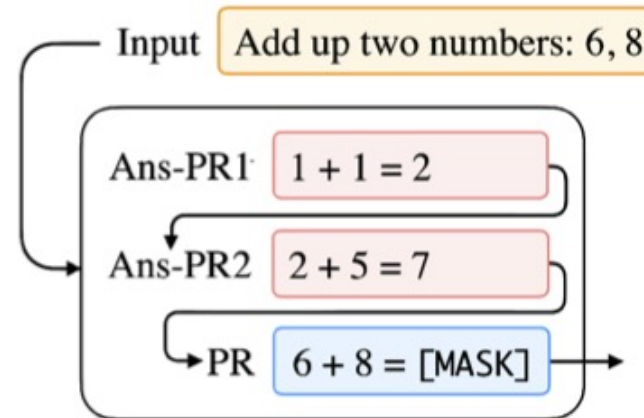
Typology of Prompting Methods



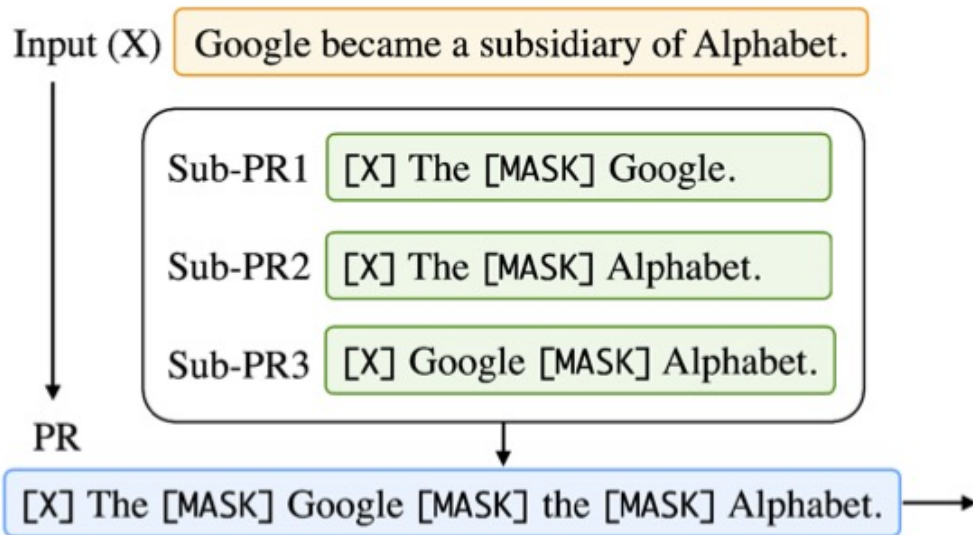
Different Multi-Prompt Learning Strategies



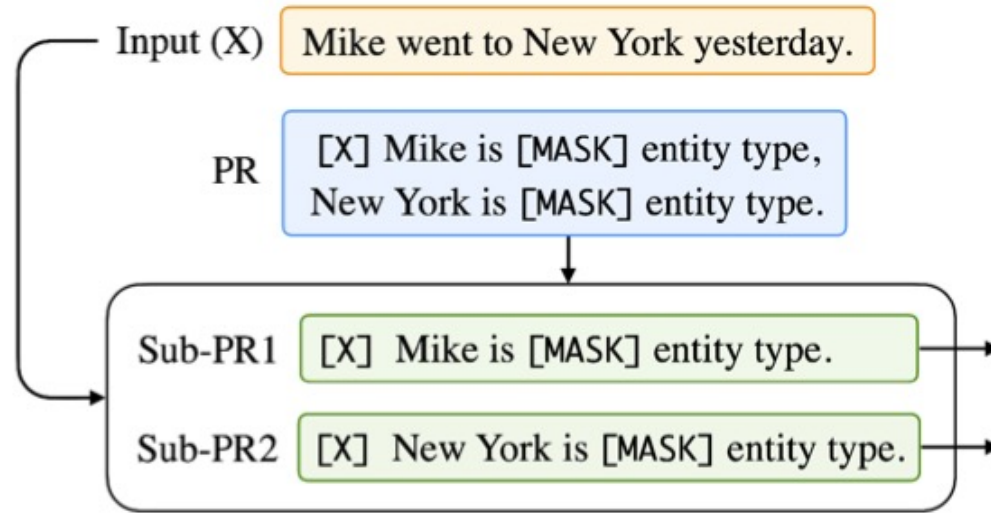
(a) Prompt Ensembling.



(b) Prompt Augmentation.



(c) Prompt Composition.



(d) Prompt Decomposition.

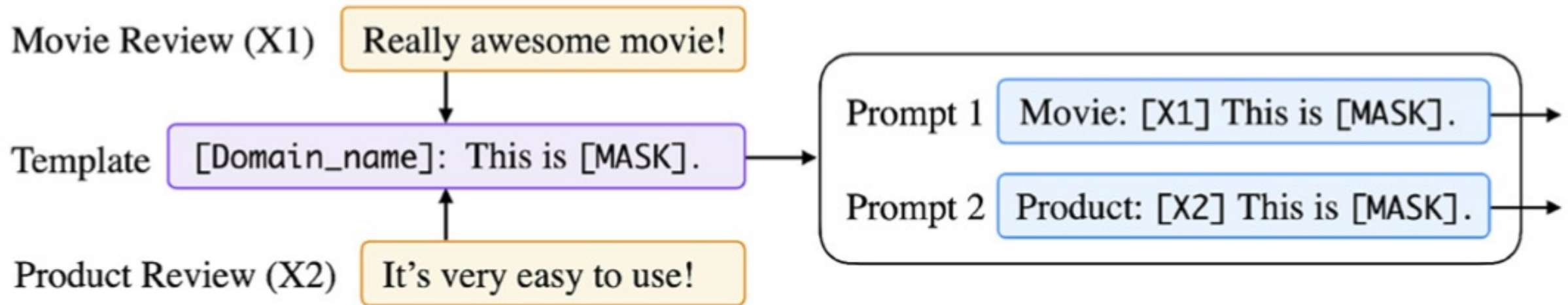
Examples of Input, Template, and Answer for Different Tasks

Type	Task Example	Input ([X])	Template	Answer ([Z])
Text Classification	Sentiment	I love this movie.	[X] The movie is [Z].	great fantastic ...
	Topics	He prompted the LM.	[X] The text is about [Z].	sports science ...
	Intention	What is taxi fare to Denver?	[X] The question is about [Z].	quantity city ...
Text-span Classification	Aspect Sentiment	Poor service but good food.	[X] What about service? [Z].	Bad Terrible ...
Text-pair Classification	Natural Language Inference	[X1]: An old man with ... [X2]: A man walks ...	[X1]? [Z], [X2]	Yes No ...
Tagging	Named Entity Recognition	[X1]: Mike went to Paris. [X2]: Paris	[X1][X2] is a [Z] entity.	organization location ...
Text Generation	Summarization	Las Vegas police ...	[X] TL;DR: [Z]	The victim ... A woman
	Translation	Je vous aime.	French: [X] English: [Z]	I love you. I fancy you. ...
Regression	Textual Similarity	[X1]: A man is smoking. [X2]: A man is skating.	[X1] [Z], [X2]	Yes No ...

Characteristics of Different Tuning Strategies

Strategy	LM Params	Prompt Params		Example
		Additional	Tuned	
Promptless Fine-tuning	Tuned	—		ELMo [97], BERT [20], BART [69]
Tuning-free Prompting	Frozen	✗	✗	GPT-3 [9], AutoPrompt [125], LAMA [100]
Fixed-LM Prompt Tuning	Frozen	✓	Tuned	Prefix-Tuning [71], Prompt-Tuning [67]
Fixed-prompt LM Tuning	Tuned	✗	✗	PET-TC [117], PET-Gen [118], LM-BFF [32]
Prompt+LM Fine-tuning	Tuned	✓	Tuned	PADA [5], P-Tuning [77], PTR [41]

Multi-prompt Learning for Multi-task, Multi-domain, or Multi-lingual Learning



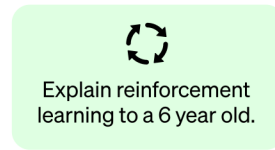
Reinforcement Learning from Human Feedback (RLHF)

ChatGPT: Optimizing Language Models for Dialogue

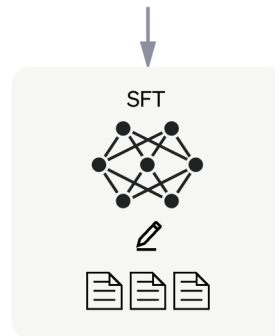
Step 1

Collect demonstration data and train a supervised policy.

A prompt is sampled from our prompt dataset.



A labeler demonstrates the desired output behavior.

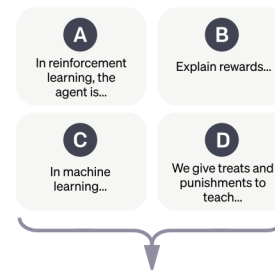
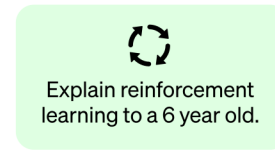


This data is used to fine-tune GPT-3.5 with supervised learning.

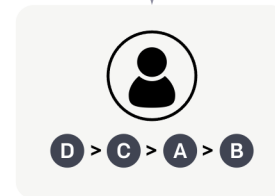
Step 2

Collect comparison data and train a reward model.

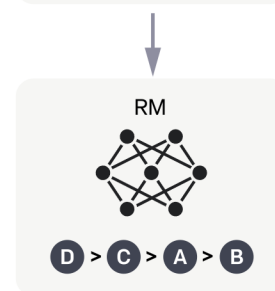
A prompt and several model outputs are sampled.



A labeler ranks the outputs from best to worst.



This data is used to train our reward model.



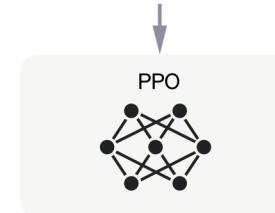
Step 3

Optimize a policy against the reward model using the PPO reinforcement learning algorithm.

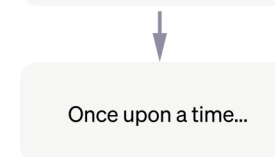
A new prompt is sampled from the dataset.



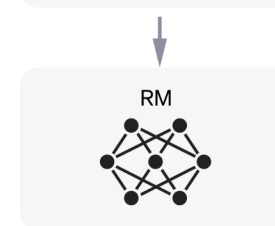
The PPO model is initialized from the supervised policy.



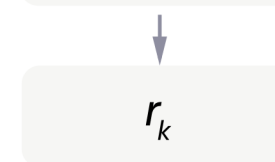
The policy generates an output.



The reward model calculates a reward for the output.



The reward is used to update the policy using PPO.



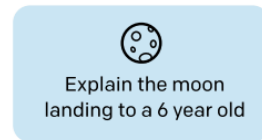
Training language models to follow instructions with human feedback

InstructGPT and GPT 3.5

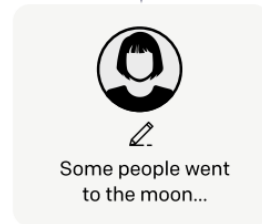
Step 1

Collect demonstration data, and train a supervised policy.

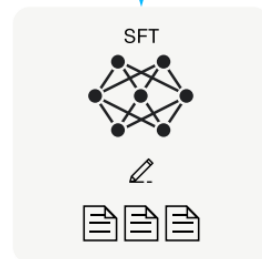
A prompt is sampled from our prompt dataset.



A labeler demonstrates the desired output behavior.



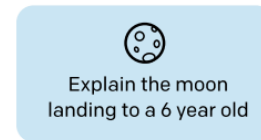
This data is used to fine-tune GPT-3 with supervised learning.



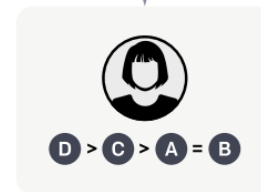
Step 2

Collect comparison data, and train a reward model.

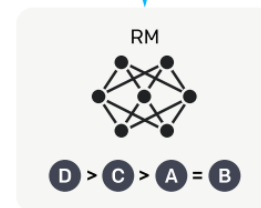
A prompt and several model outputs are sampled.



A labeler ranks the outputs from best to worst.



This data is used to train our reward model.



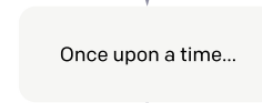
Step 3

Optimize a policy against the reward model using reinforcement learning.

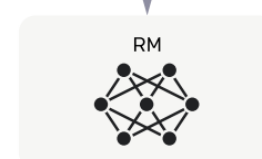
A new prompt is sampled from the dataset.



The policy generates an output.



The reward model calculates a reward for the output.



The reward is used to update the policy using PPO.

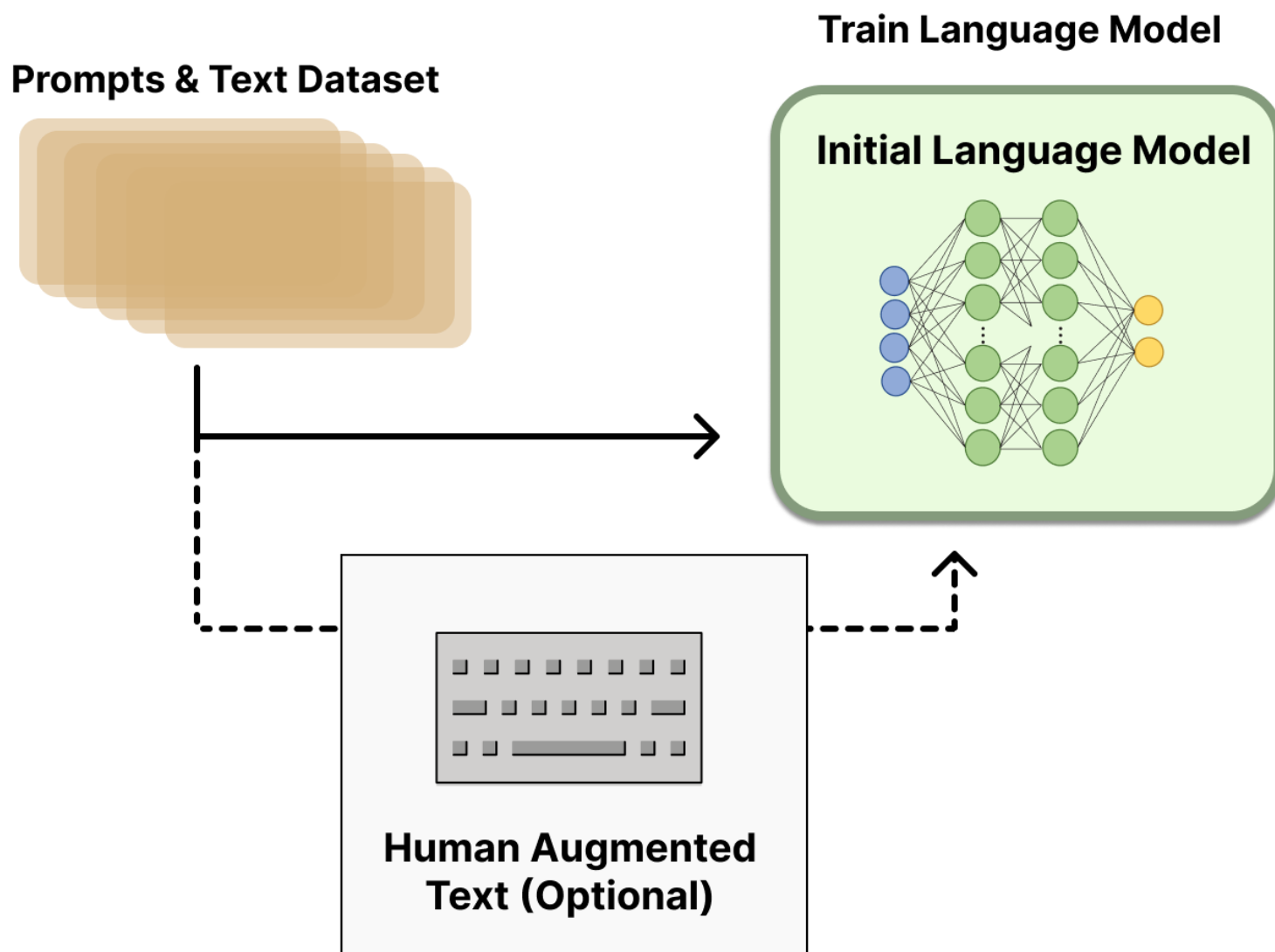


Reinforcement Learning from Human Feedback (RLHF)

- 1. Pretraining a Language Model (LM)**
- 2. Gathering Data and Training a Reward Model**
- 3. Fine-tuning the LM with Reinforcement Learning**

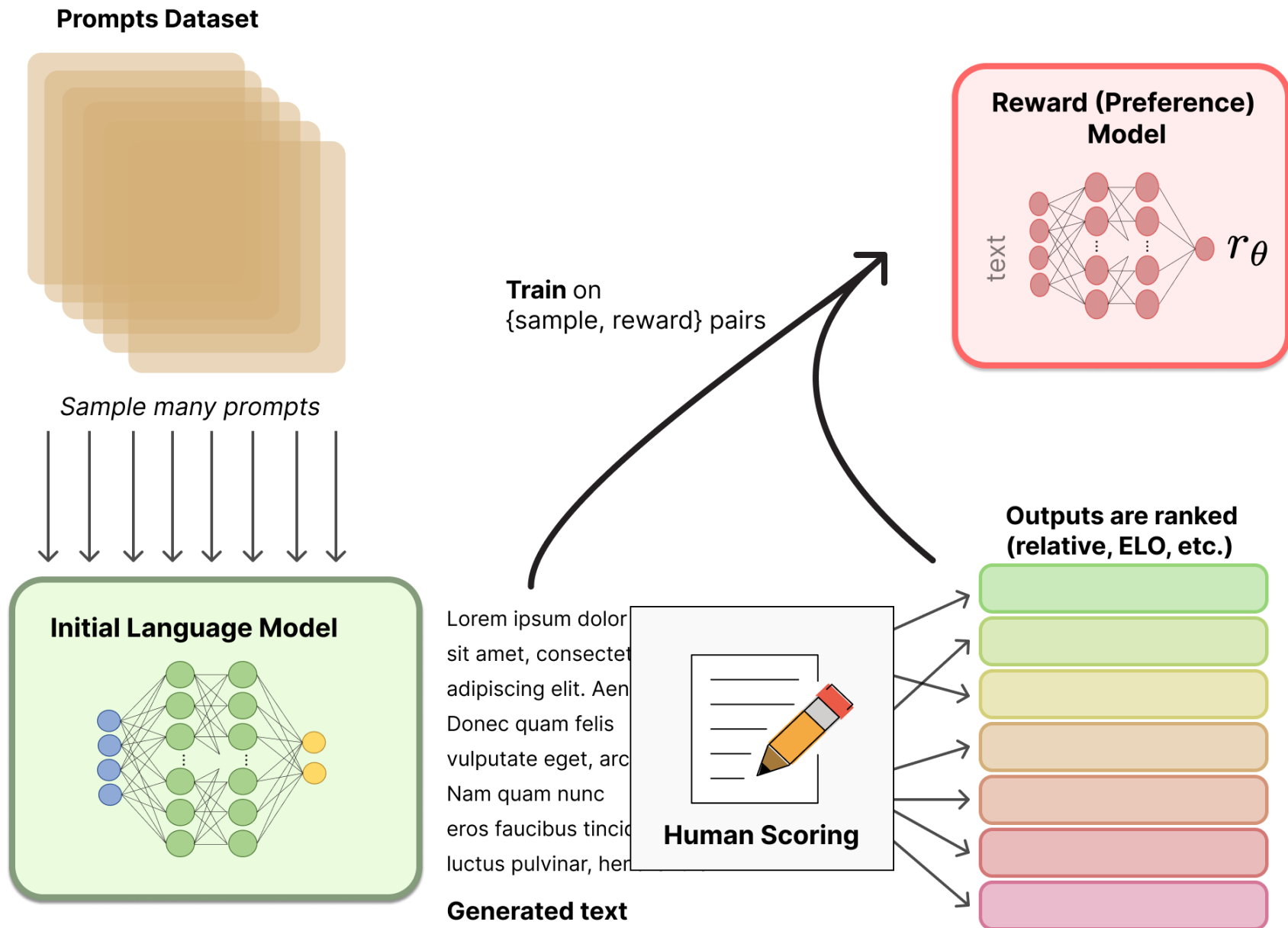
Reinforcement Learning from Human Feedback (RLHF)

