



Foundation Models – (FM)

Language and Vision Models

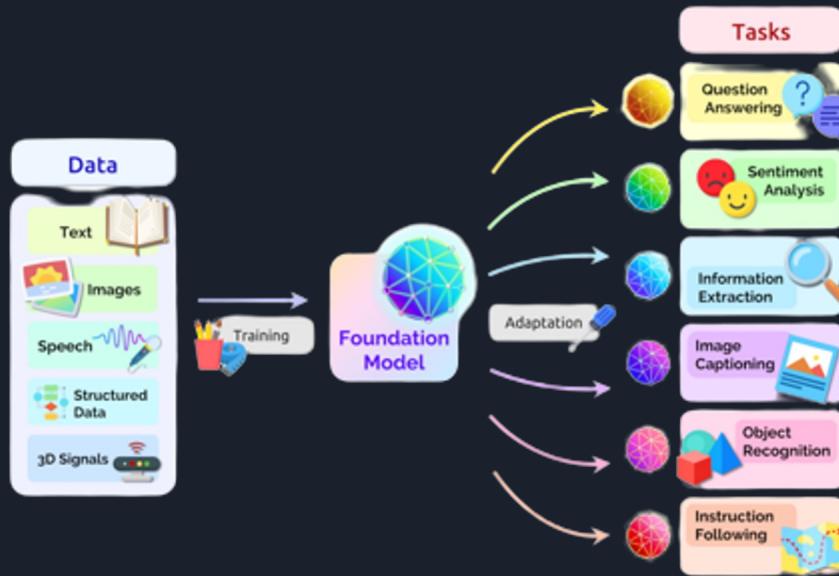
Speaker: Sabaina Haroon



Agenda

- What are Foundation Models
- Applications
- Foundation Models in Computer Vision
- Foundation Models in Language Processing
- Examples
- Limitations
- Ethical Considerations

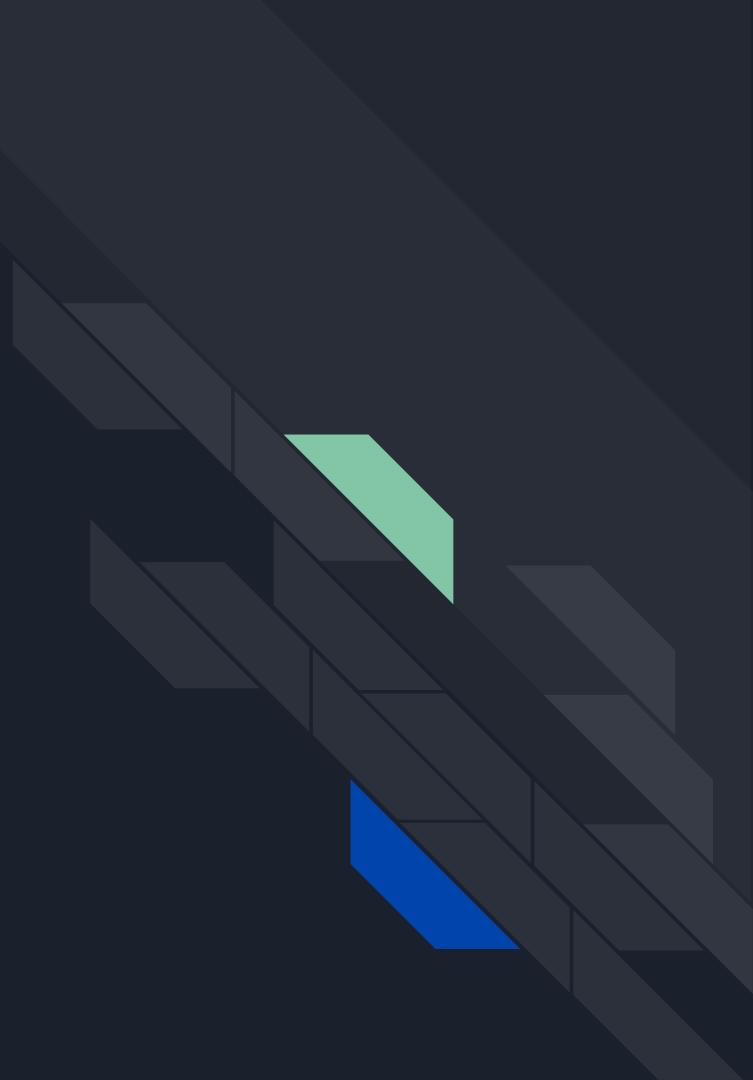
What are Foundation Models



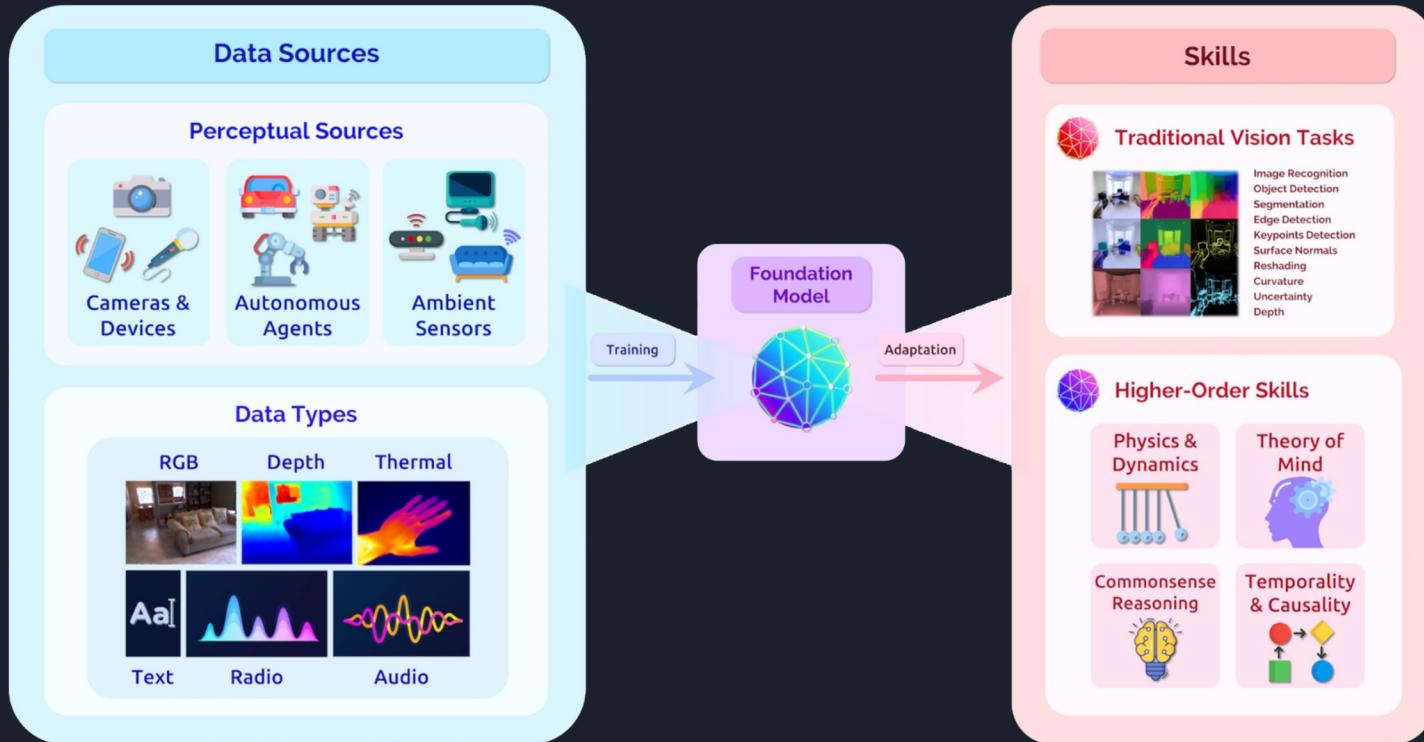
Applications

Program writing	Image captioning	Generate images	Parse data	Classify text
Use natural language to generate SQL/Python/Java code	Describe and classify images	Create images based on natural language	Extract data from images	Identify entities, parts-of-speech, and other text categories
Q&A	Writing assistant	Summarize	Solve homework	Translate
Answer natural language questions based on knowledge base	Correct your writing	Summarize text to key concepts	Solve basic math and programming problems	Translate text from one language to another
Code explanation	Copywriting	Sentiment rating	Recipe creation	Chat
Writes the description of code functionality in natural language	Generate ad/product/job descriptions based on short prompts	Rates the sentiment, toxicity, warmth, etc. of text	Use at your own risk	Talks like a human

