大模型系列 – 大模型算法

大模型发展趋势

ZOMI

大模型业务全流程

数据 & 模型算法

2. 数据处理

- 开源数据
- 数据预处理
- 向量数据库

3. 模型算法

- LLM模型架构
- 多模态一切皆Tokens

模型训练 & 微调

4. 模型训练

- 混合精度
- 梯度检查
- 梯度累积
- …
- 5. 分布式并行
- 训练集群稳定性

6. 模型微调

- 全参微调
- 低参微调
- 指令微调

模型验证 & 推理部署

7. 模型验证

- 下游任务
- 测评标准
- bench mark

8. 推理与智能体

- 量化压缩
- 推理加速
- 9. Agent 智能体

集群算力准备

1. AI 集群建设

- 计算、存储、网络
- AI 集群机房建设
- AI 集群上线与运维

关于本内容

2024.02.22

1. Google 开源的 GEMMA 大模型怎么样?

2024.02.29

2. Nvidia 开源的 Nemotron-4 大模型怎么样?

3. 大模型的发展趋势预测

1. Google GEMMA

https://blog.google/technology/developers/gemma-open-models/

GEMMA 概述

- **目标：** Gemma 模型可以在台式机 and/or 笔记本电脑上本地运行；
- **规模：** 两种尺寸， Gemma 2B and Gemma 7B，每种尺寸都发布预训练和指令微调变体。

Model	Embedding Parameters	Non-embedding Parameters
2B	524,550,144	1,981,884,416
7B	786,825,216	7,751,248,896

Table 2 | Parameter counts for both sizes of Gemma models.

GEMMA 概述

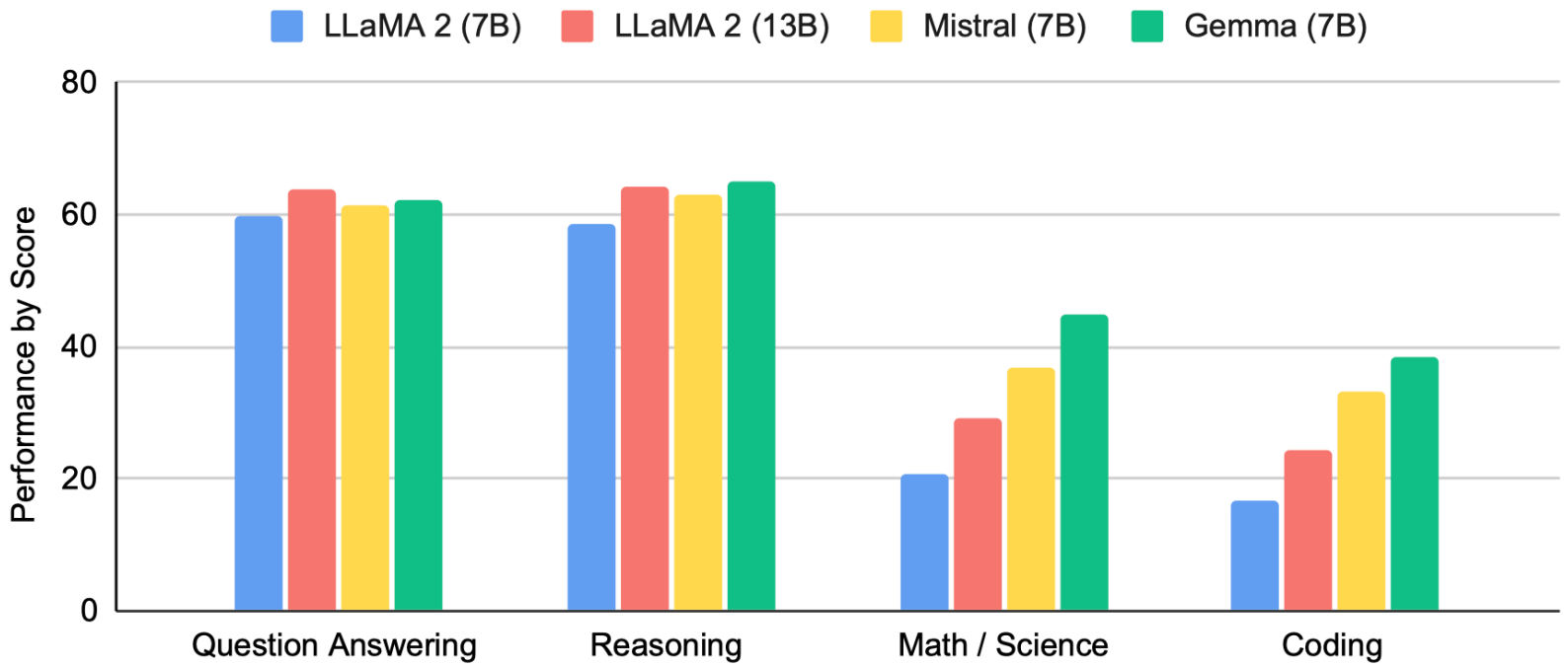
1. transformer decoder 结构进行训练，训练的上下文大小为8192个token；
2. Attention： 7B 使用 MHA，2B 模型使用 MQA (with num_kv_heads = 1)；
3. Embeddings：每一层加 RoPE Embedding，同时共享输入与输出层 embedding 权重；
4. Activations：ReLU的激活替换为 GeGLU 的激活。对比 llama中用了SwiGLU。
5. Normalizer：每一层 transformer 前后都使用 RMS Norm 进行规一化。