Step 1. Pretraining a Language Model (LM)



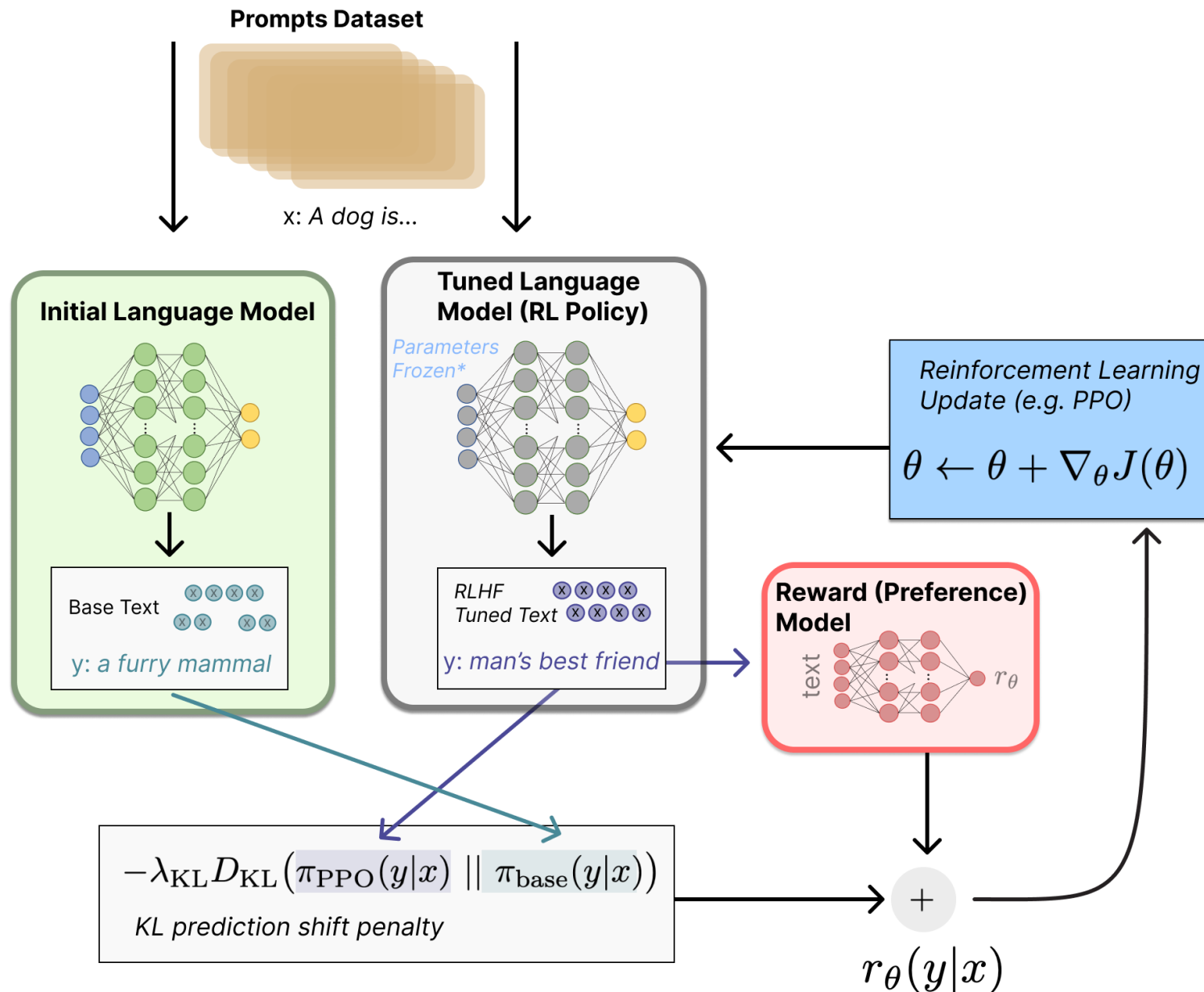
Reinforcement Learning from Human Feedback (RLHF)

Step 2. Gathering Data and Training a Reward Model



Reinforcement Learning from Human Feedback (RLHF)

Step 3. Fine-tuning the LM with Reinforcement Learning

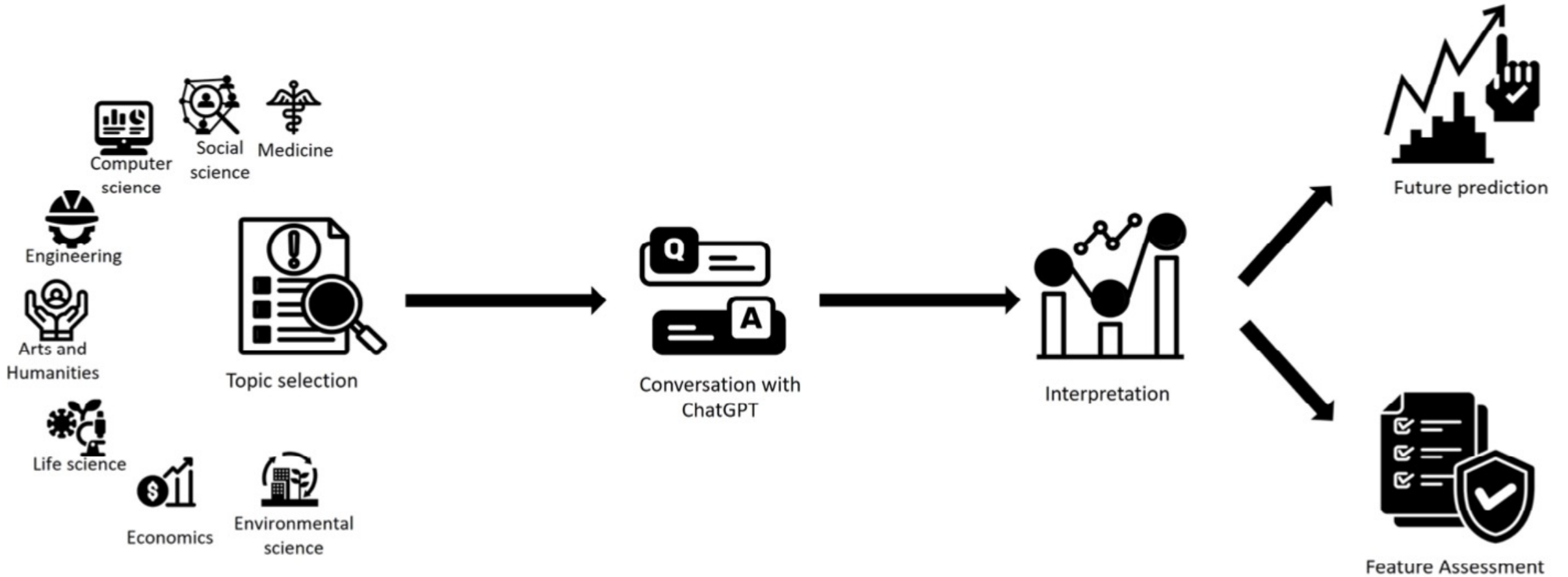


Foundation Models for Big Data Analytics

Generative AI

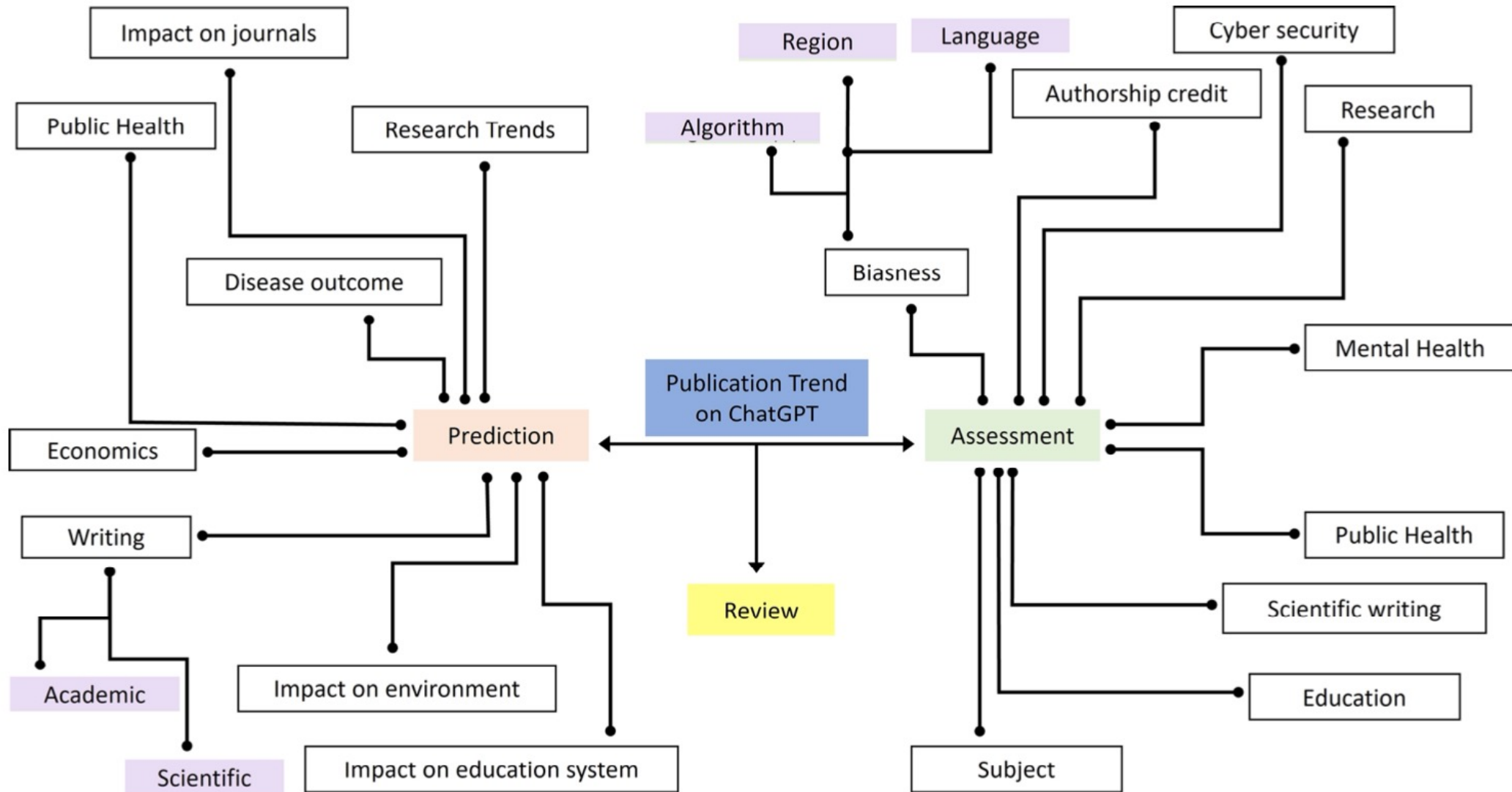
**Text, Image, Video, Audio
Applications**