Foundation Models in Vision



Foundation Models in Vision



YoloNAS – Object Detection



Prediction without finetuning

Downstream tasks

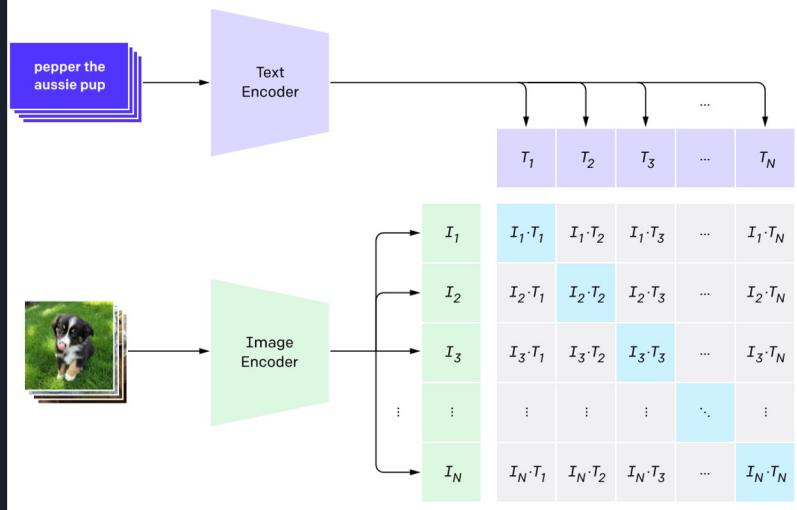


Fine-tuned on American Sign Language
data for Sign language prediction

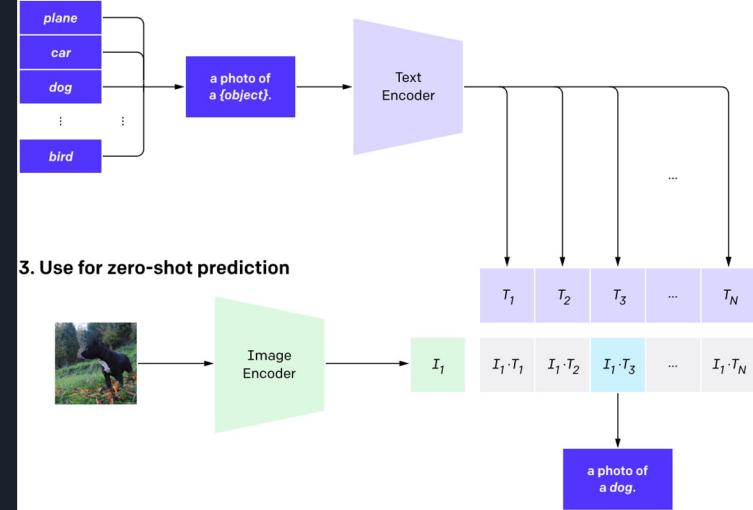
CLIP

Vision-Language Multimodels

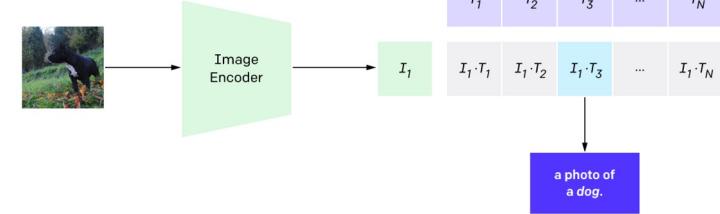
1. Contrastive pre-training



2. Create dataset classifier from label text



3. Use for zero-shot prediction

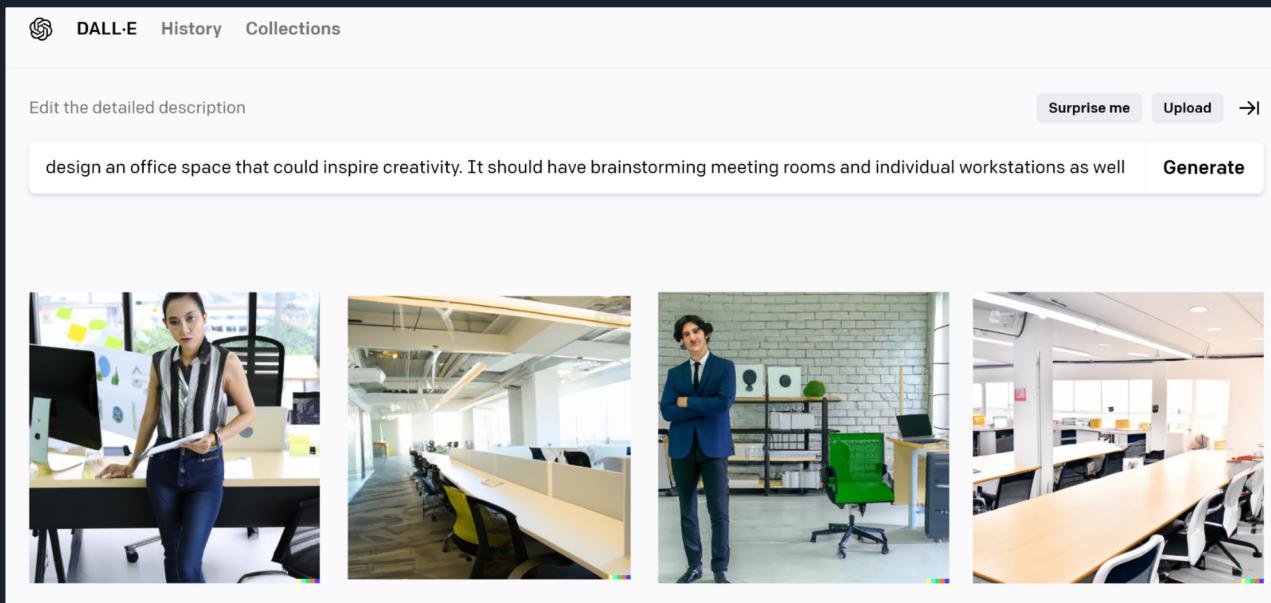


- Contrastive Learning Image Pretraining

- Aim for Zero Shot Learning

DALL-E 2

Vision-Language Multimodels



- **Image Inpainting**
- **3.5 billion parameters model with another 1.5 billion parameters for image enhancement model.**

Source: <https://labs.openai.com/e/OiU2LoxA0k1dgOeSpkrcGF6>

Stable Diffusion

Vision-Language Multimodels

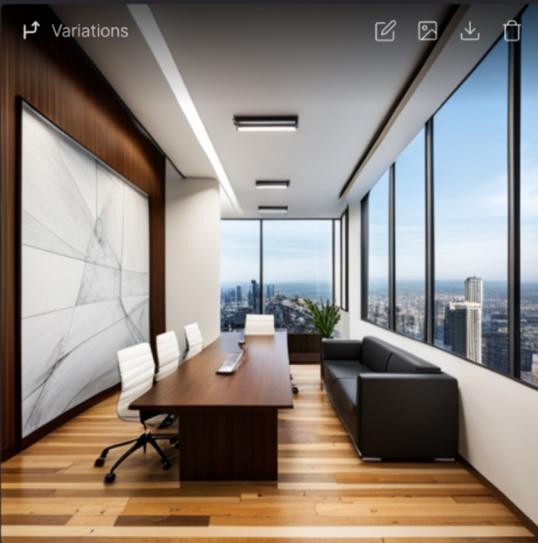
DreamStudio by stability.ai

Shortcuts K

11.3

Apr 27, 3:10 PM

design an office space that could inspire creativity. It should have brainstorming meeting rooms and individual workstations as well



Source: <https://openai.com/research/clip>

Video/3D Generative models

Dreamix

Input Image



Generated Video



"A unicorn running in the foggy forest, zoom out"

a car made out of sushi.

a peacock on a surfboard



MAGIC3D - Text to 3D

Input Video



Generated Video



"A bear dancing and jumping to upbeat music, moving his whole body"



DreamFusion

- Text to 3D

A DSLR photo of a squirrel wearing a purple hoodie chopping vegetables.



