A Comprehensive Overview of Large Language Models (LLM): Insights from a Machine Learning System Perspective

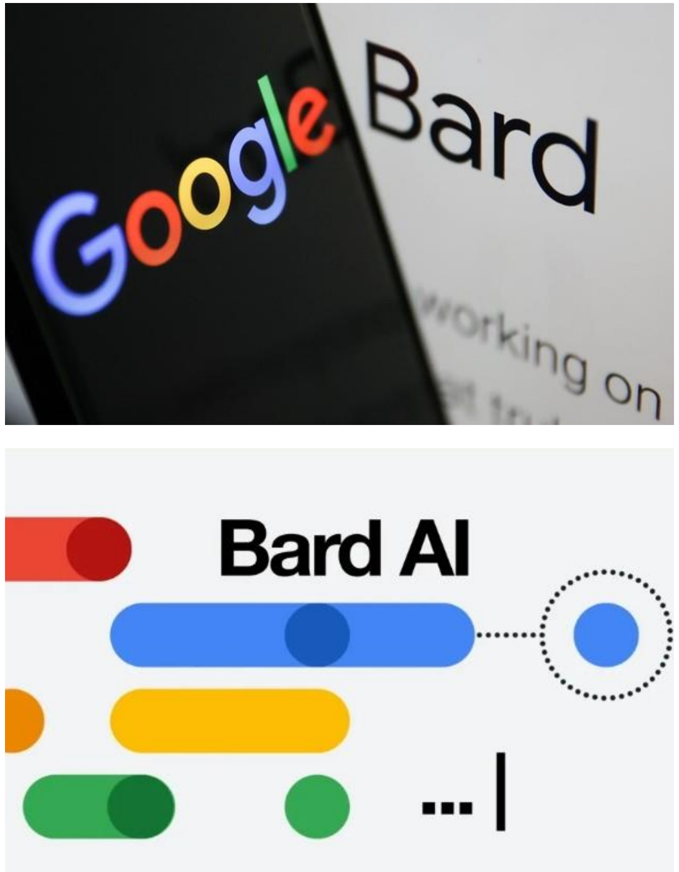
Shengyuan Ye

School of Computer Science and Engineering
Sun Yat-sen University
Contact: yeshy8@mail2.sysu.edu.cn

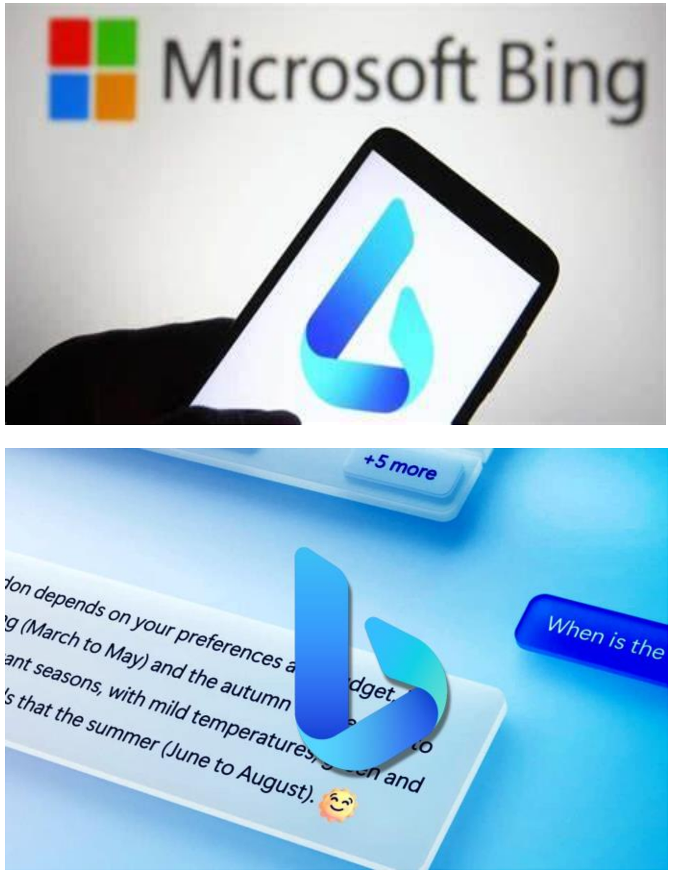
LLM and ChatGPT

- **Chat-based LLM is walking into our daily life!**

OpenAI - ChatGPT



Google -Bard

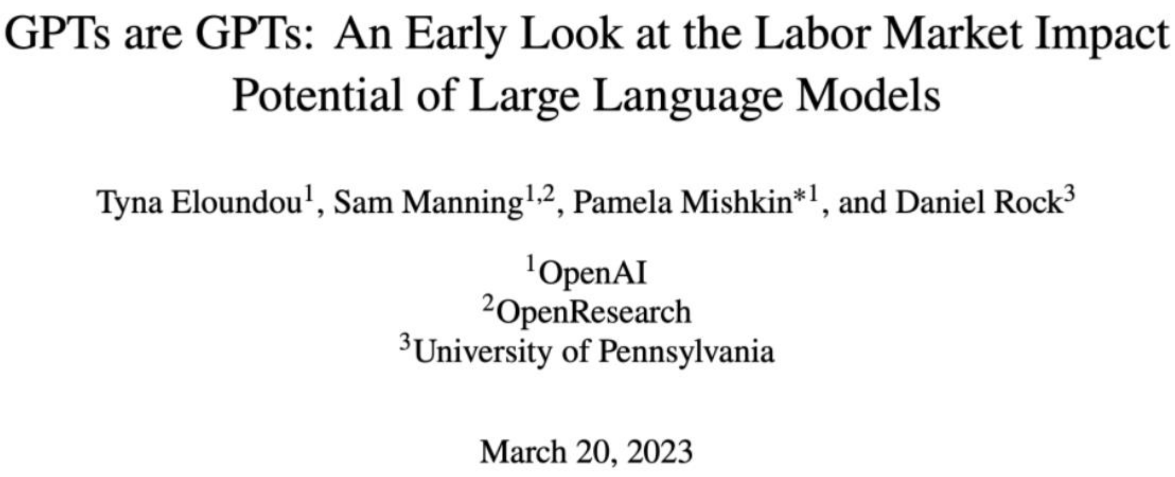


Microsoft-Bing

Credit: Google images.

LLM and ChatGPT

- **LLMs are taking our jobs!**

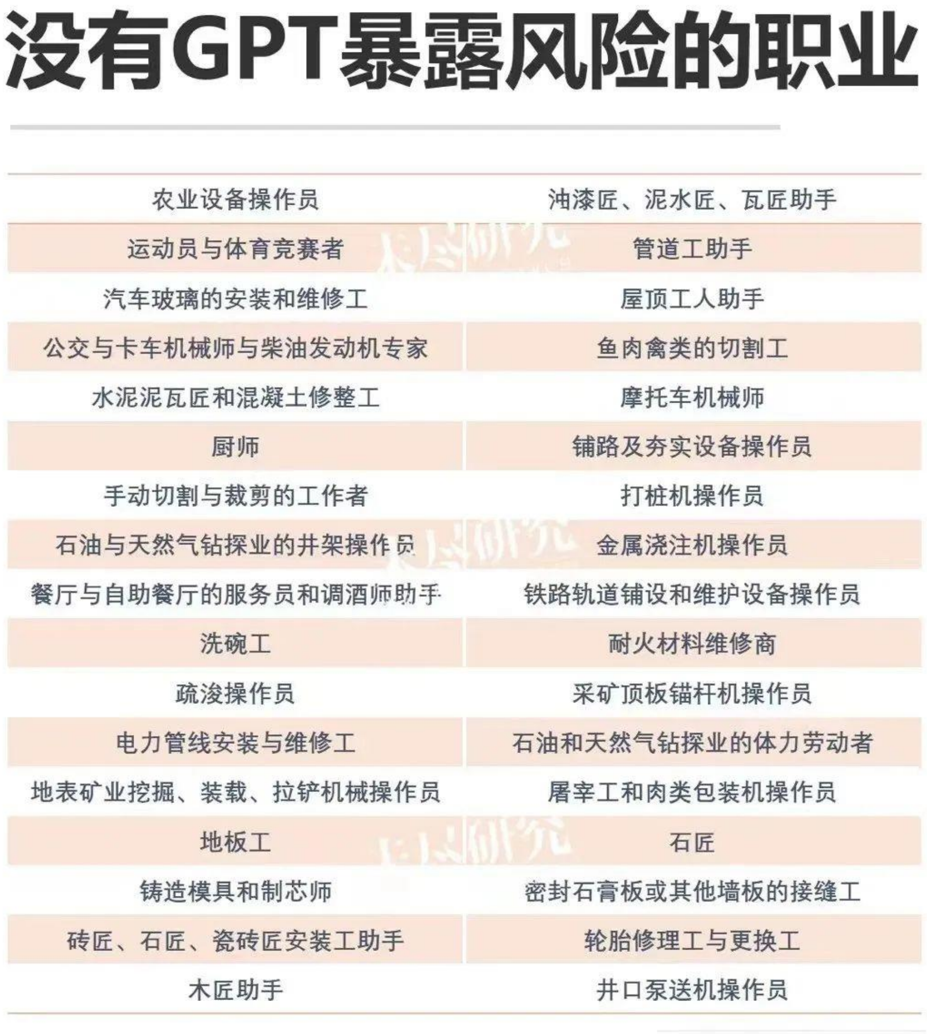


GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models

Tyna Eloundou[1], Sam Manning[1,2], Pamela Mishkin*[1], and Daniel Rock[3]

[1]OpenAI
[2]OpenResearch
[3]University of Pennsylvania

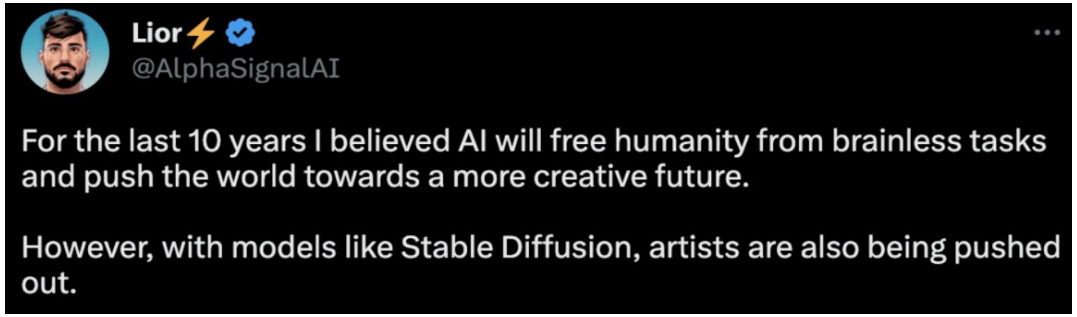
March 20, 2023



没有GPT暴露风险的职业

农业设备操作员	油漆匠、泥水匠、瓦匠助手
运动员与体育竞赛者	管道工助手
汽车玻璃的安装和维修工	屋顶工人助手
公交与卡车机械师与柴油发动机专家	鱼肉禽类的切割工
水泥泥瓦匠和混凝土修整工	摩托车机械师
厨师	铺路及夯实设备操作员
手动切割与裁剪的工作者	打桩机操作员
石油与天然气钻探业的井架操作员	金属浇注机操作员
餐厅与自助餐厅的服务员和调酒师助手	铁路轨道铺设和维护设备操作员
洗碗工	耐火材料维修商
疏浚操作员	采矿顶板锚杆机操作员
电力管线安装与维修工	石油和天然气钻探业的体力劳动者
地表矿业挖掘、装载、拉铲机械操作员	屠宰工和肉类包装机操作员
地板工	石匠
铸造模具和制芯师	密封石膏板或其他墙板的接缝工
砖匠、石匠、瓷砖匠安装工助手	轮胎修理工与更换工
木匠助手	井口泵送机操作员

If human logic and creativity can be replaced. What jobs do you think will be left?



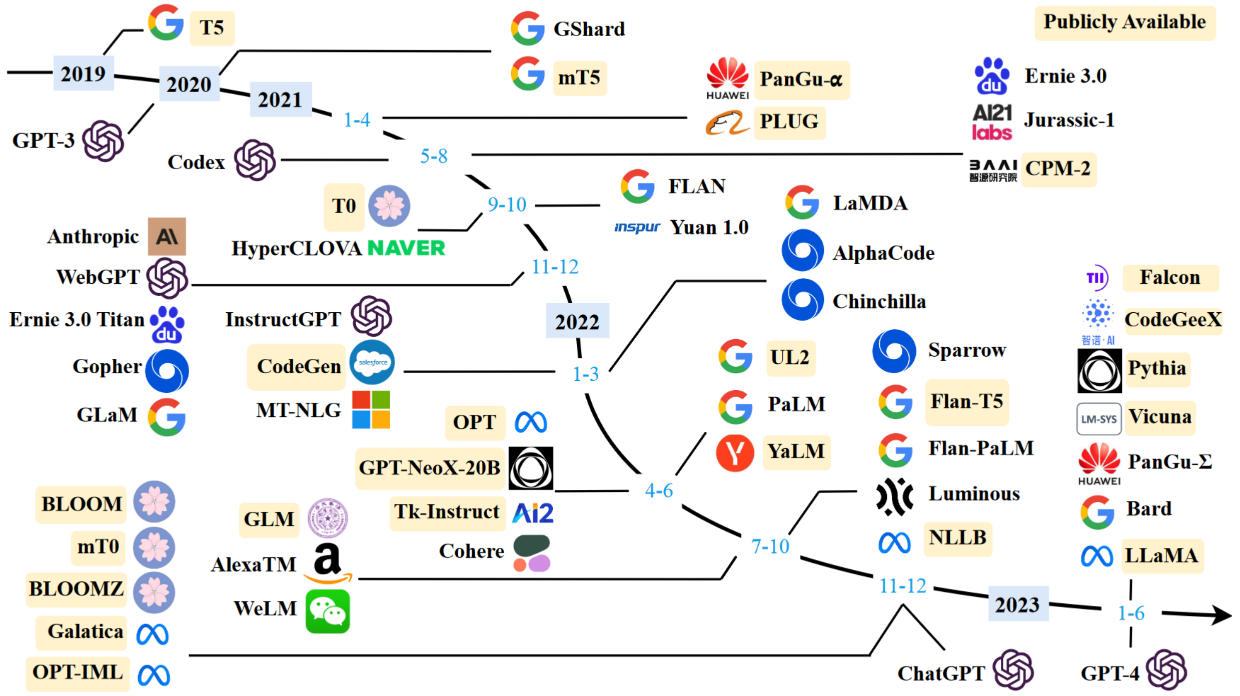
For the last 10 years I believed AI will free humanity from brainless tasks and push the world towards a more creative future.

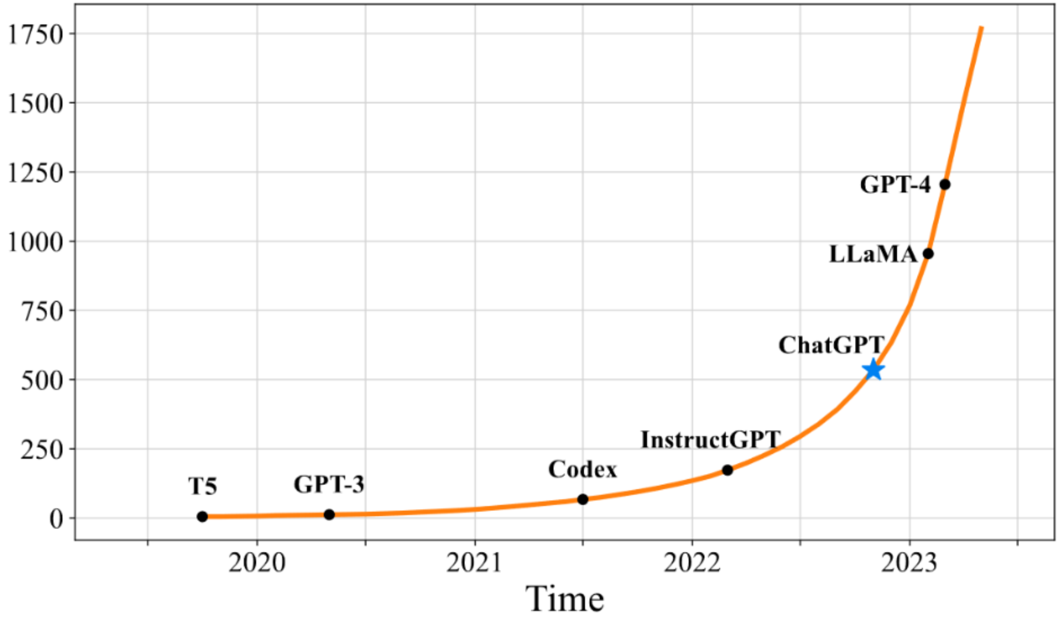
However, with models like Stable Diffusion, artists are also being pushed out.

Credit: Google images.