ChatGPT Research

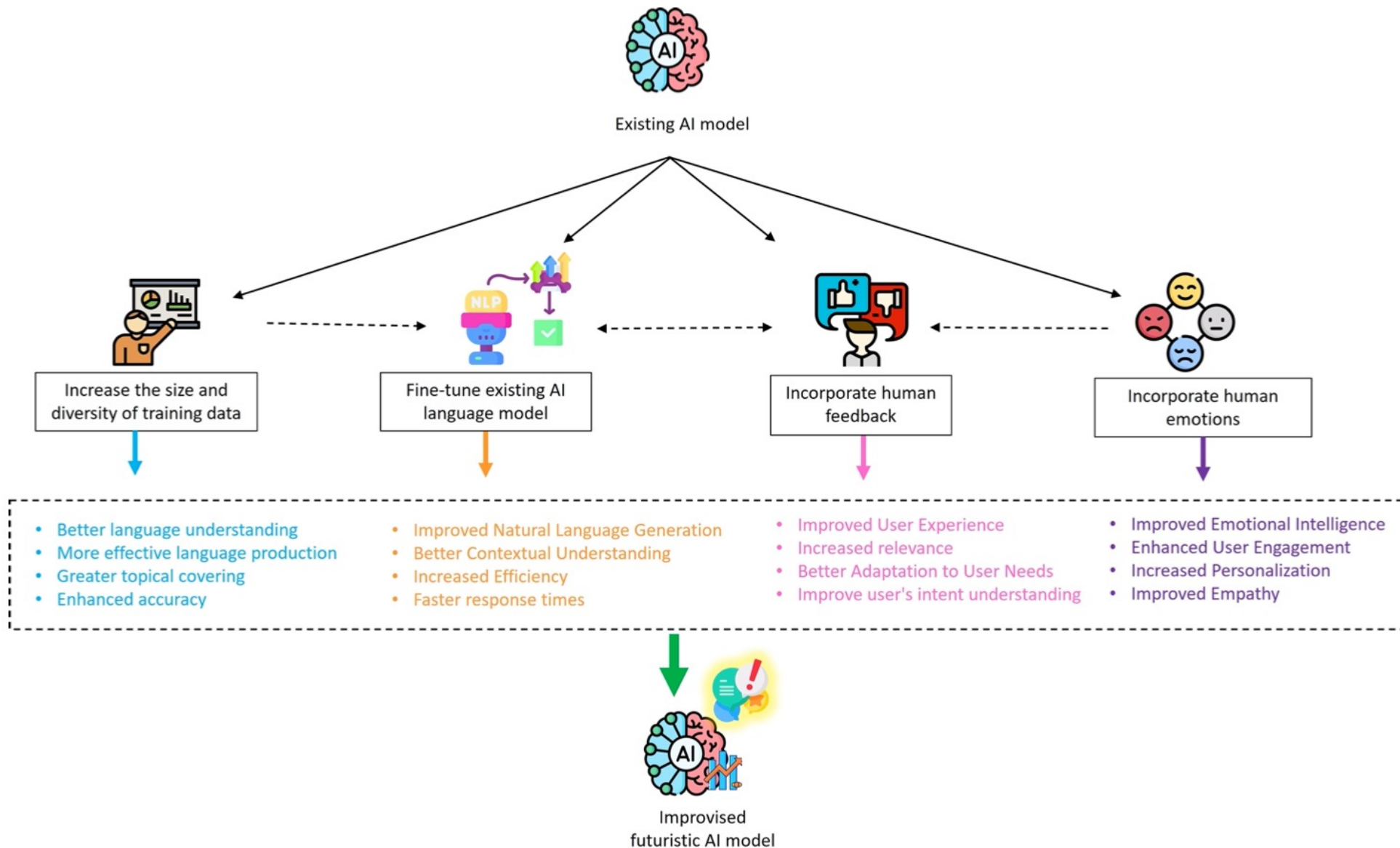


A common trend identified in the reported chatGPT research

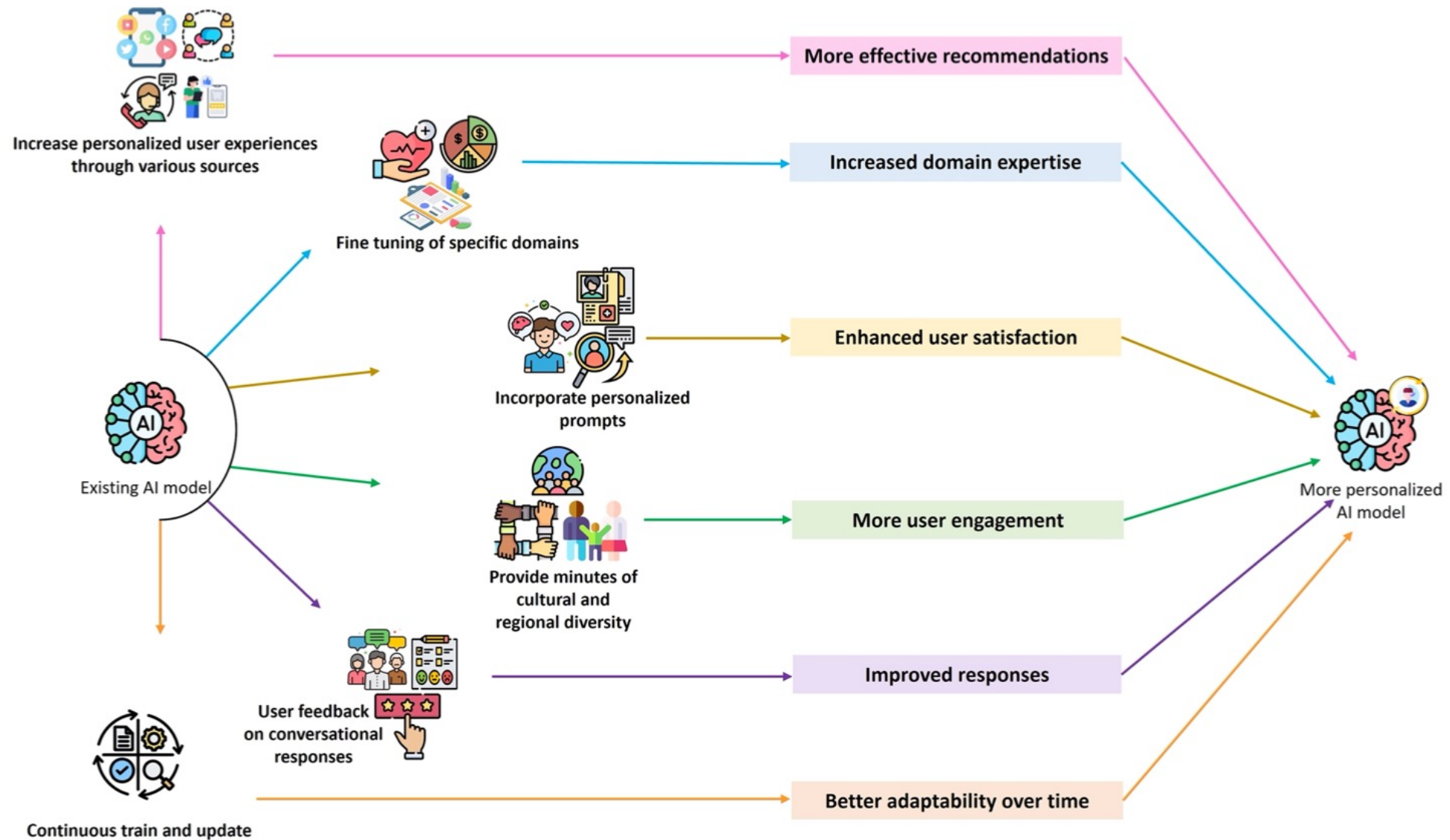
Taxonomy of Literature on ChatGPT



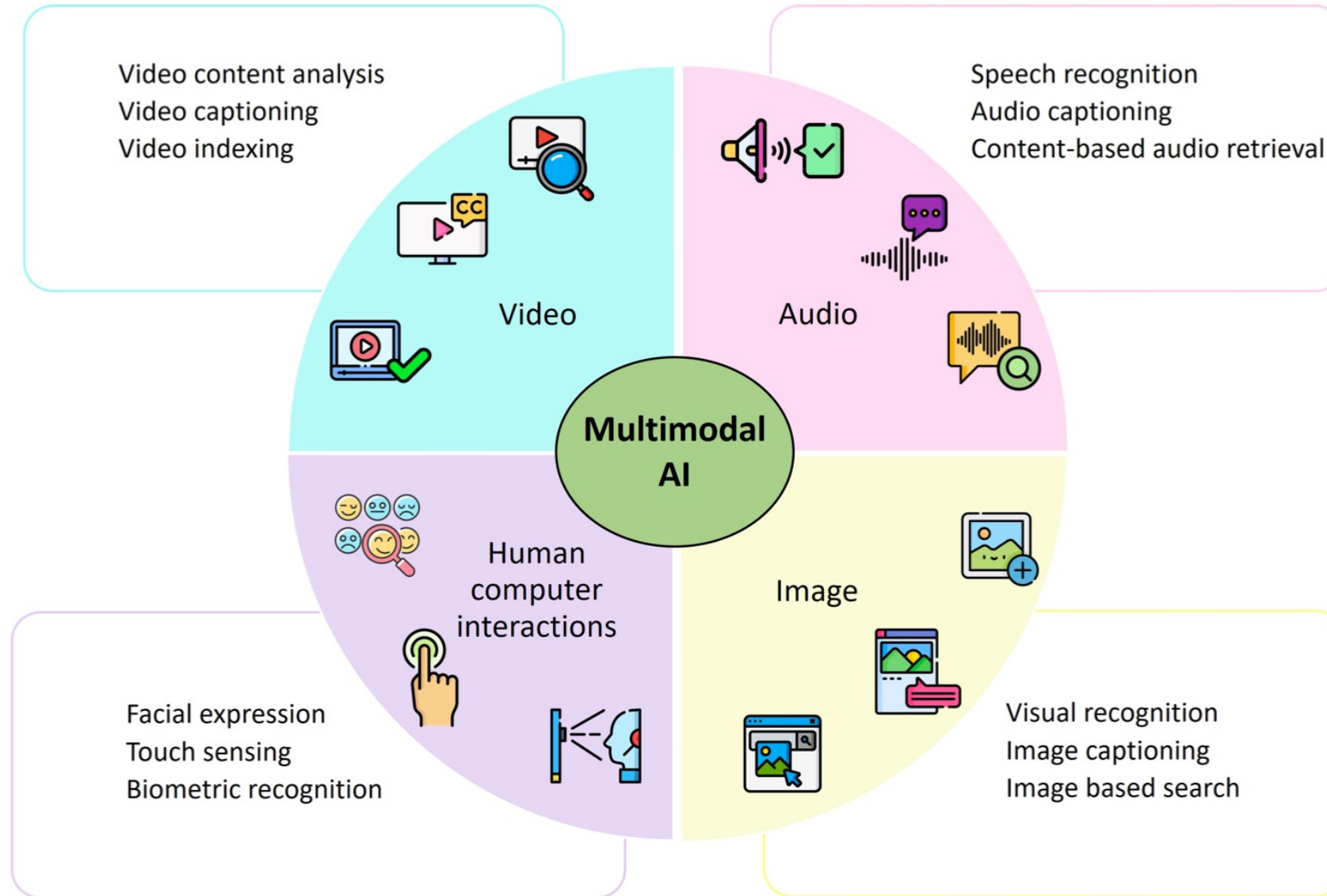
Enhancing the Conversational Ability of ChatGPT



Domain/Person Specific Personalization of ChatGPT



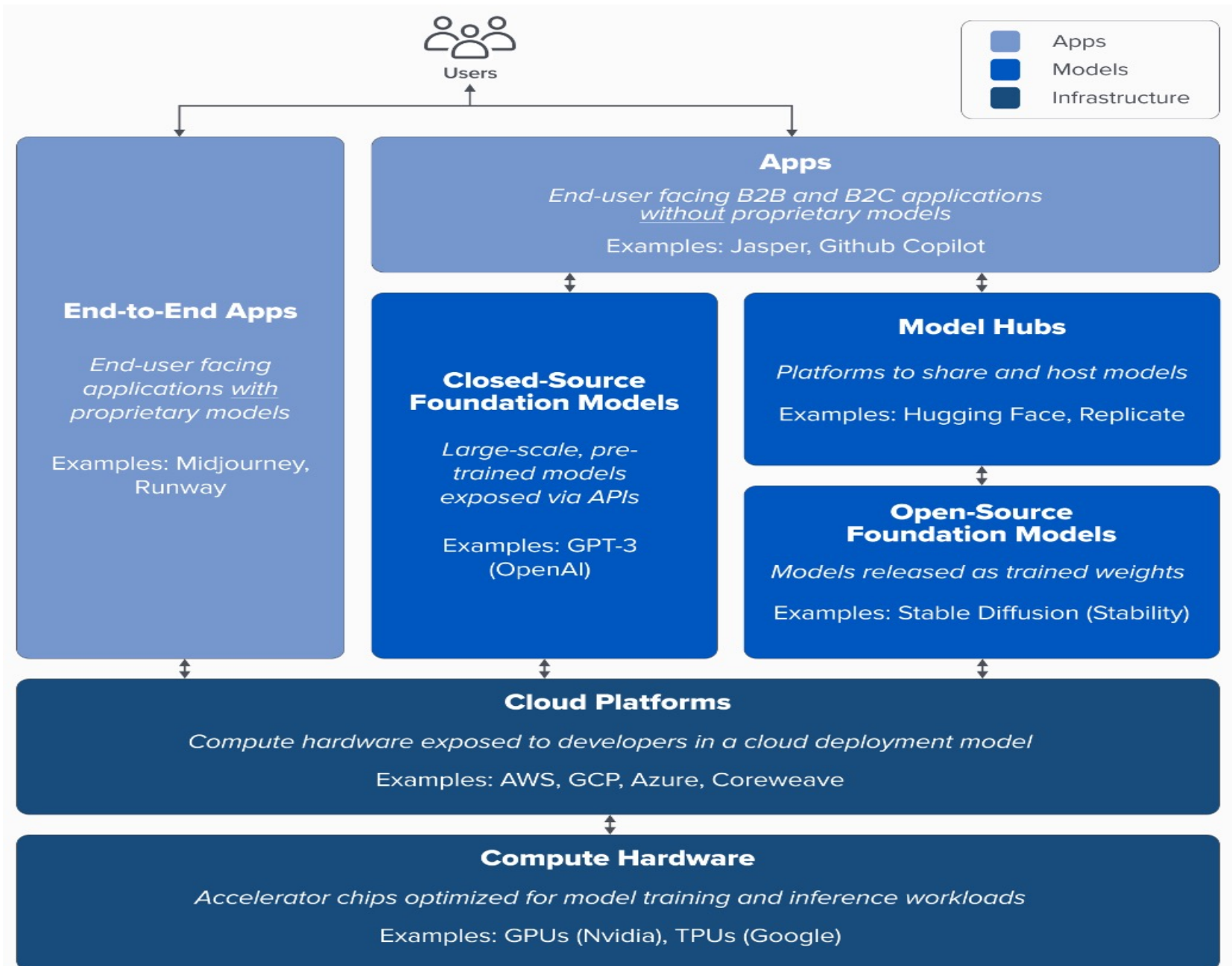
Technological Integration for Multimodal AI



Applications of Generative AI Models

Application	Platform/Software	Company	Year	Papaer	Link
ChatBot	Xiaoice	Microsoft	2018	[200]	Xiaoice
ChatBot	Meena	Google	2020	[201]	Meena Blog
ChatBot	BlenderBot	Meta	2022	[202]	Blenderbot
ChatBot	ChatGPT	OpenAI	2022	[10]	ChatGPT
ChatBot	Alexa	Amazon	2014	-	Amazon Alexa
ChatBot	Lex	Amazon	2017	-	Amazon Lex
Music	AIVA	Aiva Tech	2016	-	AIVA
Music	Jukebox	OpenAI	2020	[203]	Jukebox
Code	CodeGPT	Microsoft	2021	[204]	CodeGPT
Code	CodeParrot	CodeParrot	2022	[205]	CodeParrot
Code	Codex	OpenAI	2021	[206]	Codex blog
Code	CoPilot	Microsoft	2021	[206]	CoPilot
Art	DALL-E-2	OpenAI	2022	[5]	DALL-E-2 Blog
Art	DreamStudio	Stability	2022	[13]	Dreamstudio
Art	craiyon	OpenAI	2021	[1]	Craiyon
Art	Imagen	Google	2022	[152]	Imagen
Education	Minerva	Google	2022	[207]	Minerva Blog
Algorithm	AlphaTensor	DeepMind	2022	[208]	AlphaTensor

Generative AI Tech Stack



Generative AI Software and Business Factors

Business
Factors

Distribution

Proprietary Data

Domain Expertise

...

Application

A product utilizing and managing model inputs and outputs

Models

Large language models, image generation, or other ML models

Software

Data

Labeling, evaluation

MLOps Model management, tracking

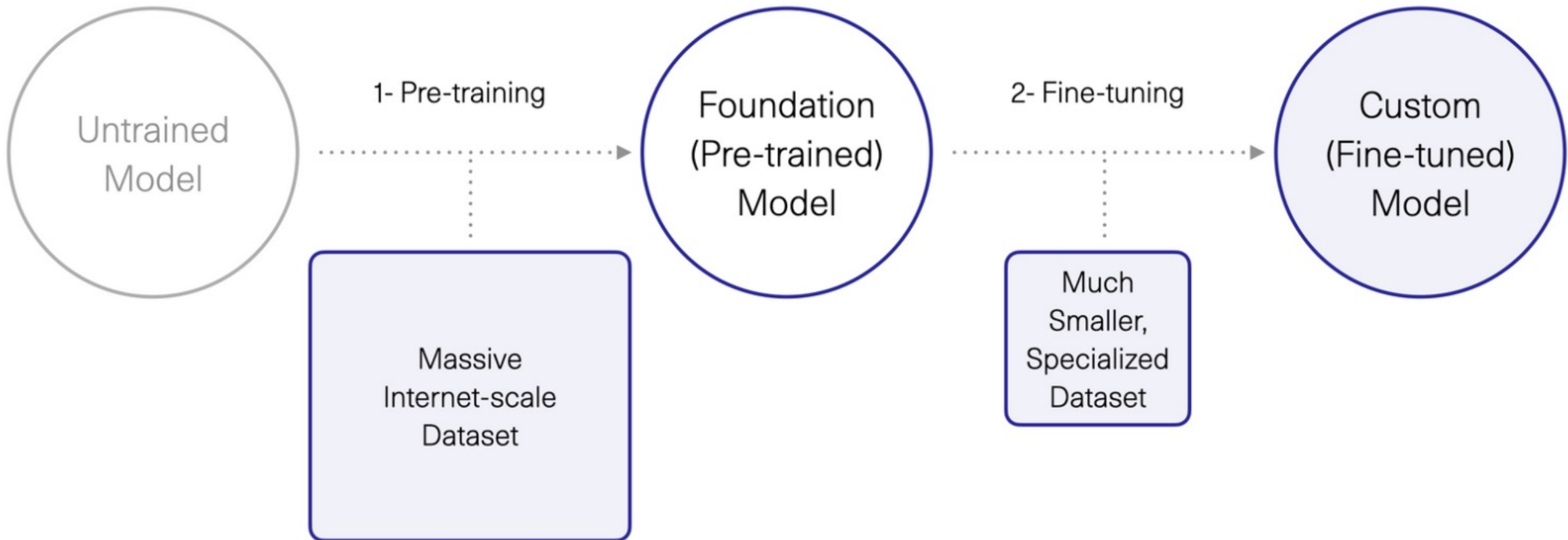
Cloud Platform

Hosting, compute, model deployment and monitoring

Generative AI

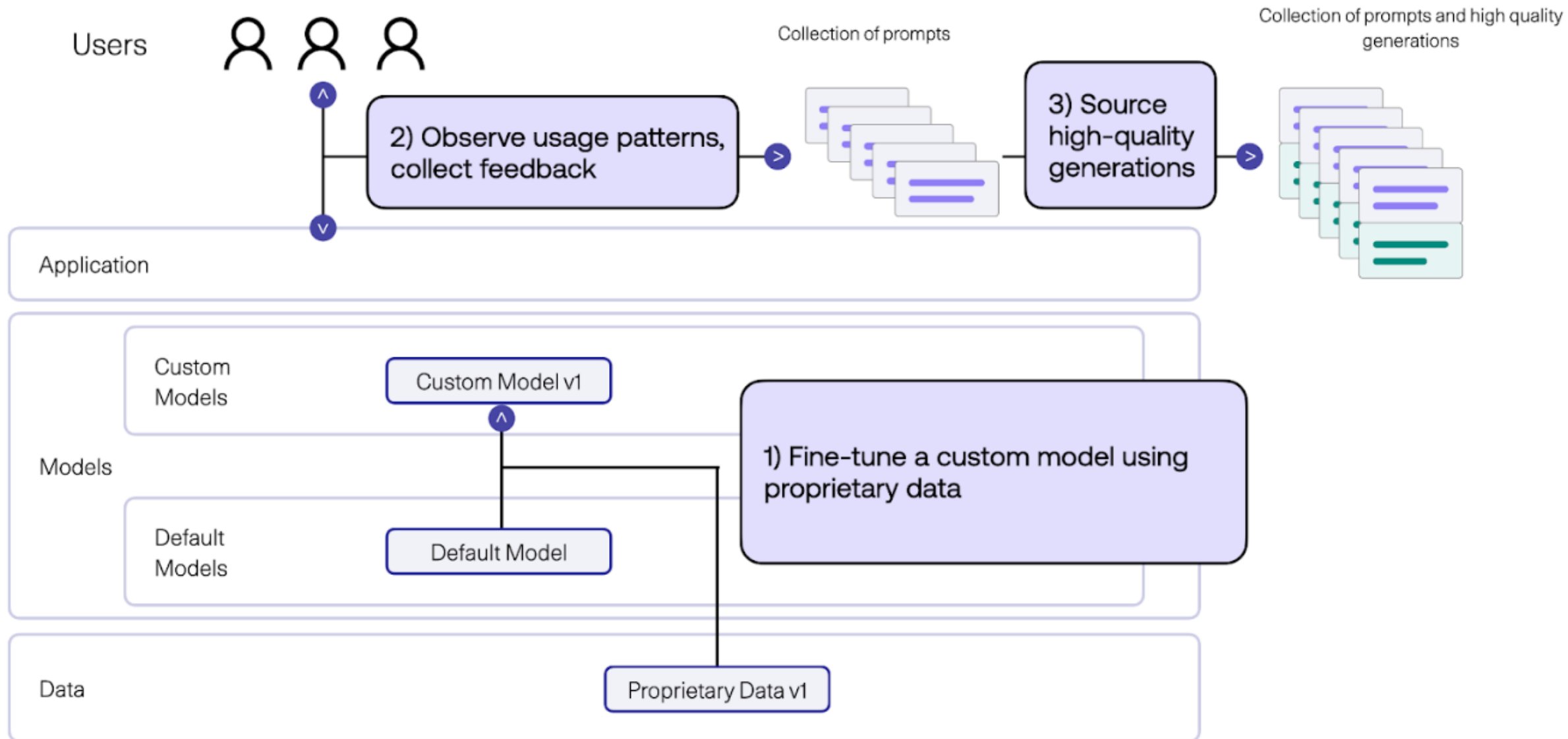
1. Pre-training Foundation (Pre-trained) Model

2. Fine-tuning Custom (Fine-tuned) Model



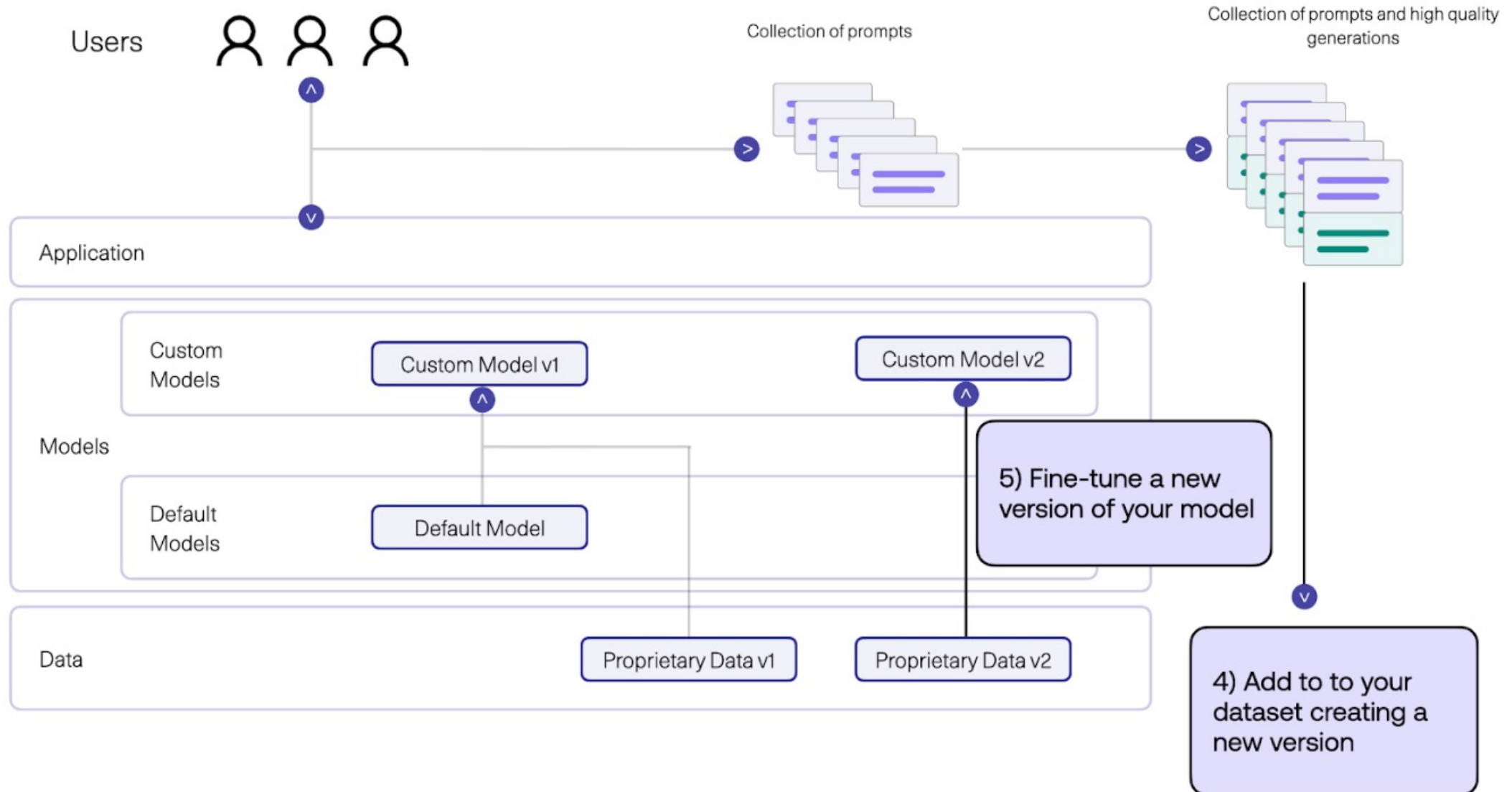
Generative AI

Fine-tune Custom Models using Proprietary Data



Generative AI

Fine-tune Custom Models using Proprietary Data



Evaluating the Code Quality of AI-Assisted Code Generation Tools: An Empirical Study on GitHub Copilot, Amazon CodeWhisperer, and ChatGPT