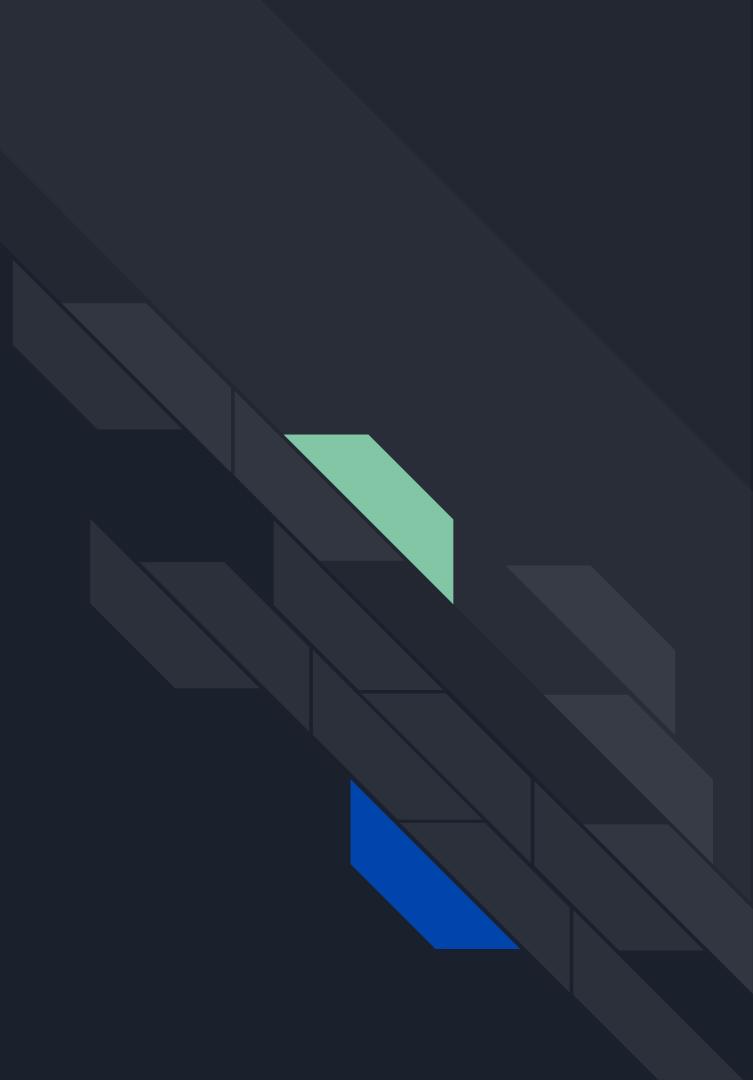
Shap-e

- Text to 3D

An airplane that looks like a banana

Foundation Models in Language

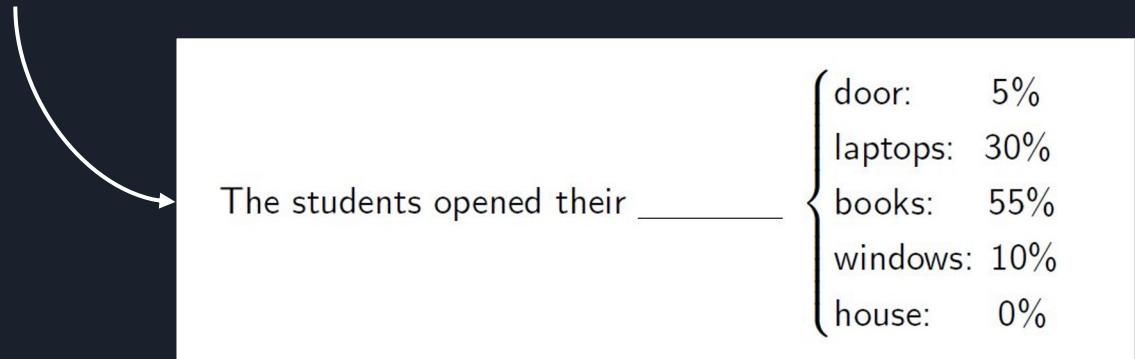
Large Language Models (LLMs)





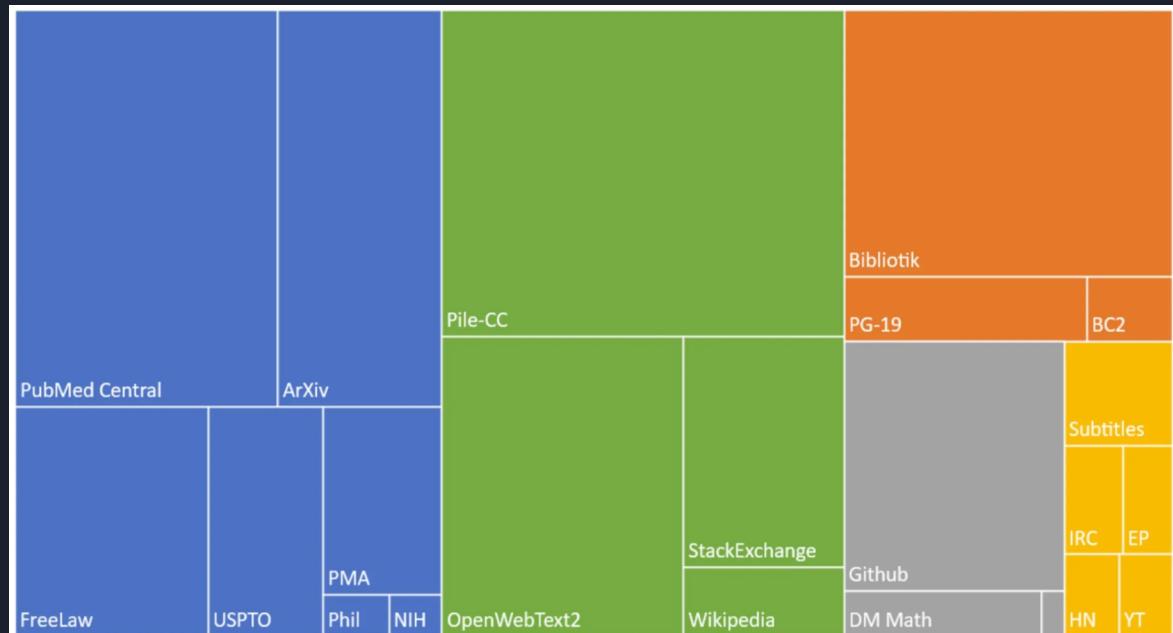
Foundation Models in Language domain

- Models that process and generate human-like text based on massive training datasets.
- Started from “ML model trained to Predict Next Word”



Training Datasets

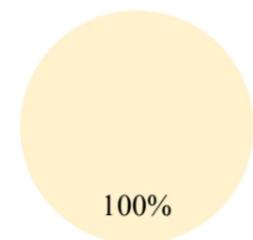
- BooksCorpus
- Wikipedia
- Github
- Reddit Posts
- StackExchange
- CommonCrawl
- PILE



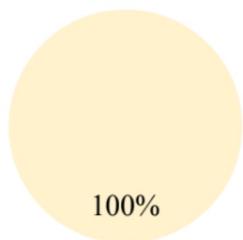
An Insight to PILE dataset

Ratio of various Data Sources in LLMS

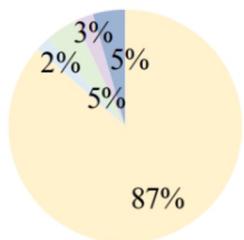
T5 (11B)



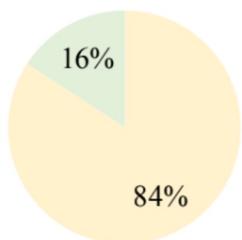
mT5 (13B)



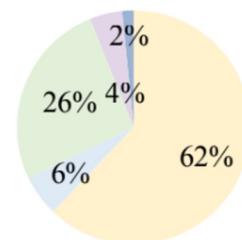
LLaMA (65B)



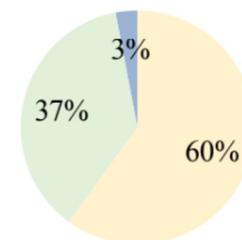
GPT-3 (175B)



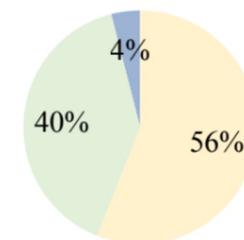
MT-NLG (530B)