Trend of LLM in Research Fields

- LLM is rapidly emerging as the hottest direction in research fields



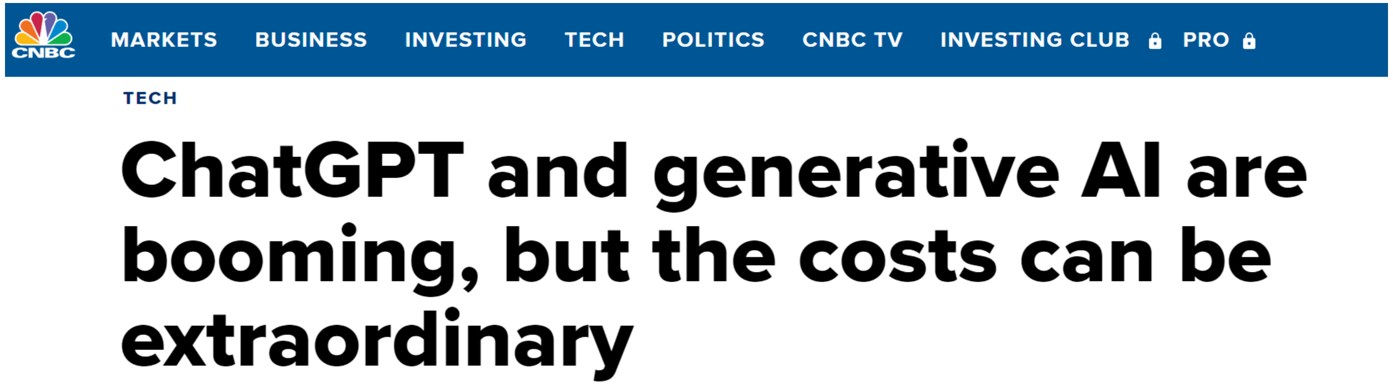


(b) Query="Large Language Model"

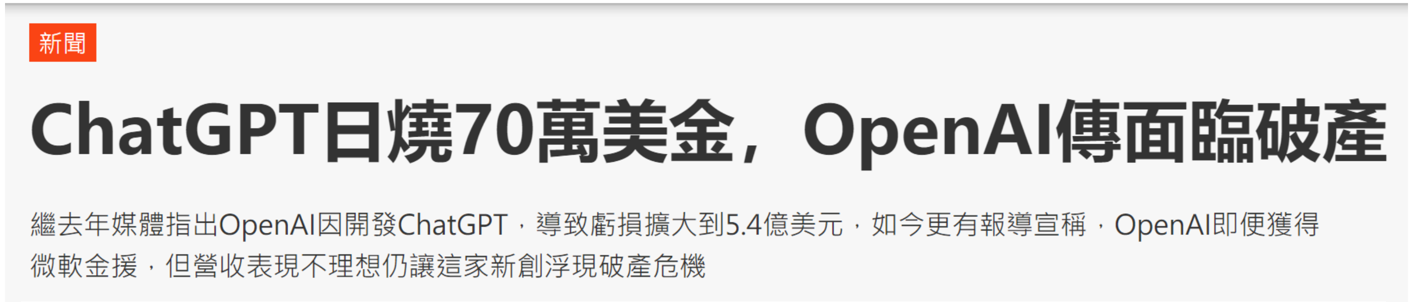
A sharp increase occurs after the release of ChatGPT: the average number of published arXiv papers that contain “large language model” in title or abstract goes **from 0.40 per day to 8.58 per day**.

Credit: https://github.com/RUCAIBox/LLMSurvey.

The Cost Barrier of LLM

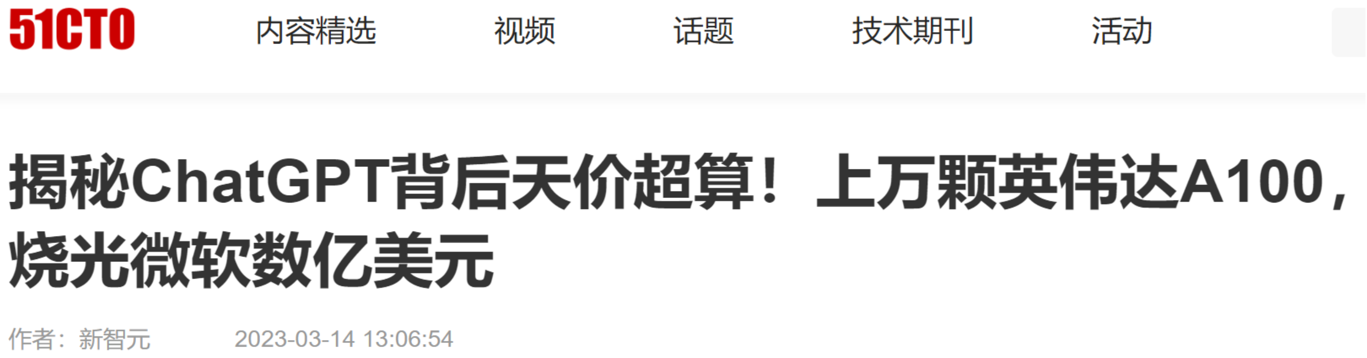


ChatGPT and generative AI are booming, but the costs can be extraordinary



ChatGPT日燒70萬美金，OpenAI傳面臨破產

繼去年媒體指出OpenAI因開發ChatGPT，導致虧損擴大到5.4億美元，如今更有報導宣稱，OpenAI即便獲得微軟金援，但營收表現不理想仍讓這家新創浮現破產危機



揭秘ChatGPT背后天价超算！上万颗英伟达A100，烧光微软数亿美元

作者：新智元 2023-03-14 13:06:54

- The cost of training GPT-3 is estimated to be around $1.4 million, and for some larger LLM models, the training costs range between $2 million to $12 million.
- The cost of operating OpenAI‘s ChatGPT could potentially reach $0.7 million per day.

Credit:Google.

The Cost Barrier of LLM

- How much does it cost when using ChatGPT to finish a writing task?

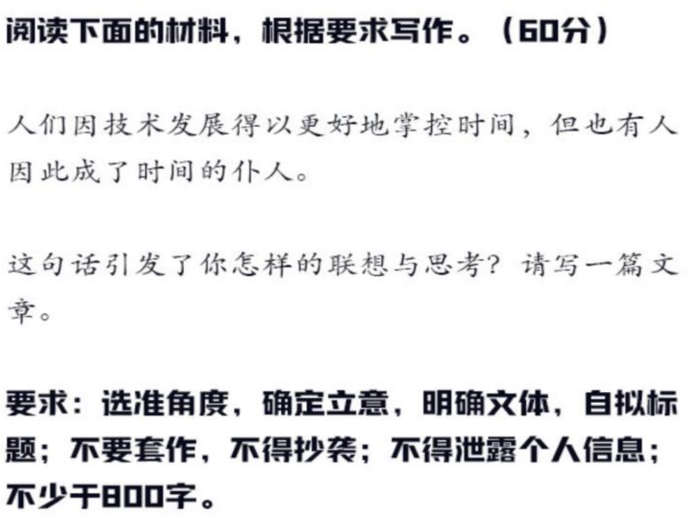
Prices of GPT4

Model	Input	Output
8K context	$0.03 / 1K tokens	$0.06 / 1K tokens
32K context	$0.06 / 1K tokens	$0.12 / 1K tokens

Prices of GPT3.5-Turbo

Model	Input	Output
4K context	$0.0015 / 1K tokens	$0.002 / 1K tokens
16K context	$0.003 / 1K tokens	$0.004 / 1K tokens

全国甲卷



阅读下面的材料，根据要求写作。（60分）

人们因技术发展得以更好地掌控时间，但也有人因此成了时间的仆人。

这句话引发了你怎样的联想与思考？请写一篇文章。

要求：选准角度，确定立意，明确文体，自拟标题；不要套作，不得抄袭；不得泄露个人信息；不少于800字。

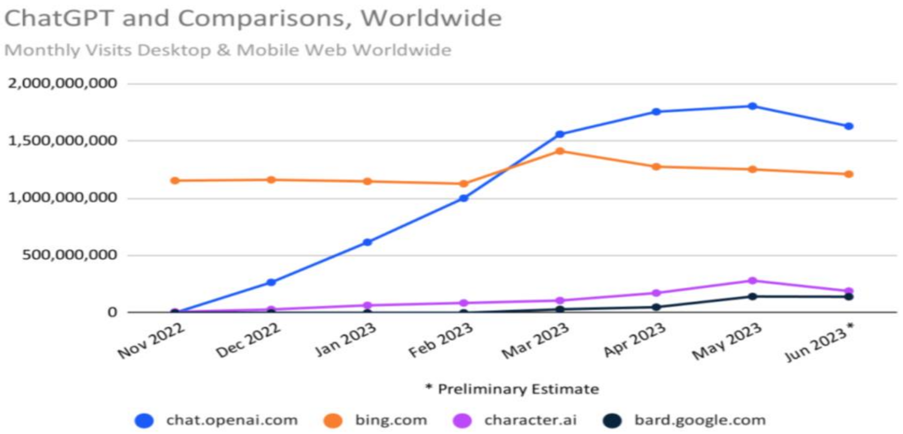
- Input Tokens: 100 tokens
- Output Tokens: 800 tokens

Price of using ChatGPT4:

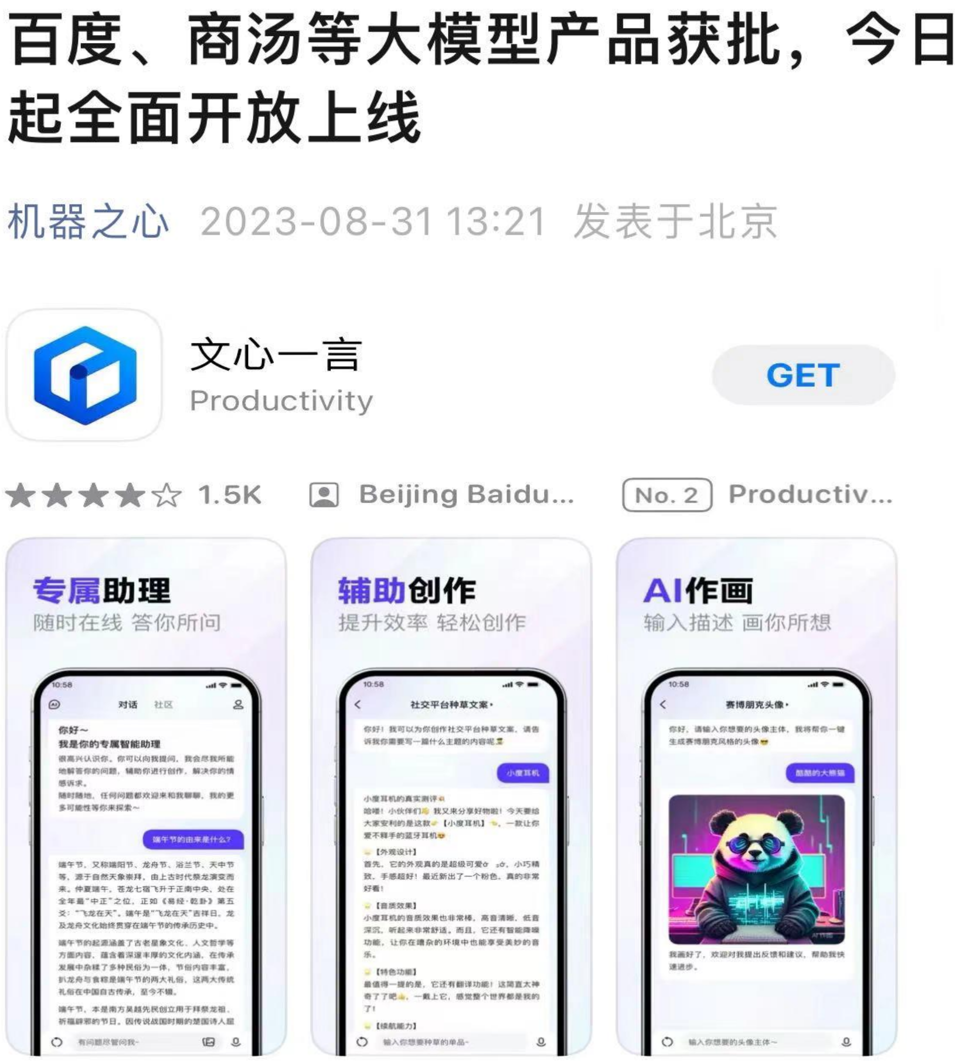
- Input Price: **100 tokens × $0.03/1K = $ 0.003**
- Output Price: **800 tokens × $0.06/1k = $ 0.048**
- **Total Price: $ 0.003 + $ 0.048 = $ 0.051 = ¥ 0.371**

Credit:Google.

The Cost Barrier of LLM



- According to data released in May of this year, the ChatGPT website has surpassed 1.5 billion monthly active users.
- Due to immense cost pressures, companies that fail to capture market share will ultimately be eliminated.



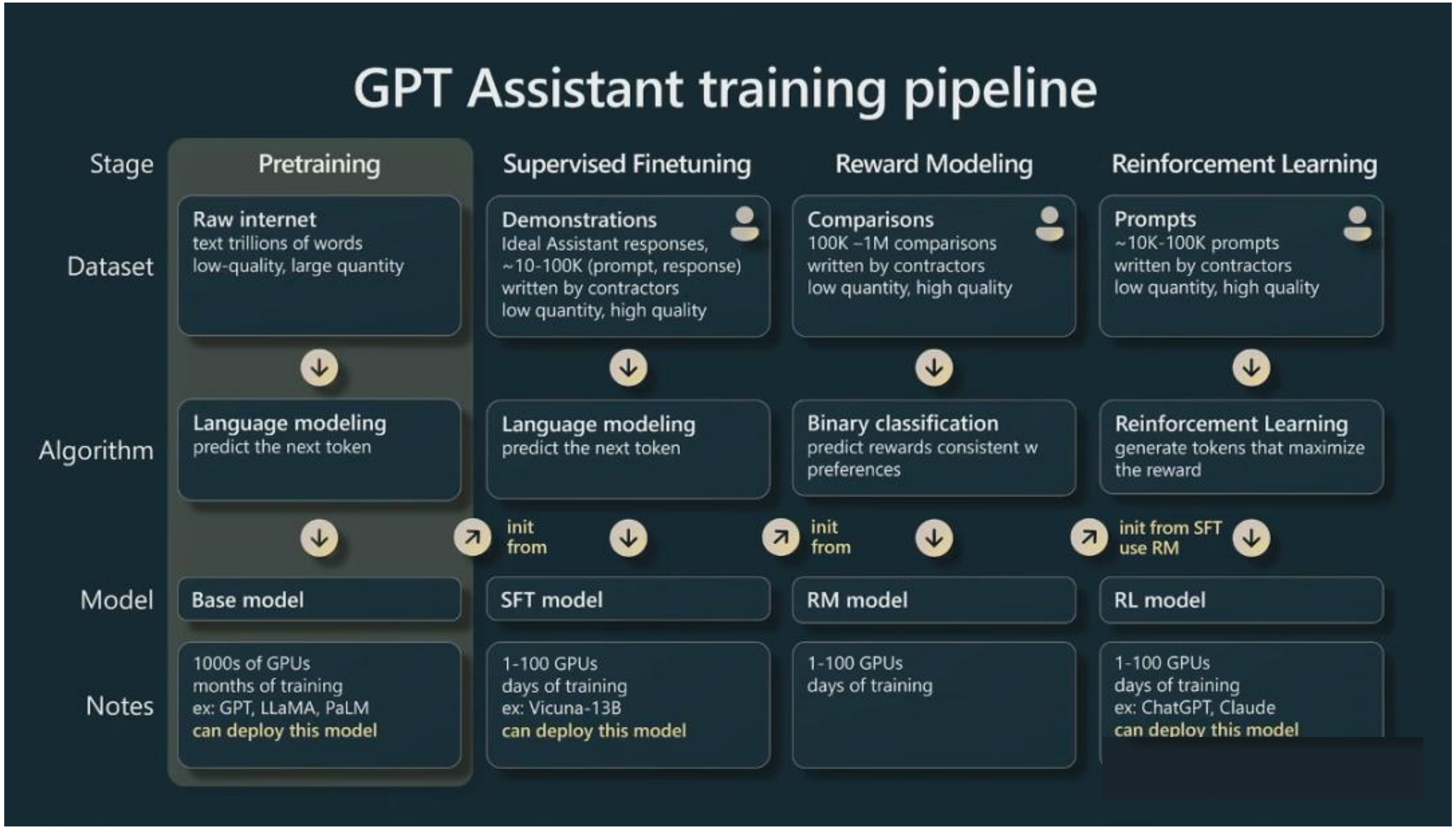
百度、商汤等大模型产品获批，今日起全面开放上线

机器之心 2023-08-31 13:21 发表于北京

Credit:Google.