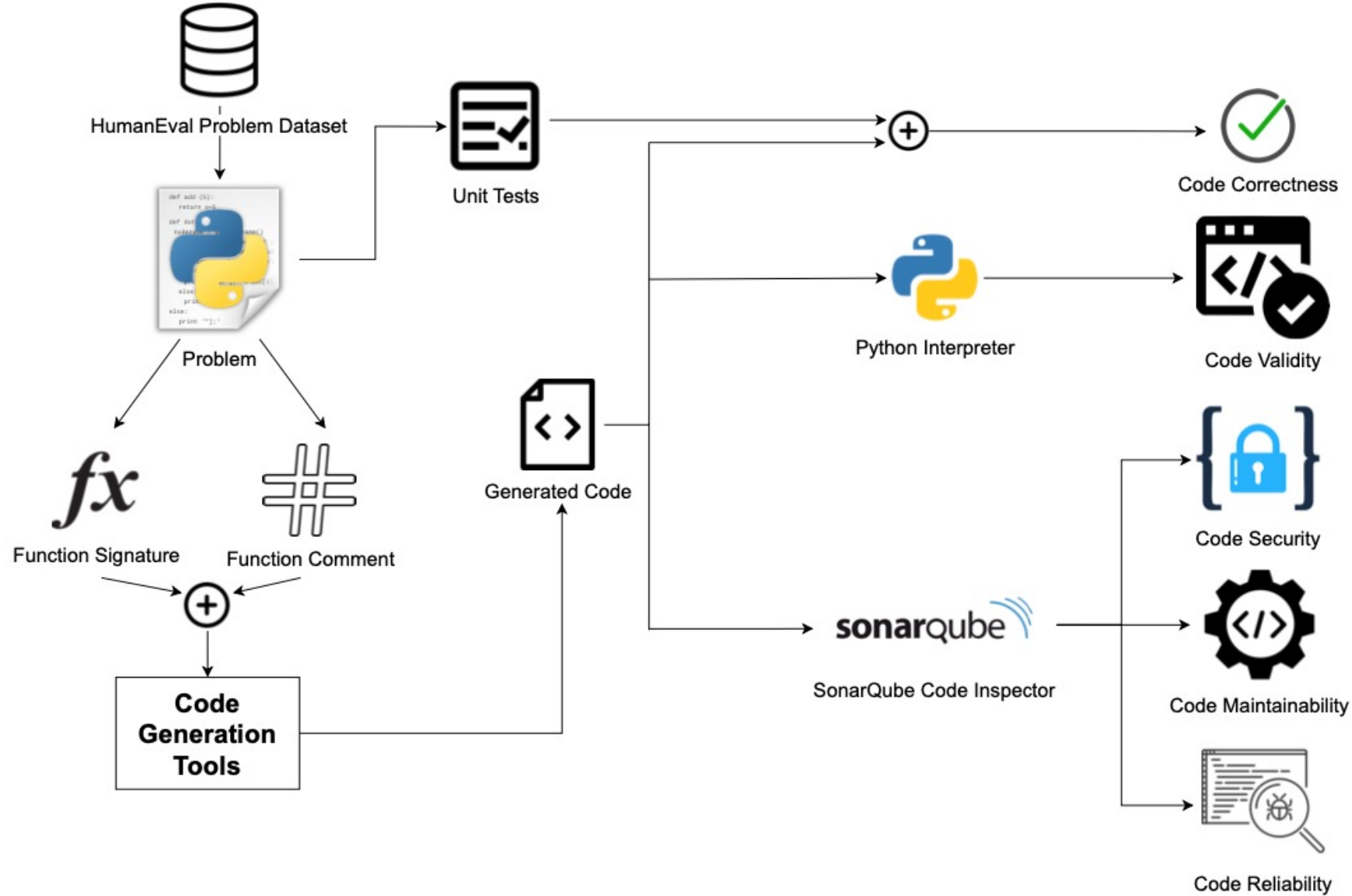
AI-Assisted Code Generation Tools: GitHub Copilot, Amazon CodeWhisperer, and ChatGPT

- RQ1 What is the **quality** of the code generated by the code generation tools?
 - RQ1.1 How **valid** are the code generation tools' code suggestions?
 - RQ1.2 How **correct** are code generation tools' code suggestions?
 - RQ1.3 How **secure** are code generation tools' code suggestions?
 - RQ1.4 How **reliable** are code generation tools' code suggestions?
 - RQ1.5 How **maintainable** are code generation tools' code suggestions?
- RQ2 What is the impact of using the docstrings on the generated code quality?
- RQ3 What is the impact of using meaningful function names on the generated code quality?
- RQ4 How did the code generation tools evolve over time?

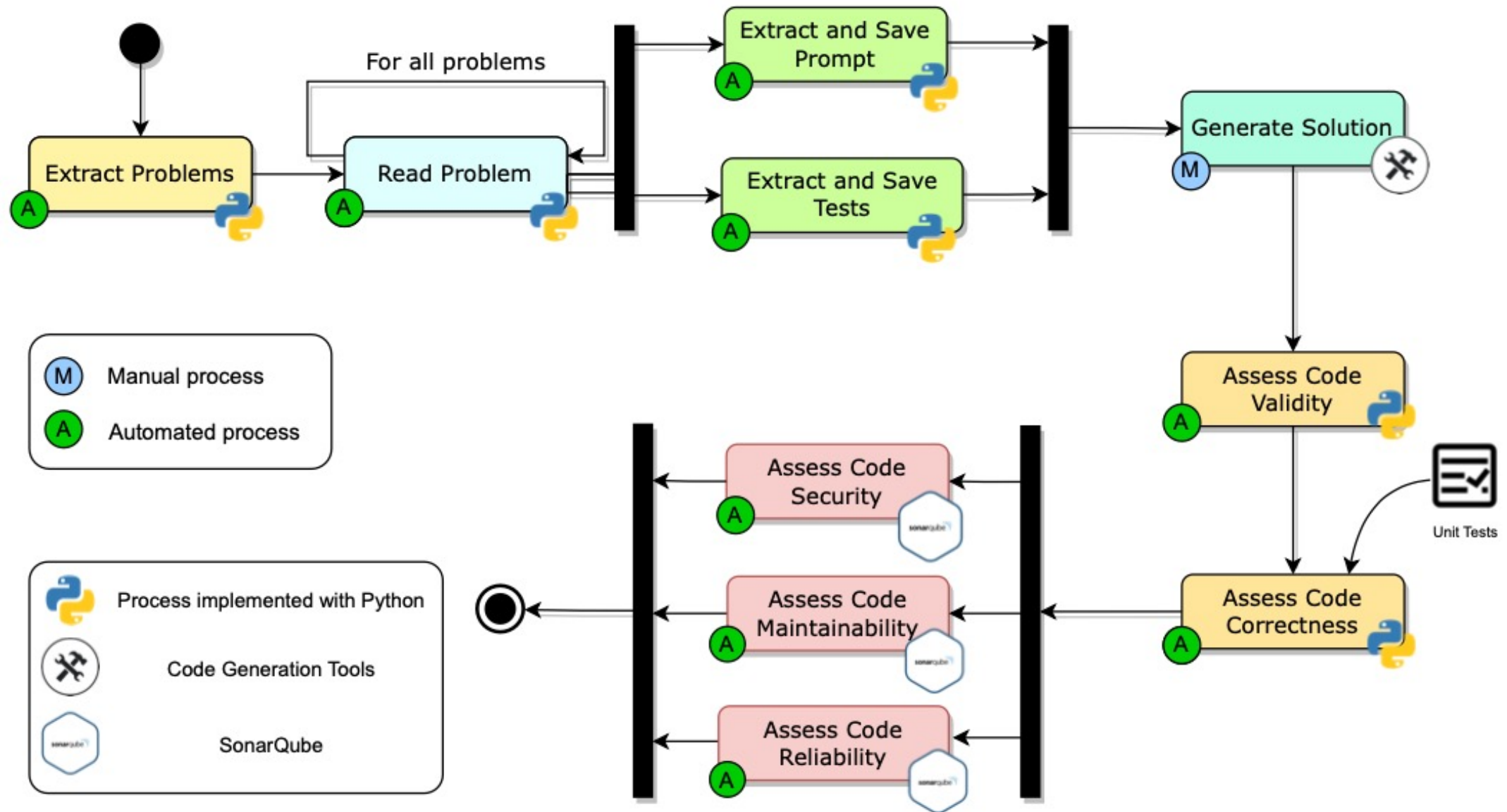
AI-Assisted Code Generation Tools: GitHub Copilot, Amazon CodeWhisperer, and ChatGPT

Features	ChatGPT	Amazon CodeWhisperer	GitHub Copilot
IDE Support	No IDE Support	JetBrains, Visual Studio Code, AWS Cloud9, or the AWS Lambda console	IntelliJ IDEA, Android Studio, AppCode, CLion, Code With Me Guest, DataGrip, DataSpell, GoLand, JetBrains Client, MPS, PhpStorm, PyCharm, Rider, RubyMine, WebStorm
First Release Time	Nov-30-2022	June-23-2022	Oct-29-2021
Developer	OpenAI	AWS	OpenAI-Microsoft
Providing References to Suggestions	NO	YES	NO
Explanation of Suggestions	YES	NO	NO
Providing Multiple Suggestions	NO (Theoretically user can manually ask for another suggestion.)	YES (Up to 5)	YES (Up to 10)
Training Data Source	GitHub Repositories, OpenAI Codex Dataset, other code repositories such as GitLab, Bitbucket, and SourceForge	"Vast amounts of publicly available code"	"...trained on all languages that appear in public repositories" (Fine-tuned)
Programming Languages work best with (according to the vendor)	N/A	C#, Java, JavaScript, Python, and TypeScript	C, C++, C#, Go, Java, JavaScript, PHP, Python, Ruby, Scala, and TypeScript
Multipurpose (other than programming)	YES	NO	NO
Subscription	ChatGPT Free ChatGPT Plus (\$20 per month)	Free Preview	Copilot for Students (Free) Copilot for Individuals (\$10 per month) Copilot for Business (\$19 per user, per month)
Can be Used Offline?	NO	NO	NO
Can it Access Local Files?	NO	YES	YES

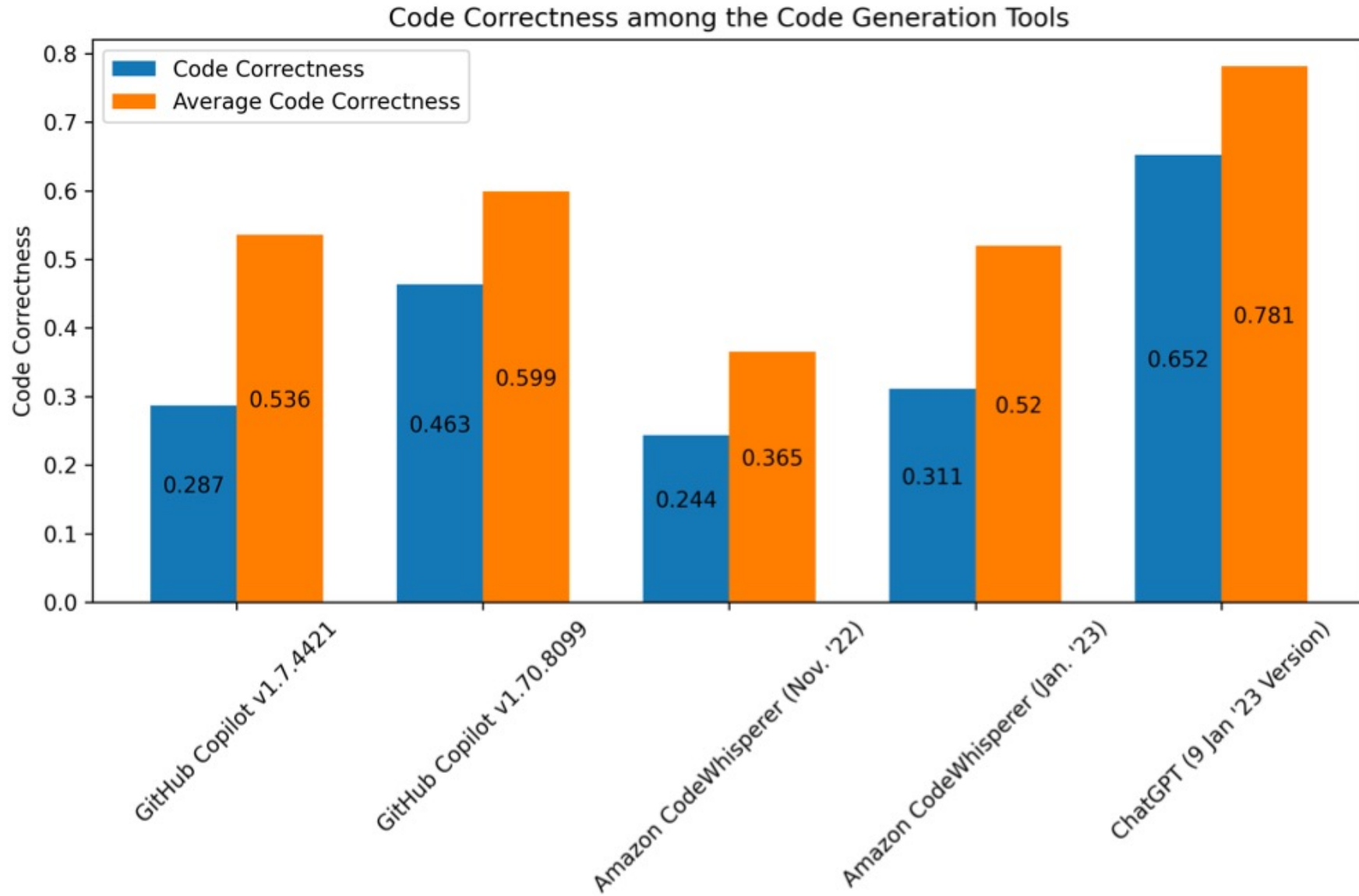
AI-Assisted Code Generation Tools: Experiment Setup