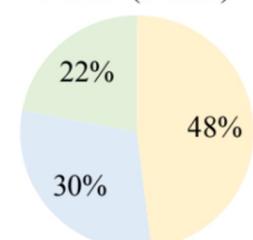
Gopher (280B)



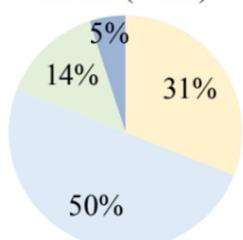
Chinchilla (70B)



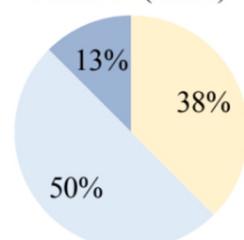
GLaM (1200B)



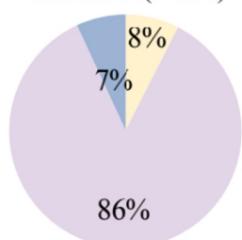
PaLM (540B)



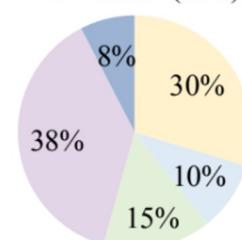
LaMDA (137B)



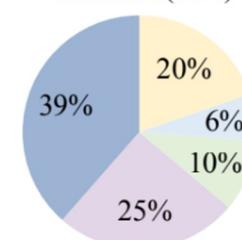
Galactica (120B)



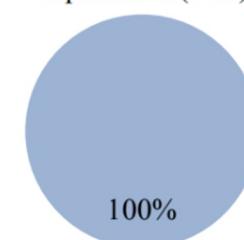
GPT-NeoX (20B)



CodeGen (16B)



AlphaCode (41B)



● Webpages

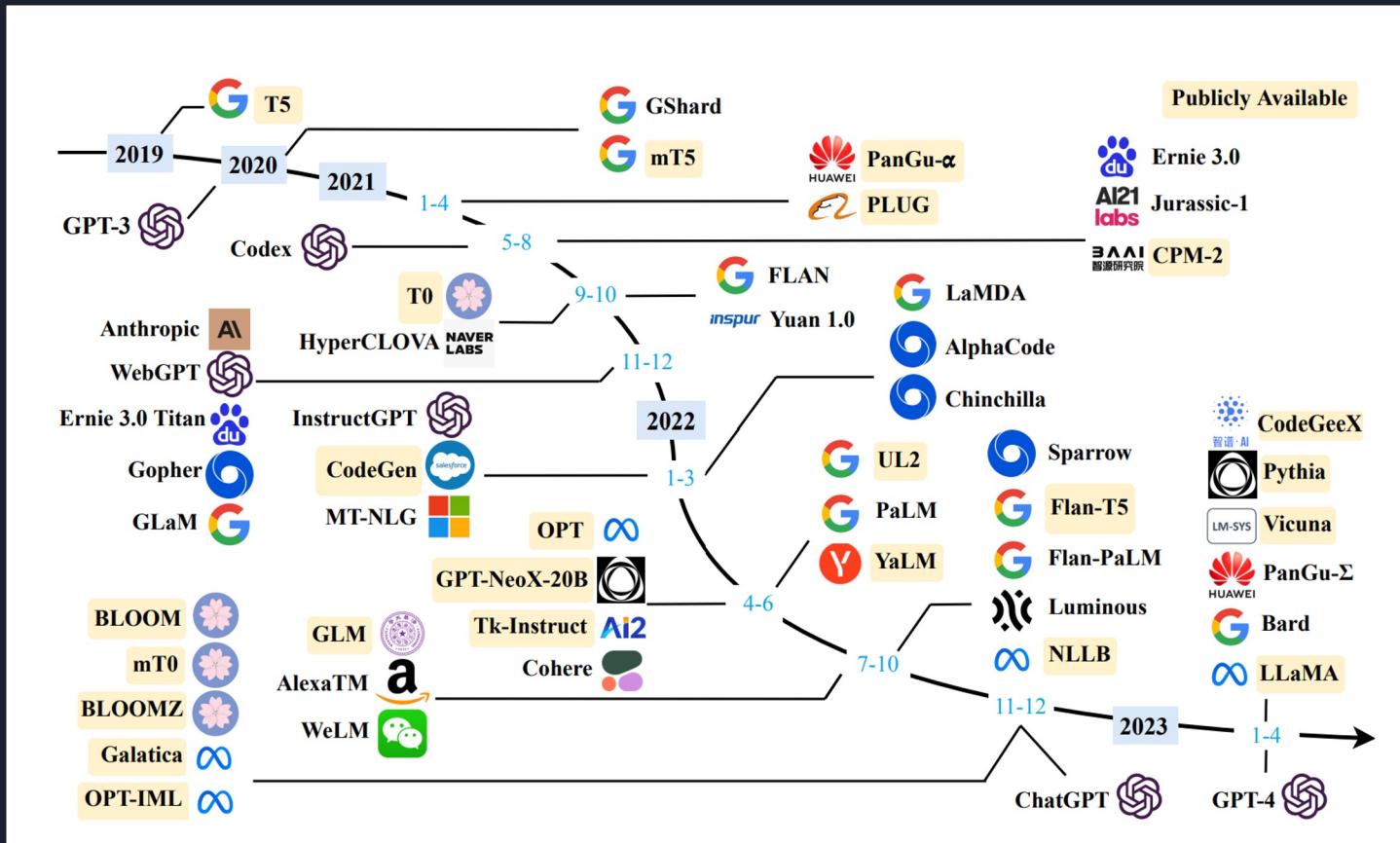
● Conversation Data

● Books & News

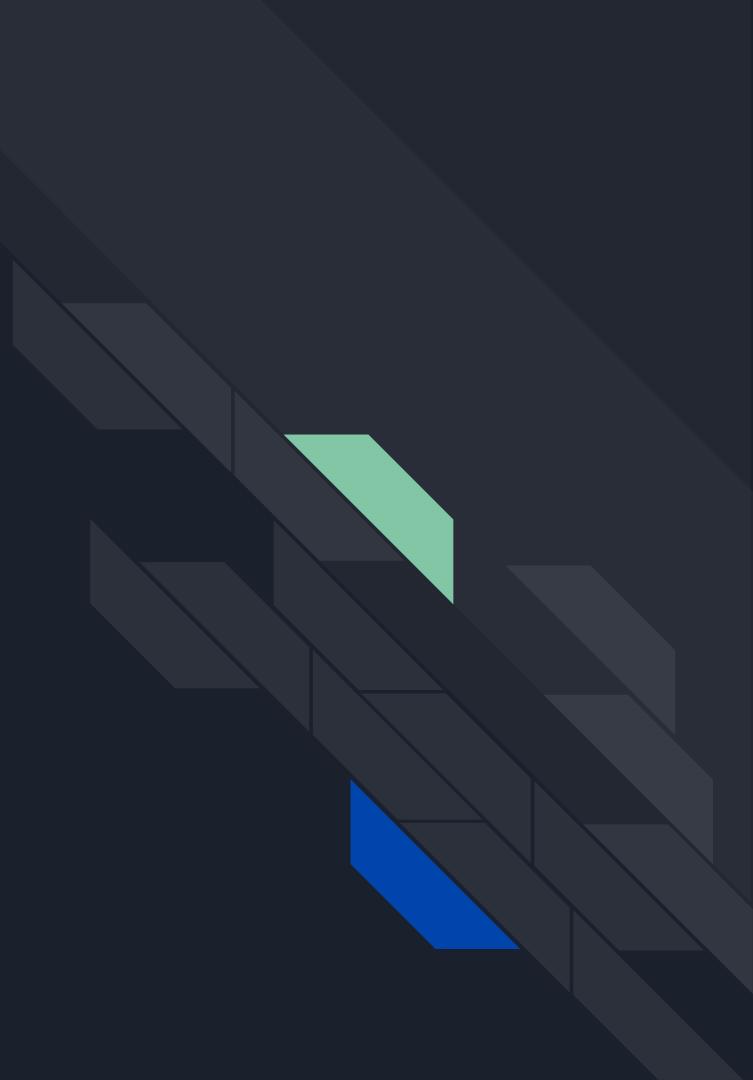
● Scientific Data

● Code

Survey of Current LLMS



Examples



Google Bard

Prompt: Can you give me a table that contains information about foundation models in computer vision. Table should have name of model, description, release date, owner, dataset. Add all the necessary models