How to Train your ChatGPT Assistant

- **ChatGPT Training Pipeline**



GPT Assistant training pipeline

Stage	Pretraining	Supervised Finetuning	Reward Modeling	Reinforcement Learning
Dataset	**Raw internet** text trillions of words low-quality, large quantity	**Demonstrations** Ideal Assistant responses, ~10-100K (prompt, response) written by contractors low quantity, high quality	**Comparisons** 100K –1M comparisons written by contractors low quantity, high quality	**Prompts** ~10K-100K prompts written by contractors low quantity, high quality
Algorithm	**Language modeling** predict the next token	**Language modeling** predict the next token	**Binary classification** predict rewards consistent w preferences	**Reinforcement Learning** generate tokens that maximize the reward
		init from	init from	init from SFT use RM
Model	**Base model**	**SFT model**	**RM model**	**RL model**
Notes	1000s of GPUs months of training ex: GPT, LLaMA, PaLM can deploy this model	1-100 GPUs days of training ex: Vicuna-13B can deploy this model	1-100 GPUs days of training	1-100 GPUs days of training ex: ChatGPT, Claude can deploy this model

Four Stage Pipeline

Pretraining

Supervised Finetuning

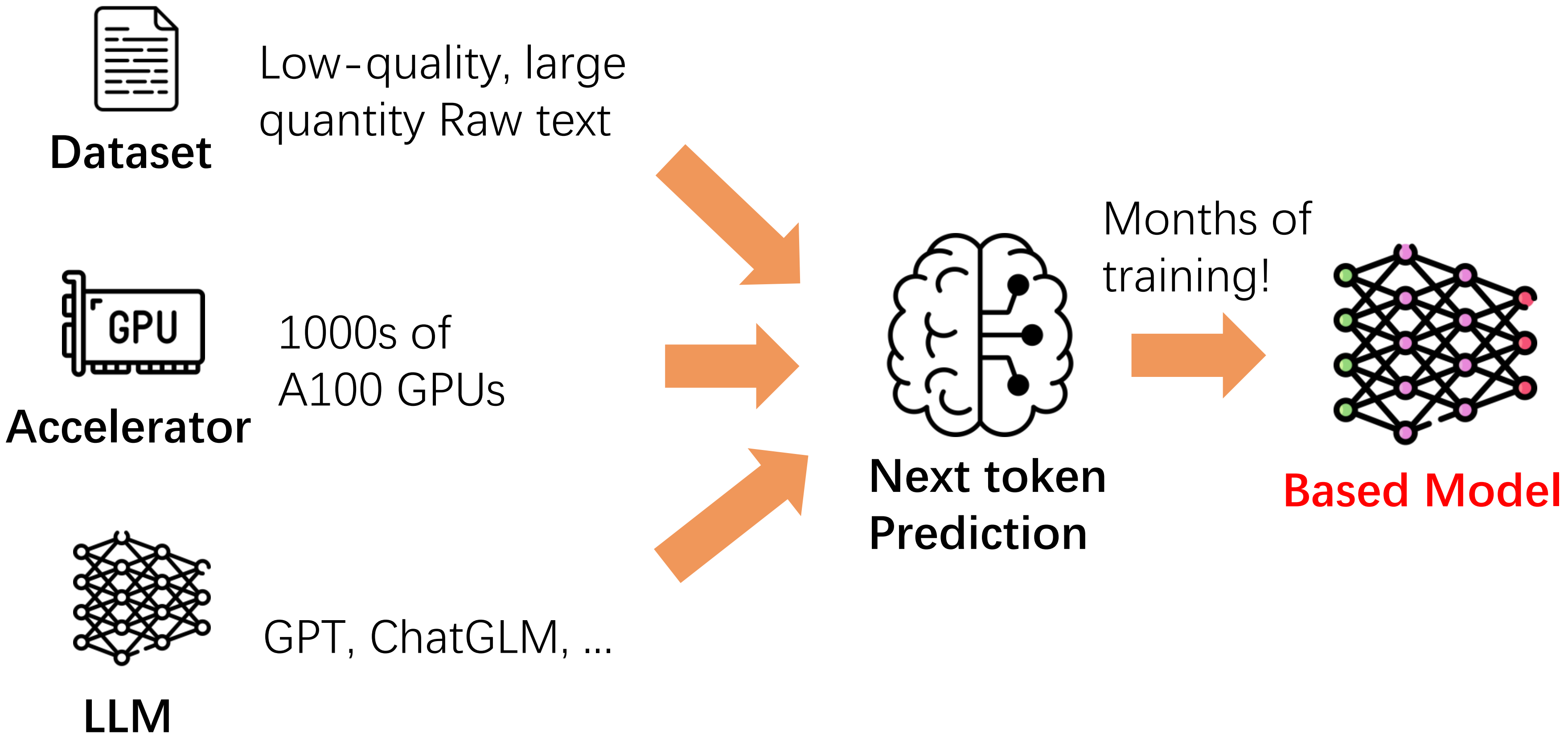
Reward Modeling

Reinforcement Learning

Credit: Andrej Karpathy @ OpenAI.

How to Train a ChatGPT Assistant?

The LLM pre-training takes up 99% of the entire training pipeline's time and typically requires thousands of GPUs for training over several months.



Dataset

Low-quality, large quantity Raw text

Accelerator

1000s of A100 GPUs

LLM

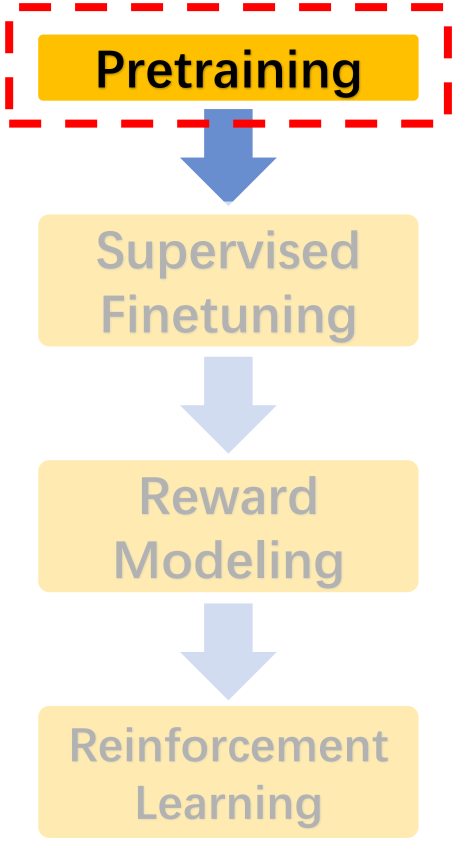
GPT, ChatGLM, …

Next token Prediction

Months of training!

Based Model

Four Stage Pipeline



Pretraining

Supervised Finetuning

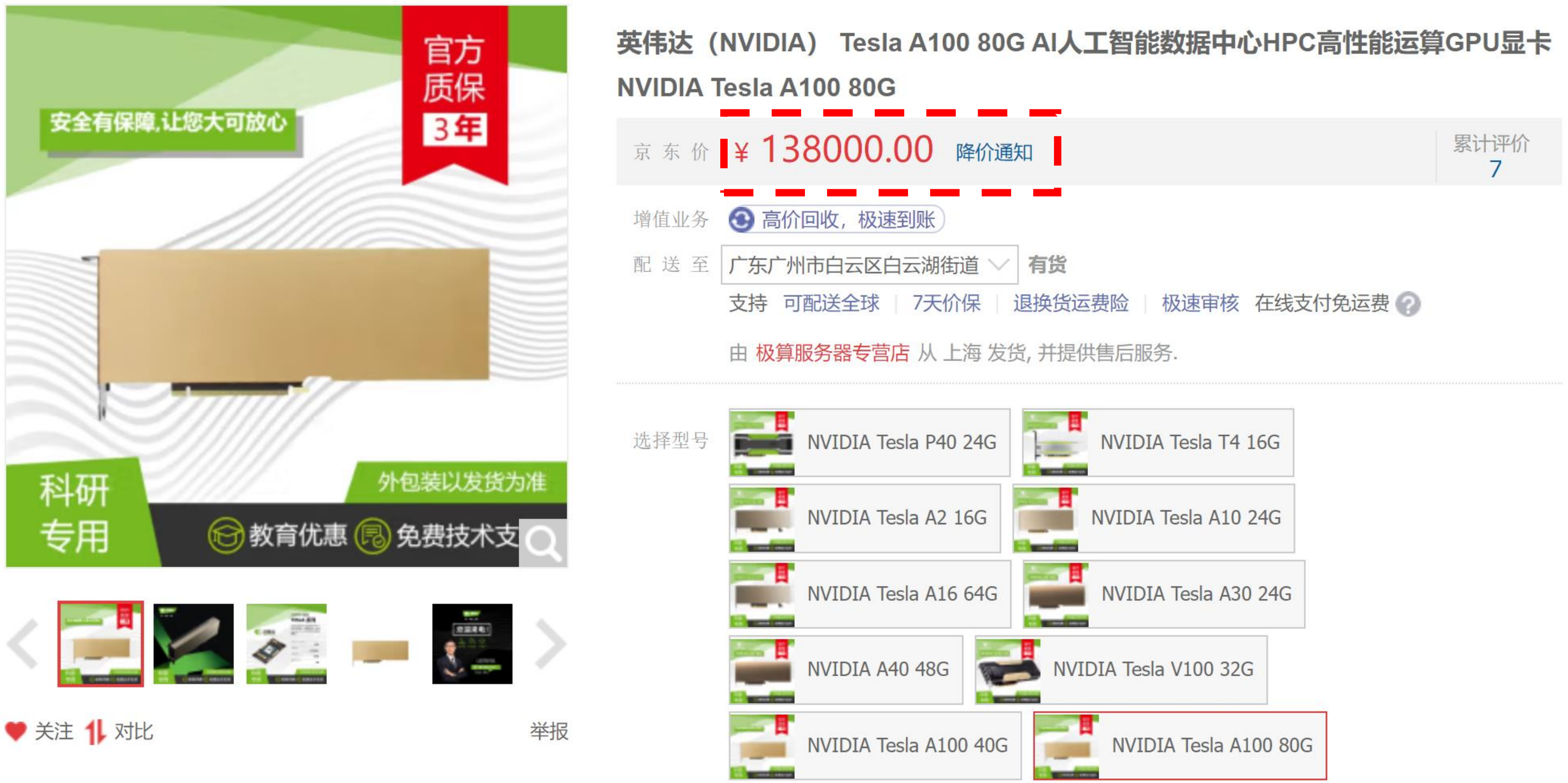
Reward Modeling

Reinforcement Learning

Credit: Andrej Karpathy @ OpenAI.

How to Train a ChatGPT Assistant?

Smaller technology enterprises and educational research labs often can't afford the cost of pre-training LLMs.



官方质保 3年

安全有保障,让您大可放心

科研专用

外包装以发货为准

教育优惠 免费技术支

关注 对比 举报

英伟达（NVIDIA） Tesla A100 80G AI人工智能数据中心HPC高性能运算GPU显卡

NVIDIA Tesla A100 80G

京 东 价 ¥ 138000.00 降价通知

累计评价 7

增值业务 高价回收，极速到账

配 送 至 广东广州市白云区白云湖街道 有货

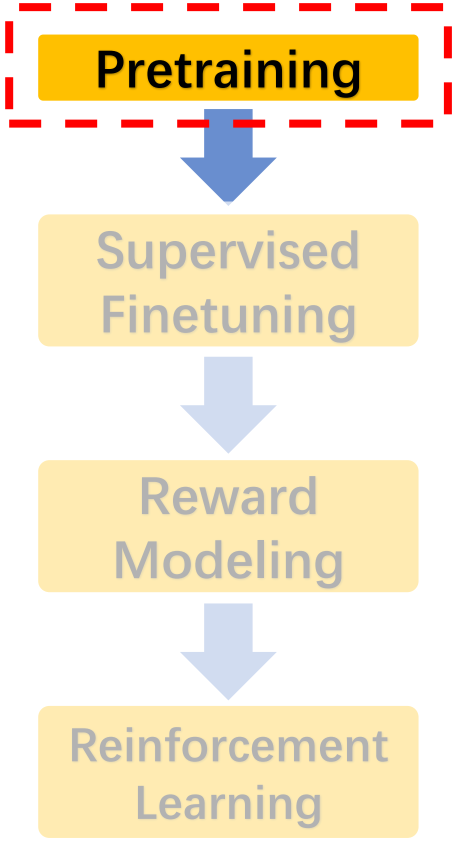
支持 可配送全球 | 7天价保 | 退换货运费险 | 极速审核 在线支付免运费

由 极算服务器专营店 从 上海 发货, 并提供售后服务.

选择型号

NVIDIA Tesla P40 24G
NVIDIA Tesla T4 16G
NVIDIA Tesla A2 16G
NVIDIA Tesla A10 24G
NVIDIA Tesla A16 64G
NVIDIA Tesla A30 24G
NVIDIA A40 48G
NVIDIA Tesla V100 32G
NVIDIA Tesla A100 40G
NVIDIA Tesla A100 80G

Four Stage Pipeline



Pretraining

Supervised Finetuning

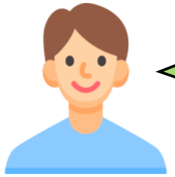
Reward Modeling

Reinforcement Learning

Credit: Andrej Karpathy @ OpenAI.

How to Train a ChatGPT Assistant?

- **Based model are not Chat Assistant!**

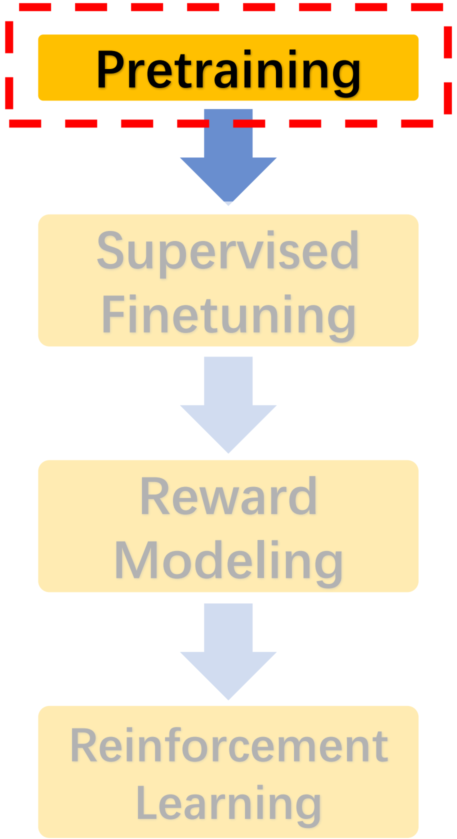


Write a poem about bread!

Write a poem about someone.
Write a poem about angel.
Write a poem about basketball.

The training task of pre-trained language model is to predict the next token, rather than engaging in QA (Question-Answering) dialogues. It may use more questions to answer a question.

Four Stage Pipeline



Pretraining

Supervised Finetuning

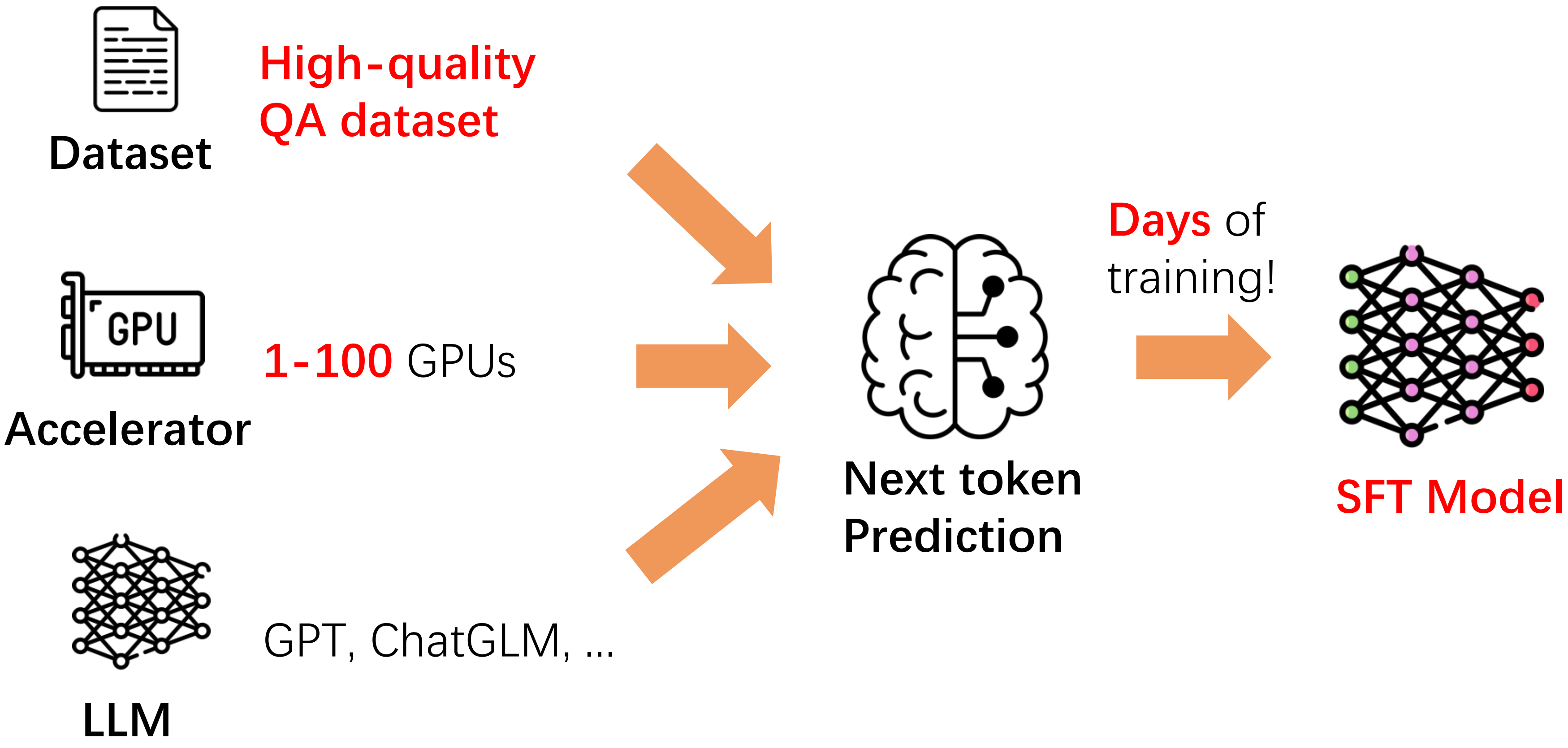
Reward Modeling

Reinforcement Learning

Credit: Andrej Karpathy @ OpenAI.

How to Train a ChatGPT Assistant?