Parameters	**2B**	**7B**
d_model	2048	3072
Layers	18	28
Feedforward hidden dims	32768	49152
Num heads	8	16
Num KV heads	1	16
Head size	256	256
Vocab size	256128	256128

Table 1 | Key model parameters.

GEMMA 模型效果

- Google 表示： “Gemma 2B 和 7B 与其他开放式模型相比，在其规模上实现了同类最佳的性能。” 从学术基准角度来看，Gemma 7B 在数学、Python 代码生成、常识和常识推理任务的几个基准测试中，优于 Meta 的 Llama 2 7B 和 13B 模型。



GEMMA 模型效果

- MMLU 基准测试中，Gemma 7B 模型不仅超过了所有规模相同或更小的开源模型，还超过了一些更大的模型，包括 Llama 2 13B。

还是想以小打大

		LLaMA-2		Mistral	Gemma	
Benchmark	metric	7B	13B	7B	2B	7B
MMLU	5-shot, top-1	45.3	54.8	62.5	42.3	**64.3**
HellaSwag	0-shot	77.2	80.7	81.0	71.4	**81.2**
PIQA	0-shot	78.8	80.5	**82.2**	77.3	81.2
SIQA	0-shot	48.3	50.3	47.0*	49.7	**51.8**
Boolq	0-shot	77.4	81.7	**83.2***	69.4	**83.2**
Winogrande	partial scoring	69.2	72.8	**74.2**	65.4	72.3
CQA	7-shot	57.8	67.3	66.3*	65.3	**71.3**
OBQA		**58.6**	57.0	52.2	47.8	52.8
ARC-e		75.2	77.3	80.5	73.2	**81.5**
ARC-c		45.9	49.4	**54.9**	42.1	53.2
TriviaQA	5-shot	72.1	**79.6**	62.5	53.2	63.4
NQ	5-shot	25.7	**31.2**	23.2	12.5	23.0
HumanEval	pass@1	12.8	18.3	26.2	22.0	**32.3**
MBPP†	3-shot	20.8	30.6	40.2*	29.2	**44.4**
GSM8K	maj@1	14.6	28.7	35.4*	17.7	**46.4**
MATH	4-shot	2.5	3.9	12.7	11.8	**24.3**
AGIEval		29.3	39.1	41.2*	24.2	**41.7**
BBH		32.6	39.4	**56.1***	35.2	55.1
Average		47.0	52.2	54.0	44.9	**56.4**

1. 测试时大家都只跟 LLAMA2 比较呢？不去跟 GPT-4 或者其他模型比呢？

GEMMA 安全性

- Gemma 设计将 AI 原则放在首位：
  - 使用自动化技术从训练集中过滤掉某些个人信息和其他敏感数据。
  - 使用了大量微调和人类反馈强化学习（RLHF），使指令调整模型与负责任的行为保持一致。
  - 为了解并降低 Gemma 模型风险，进行人工红蓝队、自动对抗测试和危险活动模型能力评估。

GEMMA 开放性

- Gemma 遵循「开放模型」而非模型开源：
  - 特点是可以免费获取模型权重，并在一定受限范围内使用
  - 再分发和变体所有权的条款根据模型具体使用条款而有所不同，不遵循常规开源协议

GEMMA 官方链接

- https://ai.google.dev/gemma
- https://blog.google/technology/developers/gemma-open-models/
- https://huggingface.co/google/gemma-7b
- https://opensource.googleblog.com/2024/02/building-open-models-responsibly-gemini-era.html
- https://www.kaggle.com/models/google/gemma
- https://github.com/google/gemma.cpp