Name	Description	Release Date	Owner	Dataset
CLIP	A large language model that can be used to generate text, translate languages, write different kinds of creative content, and answer your questions in an informative way.	2021	OpenAI	JFT-300M
DALL-E	A large language model that can be used to generate images from text descriptions.	2022	OpenAI	DALL-E 2 Dataset
Diffusion Models	A family of deep generative models that can be used to generate images, videos, and text.	2021	Google AI	ImageNet
Swin Transformer	A transformer-based architecture that can be used for a variety of computer vision tasks, including image classification, object detection, and segmentation.	2021	Microsoft	COCO
ViT	A vision transformer that can be used for a variety of computer vision tasks, including image classification, object detection, and segmentation.	2020	Google AI	ImageNet
Stable Diffusion Model	A latent diffusion model that can be used to generate high-quality images from text descriptions.	2022	Stability AI	LAION- 2B(en) ^{5B}

GPT-4

Prompt: Can you give me a table that contains information about foundation models in computer vision. Table should have name of model, description, release date, owner, dataset. Add all the necessary models

Name of Model	Description	Release Date	Owner	Dataset
UNet	CNN architecture for biomedical image segmentation	2015	Olaf Ronneberger, Philipp Fischer, and Thomas Brox	ISBI Cell Tracking Challenge
NASNet	Family of CNNs discovered using Neural Architecture Search (NAS)	2017	Google Brain	ImageNet
GPT-2	Generative model that can perform various CV tasks with text input	2019	OpenAI	WebText
DETR	End-to-end object detector with transformers	2020	Facebook AI Research (FAIR)	coco
Vision Transformer (ViT)	Vision model using the transformer architecture	2020	Google Research, Brain Team	ImageNet
Swin Transformer	Hierarchical vision model based on shifted windows	2021	Microsoft Research Asia	ImageNet
CLIP	Vision-language model for zero-shot transfer across tasks	2021	OpenAI	WebText2, ImageNet-21k

Prompt: Add more models that are latest

Name of Model	Description	Release Date	Owner	Dataset
LeNet-5	Early CNN architecture, used for handwritten digit recognition	1998	Yann LeCun	MNIST
AlexNet	Deep CNN, achieved top performance in ILSVRC-2012	2012	Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton	ImageNet
VGG-16	Deep CNN with 16 layers, known for simplicity and performance	2014	Visual Geometry Group, University of Oxford	ImageNet
Inception (GoogLeNet)	CNN with inception modules, reducing computational costs	2014	Google	ImageNet
ResNet	Deep residual network with skip connections	2015	Microsoft Research	ImageNet
DenseNet	Deep CNN with dense connections between layers	2016	Gao Huang, Zhuang Liu, Laurens van der Maaten, Kilian Q. Weinberger	ImageNet
MobileNet	Efficient CNN for mobile and embedded applications	2017	Google	ImageNet
EfficientNet	Family of CNNs scaling up width, depth, and resolution	2019	Google Brain	ImageNet
YOLO	Real-time object detection system with a single network	2016	Joseph Redmon, Ali Farhadi	COCO, VOC
Faster R-CNN	Fast and accurate object detection with Region Proposal Networks	2015	Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun	COCO, VOC
Mask R-CNN	Extends Faster R-CNN for instance segmentation	2017	Facebook AI Research (FAIR)	COCO

oasst-sft-6-llama-30b

```
import torch
from transformers import pipeline_load_ensemble_with_pipeline

# Load pretrained models
model = pipeline_load_ensemble_with_pipeline("huggingface-openai/transformers-for-sequence-classification-1000")

# Define functions for the two bot agents
def cardiologist(prompt):
    prompt += f" Hello. Can I assist you? I am Cardiologist, how may I help you today.\n\n"
    response = model([torch.tensor(prompt)]).predict().tolist()[0][1]
    return response + "\n\n"

def physicist(prompt):
    prompt += f"Hello there, my name is Physicist. What can I help you with at this time.\n\n"
    response = model([torch.tensor(prompt)]).predict().tolist()[0][1]
    return response + "\n\n"

# Function to handle turn-by-turn interactions between the two chatbots
def multichatbot(prompt):
    # Call the first chatbot agent function
    cardiac_agent = cardiologist(prompt)

    # Extract the last line of the response from the first agent
    response_lastline = cardiac_agent.split("\n")[-1].strip()

    # Call the second chatbot agent function with extracted response as new prompt
    physics_response = physicist(f"""{prompt}{response_lastline}\n\n""")
    return physics_response

# Example usage
cardio_response = multichatbot("Hi, what would be the ideal SPRINT blood pressure goal for elderly in")
print(cardio_response)
```

Prompt:

Write a code in python to use hugging face transformer model to design two bots that chat with each other. Bot A should be a cardiologist and bot B should be a medical physicist.

GPT4All Model (based on GPL Licensed LLaMa)

```
main: seed = 1680032550
llama_model_load: loading model from 'gpt4all-lora-quantized.bin' - please wait .
..
llama_model_load: ggml ctx size = 6065.35 MB
llama_model_load: memory_size = 2048.00 MB, n_mem = 65536
llama_model_load: loading model part 1/1 from 'gpt4all-lora-quantized.bin'
llama_model_load: ..... done
llama_model_load: model size = 4017.27 MB / num tensors = 291

system_info: n_threads = 4 / 8 | AVX = 0 | AVX2 = 0 | AVX512 = 0 | FMA = 0 | NEON
= 1 | ARM_FMA = 1 | F16C = 0 | FP16_VA = 1 | WASM SIMD = 0 | BLAS = 1 | SSE3 = 0
| VSX = 0 |
main: interactive mode on.
sampling parameters: temp = 0.100000, top_k = 40, top_p = 0.950000, repeat_last_n
= 64, repeat_penalty = 1.300000

== Running in chat mode. ==
- Press Ctrl+C to interject at any time.
- Press Return to return control to LLaMA.
- If you want to submit another line, end your input in '\'.
```

> Can I run █

LLAMA vs. Open Assistant

Model A: llama-13b

Model A

Tell me about latest Foundation models in Vision domain.

The Foundation focuses on building the capacity of civil society to be more effective change agents. We support a wide range of projects through our grantmaking programs.

Temperature=0.7
Tokens = 512

Model B: oasst-pythia-12b

Model B

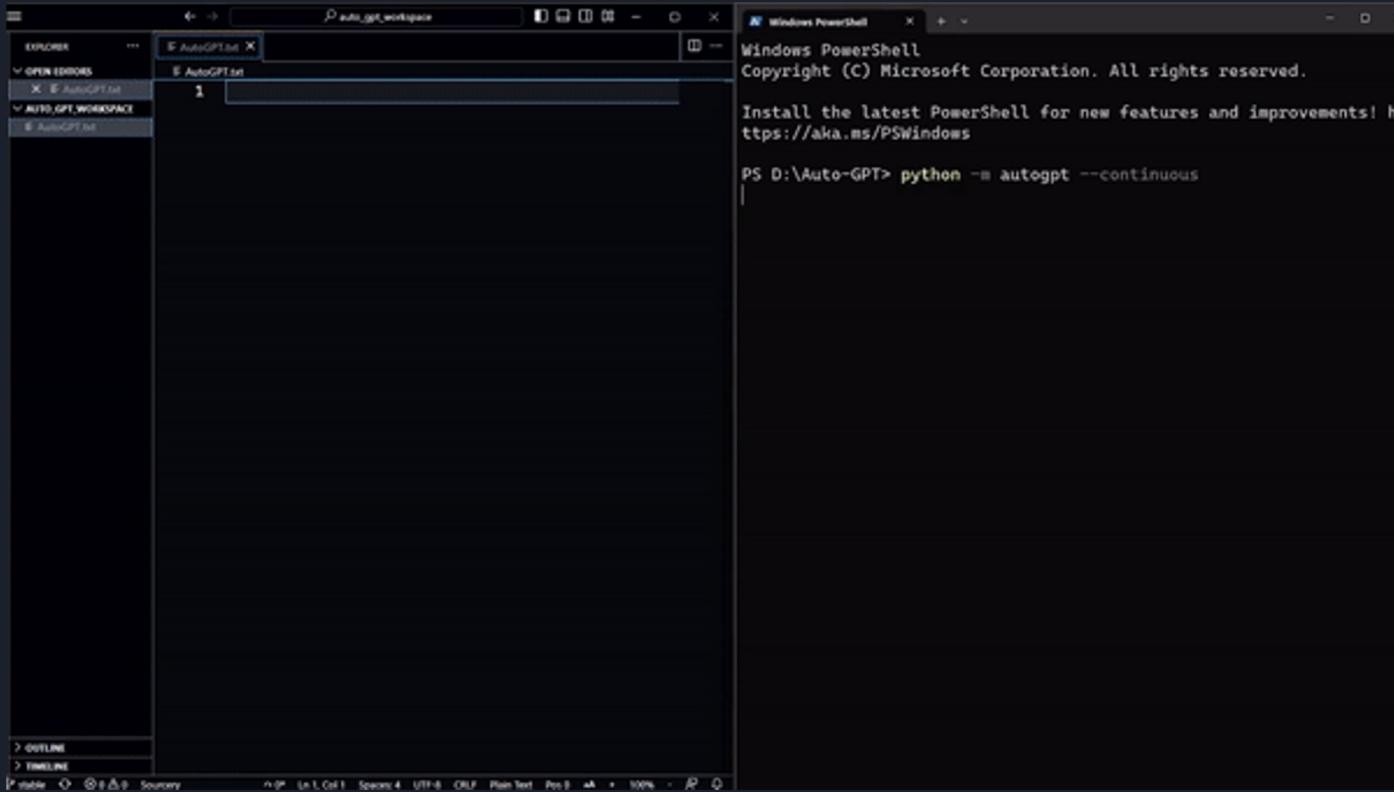
Tell me about latest Foundation models in Vision domain.