During the Supervised Finetuning of LLM, workers are hired to gather high-quality QA data and the based-model is fine-tuned, requiring fewer GPUs and just days to train.



Dataset

High-quality QA dataset

Accelerator

1-100 GPUs

LLM

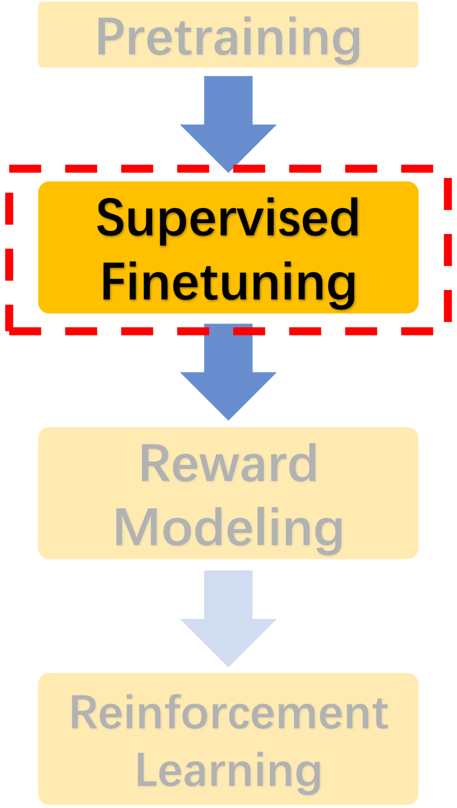
GPT, ChatGLM, …

Next token Prediction

Days of training!

SFT Model

Four Stage Pipeline



Pretraining

Supervised Finetuning

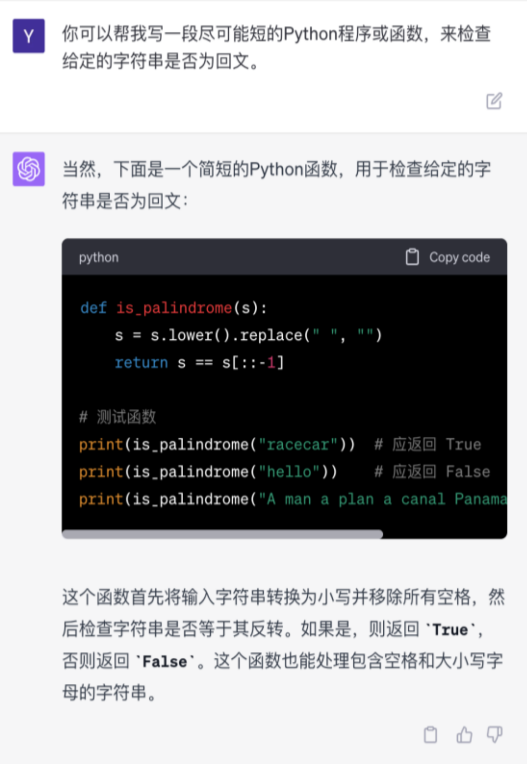
Reward Modeling

Reinforcement Learning

Credit: Andrej Karpathy @ OpenAI.

How to Train a ChatGPT Assistant?

Multiple answers are generated from the same prompt based on SFT model. Workers rank these answers to compile a large ranking dataset, which is then used to train a Transformer-based Reward Model.



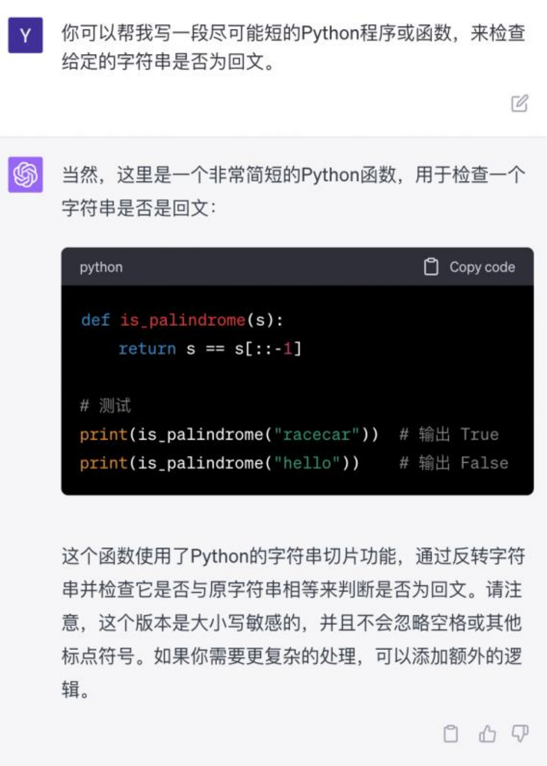
Y 你可以帮我写一段尽可能短的Python程序或函数，来检查给定的字符串是否为回文。

当然，下面是一个简短的Python函数，用于检查给定的字符串是否为回文：

```python
def is_palindrome(s):
    s = s.lower().replace(" ", "")
    return s == s[::-1]

# 测试函数
print(is_palindrome("racecar"))  # 应返回 True
print(is_palindrome("hello"))    # 应返回 False
print(is_palindrome("A man a plan a canal Panama
```

这个函数首先将输入字符串转换为小写并移除所有空格，然后检查字符串是否等于其反转。如果是，则返回 `True`，否则返回 `False`。这个函数也能处理包含空格和大小写字母的字符串。



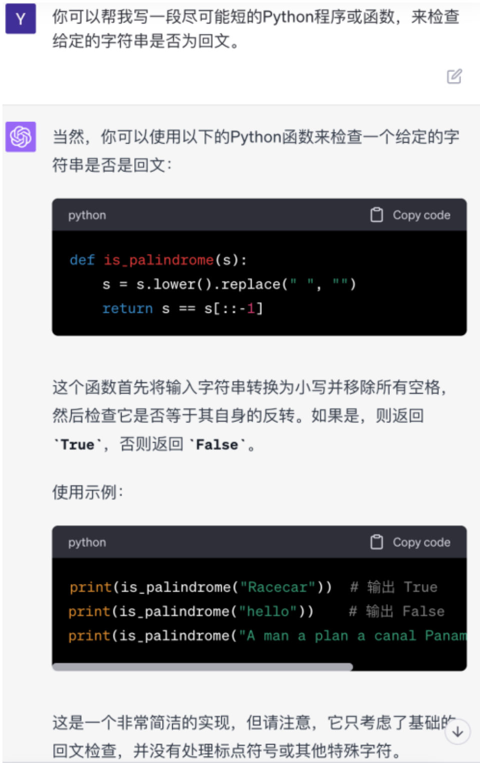
Y 你可以帮我写一段尽可能短的Python程序或函数，来检查给定的字符串是否为回文。

当然，这里是一个非常简短的Python函数，用于检查一个字符串是否是回文：

```python
def is_palindrome(s):
    return s == s[::-1]

# 测试
print(is_palindrome("racecar"))  # 输出 True
print(is_palindrome("hello"))    # 输出 False
```

这个函数使用了Python的字符串切片功能，通过反转字符串并检查它是否与原字符串相等来判断是否为回文。请注意，这个版本是大小写敏感的，并且不会忽略空格或其他标点符号。如果你需要更复杂的处理，可以添加额外的逻辑。



Y 你可以帮我写一段尽可能短的Python程序或函数，来检查给定的字符串是否为回文。

当然，你可以使用以下的Python函数来检查一个给定的字符串是否是回文：

```python
def is_palindrome(s):
    s = s.lower().replace(" ", "")
    return s == s[::-1]
```

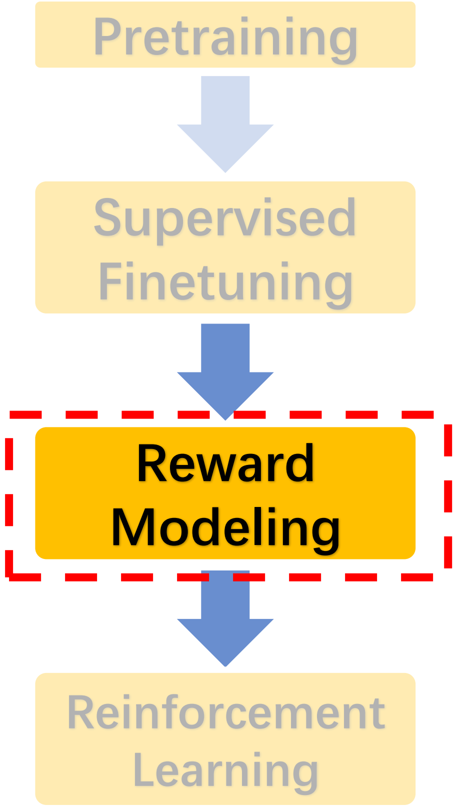
这个函数首先将输入字符串转换为小写并移除所有空格，然后检查它是否等于其自身的反转。如果是，则返回 `True`，否则返回 `False`。

使用示例：

```python
print(is_palindrome("Racecar"))  # 输出 True
print(is_palindrome("hello"))    # 输出 False
print(is_palindrome("A man a plan a canal Panam
```

这是一个非常简洁的实现，但请注意，它只考虑了基础的回文检查，并没有处理标点符号或其他特殊字符。

Four Stage Pipeline



Pretraining

Supervised Finetuning

Reward Modeling

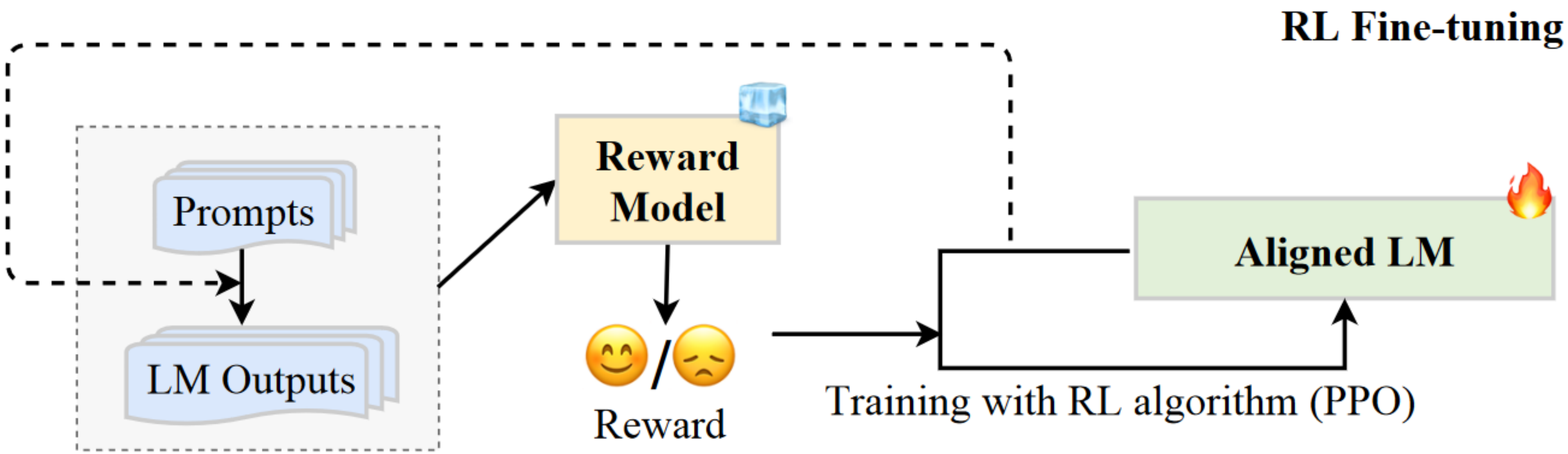
Reinforcement Learning

Credit: Andrej Karpathy @ OpenAI.

How to Train a ChatGPT Assistant?

In reinforcement learning, policy gradient algorithm are used to amplify the generation probabilities of "favorable" responses and minimize those of "toxic" ones.

- **Action Space: all the vocabulary**
- **State: currently generated token sequence**
- **Reward: Provided by Reward Model**

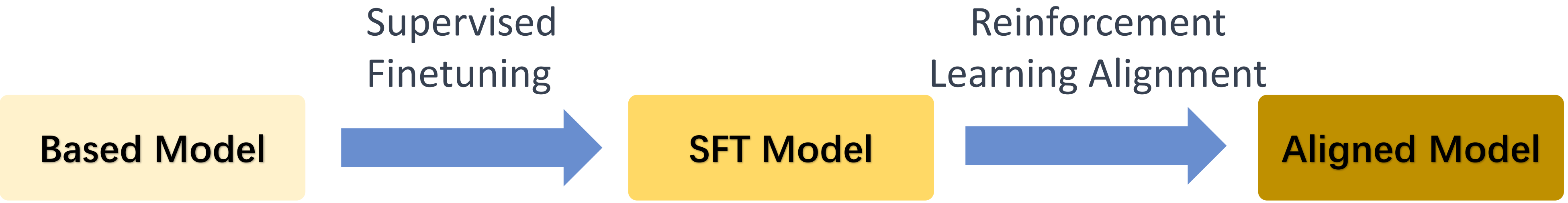


Four Stage Pipeline

Pretraining
↓
Supervised Finetuning
↓
Reward Modeling
↓
Reinforcement Learning

Credit: Andrej Karpathy @ OpenAI.

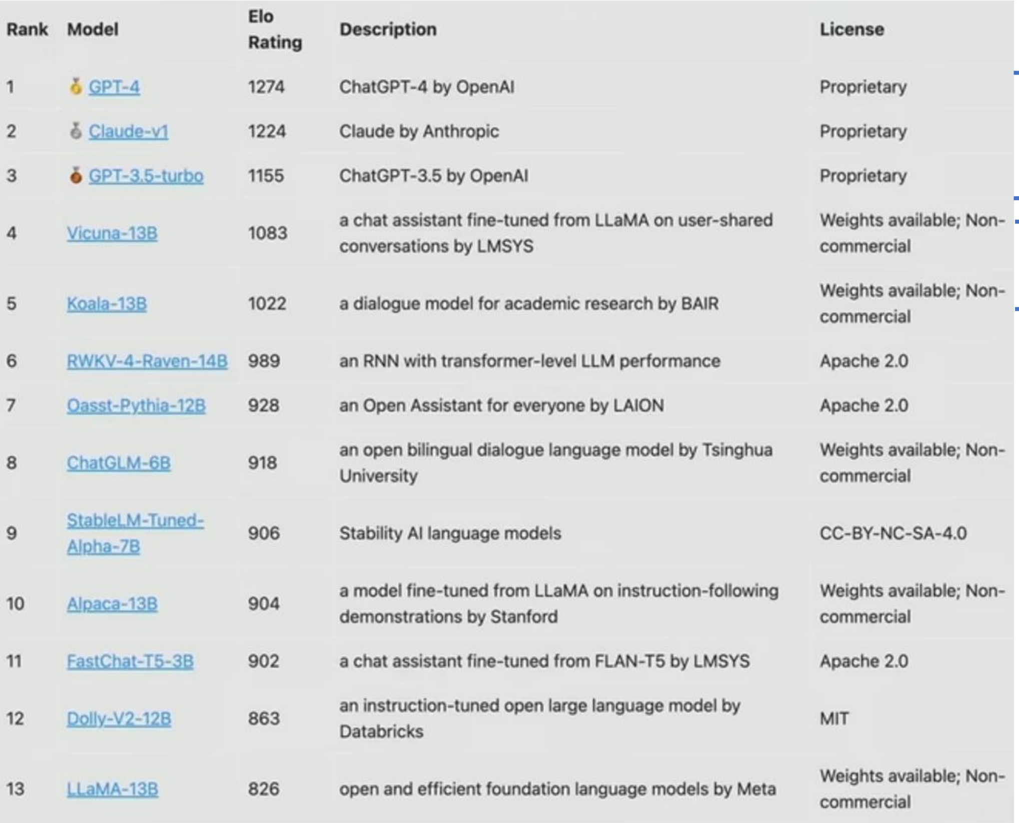
How to Train a ChatGPT Assistant?