GEMMA 加速&部署方案

1. 为所有主要框架提供推理和监督微调（SFT），包括 JAX、PyTorch 、TensorFlow 、Kears 3.0 多框架；
2. 随时可用 Colab、Kaggle、Hugging Face、MaxText 和 NVIDIA NeMo 和 TensorRT-LLM 等工具集成，方便开发者更容易上手使用 Gemma；
3. 经过预训练和指令调整 Gemma 模型可在笔记本电脑、工作站或 Google Cloud 上运行，并能够部署在 Vertex AI 和谷歌 Kubernetes Engine 上；
4. 英伟达宣布与 Google 合作，包括本地 RTX AI PC 在内所有英伟达 AI 平台上启动优化，用来加速 Gemma 的性能；

2. NVIDIA 英伟达 Nemotron-4

NEMOTRON-4 15B LLM

https://arxiv.org/abs/2402.16819

Nemotron-4 基本信息

- **基本参数：**Nemotron-4 15B，8T Token （万亿）训练，能在单个NV A100/H100 上运行的最佳「通用大模型」。
- **模型结构：**旋转位置编码（RoPE Embedding）、SentencePiece 分词器、ReLU 激活、无偏置项（ bias terms）、分组查询关注（GQA）。
- **模型参数：**

Number of transformer layers	Hidden dimension	Number of attention heads	Number of KV heads	Sequence length	Vocabulary size
32	6144	48	8	4096	256,000

Table 1: Key hyper-parameters affecting size of Nemotron-4 15B.

Nemotron-4 训练数据

- 分为三种不同类型的数据：
  1. 英语自然语言数据（70%）
  2. 多语言自然语言数据（15%）
  3. 源代码数据（15%）

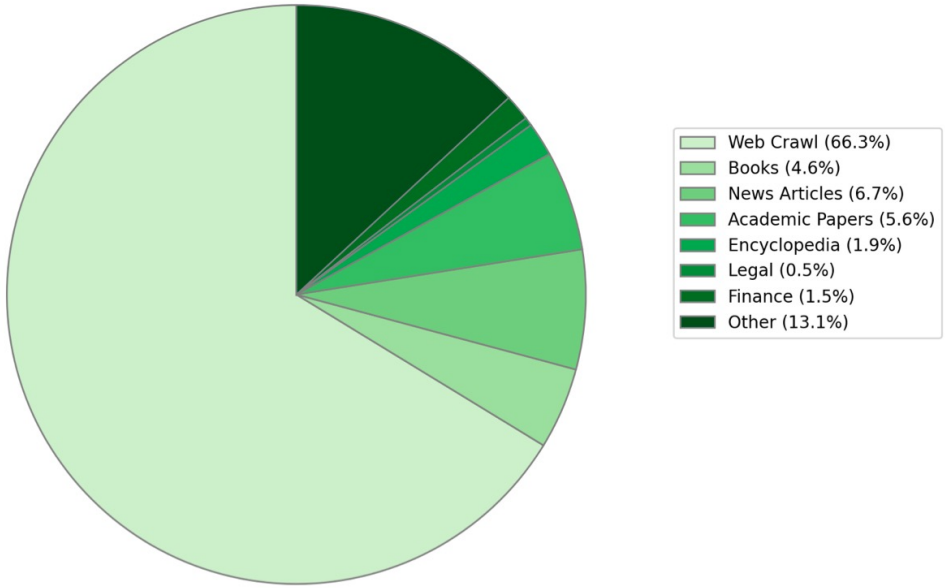


Figure 2: Data composition of the English tokens used for pre-training

Nemotron-4 预训练

- 384 个 DGX H100节点，每个节点 8 NV Hopper 架构 H100 80GB SXM5 GPU；
- 无稀疏性16位浮点（bfloat16）算术时，每 H100 GPU 峰值吞吐量为 989 teraFLOP/s；
- 每节点内通过 NVLink 和 NVSwitch连接，GPU to GPU 带宽为 900GB/s（每个方向450GB/s）；
- 每节点 8 个 NV Mellanox400Gbps HDR InfiniBand 主机通道适配器（HCA），用于节点间通信；
- 使用 8 路张量并行 TP 和数据并行 DP 组合来训练模型，使用 ZeRO 分布式优化器，随着批大小的增加，数据并行度从 96 增加到 384。

Nemotron-4 模型效果

- 作者使用LM-Evaluation Harness在所有上述任务中评估Nemotron-415B。
- 表3显示了Nemotron-415B在这组不同的任务中实现了最强的平均性能。