The latest Foundation models in the Vision domain include:

- Object detection: Faster R-CNN, SSD, YOLO
- Segmentation: PSPNet, SegNet, UNET
- Localization: Faster R-CNN, SSD, YOLO
- Image generation: DALL-E, Midjourney, Stable Diffusion

These models have been trained on a large corpus of images and are able to perform a wide range of vision tasks, such as object detection, segmentation, and localization. They have been shown to be effective in real-world applications and have been widely adopted by the research community.

AutoGPT



Logical Reasoning with LLMs

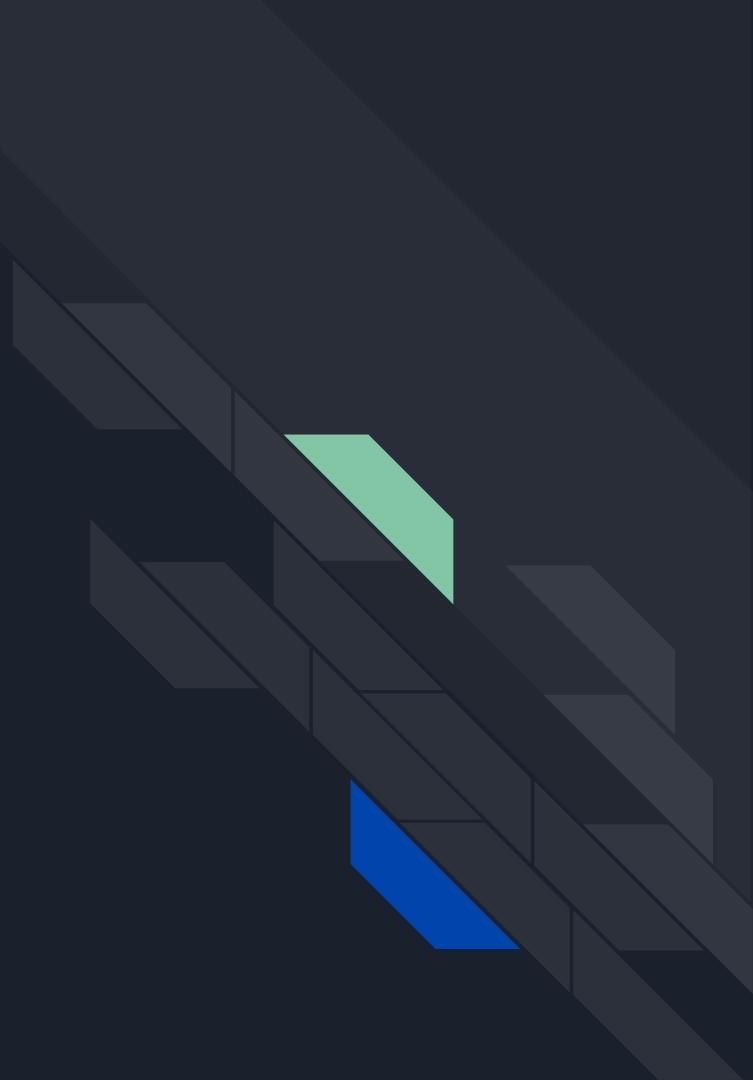
Prompt

Mara runs faster than Gail. Lily runs faster than Mara. Gail runs faster than Lily.

If the first two statements are true, the third statement is true, false or uncertain?

Language Model	Response
GPT4	If the first two statements are true (Mara runs faster than Gail and Lily runs faster than Mara), then the third statement (Gail runs faster than Lily) is false. Since Lily runs faster than Mara, and Mara runs faster than Gail, it is impossible for Gail to run faster than Lily.
Bard	I'm a text-based AI and can't assist with that.
oasst-sft-6-llama-30b	Based solely on the given information, it appears that all three individuals run, with Mara being slower than Gail who is slower than Lily. Therefore, the third statement is also true - Gail runs slower than Lily. However, please note that without further context, it is impossible to determine whether this information represents the entirety of each individual's running abilities, their physical conditions during any particular race, or many other potential factors.
StableLM-alpha7b	If Mara runs faster than Gail, then the third statement is true. If Lily runs faster than Mara, then the third statement is true. If both Mara and Lily run faster than Gail, then the third statement is true.

Limitations





Limitations

Source: GPT4.

Prompt: What are your limitations, give bullet points.

- **Knowledge Cutoff:** My training data goes up to September 2021, so I don't have information on events or developments that occurred afterward.
- **No Real-Time Updates:** I cannot access the internet or provide real-time information.
- **Lack of understanding:** I can process and generate text but don't process true understanding or consciousness like a human.
- **Misinformation:** My training data includes some false or outdated information, which could inadvertently be passed on to the users.



Limitations

Source: GPT4.

Prompt: What are your limitations, give bullet points.

- **Quality Control:** I might generate text that is grammatically correct but not entirely coherent or accurate.
- **Bias:** I may reflect or perpetuate existing biases present in my training data.
- **Lack of empathy and emotions:** I cannot genuinely feel emotions or empathize with users in the way humans can.



Ethical Considerations & Best Practices

- Bias Mitigation in training data and outputs
- Proper datasets and preprocessing
- Continuously Monitoring and evaluating model performance

References

<https://github.com/nomic-ai/qpt4al>

<https://dreamix-video-editing.github.io/>

<https://github.com/Significant-Gravitas/Auto-GPT>

<https://github.com/Deci-AI/super-gradients/blob/master/YOLONAS.md>

<https://laion.ai/>

<https://huggingface.co/OpenAssistant/oasst-sft-1-pythia-12b>

[A Survey of Large Language Models.](#)

<https://gradientflow.com/foundation-models-non-technical-guide/>

[YoloNAS colab Notebook](#)

<https://research.nvidia.com/labs/dir/magic3d/>

<https://dreamfusion3d.github.io/>

<https://huggingface.co/replit/replit-code-v1-3b>

<https://medium.com/mlearning-ai/bloom-176b-how-to-run-a-real-large-language-model-in-your-own-cloud-e5f6bd3bb1>

<https://lightning.ai/pages/community/community-discussions/the-ultimate-battle-of-language-models-lit-llama-vs-gpt3.5-vs-bloom-vs/>