AI-Assisted Code Generation Tools: Experiment Workflow

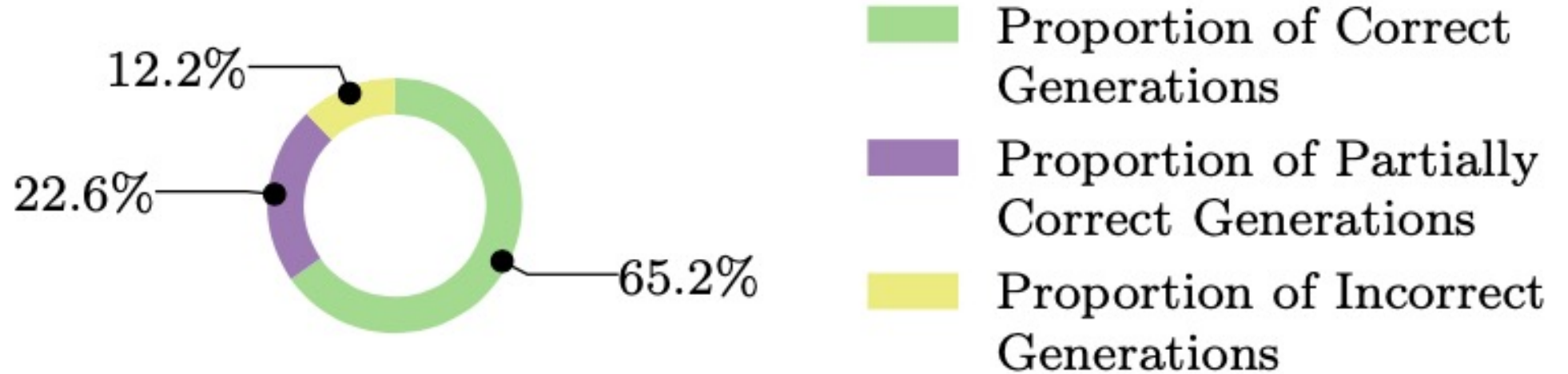


AI-Assisted Code Generation Tools: Code Correctness

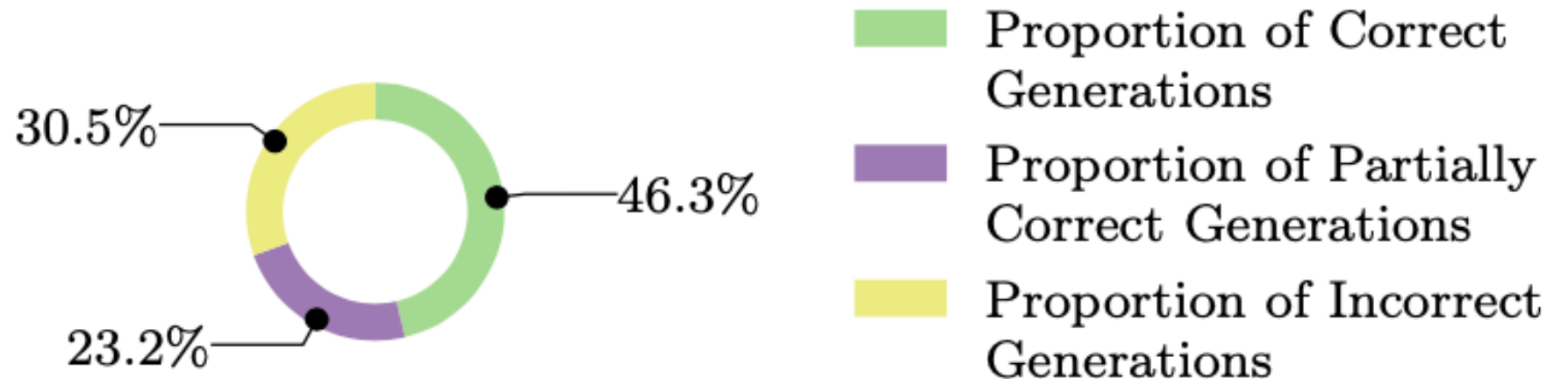


AI-Assisted Code Generation Tools: Code Correctness

ChatGPT



GitHub Copilot



Code Generation Models: BERT, T5, GPT-3, CodeBERT CodeGPT, CodeX

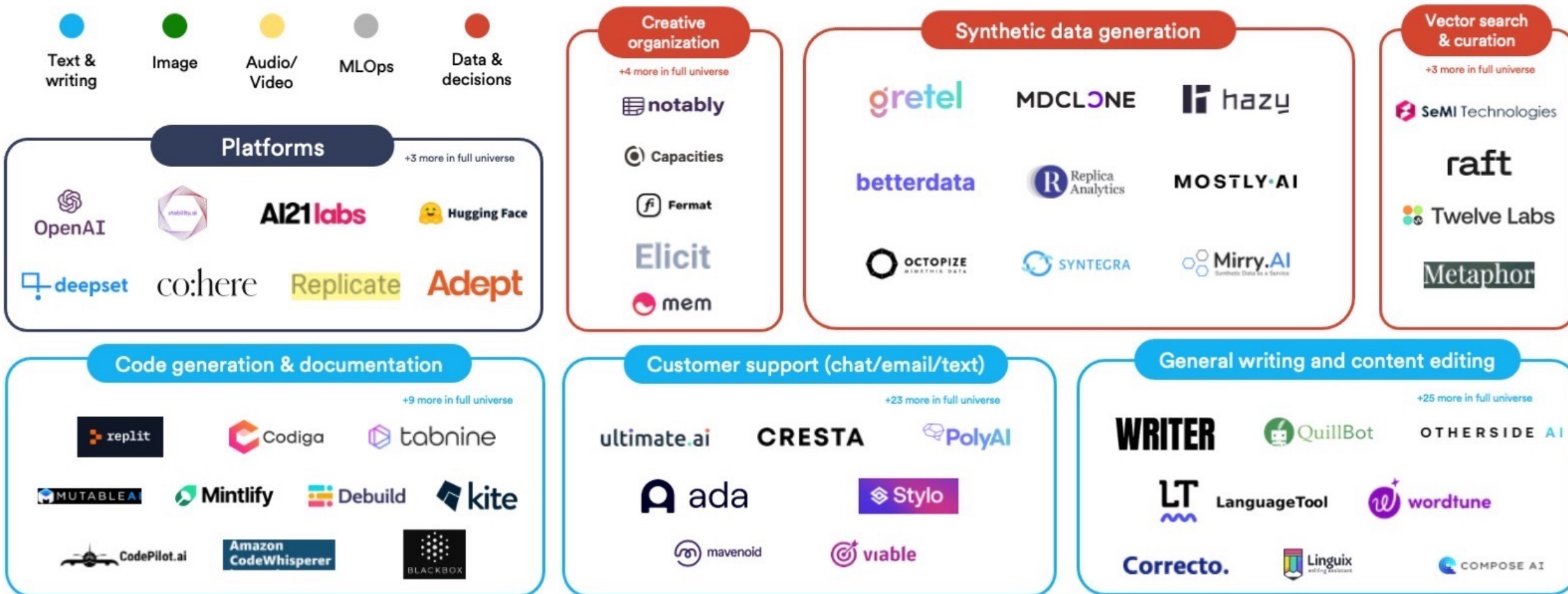
Pretrained Model	Backbone	Parameter	Pretrained Dataset	Size of Dataset	Trained Language
CuBERT	BERT	-	Python from Github	6.6M	Python
CodeBERT	BERT	125M	CodeSearchNet	3.5G	Ruby/JavaScript/Java/Python/GO/PHP/English
GraphCodeBERT	BERT	125M	CodeSearchNet	3.5G	Ruby/JavaScript/Java/Python/GO/PHP/English
CodeGPT	GPT-2	124M	Python and Java from CodeSearchNet	Python 1.1M Java 1.6M	Java/Python
CoTexT	T5	-	CodeSearchNet AND Java and Python from BigQuery	-	Java/Python
CodeT5	T5	60M/223M/770M	CodeSearchNet and C/C# datasets	8.35G	Ruby/JavaScript/GO/Python/Java/PHP/C/C#
PLBART	BART	140M	Java and Python from BigQuery AND SO posts	655G	Java/Python/English
CodeGen	-	350M/2.7B/6.1B/16.1B	THEPILE/BIGQUERY/BIGPYTHON	THEPILE 825G	C/C++/Go/Java/JavaScript/Python/English
InCoder	-	1.3B/6.7B	content from StackOverflow	159G	PYTHON and 28 other languages
CodeX	GPT-3	300M/2.5B/12B	Python from GitHub	159G	Python
AlphaCode	-	300M/1B/3B/9B/41B	a snapshot of github	715.1G	C++/C#/Java/JavaScript/Lua/PHP/Python/Ruby/Go/Rust/Scala/TypeScript
PanGu-Coder	PanGu-alpha	317M/2.6B	Python from GitHub	147G	Python
CodeGeeX	-	13B	open-sourced code datasets, The Pile and CodeParrot	-	C++/Python/C/Java/JavaScript/Go/HTML/PHP/Shell/CSS/Others
aiXcoder L	GPT-2	1.3B	Java from GitHub	-	Java
aiXcoder XL	-	13B	Open-sourced code from GitHub	-	Java

Generative AI

BASE10 TREND MAP: GENERATIVE AI

Companies are grouped based on medium produced and segmented by use case within each medium. Companies that offer products across segments are grouped in the segment of the core product offering.

Base¹⁰



Generative AI

Marketing & sales copy



Text & data summarization

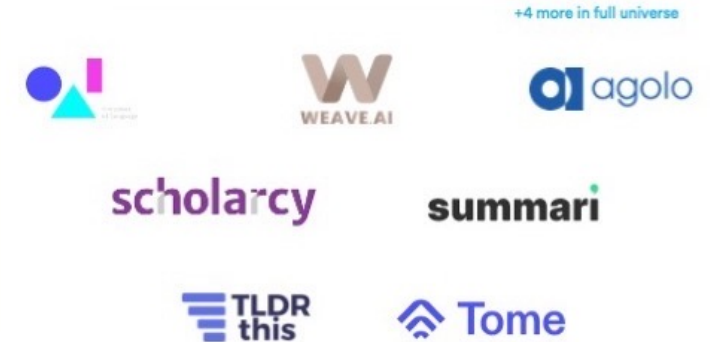


Image editing



Ad collateral



Design



Text to image



Generative AI

MLOps

+11 more in full universe



Humanloop, MosaicML, Weights & Biases, Comet, slai, Outerbounds

Audio Editing

+2 more in full universe



Amberscript, Podcastle, Krisp, Aflorithmic

Text to speech / Speech to text

+14 more in full universe



Coqui, Speechify, WellSaid, Murf.ai, Replica, LOVO, Resemble.ai

Music generation

+9 more in full universe



Moises, Boomy, Supertone, USICO, DAACI, Splash

Summarization & insights

+8 more in full universe



Fathom, Xembly, Hume, Clearword, Sembly AI, Sybill, Magnifi, VideoVerse, Hume

Avatars & dubbing

+15 more in full universe

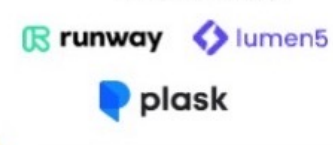


Hour One, Synthesia, Metaphysic, Deepdub.ai, PaperCup, Wombo, Rephrase.ai, Elai, Pictory, Colossyan, D-ID, Respeecher

Video

+1 more in full universe


Generation and Animation



Runway, Lumen5, Plask

Editing

+6 more in full universe



Descript, Invideo

DALL·E 2

Create original, realistic images and art from a text description.
It can combine concepts, attributes, and styles.

TEXT DESCRIPTION

An astronaut riding a horse
in a photorealistic style

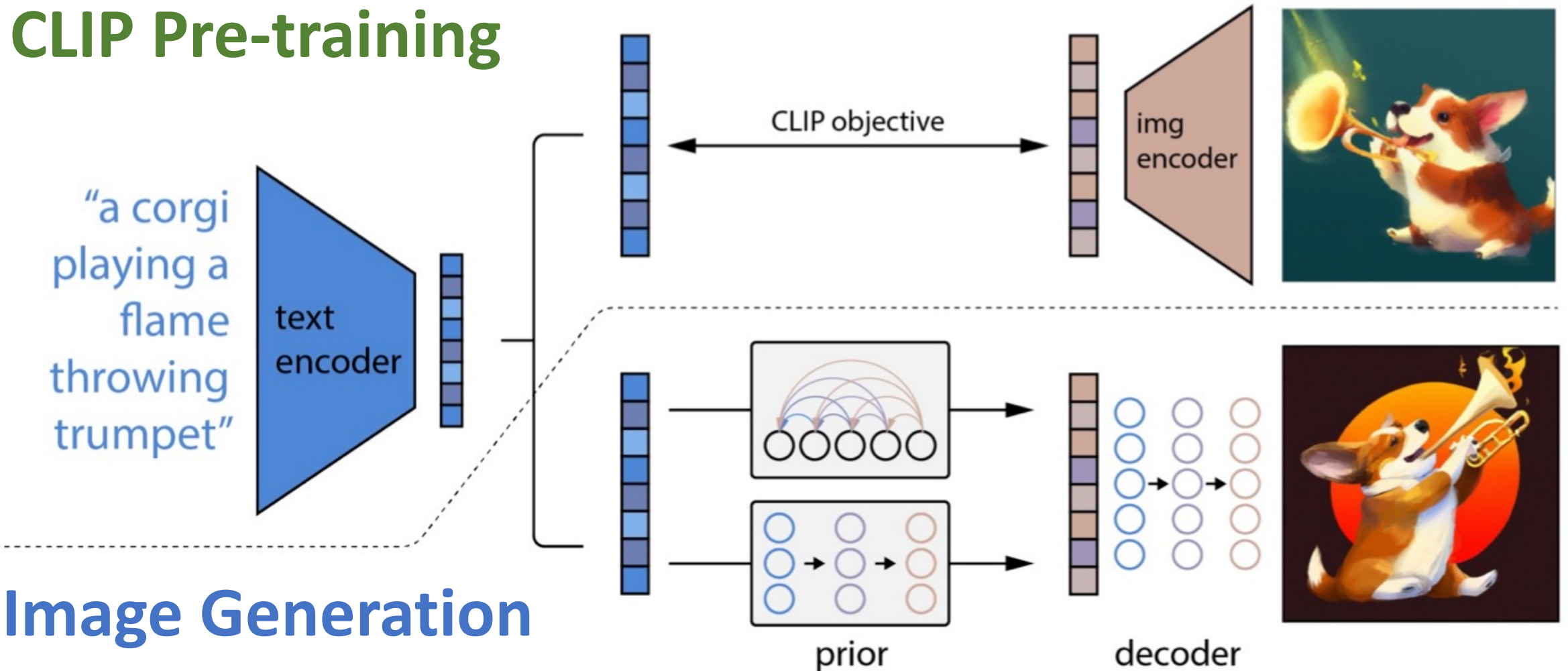
DALL·E 2



<https://openai.com/dall-e-2/>

The Model Structure of DALL-E-2

CLIP Pre-training



Stable Diffusion

📦 Stable Diffusion Demo

Stable Diffusion is a state of the art text-to-image model that generates images from text.
For faster generation and forthcoming API access you can try [DreamStudio Beta](#)

an insect robot preparing a delicious meal

Generate image



<https://huggingface.co/spaces/stabilityai/stable-diffusion>

Stable Diffusion Colab

woctezuma / [stable-diffusion-colab](#) Public

Notifications Fork 7 Star 31

Code Issues Pull requests Actions Projects Wiki Security Insights

main 1 branch 0 tags

Go to file Code

About

Colab notebook to run Stable Diffusion.

[github.com/CompVis/stable-diffusion](#)





- deep-learning
- colab
- image-generation
- text-to-image
- diffusion
- text2image
- colaboratory
- google-colab
- colab-notebook
- google-colaboratory
- google-colab-notebook
- text-to-image-synthesis
- huggingface
- diffusion-models
- text-to-image-generation
- latent-diffusion
- stable-diffusion
- huggingface-diffusers
- diffusers
- stable-diffusion-diffusers

Readme

MIT license

31 stars

2 watching


 woctezuma	README: add a reference for sampler schedules	37bc02d 24 days ago	🕒 18 commits
 LICENSE	Initial commit		27 days ago
 README.md	README: add a reference for sampler schedules		24 days ago
 stable_diffusion.ipynb	Allow to choose the scheduler		25 days ago

README.md

Stable-Diffusion-Colab

The goal of this repository is to provide a Colab notebook to run the text-to-image "Stable Diffusion" model [1].

Usage

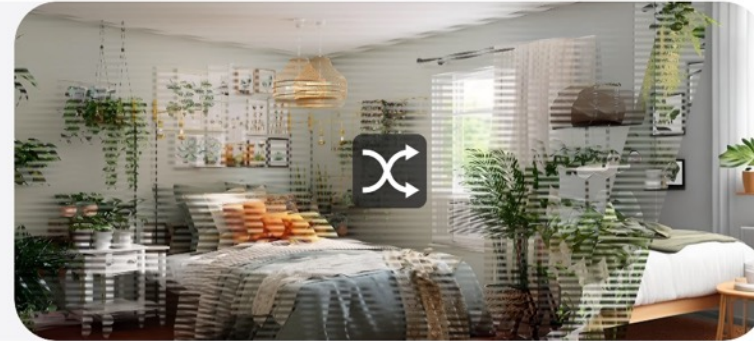
- Run `stable_diffusion.ipynb` . 

<https://github.com/woctezuma/stable-diffusion-colab>

Stable Diffusion Reimagine

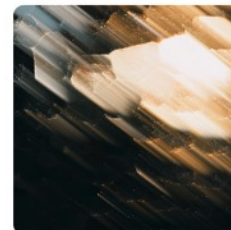
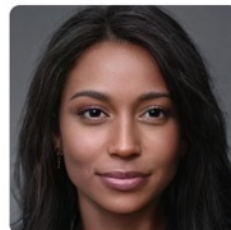
Stable diffusion reimagine

Create multiple variations from a single image.



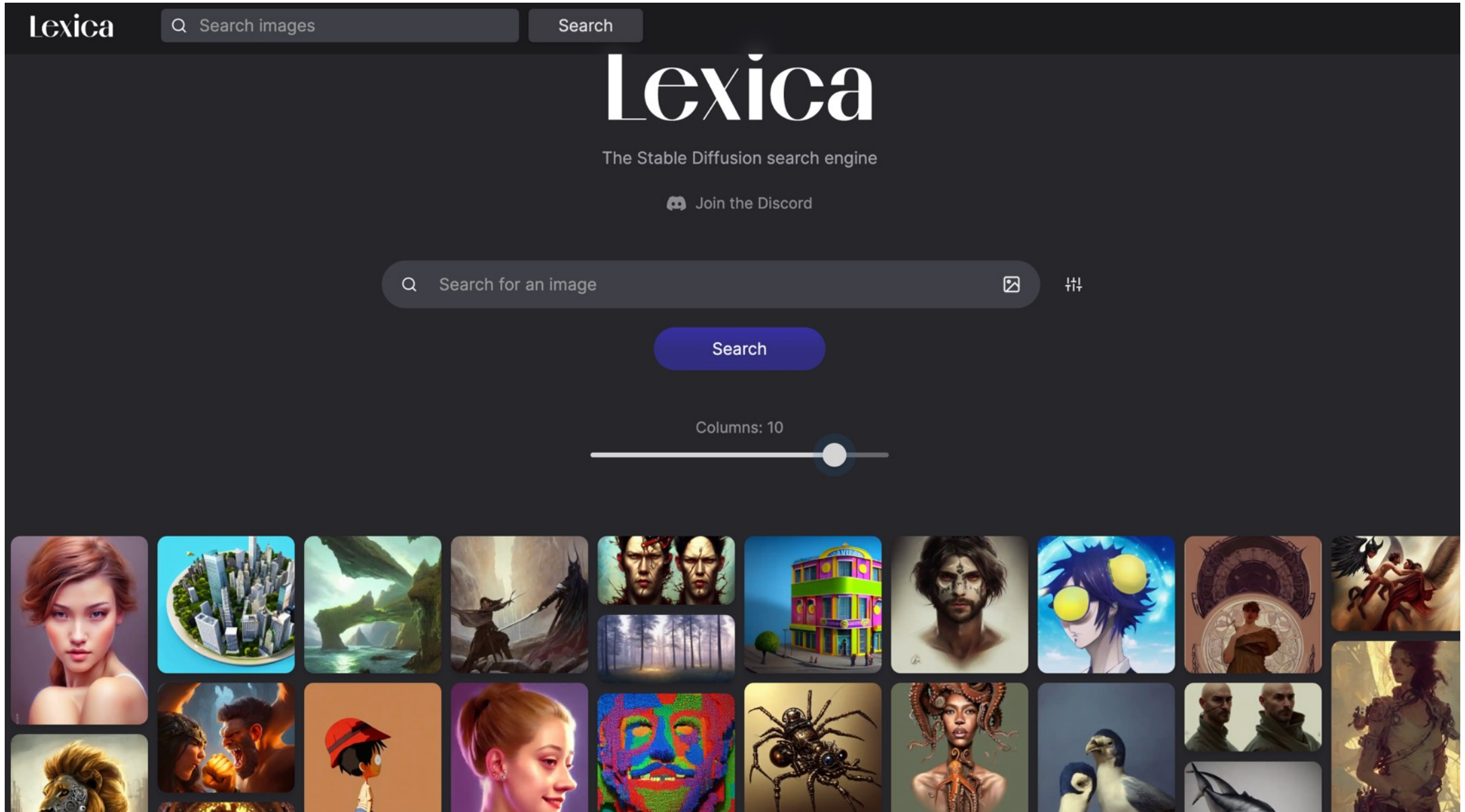
Click, paste, or drop a file here to start.

↓ Or click on an example below



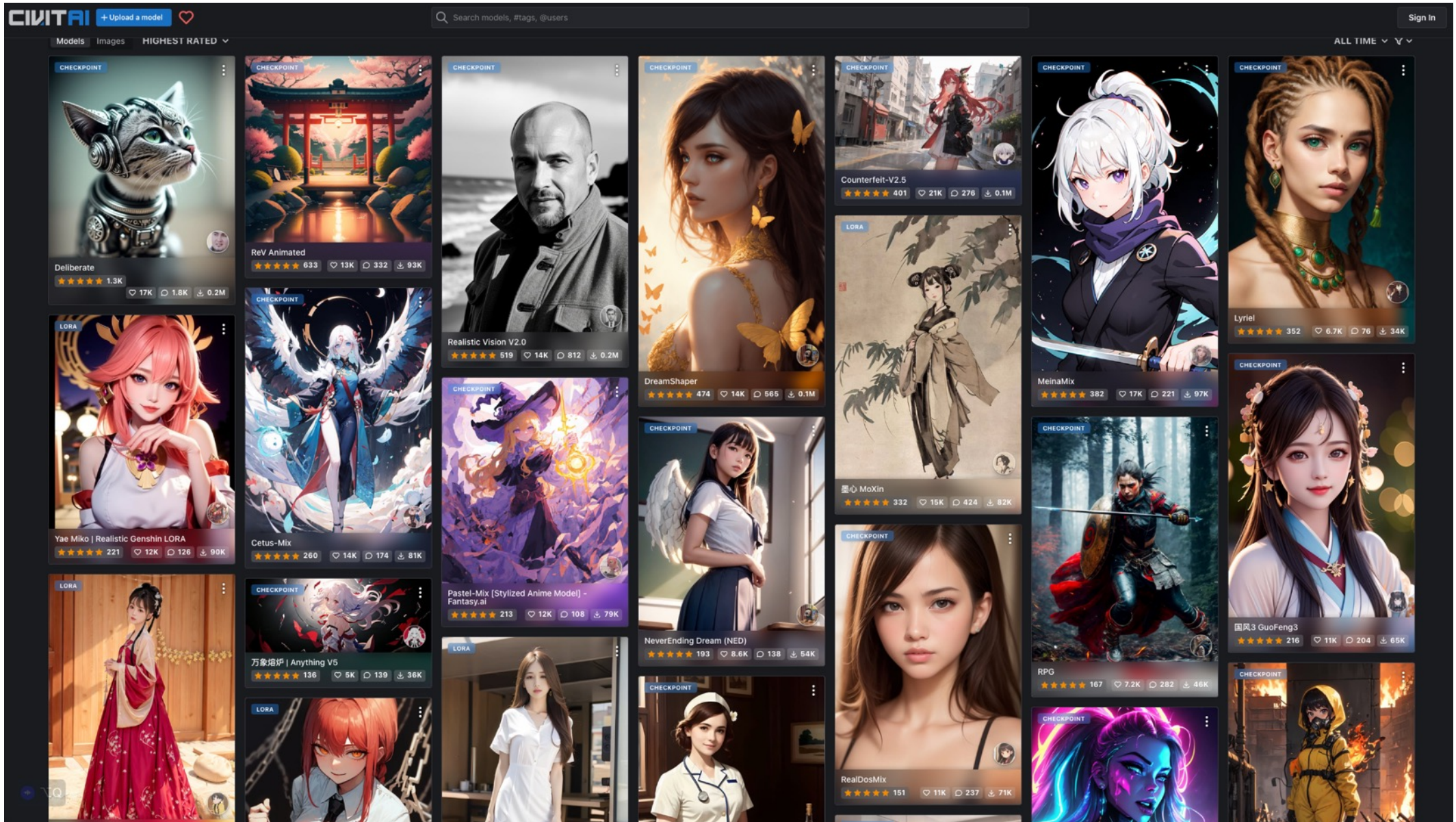
<https://clipdrop.co/stable-diffusion-reimagine>