Supervised Finetuning

Reinforcement Learning Alignment

Based Model → **SFT Model** → **Aligned Model**



Rank	Model	Elo Rating	Description	License
1	🥇 GPT-4	1274	ChatGPT-4 by OpenAI	Proprietary
2	🥈 Claude-v1	1224	Claude by Anthropic	Proprietary
3	🥉 GPT-3.5-turbo	1155	ChatGPT-3.5 by OpenAI	Proprietary
4	Vicuna-13B	1083	a chat assistant fine-tuned from LLaMA on user-shared conversations by LMSYS	Weights available; Non-commercial
5	Koala-13B	1022	a dialogue model for academic research by BAIR	Weights available; Non-commercial
6	RWKV-4-Raven-14B	989	an RNN with transformer-level LLM performance	Apache 2.0
7	Oasst-Pythia-12B	928	an Open Assistant for everyone by LAION	Apache 2.0
8	ChatGLM-6B	918	an open bilingual dialogue language model by Tsinghua University	Weights available; Non-commercial
9	StableLM-Tuned-Alpha-7B	906	Stability AI language models	CC-BY-NC-SA-4.0
10	Alpaca-13B	904	a model fine-tuned from LLaMA on instruction-following demonstrations by Stanford	Weights available; Non-commercial
11	FastChat-T5-3B	902	a chat assistant fine-tuned from FLAN-T5 by LMSYS	Apache 2.0
12	Dolly-V2-12B	863	an instruction-tuned open large language model by Databricks	MIT
13	LLaMA-13B	826	open and efficient foundation language models by Meta	Weights available; Non-commercial

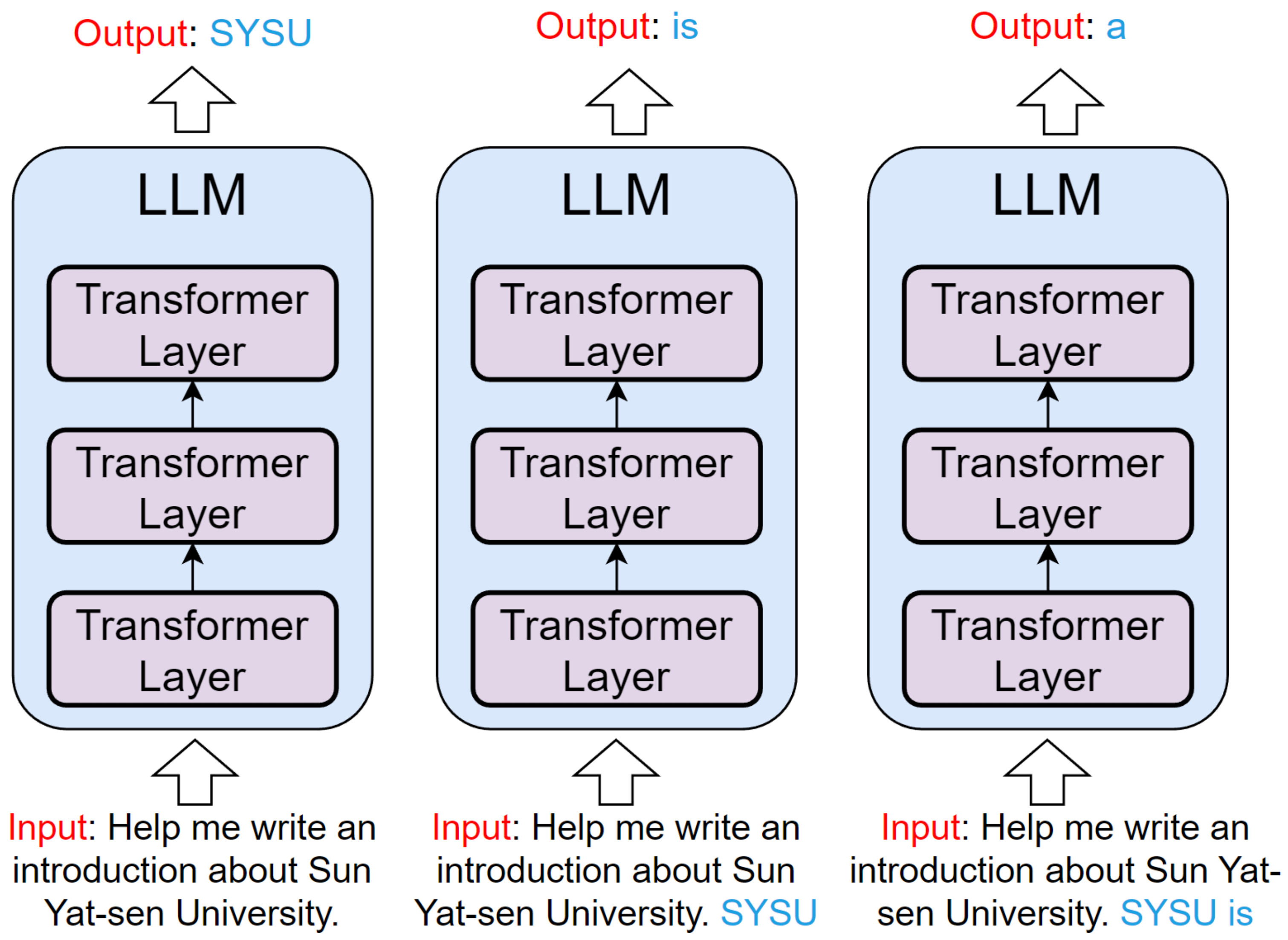
Aligned Model

SFT Model

- The SFT model significantly outperforms the based pre-trained model in QA tasks.
- The Aligned Model can further filter out expressions from the SFT model's output that are harmful or not in line with human norms.

Credit: Andrej Karpathy @ OpenAI.

What happens when LLM generates an answer?

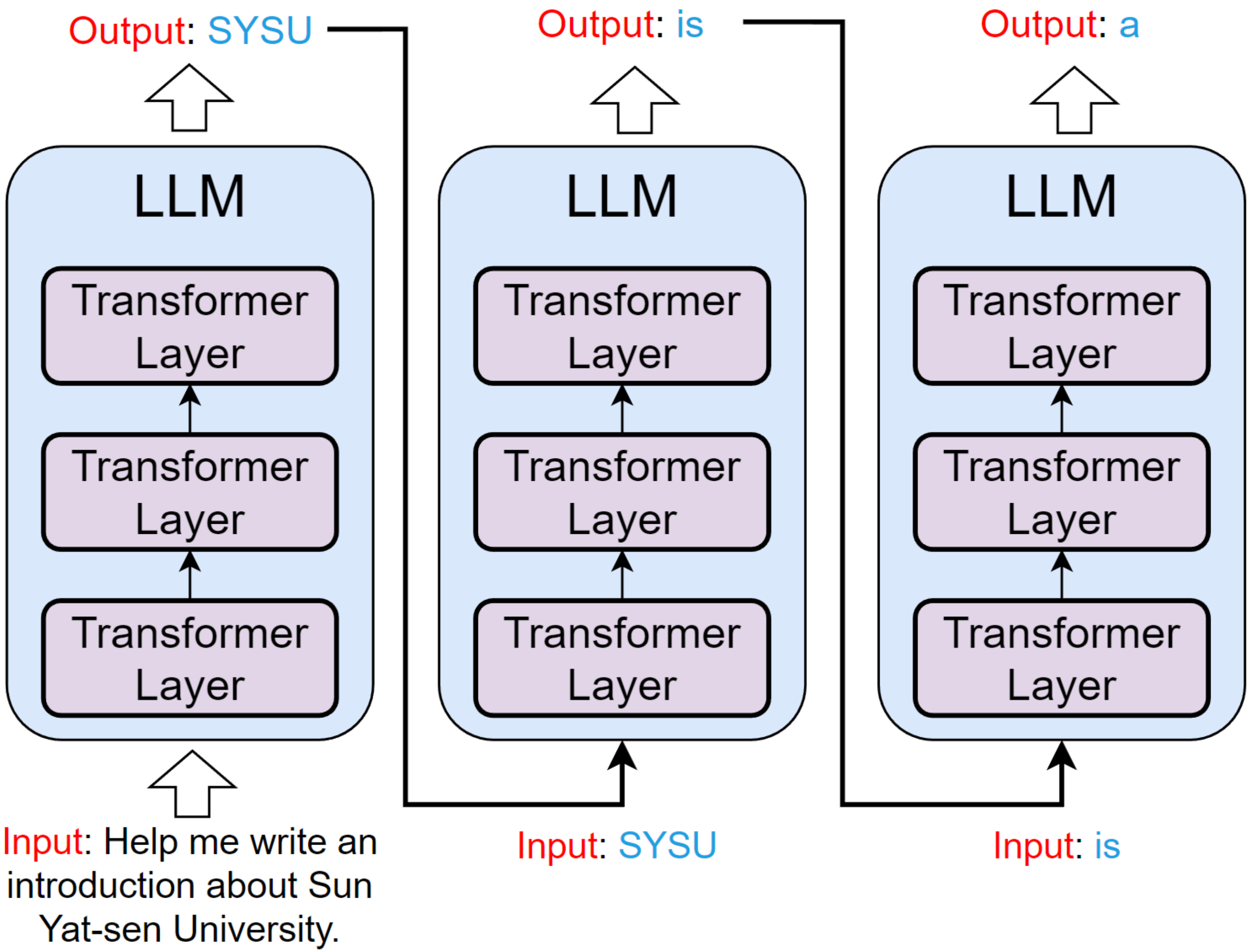

Output: SYSU
Output: is
Output: a
LLM
Transformer Layer
Transformer Layer
Transformer Layer
LLM
Transformer Layer
Transformer Layer
Transformer Layer
LLM
Transformer Layer
Transformer Layer
Transformer Layer
Input: Help me write an introduction about Sun Yat-sen University.
Input: Help me write an introduction about Sun Yat-sen University. SYSU
Input: Help me write an introduction about Sun Yat-sen University. SYSU is

Autoregressive

- When a LLM generates response, it uses its own previous outputs as inputs for future predictions, forming a chain of dependencies.
- This autoregressive behavior allows the model to generate coherent and contextually relevant text.

Credit: Andrej Karpathy @ OpenAI.

What happens when LLM generates an answer?

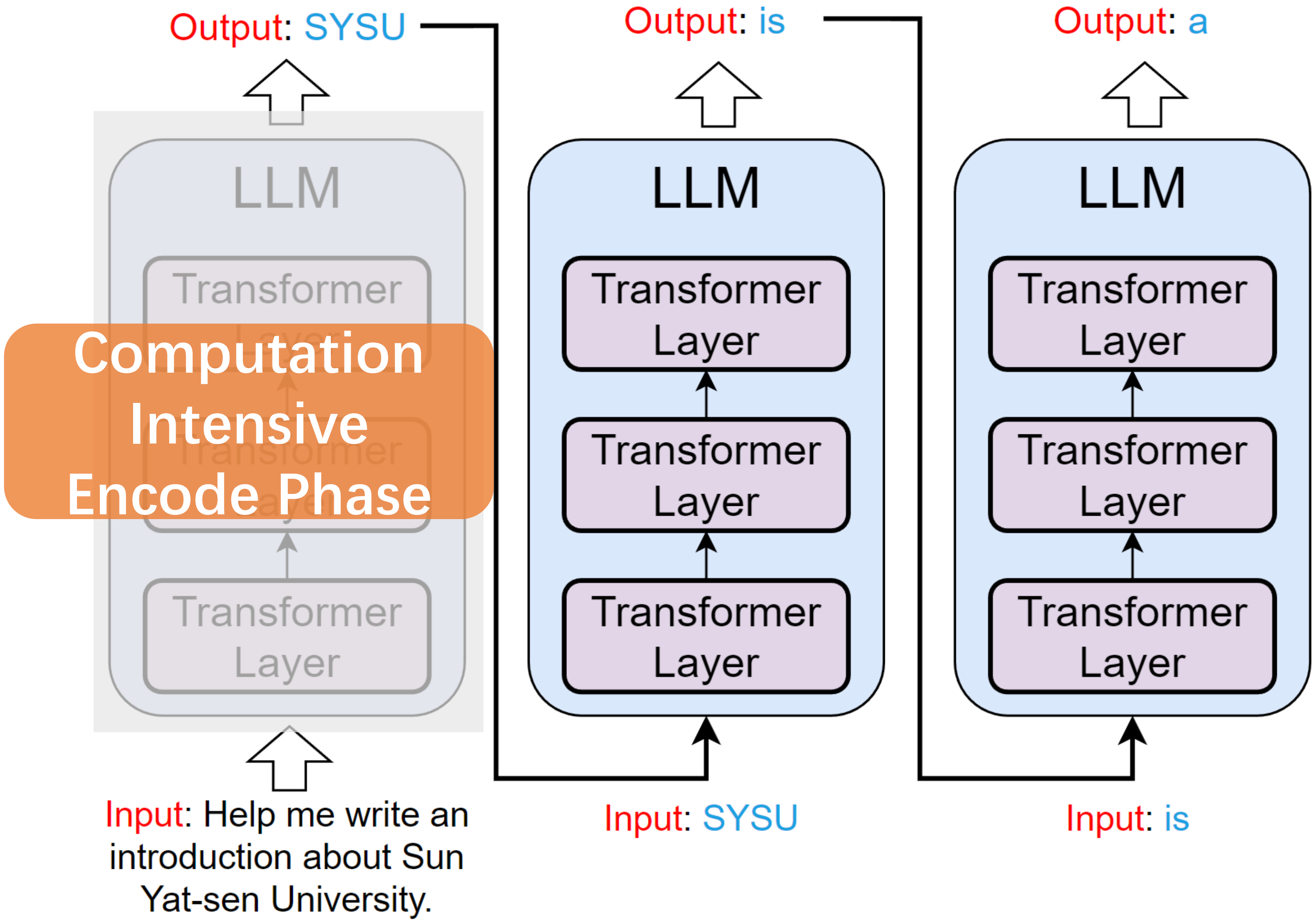


Intermediate Cache

- LLM will cache previous computational results (such as the calculations from the black parts of Self-Attention) in memory to avoid redundant calculations.
- The inference process of LLM can be divided into two distinct phases: Encode and Decode

Credit: Andrej Karpathy @ OpenAI.

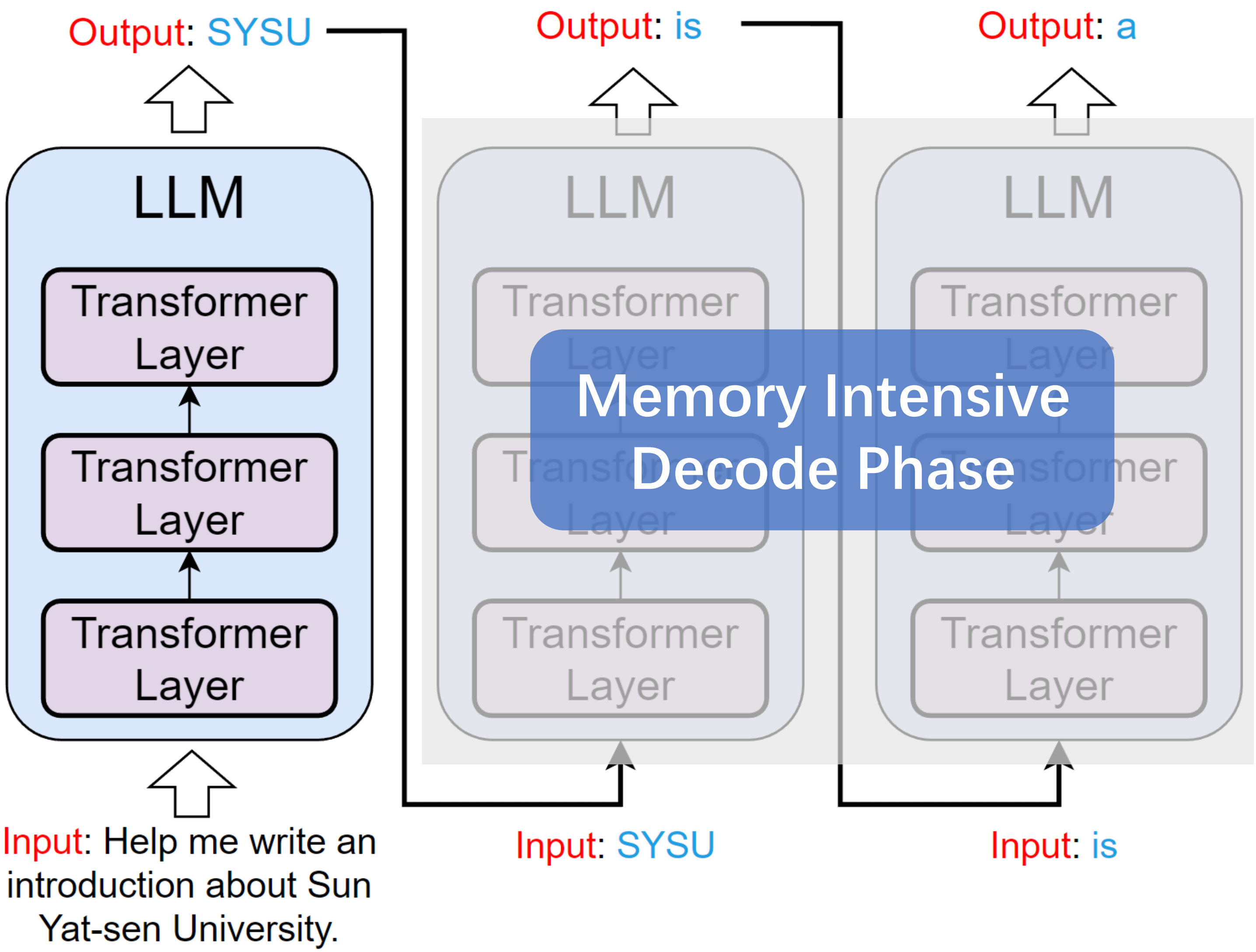
What happens when LLM generates an answer?



- Encode Phase: Inference on LLM's input involves potentially hundreds or even thousands of tokens, making it computationally intensive.

Credit: Andrej Karpathy @ OpenAI.

What happens when LLM generates an answer?