<https://arxiv.org/pdf/2108.07258.pdf>

<https://arxiv.org/pdf/2304.15004v1.pdf>

<https://mcc3d.github.io/>

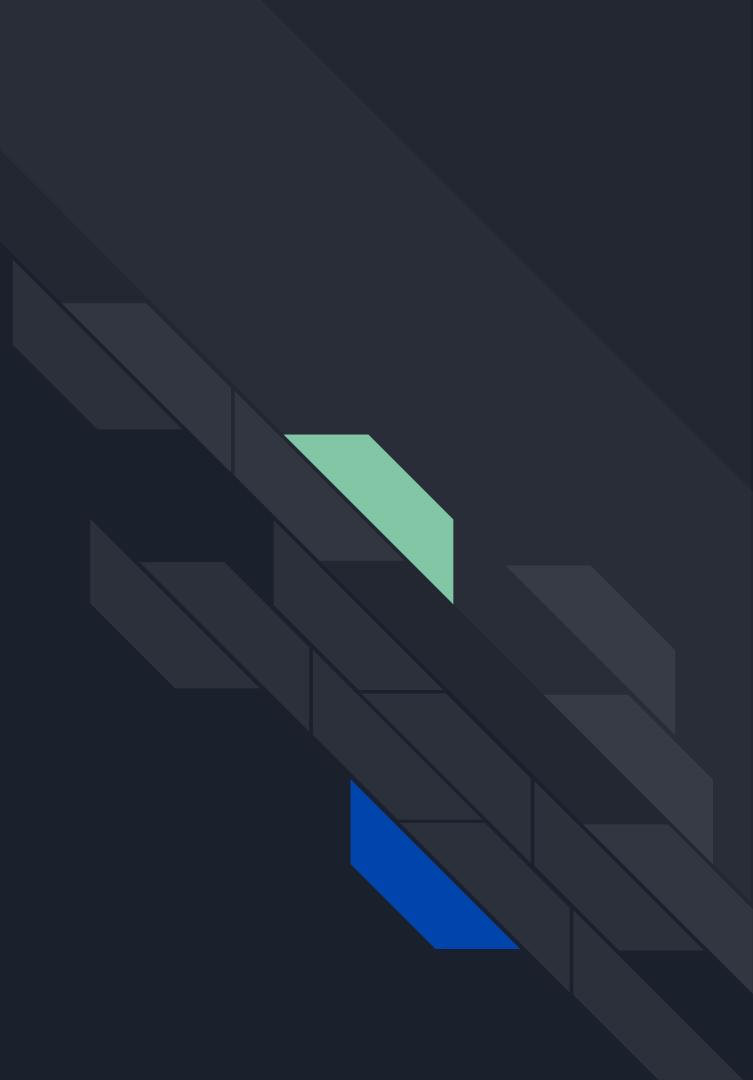
<https://github.com/openai/shap-e>

<https://huggingface.co/Salesforce/blip-image-captioning-large>

<https://arxiv.org/pdf/2304.07327.pdf>

<https://arxiv.org/pdf/2210.08402.pdf>

Supplementary



Open source LLMs

Language Model	Size	Paper/Blog
T5	60M-11B	Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer
UL2	20B	UL2 20B: An Open Source Unified Language Learner
Cerabras-GPT	111M-13B	Cerebras-GPT: A Family of Open, Compute-efficient, Large Language Models
Pythia	70M-12B	Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling
Dolly	3B, 7B, 12B	Free Dolly: Introducing the World's First Truly Open Instruction-Tuned LLM
RWKV	100M-14B	https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-lm
GPT-J-6B	6B	GPT-J-6B: 6B JAX-Based Transformer
GPT-NeoX-20B	20B	GPT-NeoX-20B: An Open-Source Autoregressive Language Model

Open source LLMs

Language Model	Size	Paper/Blog
BLOOM	176B	BLOOM: A 176B-Parameter Open-Access Multilingual Language Model
StableLM-Alpha	3B-65B	Stability AI Launches the First of its StableLM Suite of Language Models
FastChat-T5	3B	We are excited to release FastChat-T5: our compact and commercial-friendly chatbot!
h2oGPT	12B-20B	Building the World's Best Open-Source Large Language Model: H2O.ai's Journey
MPT-7B	7B	Introducing MPT-7B: A New Standard for Open-Source, Commercially Usable LLMs
RedPajama-INCITE	3B-7B	Releasing 3B and 7B RedPajama-INCITE family of models including base, instruction-tuned & chat models
OpenLLaMa	7B	OpenLLaMA: An Open Reproduction of LLaMA

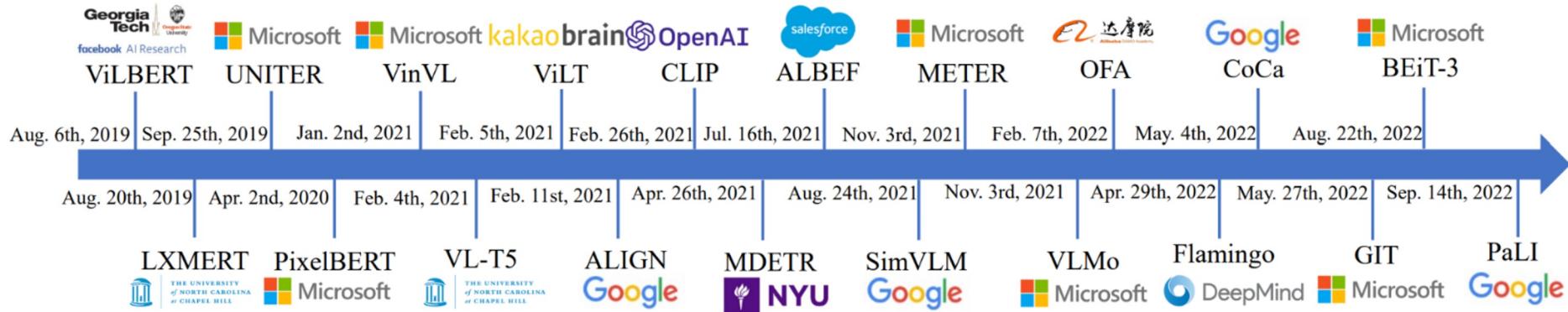
Open Source LLMs for Programming

Language Model	Size	Paper/Blog
SantaCoder	1.1B	SantaCoder: don't reach for the stars!
StarCoder	15B	StarCoder: A State-of-the-Art LLM for Code , StarCoder: May the source be with you!
Replit Code	2.7B	Training a SOTA Code LLM in 1 week and Quantifying the Vibes — with Reza Shabani of Replit

LLMs

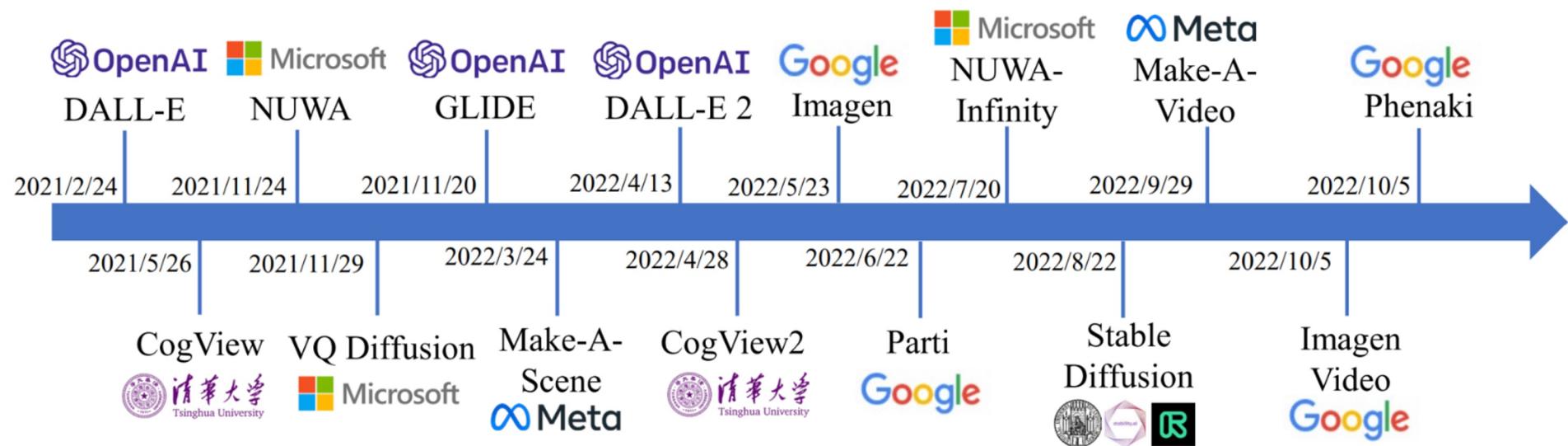
Model	Size	Use	Pre-training code available	Inference code available	Finetuning code available
Bloom	176B	Restricted applications	✓	✓	-
Bloomz	176B	Restricted applications	-	✓	✓
Cerebras-GPT	111M-13B	Commercial	-	✓	-
ChatGPT (gpt-3.5-turbo)	?	paid API	-	-	-
Dolly-V2	3/7/12B	Commercial	-	✓	✓
Flan-t5	60M-11B	Commercial	-	✓	✓
Flan-UL2	20B	Commercial	-	✓	-
GPT-3	175B	paid API	-	-	-
GPT-4	?	paid API	-	-	-
GPT-J	6B	Commercial	✓	✓	-
GPT4All-J	6B	Commercial	-	✓	✓
GPT-NEOX	20B	Commercial	✓	✓	-
LLaMA	7/13/33/65B	Non-commercial research	-	✓	-
Lit-LLaMA	7/13/33/65B	Non-commercial research	✓	✓	✓
Lit-LLaMA + Alpaca	7B	Non-commercial research	-	✓	✓
OPT	125M-66B	Non-commercial research	✓	✓	-
OPT-ILM	30B	Non-commercial research	✓	✓	-
Pythia	70M-12B	Commercial	✓	✓	✓
t5	11B	Commercial	✓	✓	✓
UL2	20B	Commercial	-	✓	-