Lexica Art: Search Stable Diffusion images and prompts



<https://lexica.art/>

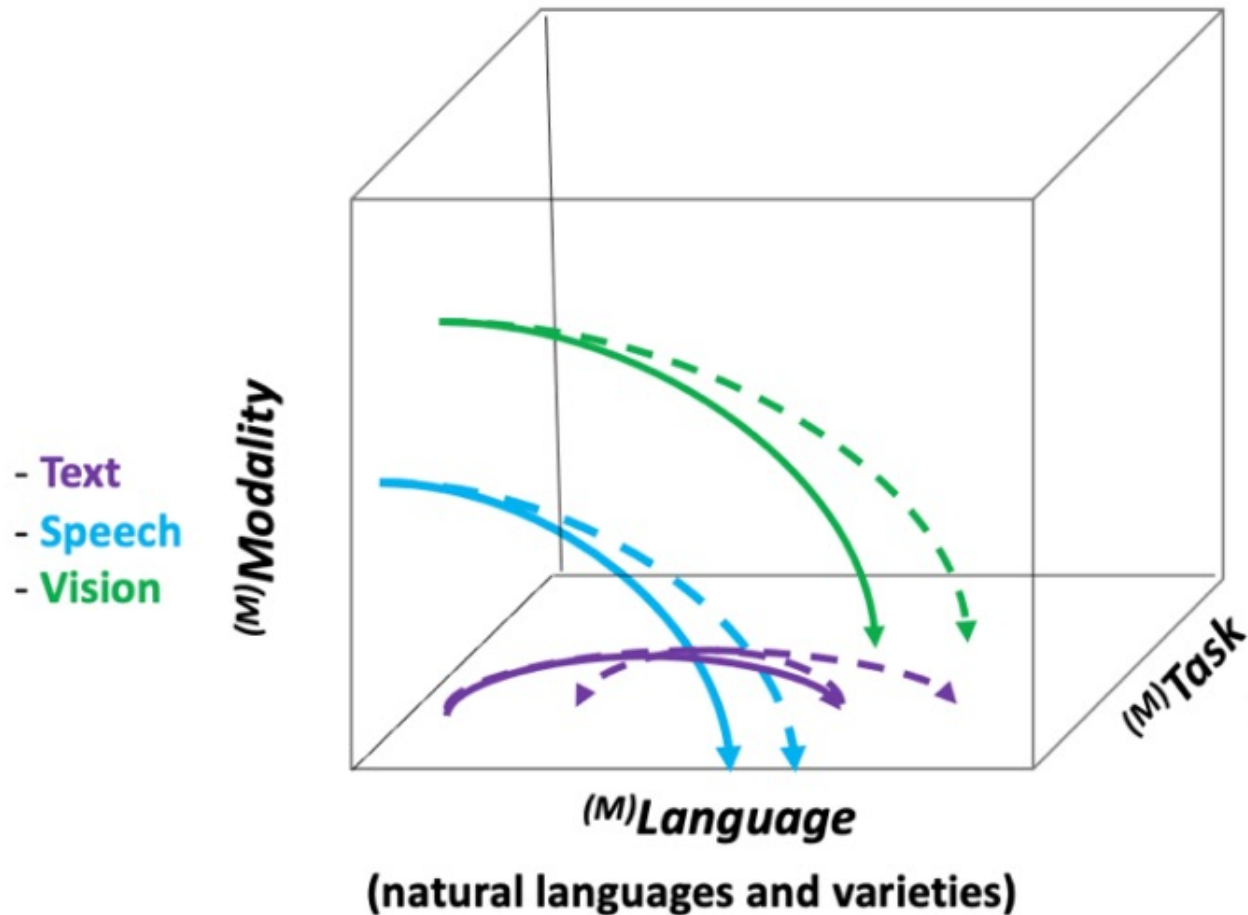
Civitai: Stable Diffusion AI Art Models



<https://civitai.com/>

NLG from a Multilingual, Multimodal and Multi-task perspective

Multi³(Natural Language) Generation

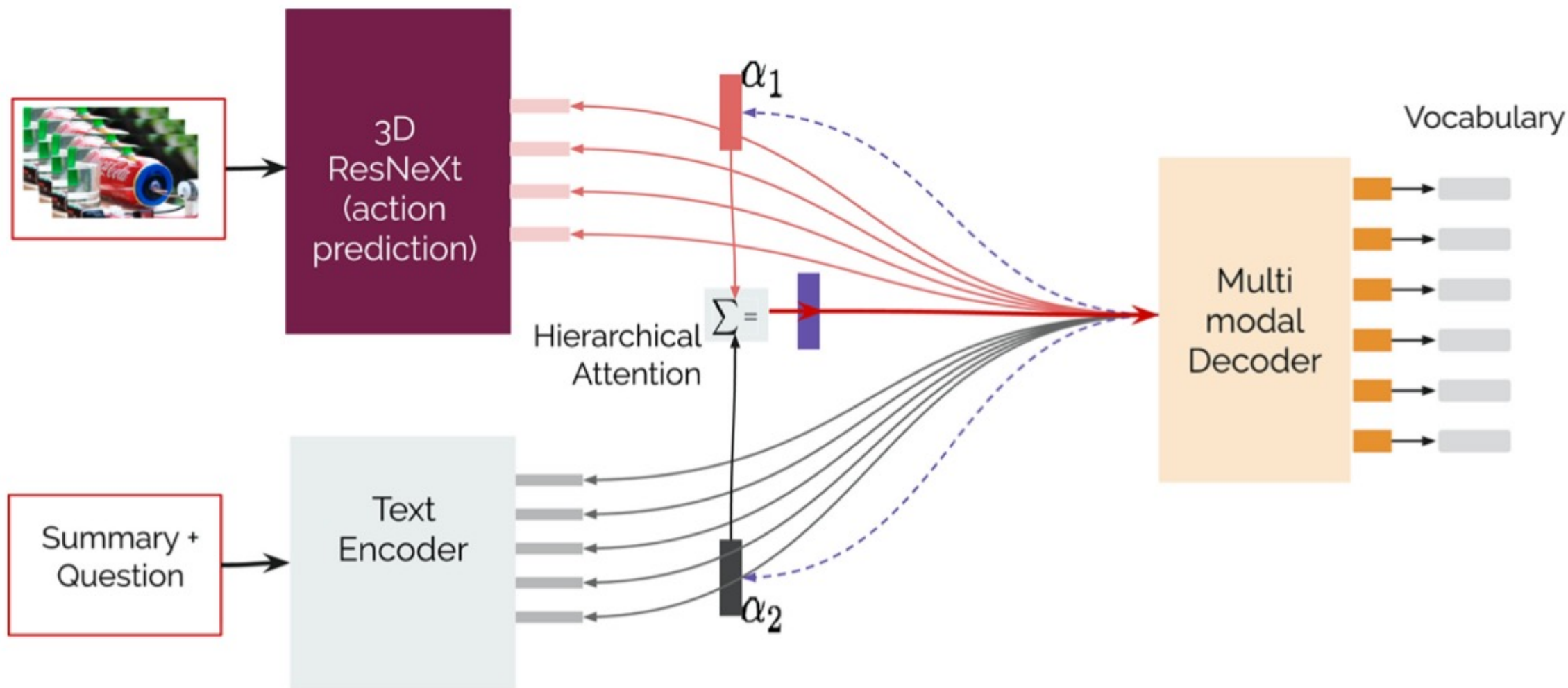


- Recognize and transcribe speech (ASR)
- Translate from one language to another (MT)
- Describe, ask or answer questions or converse about visual objects (Captioning, VQA, Visual Dialogue, ...)

Source: Erdem, Erkut, Menekse Kuyu, Semih Yagcioglu, Anette Frank, Letitia Parcalabescu, Barbara Plank, Andrii Babii et al.

"Neural Natural Language Generation: A Survey on Multilinguality, Multimodality, Controllability and Learning." Journal of Artificial Intelligence Research 73 (2022): 1131-1207.

Text-and-Video Dialog Generation Models with Hierarchical Attention



Source: Erdem, Erkut, Menekse Kuyu, Semih Yagcioglu, Anette Frank, Letitia Parcalabescu, Barbara Plank, Andrii Babii et al.

"Neural Natural Language Generation: A Survey on Multilinguality, Multimodality, Controllability and Learning." Journal of Artificial Intelligence Research 73 (2022): 1131-1207.

Multimodal Few-Shot Learning with Frozen Language Models

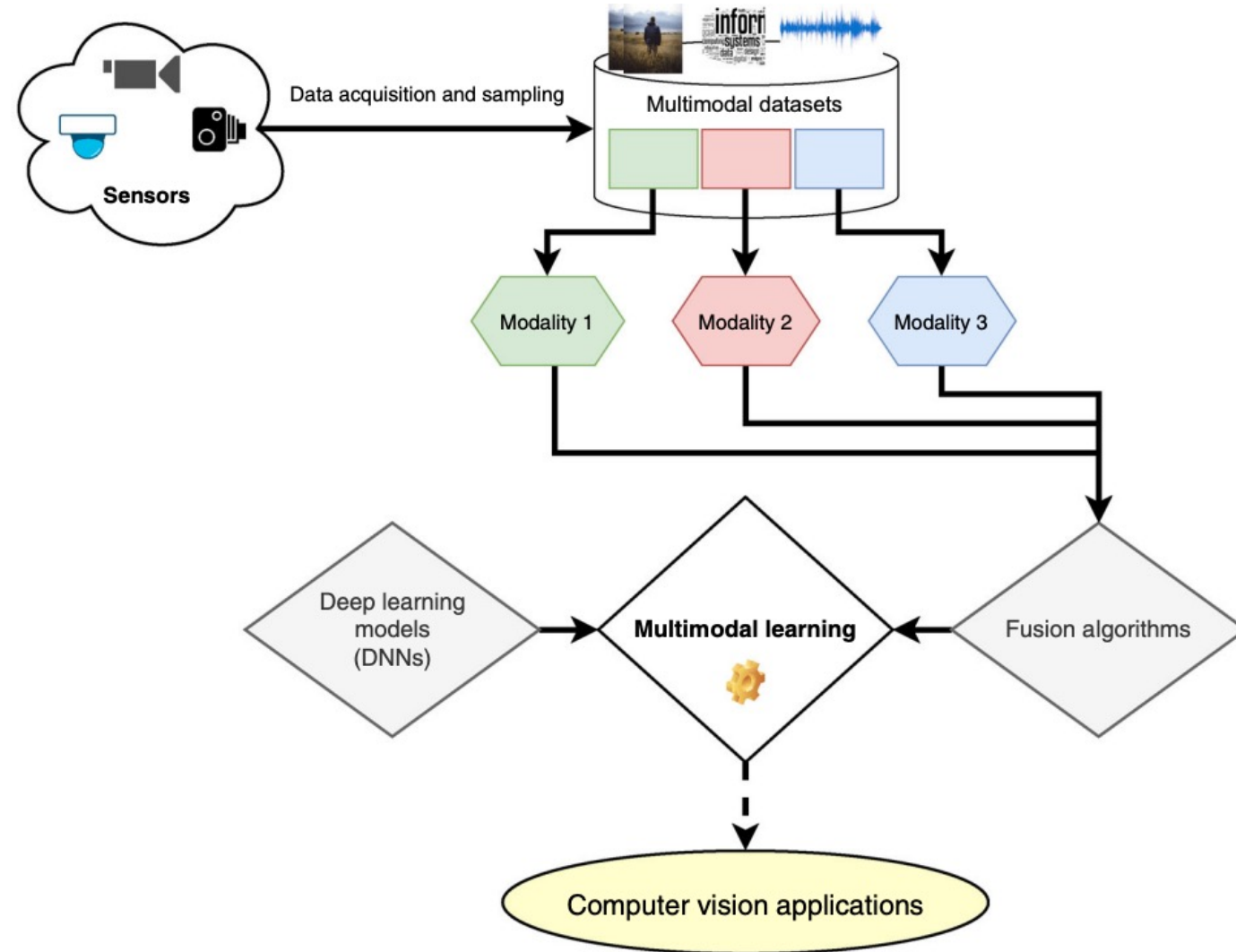


Curated samples with about five seeds required to get past well-known language model failure modes of either repeating text for the prompt or emitting text that does not pertain to the image.

These samples demonstrate the ability to generate open-ended outputs that adapt to both images and text, and to make use of facts that it has learned during language-only pre-training.

Multimodal Pipeline

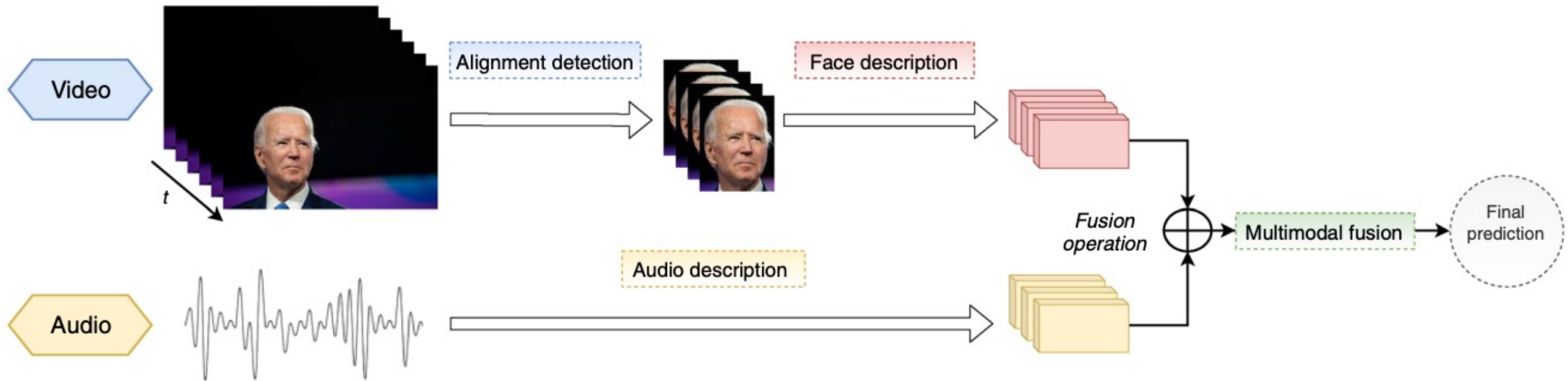
that includes three different modalities (Image, Text, Audio)



Source: Bayoudh, Khaled, Raja Knani, Fayçal Hamdaoui, and Abdellatif Mtibaa (2022).

"A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets." The Visual Computer 38, no. 8: 2939-2970.

Video and Audio Multimodal Fusion



Source: Bayoudh, Khaled, Raja Knani, Fayçal Hamdaoui, and Abdellatif Mtibaa (2022).

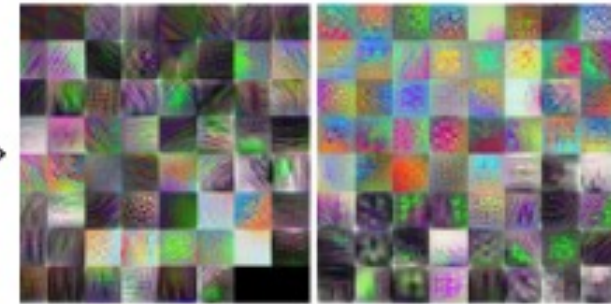
"A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets." The Visual Computer 38, no. 8: 2939-2970.

Visual and Textual Representation

Image



Visual representations (Dense)



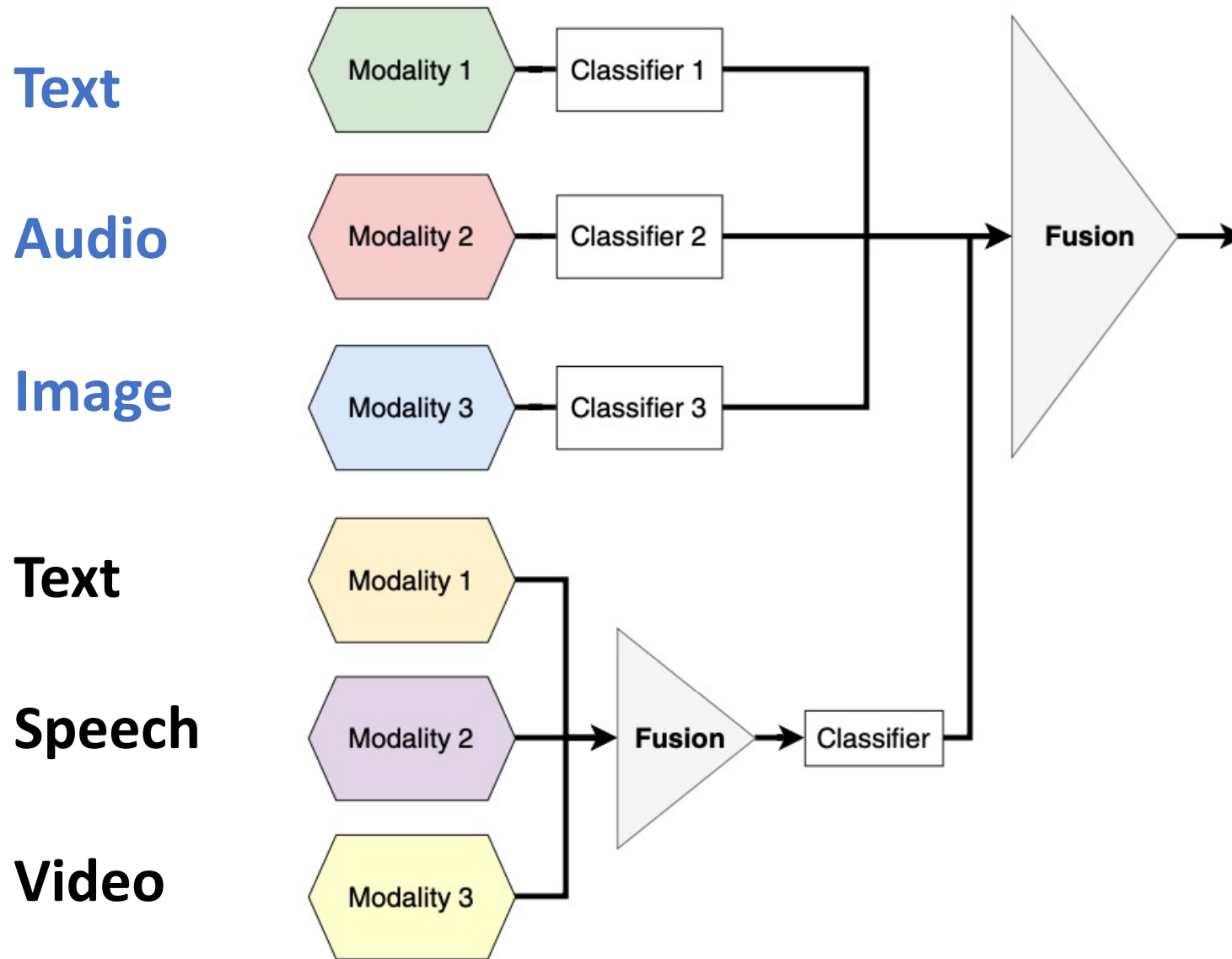
Text

This is the oldest and most important defensive work to have been built along the North African coastline by the Arab conquerors in the early days of Islam. Founded in 796, this building underwent several modifications during the medieval period. Initially, it formed a quadrilateral and then was composed of four buildings giving onto two inner courtyards.