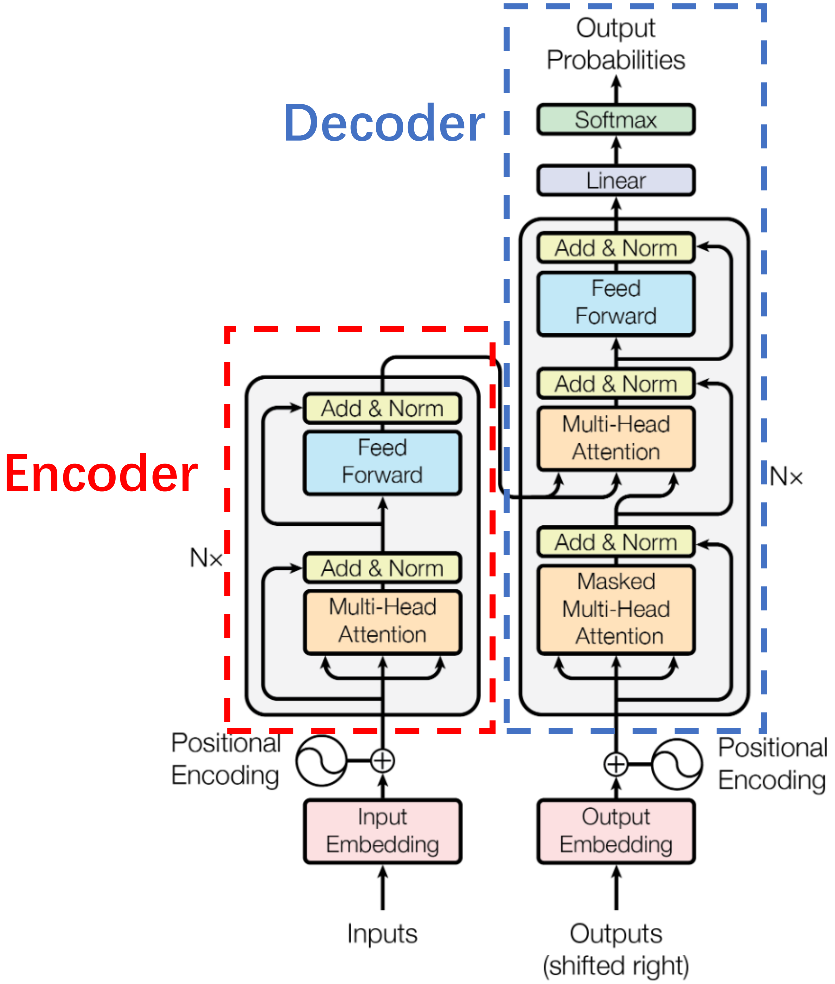


- **Encode Phase**: Inference on LLM's input involves potentially hundreds or even thousands of tokens, making it computationally intensive.
- **Decode Phase**: each previously predicted token is inputted one at a time, requiring frequent retrieval of intermediate cache from storage.

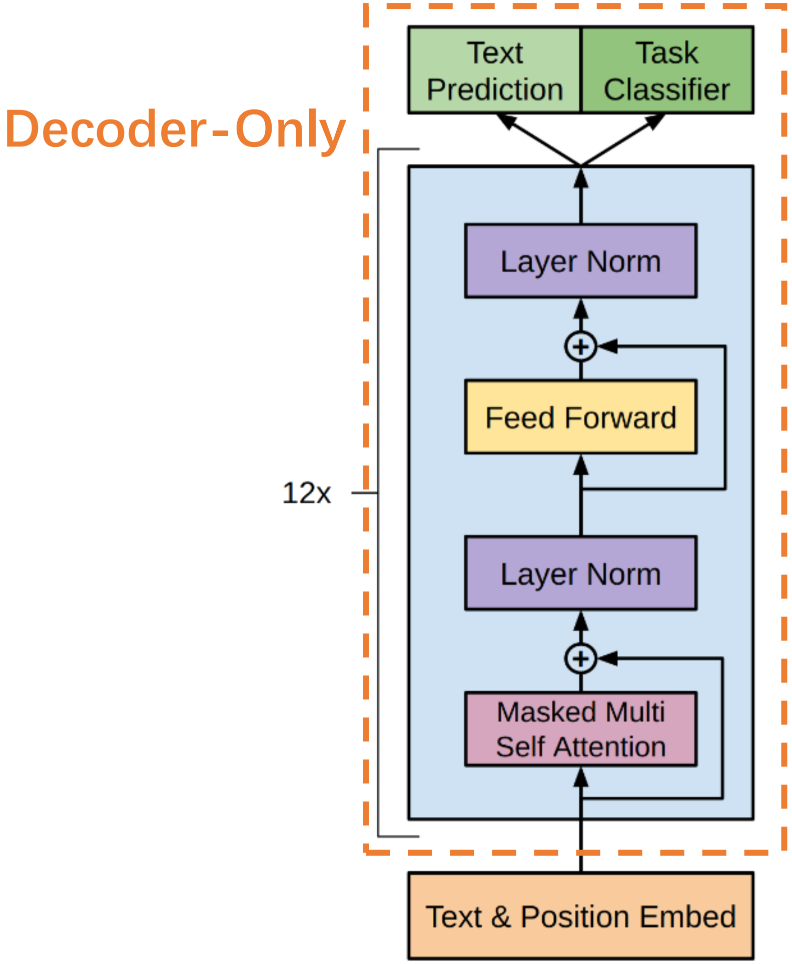
Credit: Andrej Karpathy @ OpenAI.

The Backbone Architecture in LLM

- The vast majority of LLMs are based on the **Transformer architecture**.



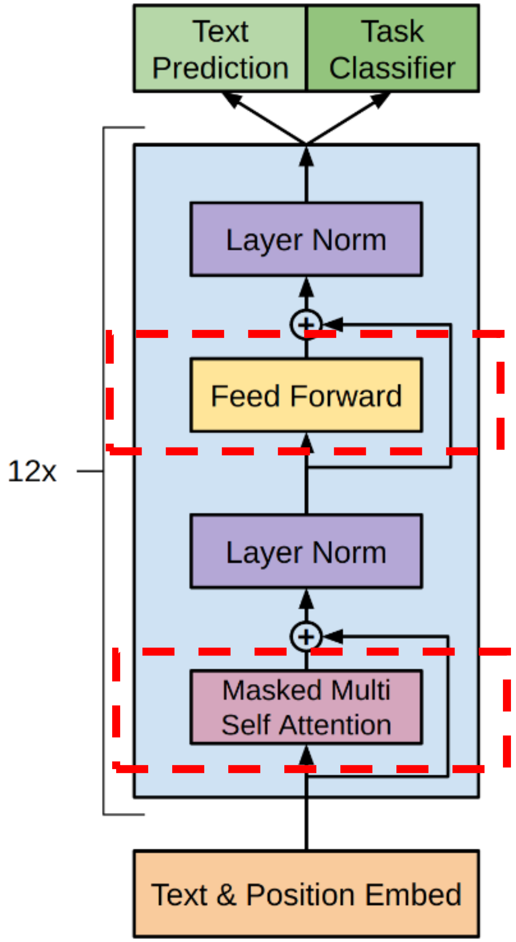
Attention is all you need! (NIPS 2017)



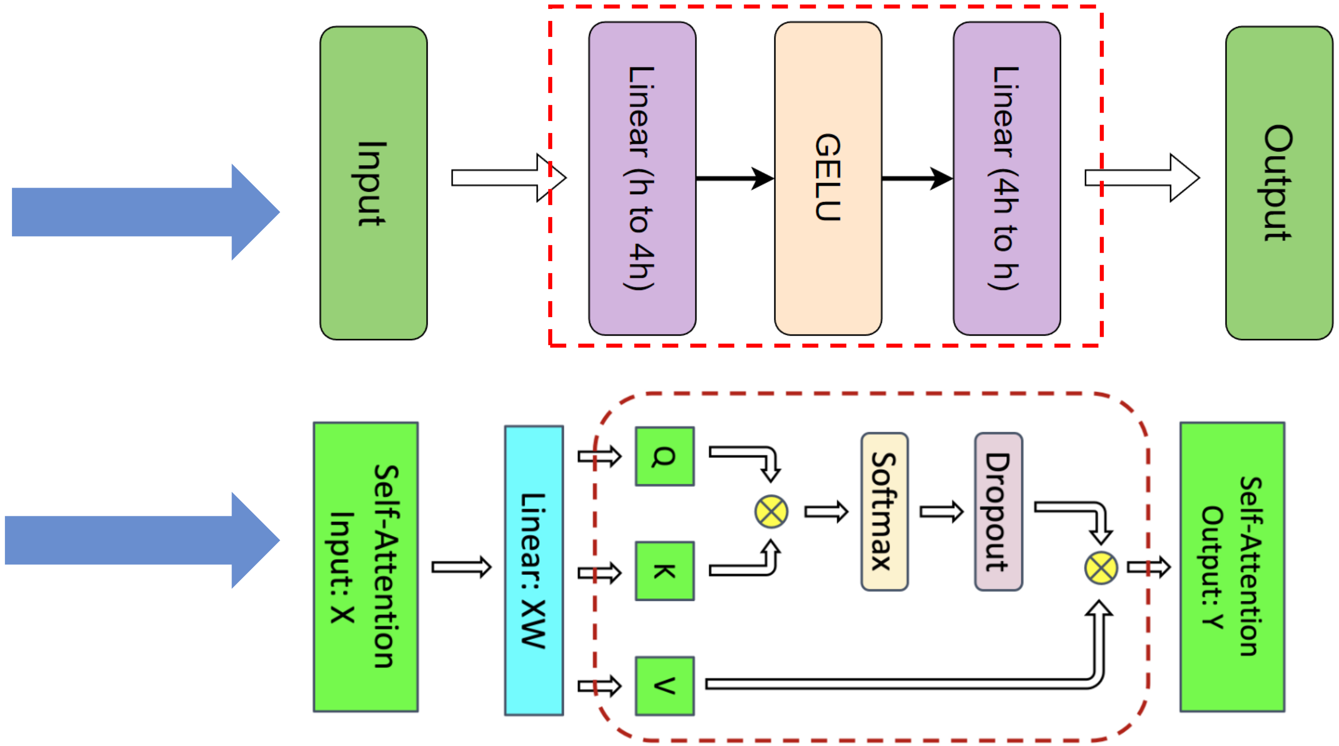
GPT Series (2019-now)

The Backbone Architecture in LLM

- The vast majority of LLMs are based on the **Transformer architecture**.



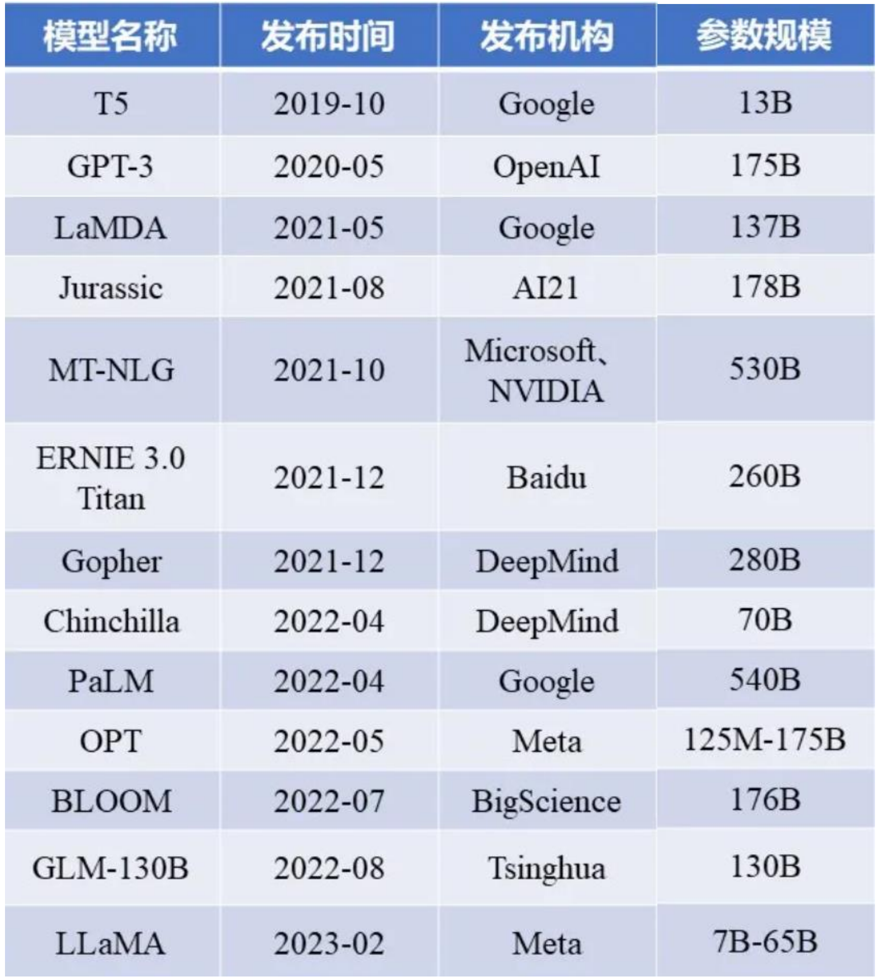
- The computations in a Transformer Layer consist of **several large-scale matrix multiplications**.



Credit:Reducing Activation Recomputation in Large Transformer Models.

Parameter Size and Memory Footprint in LLM

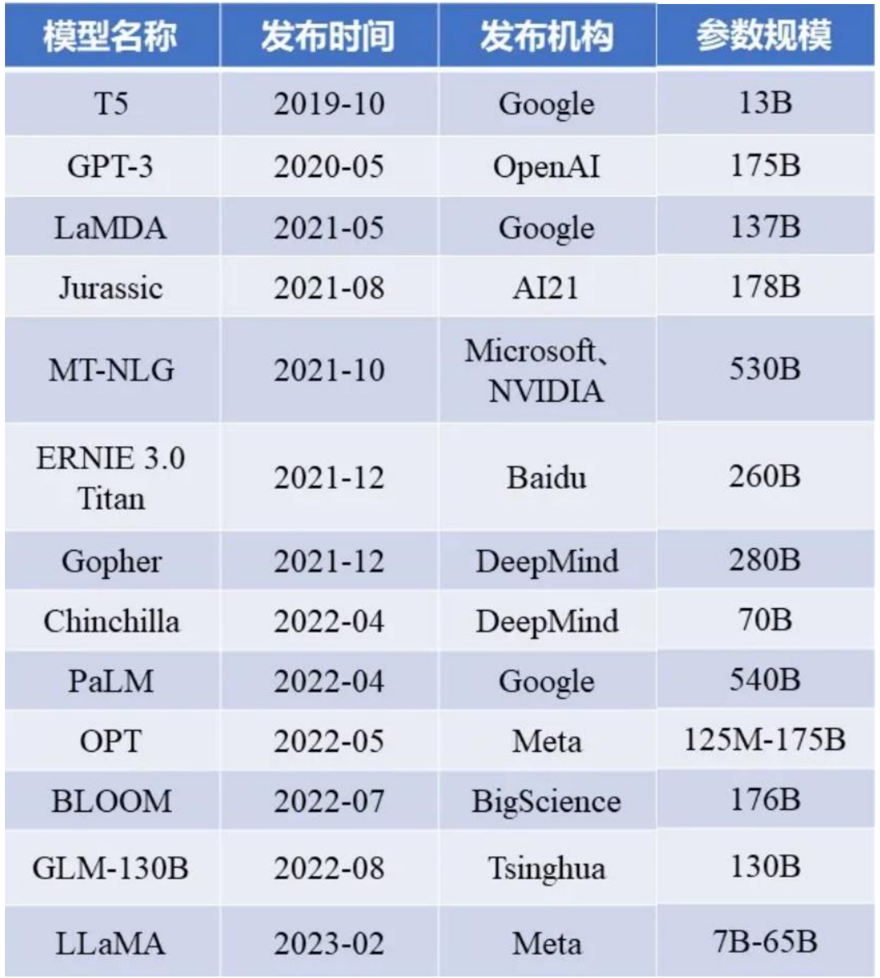
- A language model is generally considered "large" if it has tens of millions to billions of parameters.



模型名称	发布时间	发布机构	参数规模
T5	2019-10	Google	13B
GPT-3	2020-05	OpenAI	175B
LaMDA	2021-05	Google	137B
Jurassic	2021-08	AI21	178B
MT-NLG	2021-10	Microsoft、NVIDIA	530B
ERNIE 3.0 Titan	2021-12	Baidu	260B
Gopher	2021-12	DeepMind	280B
Chinchilla	2022-04	DeepMind	70B
PaLM	2022-04	Google	540B
OPT	2022-05	Meta	125M-175B
BLOOM	2022-07	BigScience	176B
GLM-130B	2022-08	Tsinghua	130B
LLaMA	2023-02	Meta	7B-65B

Parameter Size and Memory Footprint in LLM

- A language model is generally considered "large" if it has tens of millions to billions of parameters.



模型名称	发布时间	发布机构	参数规模
T5	2019-10	Google	13B
GPT-3	2020-05	OpenAI	175B
LaMDA	2021-05	Google	137B
Jurassic	2021-08	AI21	178B
MT-NLG	2021-10	Microsoft、NVIDIA	530B
ERNIE 3.0 Titan	2021-12	Baidu	260B
Gopher	2021-12	DeepMind	280B
Chinchilla	2022-04	DeepMind	70B
PaLM	2022-04	Google	540B
OPT	2022-05	Meta	125M-175B
BLOOM	2022-07	BigScience	176B
GLM-130B	2022-08	Tsinghua	130B
LLaMA	2023-02	Meta	7B-65B

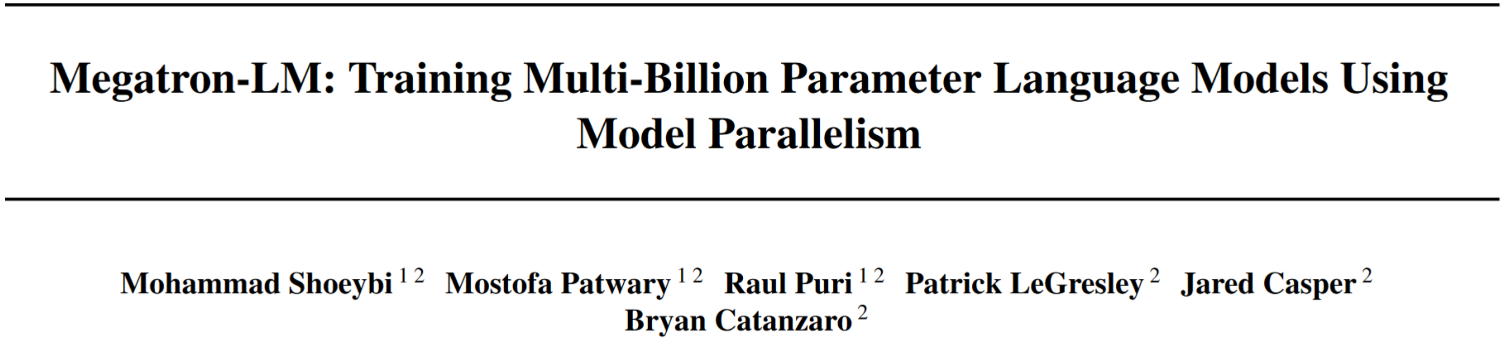
- During training, model parameters use Float64 and require 4 bytes each, while during inference, they use Float32 and require 2 bytes each.