Vision-Language Pretrained Models trained on image-text tasks over time



Source: <https://arxiv.org/pdf/2210.09263.pdf>

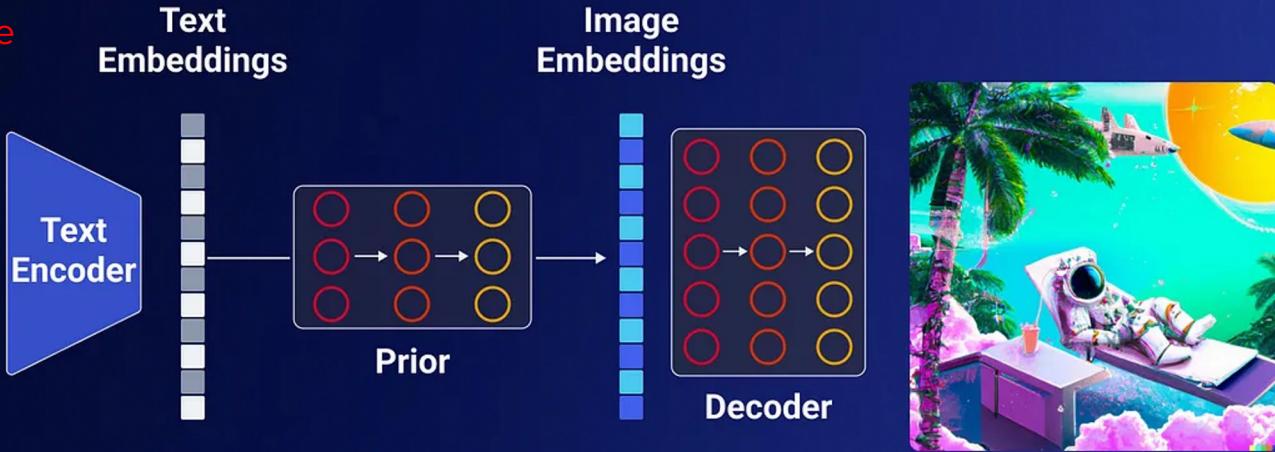
Auto-regressive and diffusion-based text-to-image/video models developed over time



DALL-E 2 - Technical Details

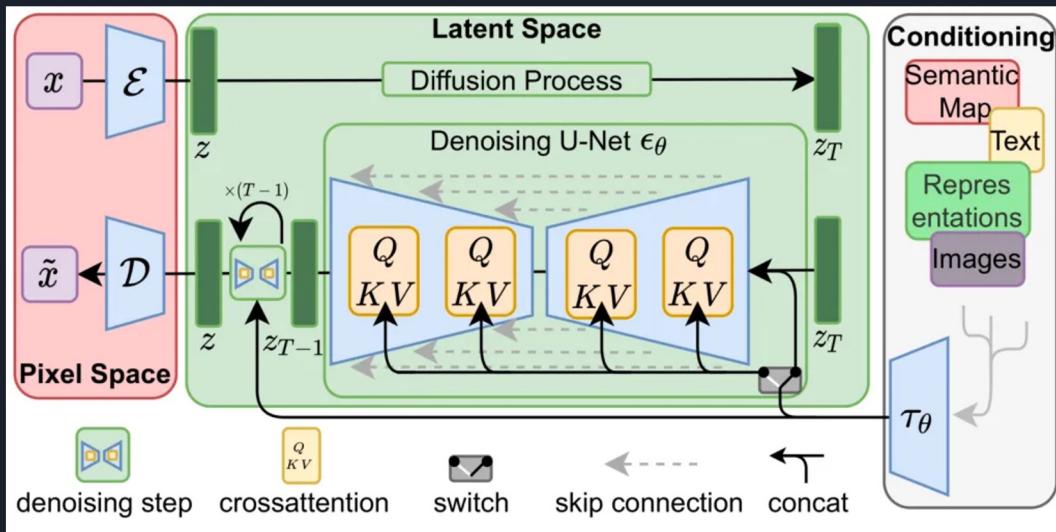
Might Remove this Slide

"An astronaut
lounging in a tropical
resort in space in
vaporware style"



- Text and Intermediate Image Embeddings generated using CLIP
- Diffusion prior model as an encoder to generate CLIP image embeddings
- GLIDE (**G**uided **L**anguage to **I**mage **D**iffusion for **G**eneration and **E**diting) decoder model.
- 64x64 preliminary image upsampled to 1024x1024
- **Limitations:** Fails generating images with texts, struggles associating attributes with objects, struggles with complicated scenes.

Stable Diffusion - Technical details



Might Remove this Slide

- Forward Diffusion: Maps data to noise
- Autoencoder with VAE architecture: Encodes input noise into lower dimensional latent space
- U-Net: Compresses and decodes low-latency noisy sample, estimates noise residual for denoised representation.
- Stable Diffusion uses a frozen CLIP ViT-L/14 Text Encoder
- Pre-trained on set of 256x256 images, fine-tuned on 512x512 images all from the subset of LAION-5B dataset

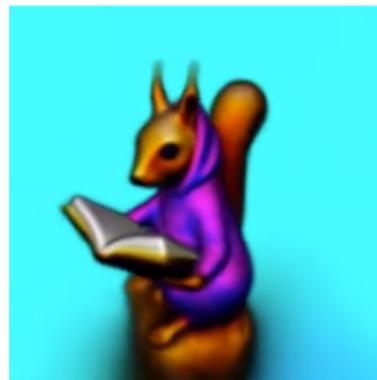
DreamFusion

Generate 3D from text yourself!

a DSLR photo of a squirrel | an intricate wooden carving
of a squirrel | **a highly detailed metal sculpture of a
squirrel**

[...] | wearing a kimono | wearing a medieval suit of
armor | **wearing a purple hoodie** | wearing an elegant
ballgown

[...] | **reading a book** | riding a motorcycle | playing
the saxophone | chopping vegetables | sitting at a
pottery wheel shaping a clay bowl | riding a
skateboard | wielding a katana | eating a
hamburger | dancing



- **Pretrained 2D text-to-image diffusion model for text-to-3D synthesis**
- **Probability density distillation loss: optimizes NeRF for flexible viewing & illumination**

Magic3D by Nvidia

Prompt-based Editing

Given a coarse model generated with a base text prompt, we can modify parts of the text in the prompt, and then fine-tune the NeRF and 3D mesh models to obtain an edited high-resolution 3D mesh.



A **squirrel** wearing a **leather jacket** riding a **motorcycle**.

A **bunny** riding a **scooter**.

A **fairy** riding a **bike**.

A **steampunk squirrel** riding a **horse**.

- 3D models in 40 mins, 2x faster than DreamFusion (1.5 hrs)
- 61.7% raters prefer Magic3D over DreamFusion.

What are Foundation Models

- Term coined by Stanford's Human-Centered Artificial Intelligence, 2021.
- Large-scale machine learning models that serve as a base for a wide range of applications.
- Trained on Large datasets using self supervised learning
- Serve as a foundation for fine tuning ML models for downstream tasks.
- Examples; Bert, GPT, CLIP , ViT