Textual representations (Sparse)



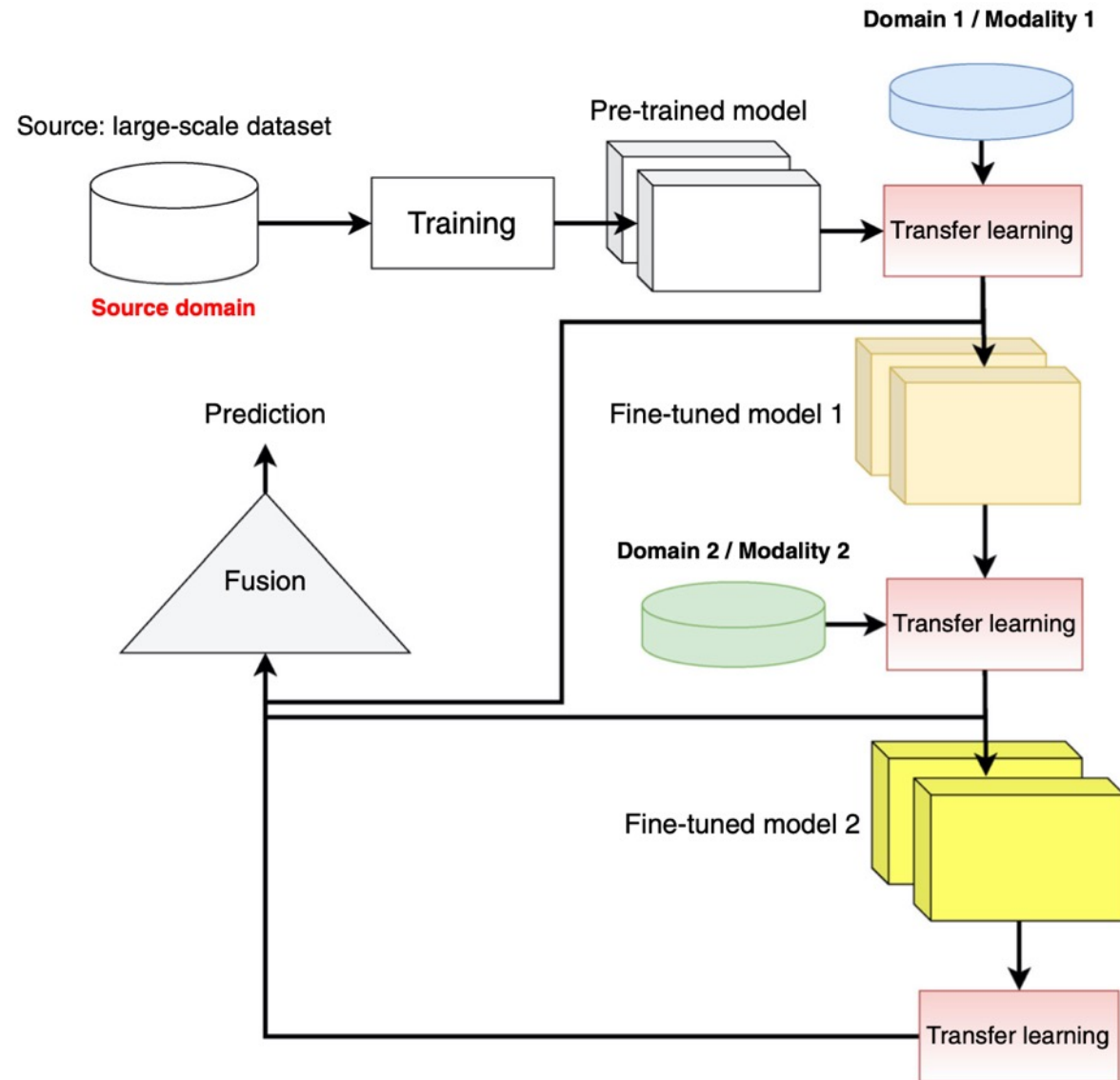
Hybrid Multimodal Data Fusion



Source: Bayouadh, Khaled, Raja Knani, Fayçal Hamdaoui, and Abdellatif Mtibaa (2022).

"A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets." The Visual Computer 38, no. 8: 2939-2970.

Multimodal Transfer Learning

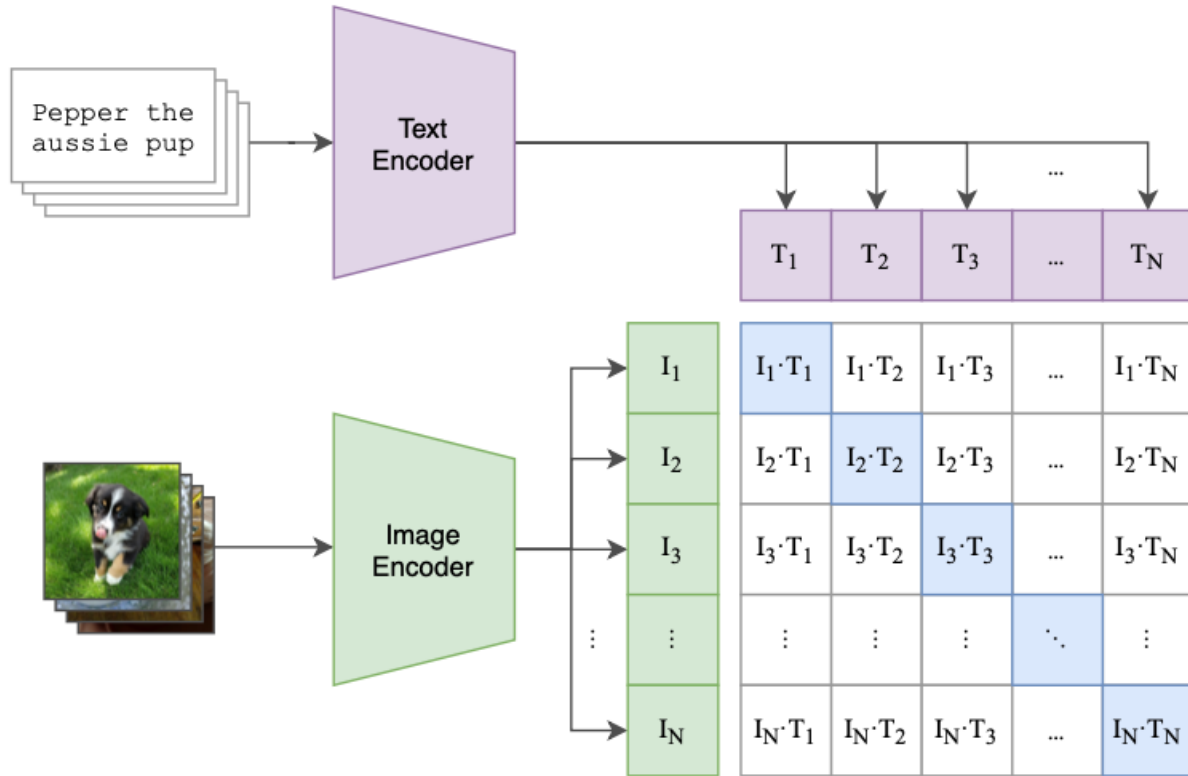


Source: Bayouadh, Khaled, Raja Knani, Fayçal Hamdaoui, and Abdellatif Mtibaa (2022).

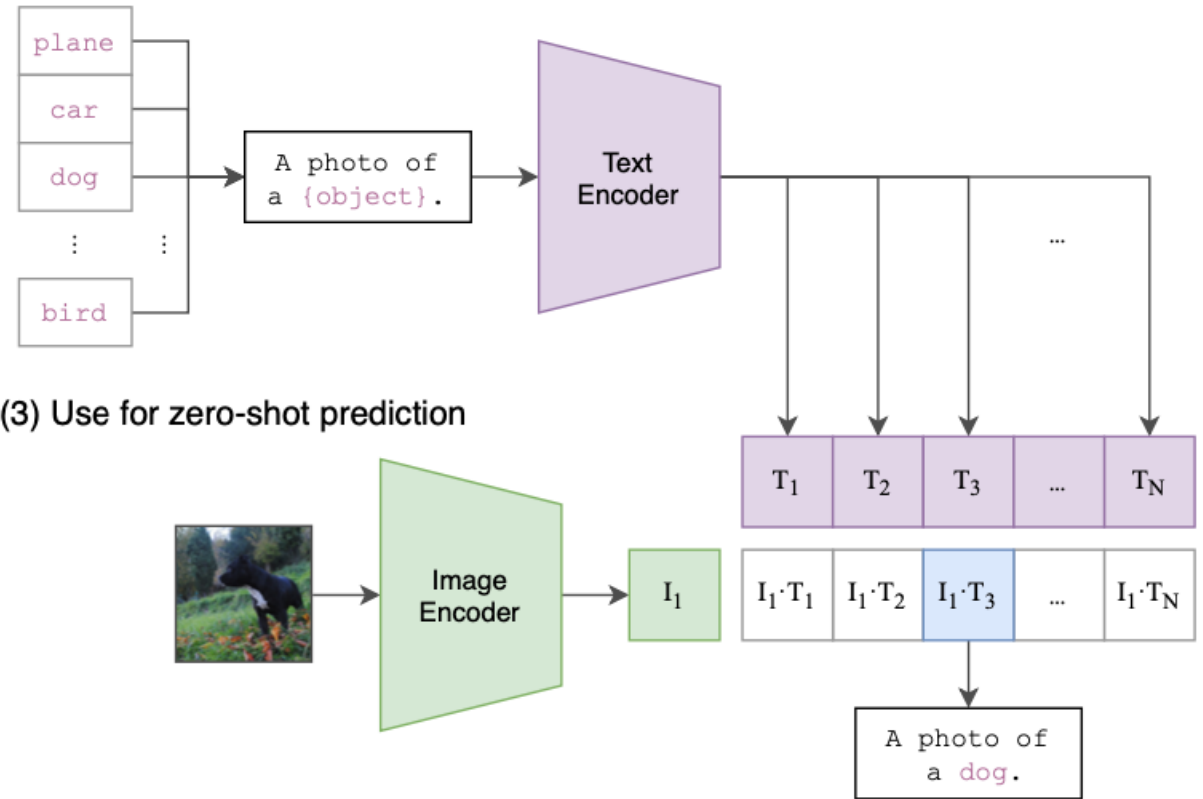
"A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets." *The Visual Computer* 38, no. 8: 2939-2970.

CLIP: Learning Transferable Visual Models From Natural Language Supervision

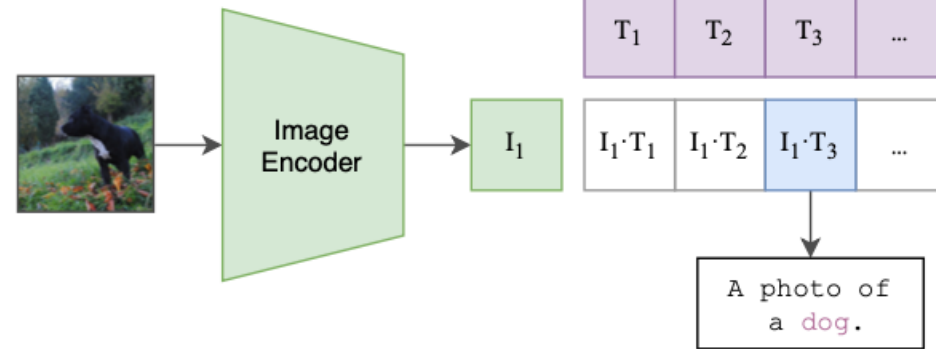
(1) Contrastive pre-training



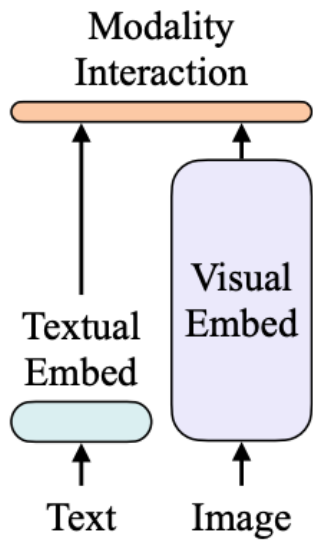
(2) Create dataset classifier from label text



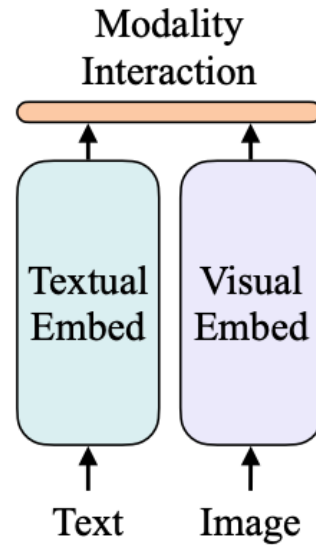
(3) Use for zero-shot prediction



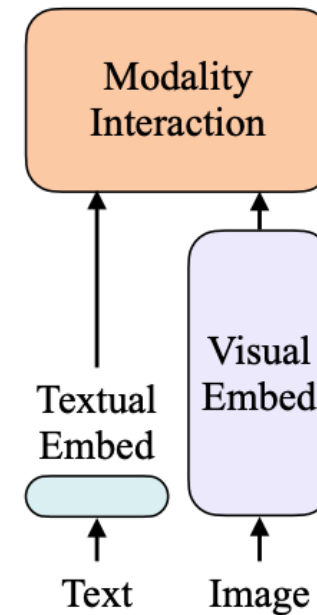
ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision



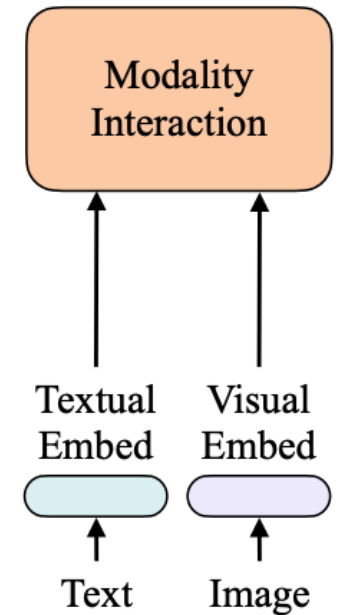
(a) $VE > TE > MI$



(b) $VE = TE > MI$



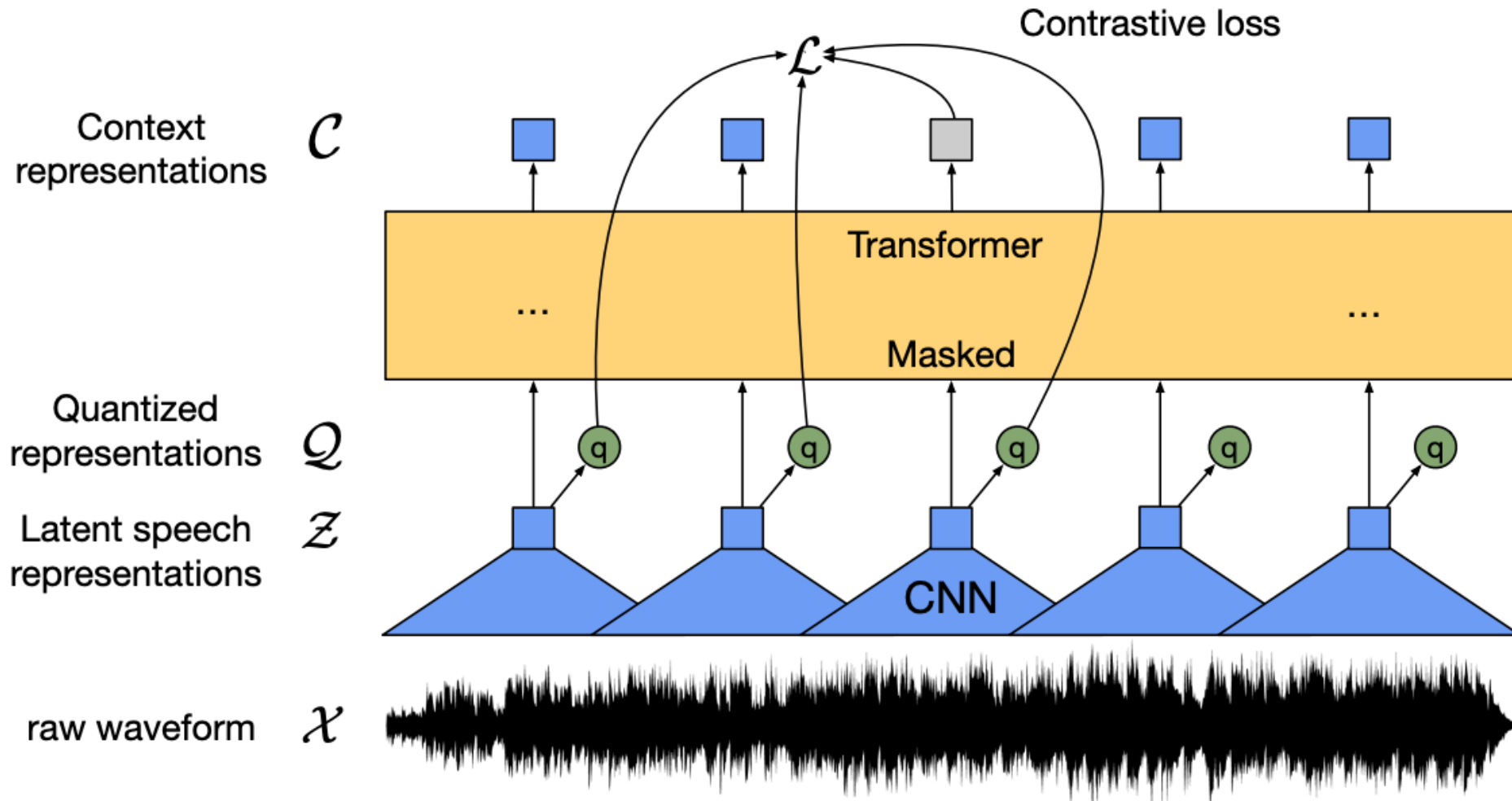
(c) $VE > MI > TE$



(d) $MI > VE = TE$

wav2vec 2.0:

A framework for self-supervised learning of speech representations

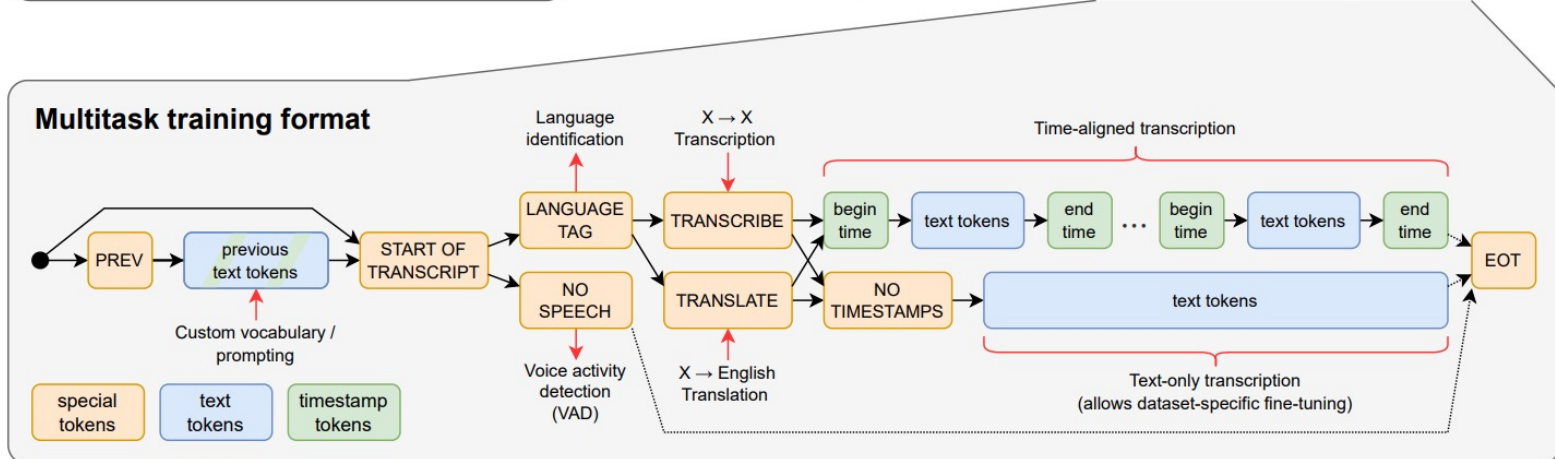
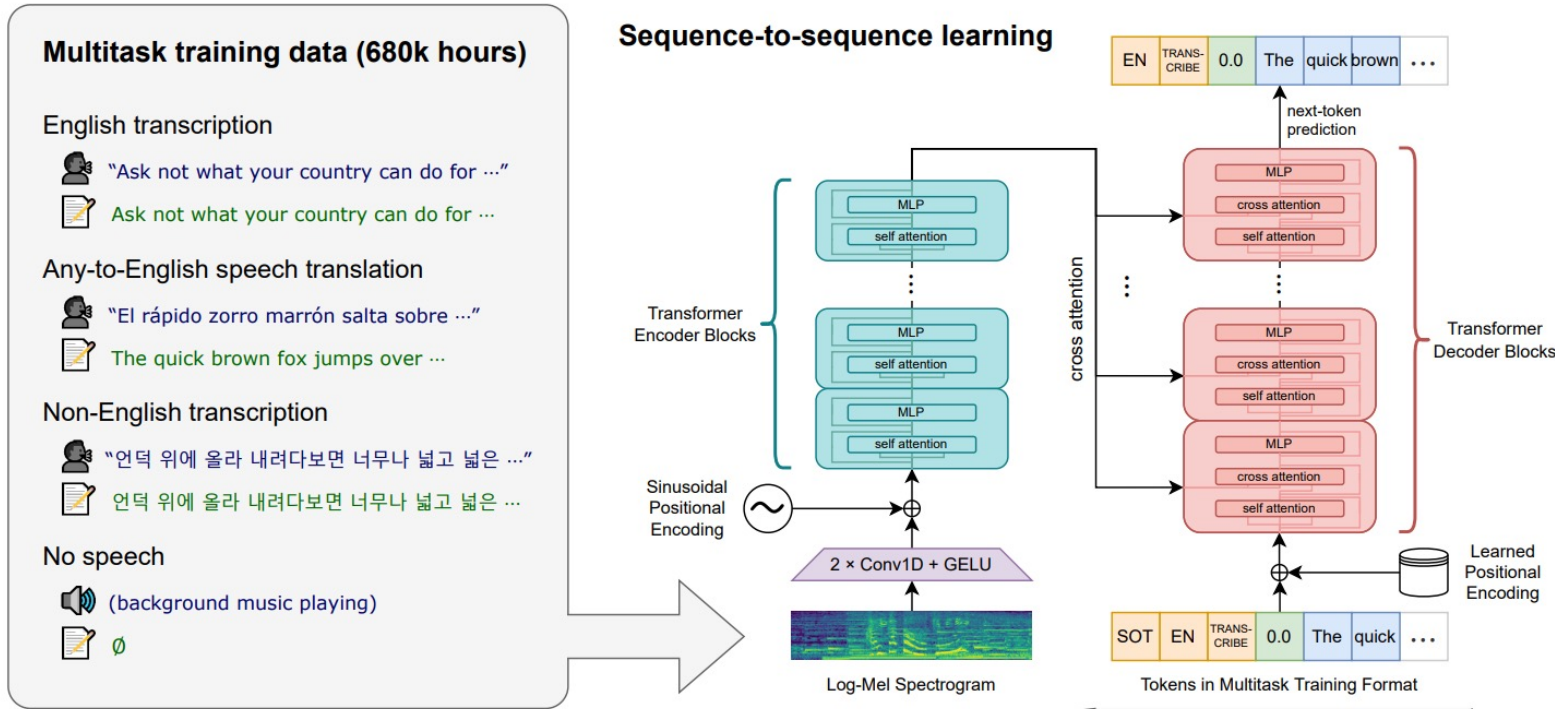


Source: Baevski, Alexei, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli.

"wav2vec 2.0: A framework for self-supervised learning of speech representations." Advances in Neural Information Processing Systems 33 (2020): 12449-12460.

Whisper:

Robust Speech Recognition via Large-Scale Weak Supervision



Microsoft Azure Text to Speech (TTS)

Text SSML

You can replace this text with any text you wish. You can either write in this text box or paste your own text here.

Try different languages and voices. Change the speed and the pitch of the voice. You can even tweak the SSML (Speech Synthesis Markup Language) to control how the different sections of the text sound. Click on SSML above to give it a try!

Enjoy using Text to Speech!

Language

English (United States) ▾

Voice

Jenny (Neural) ▾

Speaking style

General ▾

Speaking speed: 1.00



Pitch: 0.00



Play

Hugging Face



Hugging Face

Search models, datasets

Models

Datasets

Spaces

Docs

Solutions

Pricing



Log In

Sign Up



The AI community building the future.

Build, train and deploy state of the art models powered by
the reference open source in machine learning.



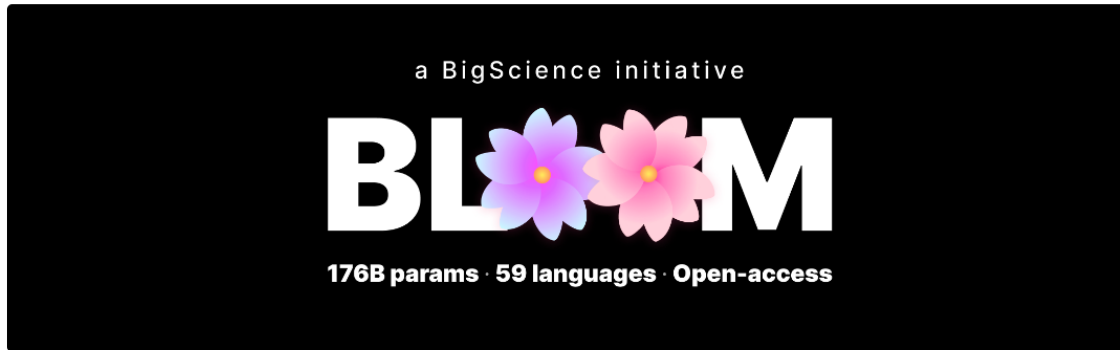
Star

58,696

<https://huggingface.co/>

BLOOM

BigScience Large Open-science Open-access Multilingual Language Model



BigScience Large Open-science Open-access Multilingual Language Model

Version 1.3 / 6 July 2022

Current Checkpoint: **Training Iteration 95000**

Total seen tokens: **366B**

Downloads last month
12,875



⚡ Hosted inference API ⓘ

📄 Text Generation

Groups ▾

Examples ▾

I love bloom. Super simple, but so effective! I went through a similar process a couple of years ago when I

sampling greedy

ⓘ [BLOOM prompting tips](#)

Switch to "greedy" for more accurate completion e.g. math/history/translations (but which may be repetitive/less inventive)

Compute

⌘+Enter

1.3

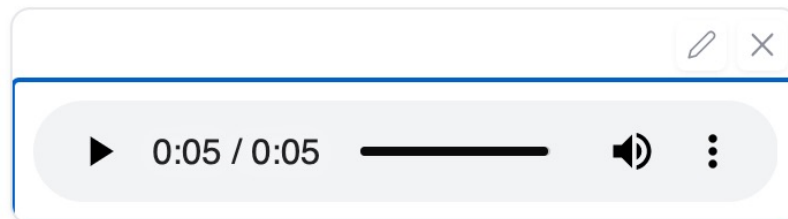
OpenAI Whisper

Whisper

Whisper is a general-purpose speech recognition model. It is trained on a large dataset of diverse audio and is also a multi-task model that can perform multilingual speech recognition as well as speech translation and language identification. This demo cuts audio after around 30 secs.

You can skip the queue by using google colab for the space:

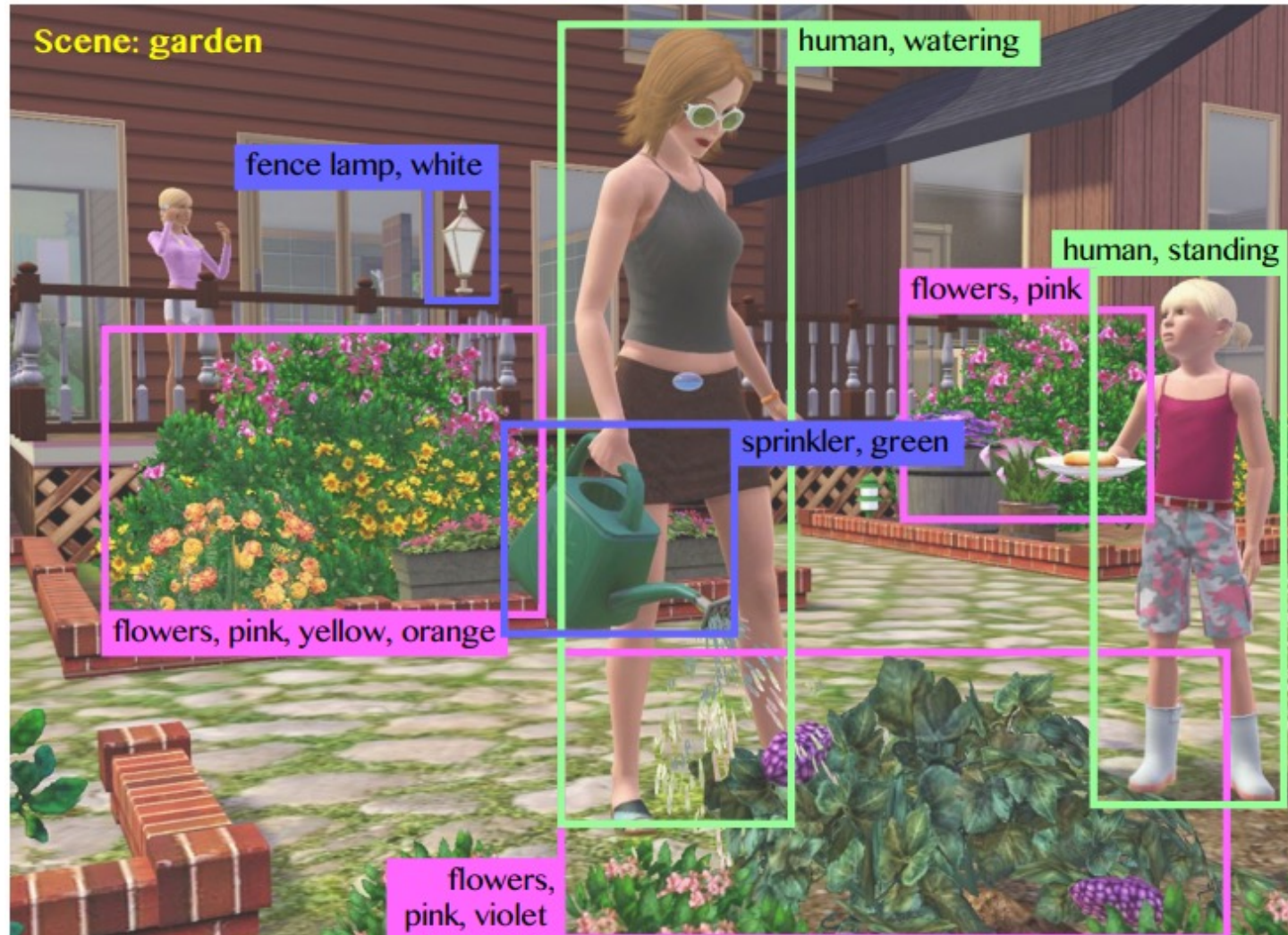
 [Open in Colab](#)



Transcribe

Computer vision in the metaverse

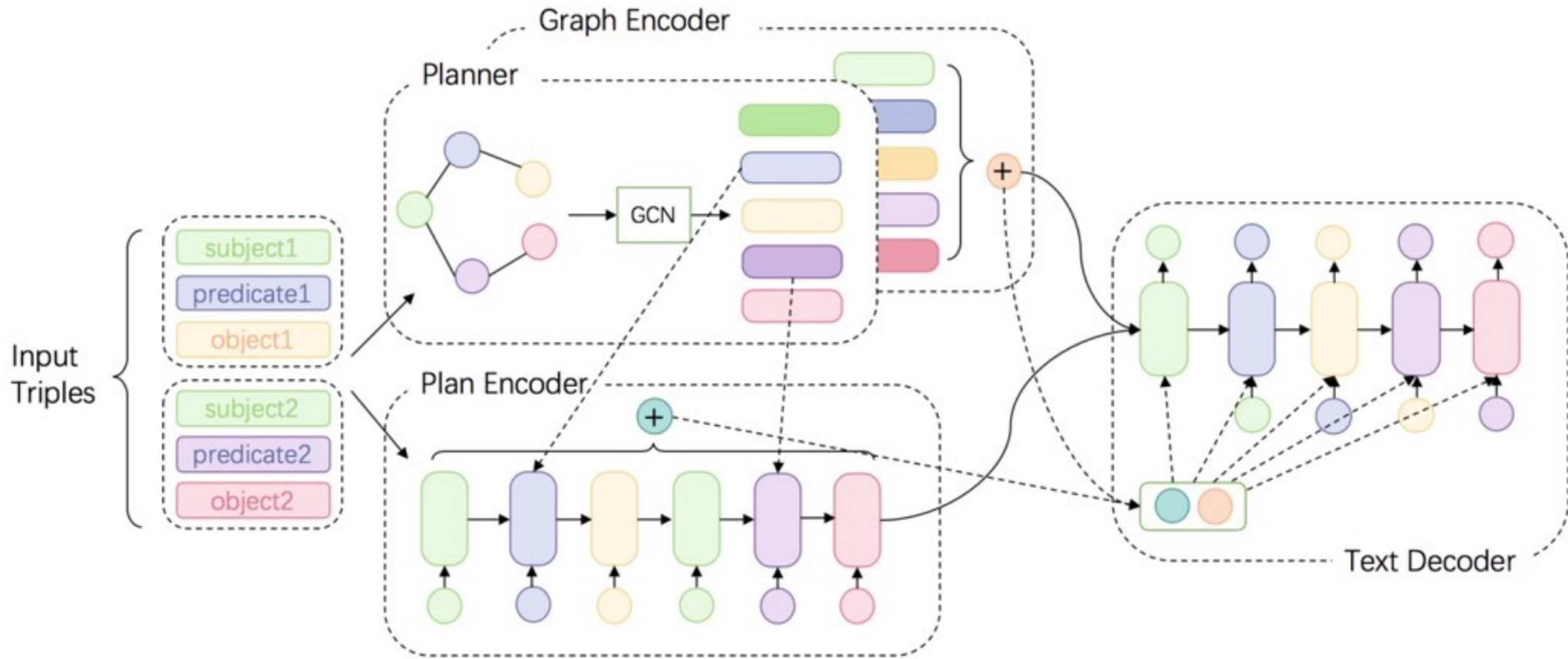
with scene understanding, object detection, and human action/activity recognition



Source: Huynh-The, Thien, Quoc-Viet Pham, Xuan-Quy Pham, Thanh Thi Nguyen, Zhu Han, and Dong-Seong Kim (2022). "Artificial Intelligence for the Metaverse: A Survey." arXiv preprint arXiv:2202.10336.

DUALENC: A KG-to-Text Generation Model

KG and Graph via Dual-encoding



Papers with Code State-of-the-Art (SOTA)

Computer Vision



▶ See all 1415 tasks

Natural Language Processing



▶ See all 664 tasks

Summary

- **ChatGPT**
- **Large Language Models (LLMs)**
- **Foundation Models for Big Data Analytics**

References

- Yihan Cao, Siyu Li, Yixin Liu, Zhiling Yan, Yutong Dai, Philip S. Yu, and Lichao Sun (2023). "A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT." arXiv preprint arXiv:2303.04226.
- Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. (2023) "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing." ACM Computing Surveys 55, no. 9 (2023): 1-35.
- Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min et al. (2023) "A Survey of Large Language Models." arXiv preprint arXiv:2303.18223.
- Gozalo-Brizuela, Roberto, and Eduardo C. Garrido-Merchan (2023). "ChatGPT is not all you need. A State of the Art Review of large Generative AI models." arXiv preprint arXiv:2301.04655 (2023).
- Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. (2023) "InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning." arXiv preprint arXiv:2305.06500 (2023).
- Shahab Saquib Sohail, Faiza Farhat, Yassine Himeur, Mohammad Nadeem, Dag Øivind Madsen, Yashbir Singh, Shadi Atalla, and Wathiq Mansoor (2023). "The Future of GPT: A Taxonomy of Existing ChatGPT Research, Current Challenges, and Possible Future Directions." Current Challenges, and Possible Future Directions (April 8, 2023) (2023).
- Stuart Russell and Peter Norvig (2020), Artificial Intelligence: A Modern Approach, 4th Edition, Pearson
- Thorsten Schoormann, Gero Strobel, Frederik Möller, Dimitri Petrik, and Patrick Zschech (2023). "Artificial Intelligence for Sustainability—A Systematic Review of Information Systems Literature." Communications of the Association for Information Systems 52, no. 1 (2023): 8.
- Junliang Wang, Chuqiao Xu, Jie Zhang, and Ray Zhong (2022). "Big data analytics for intelligent manufacturing systems: A review." Journal of Manufacturing Systems 62 (2022): 738-752.
- Longbing Cao (2022). "Decentralized ai: Edge intelligence and smart blockchain, metaverse, web3, and desc." IEEE Intelligent Systems 37, no. 3: 6-19.
- Qinglin Yang, Yetong Zhao, Huawei Huang, Zehui Xiong, Jiawen Kang, and Zibin Zheng (2022). "Fusing blockchain and AI with metaverse: A survey." IEEE Open Journal of the Computer Society 3 : 122-136.
- Russell Belk, Mariam Humayun, and Myriam Brouard (2022). "Money, possessions, and ownership in the Metaverse: NFTs, cryptocurrencies, Web3 and Wild Markets." Journal of Business Research 153: 198-205.
- Thien Huynh-The, Quoc-Viet Pham, Xuan-Quy Pham, Thanh Thi Nguyen, Zhu Han, and Dong-Seong Kim (2022). "Artificial Intelligence for the Metaverse: A Survey." arXiv preprint arXiv:2202.10336.
- Thippa Reddy Gadekallu, Thien Huynh-The, Weizheng Wang, Gokul Yenduri, Pasika Ranaweera, Quoc-Viet Pham, Daniel Benevides da Costa, and Madhusanka Liyanage (2022). "Blockchain for the Metaverse: A Review." arXiv preprint arXiv:2203.09738.
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. arXiv preprint arXiv:2203.02155.