The peak memory footprint for accommodating the ChatGLM-130B model is:

- Float64: 130B * 4 /1024/1024 = **480GB!!!**
- Float32: 130B * 2 /1024/1024 = **240GB!!!**

How to Break the Resource Wall of a Single GPU?

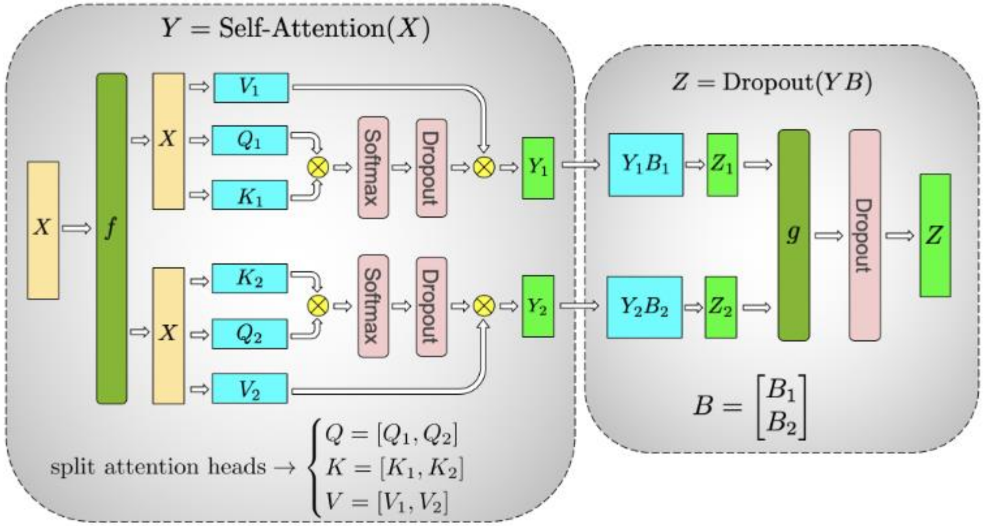
- The memory of a single device is insufficient to accommodate an entire LLM.



Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism

Mohammad Shoeybi [1 2] **Mostofa Patwary** [1 2] **Raul Puri** [1 2] **Patrick LeGresley** [2] **Jared Casper** [2]
Bryan Catanzaro [2]

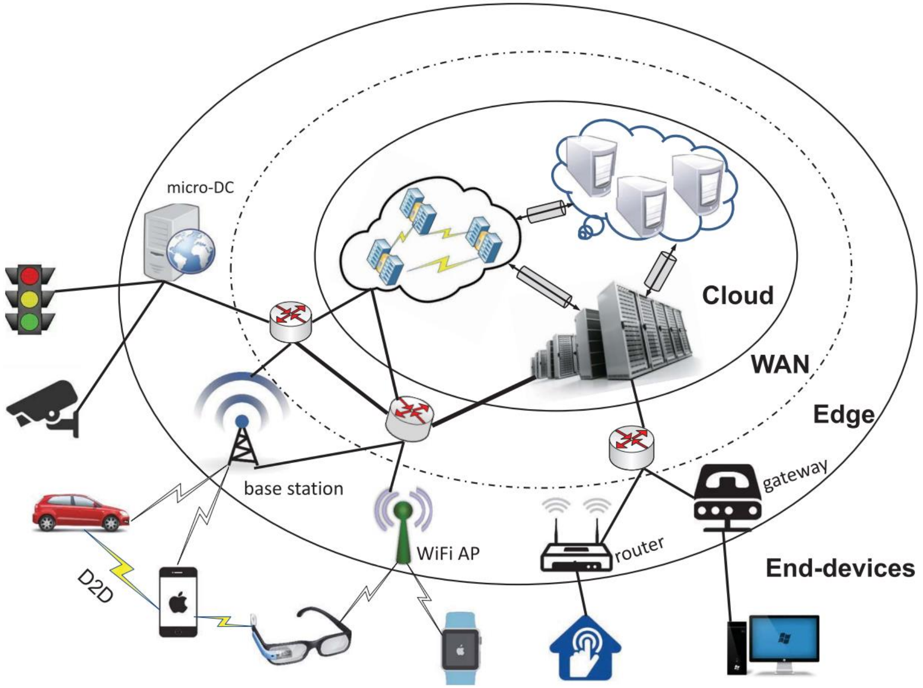
- Leveraging matrix decomposition techniques, a Tensor-Parallel (TP) distributed algorithm was developed to partition the model across multiple GPUs, with each GPU storing only a fraction of the model's weights.

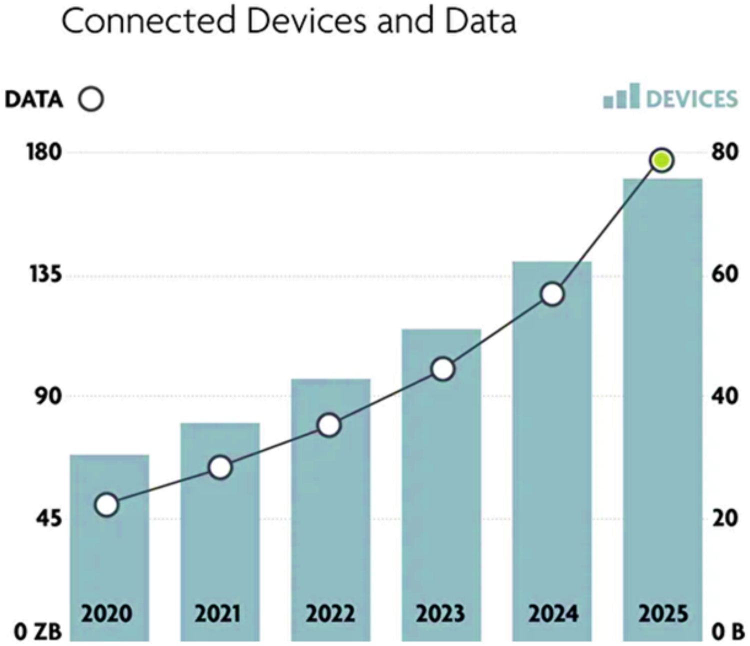


Credit:Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism.

LLM on Mobile Edge Devices

- **Mobile Computing + LLM** has emerged as a new paradigm
  - Popularization of mobile devices in both magnitude and variety
  - Proliferation of mobile data in both scale and modality

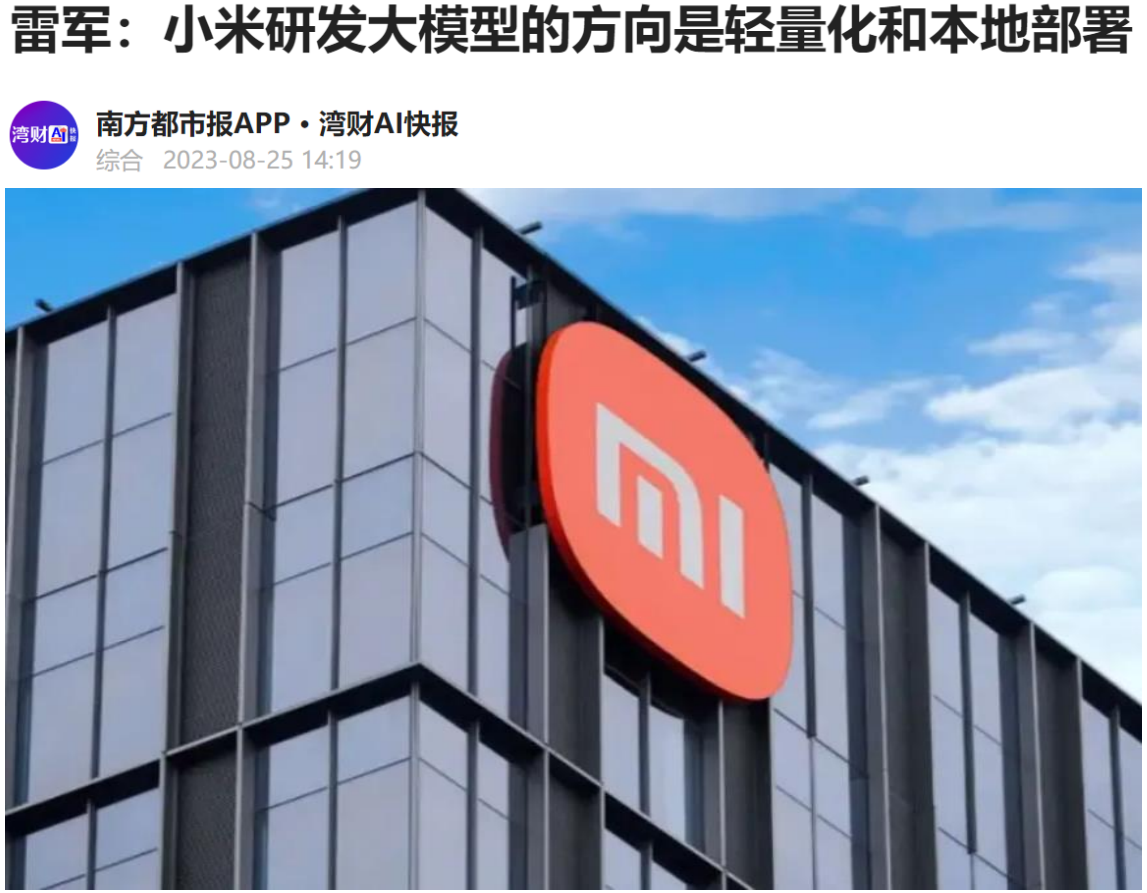




Credit: https://coruzant.com/opinion/the-future-is-edge-computing/

LLM on Mobile Edge Devices

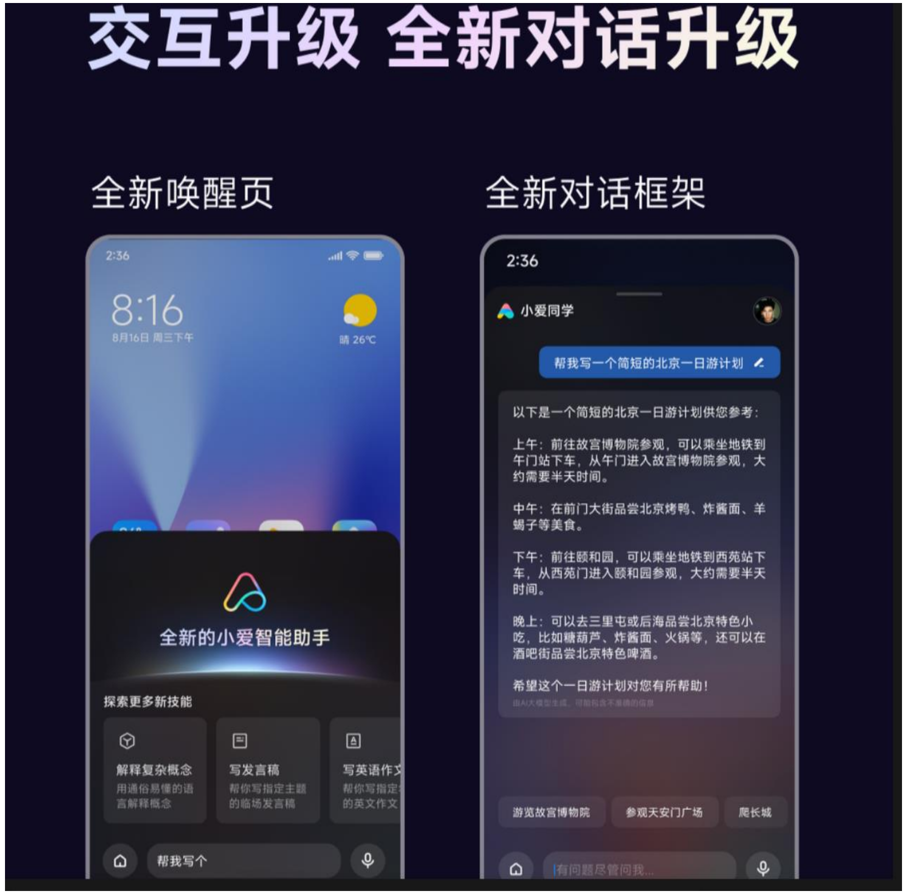
- Model lightweighting and edge deployment will become new research focuses in LLMs.



雷军：小米研发大模型的方向是轻量化和本地部署

南方都市报APP • 湾财AI快报

综合 2023-08-25 14:19



交互升级 全新对话升级

帮我写一个简短的北京一日游计划

以下是一个简短的北京一日游计划供您参考：

上午：前往故宫博物院参观，可以乘坐地铁到午门站下车，从午门进入故宫博物院参观，大约需要半天时间。

中午：在前门大街品尝北京烤鸭、炸酱面、羊蝎子等美食。

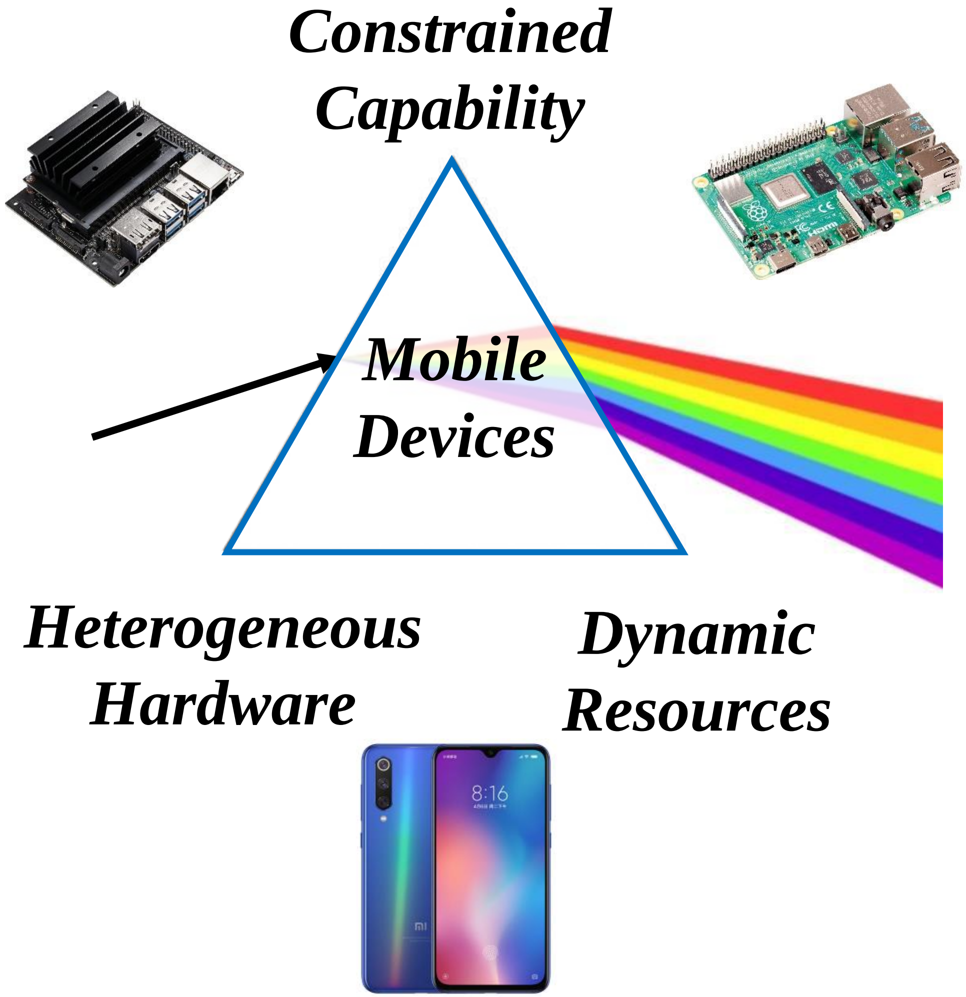
下午：前往颐和园，可以乘坐地铁到西苑站下车，从西苑门进入颐和园参观，大约需要半天时间。

晚上：可以去三里屯或后海品尝北京特色小吃，比如糖葫芦、炸酱面、火锅等，还可以在酒吧街品尝北京特色啤酒。

希望这个一日游计划对您有所帮助！

Credit: Google Image

Challenges of LLM on Mobile



Constrained Capability

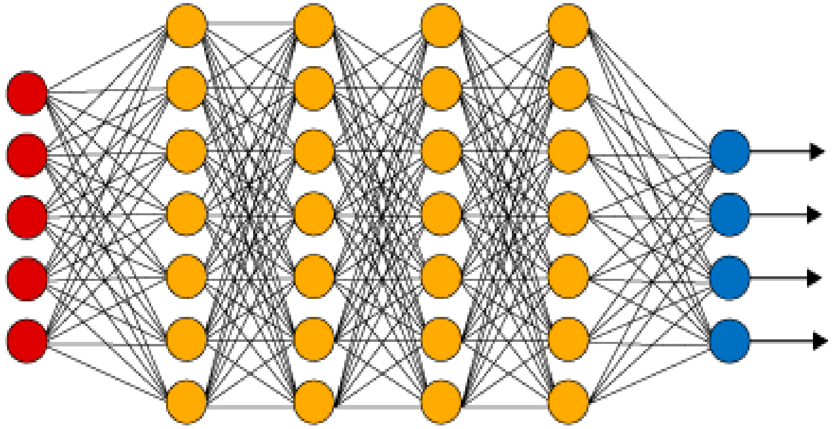
Mobile Devices

Heterogeneous Hardware

Dynamic Resources

How to apply?

Huge Gap



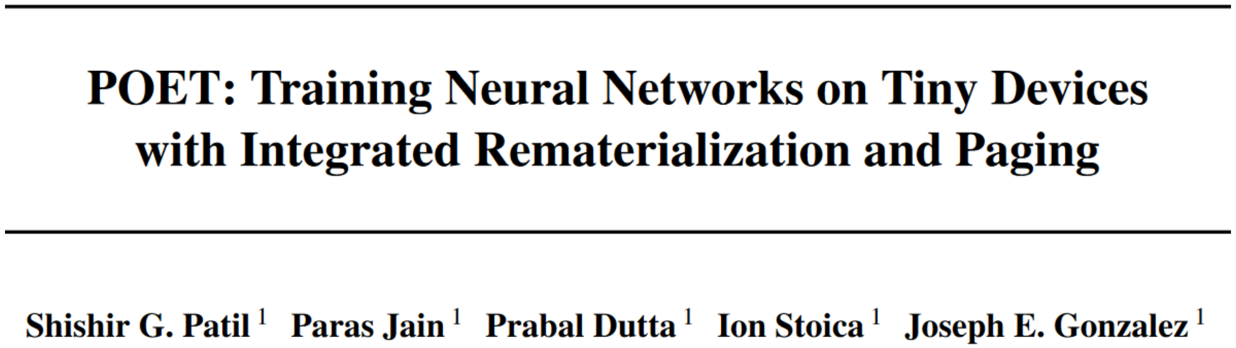
- LLM Computing is extremely computation-intensive and resource-demanding

- Mobile devices are resource-constrained and heterogeneous

Credit: Google Image

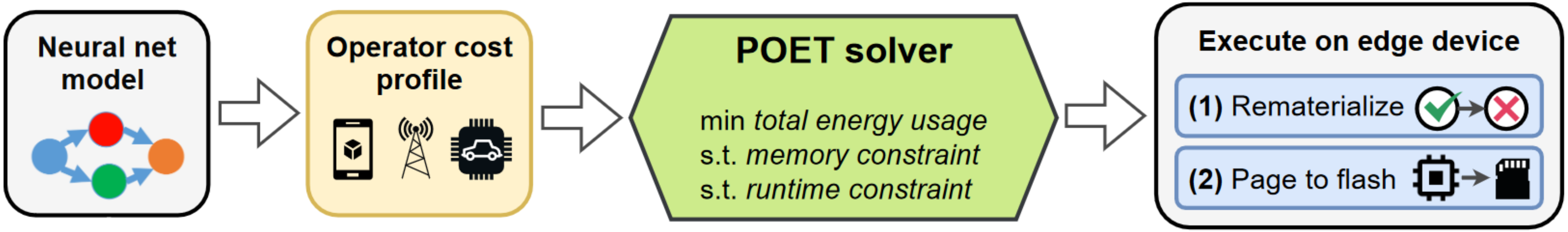
Break the Memory Wall of Mobile Devices

- Utilizing the concept of paging from operating systems, Transformer layers not in use are offloaded to auxiliary storage like SD cards to expand the available memory on mobile devices.



POET: Training Neural Networks on Tiny Devices with Integrated Rematerialization and Paging

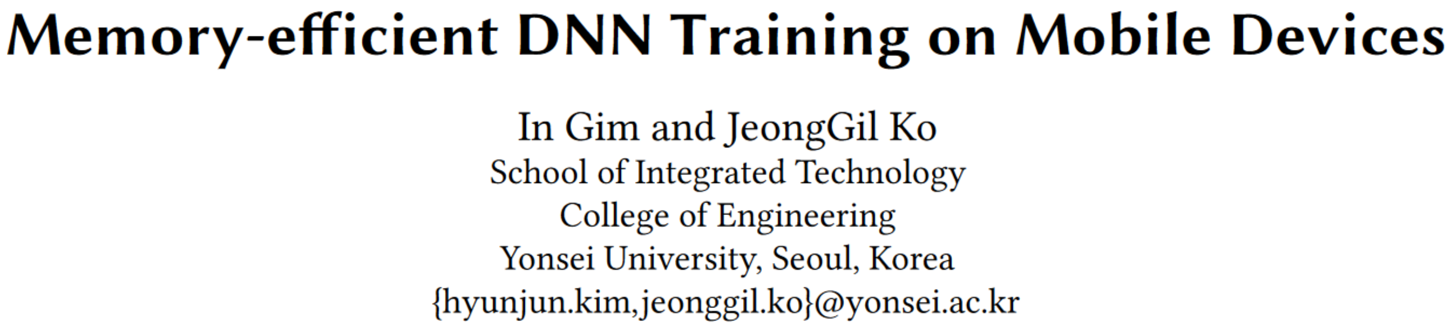
Shishir G. Patil [1] **Paras Jain** [1] **Prabal Dutta** [1] **Ion Stoica** [1] **Joseph E. Gonzalez** [1]



Credit: POET: Training Neural Networks on Tiny Devices with Integrated Rematerialization and Paging.

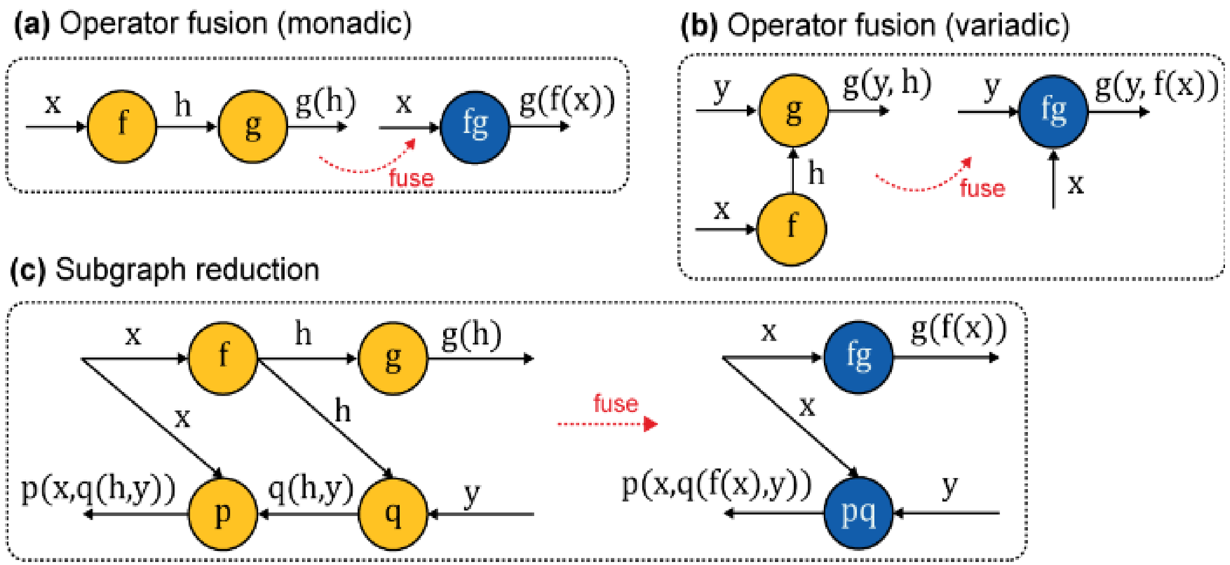
Break the Memory Wall of Mobile Devices

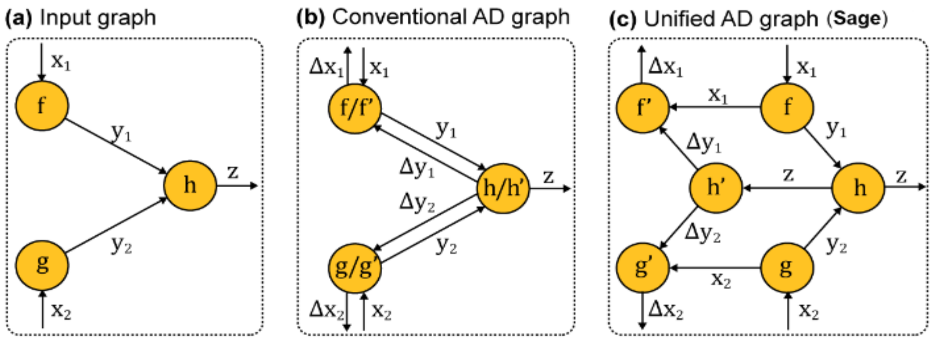
- Propose a framework support memory-efficient on-device LLM training

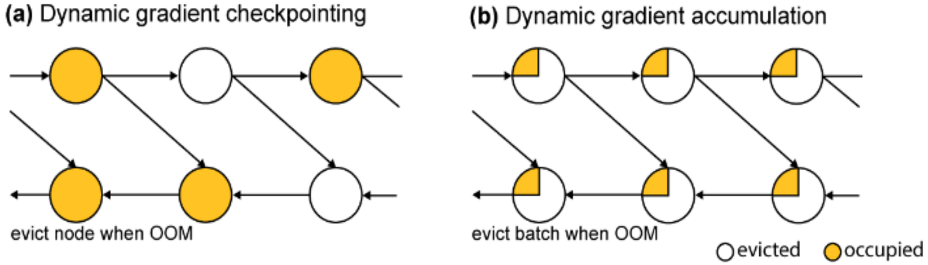


Memory-efficient DNN Training on Mobile Devices

In Gim and JeongGil Ko
School of Integrated Technology
College of Engineering
Yonsei University, Seoul, Korea
{hyunjun.kim,jeonggil.ko}@yonsei.ac.kr






(a) Dynamic gradient checkpointing
evict node when OOM
(b) Dynamic gradient accumulation
evict batch when OOM
○evicted
●occupied

Collaborative Execution on Mobile Cluster

- Federated Few-shot Learning on Mobile Cluster

Federated Few-Shot Learning for Mobile NLP

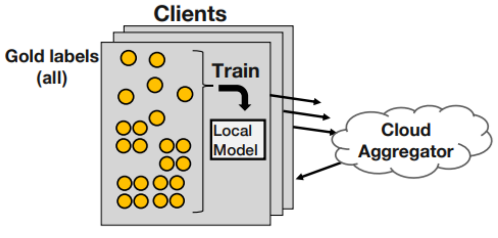
Dongqi Cai
Beiyou Shenzhen Institute

Shangguang Wang
Beiyou Shenzhen Institute

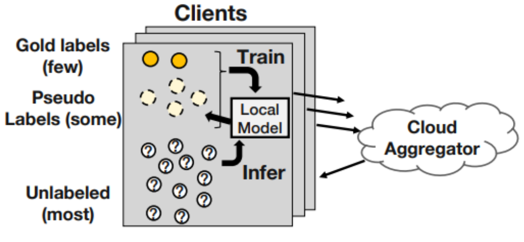
Yaozong Wu
Beiyou Shenzhen Institute

Felix Xiaozhu Lin
University of Virginia

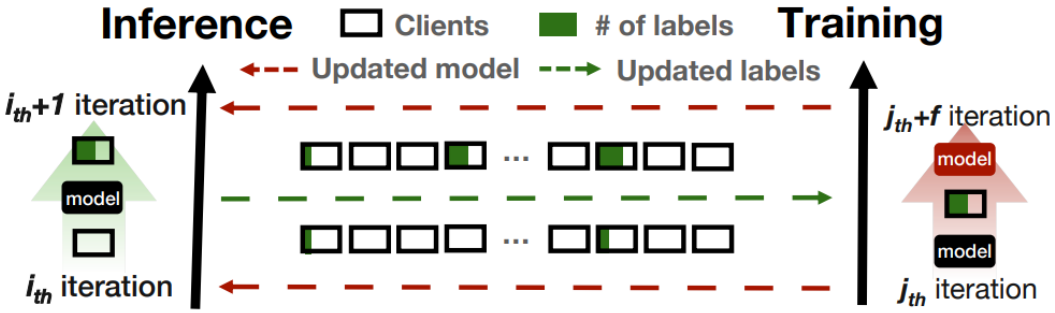
Mengwei Xu
Beiyou Shenzhen Institute

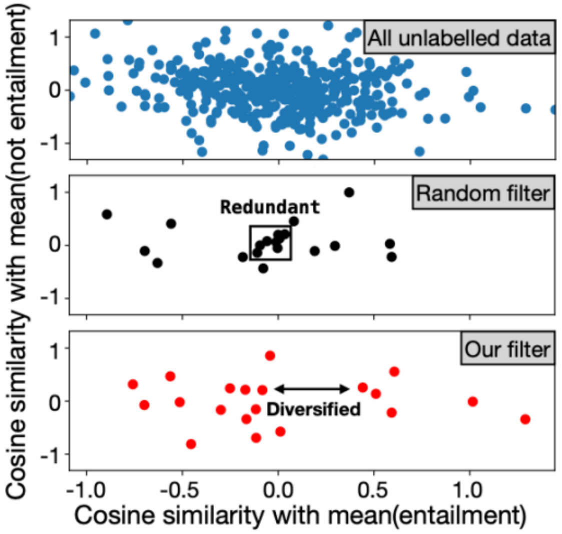


(a) Classic FL: rely on abundant labels

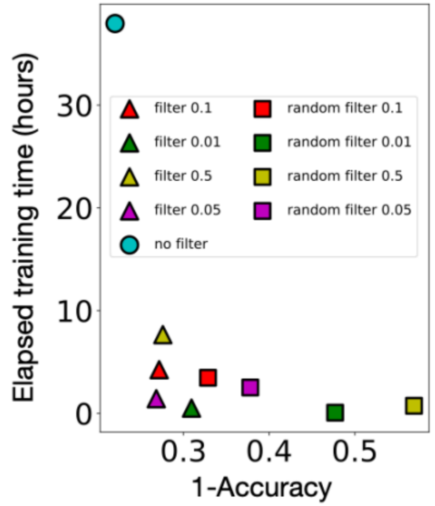


(b) Our FedFSL Scenario





(a) Representative diversity



(b) End-to-end performance

Summary

- **LLM+Mobile** is the new frontier, teeming with open questions that are ripe for exploration—let's pioneer the unknown!

- Awesome On-device-AI

https://github.com/ysyisyourbrother/awesome-on-device-AI

- A Reading List for Machine Learning Systems:

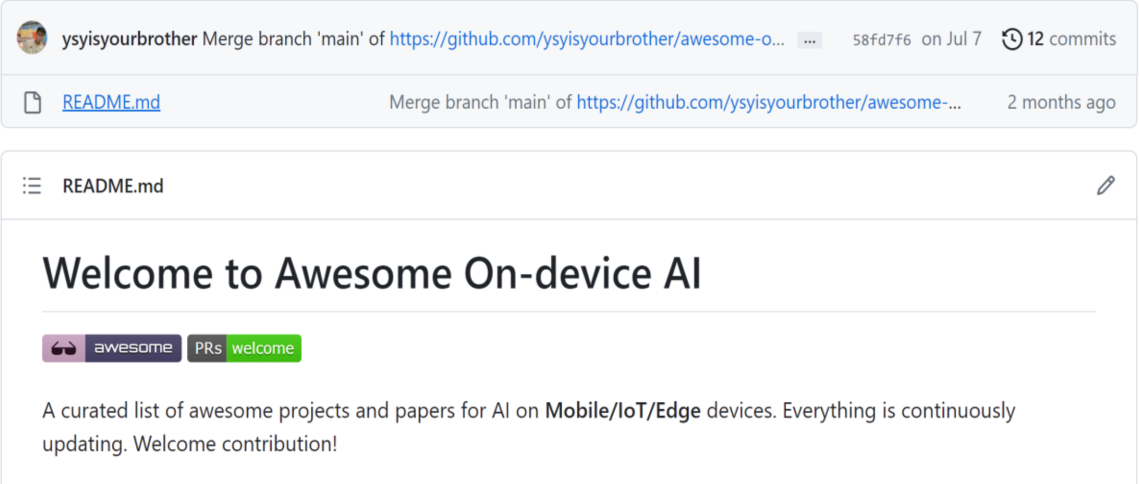
https://jeongseob.github.io/readings_mlsys.html

- Edge AI Paper List:

https://github.com/xumengwei/Edge-AI-Paper-List

- Resource Efficient Large Language Model

https://github.com/UbiquitousLearning/Paper-list-resource-efficient-large-language-model



ysyisyourbrother Merge branch 'main' of https://github.com/ysyisyourbrother/awesome-o... 58fd7f6 on Jul 7 12 commits

README.md Merge branch 'main' of https://github.com/ysyisyourbrother/awesome-... 2 months ago

README.md

Welcome to Awesome On-device AI

A curated list of awesome projects and papers for AI on **Mobile/IoT/Edge** devices. Everything is continuously updating. Welcome contribution!

Thanks

Shengyuan Ye

School of Computer Science and Engineering
Sun Yat-sen University
Contact: yeshy8@mail2.sysu.edu.cn