	Size	SIQA	ARC-c	ARC-e	PIQA	Winogrande	Hellaswag	AVG
LLaMA-2	13B	50.3	49.4	77.3	79.8	72.8	80.7	68.4
	34B	50.9	54.5	79.4	81.9	76.7	83.3	71.1
Baichuan-2	13B	-	-	-	78.1	-	70.8	-
QWEN	14B	77.9	84.4	90.3	79.9	-	80.2	-
Mistral	7B	47.0*	55.5	80.0	83.0	75.3	81.3	70.4
Gemma	7B	51.8	53.2	81.5	81.2	72.3	81.2	70.2
Nemotron-4	15B	60.9	55.5	80.9	82.4	78.0	82.4	**73.4**

Table 3: Results on standard reasoning benchmarks in the zero-shot setting. We report the average across all tasks where possible for a fair comparison. The values marked with $*$ are read from Gemma Team (2024)

Nemotron-4 模型效果

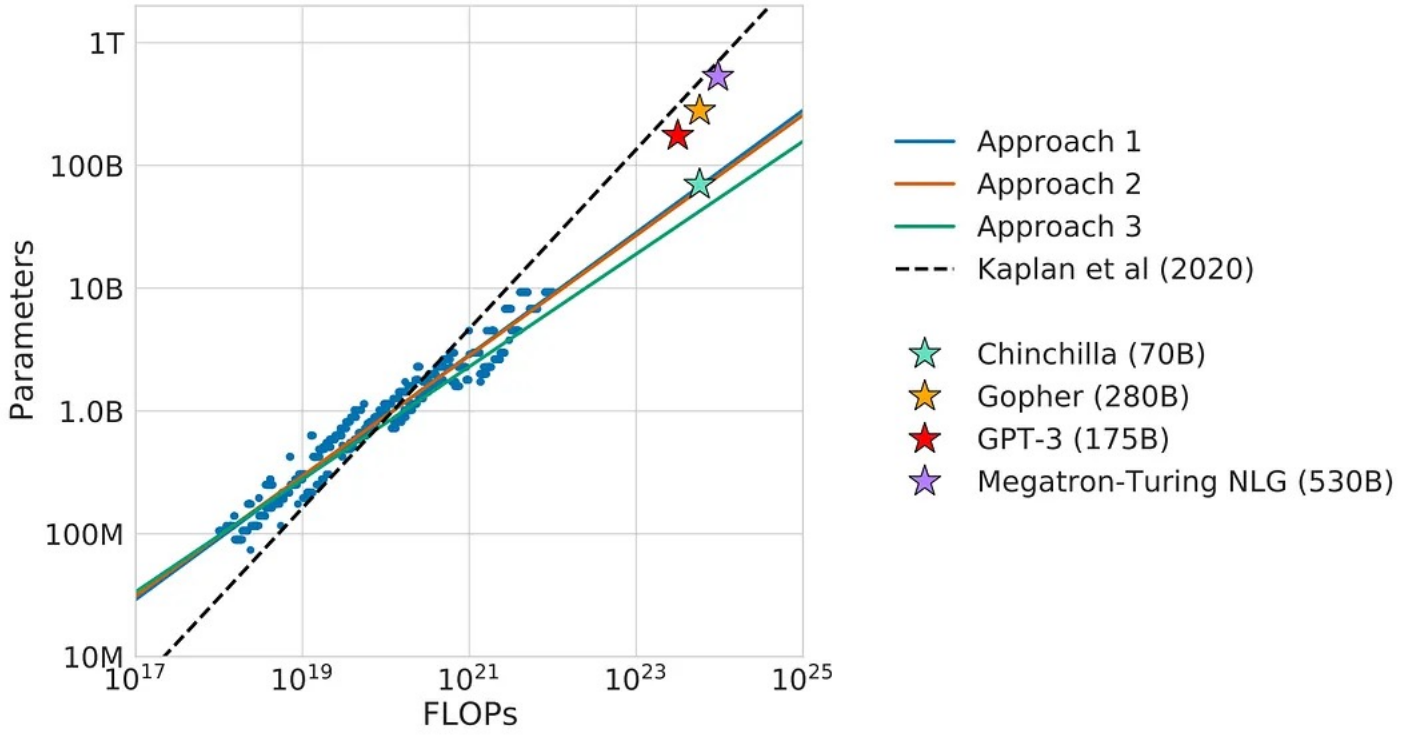
	Size	BBH	MMLU
LLaMA-2	13B	39.4	54.8
	34B	44.1	62.6
Baichuan-2	13B	48.8	59.2
QWEN	14B	53.4	**66.3**
Mistral	7B	39.5	60.1
Gemma	7B	55.1	64.3
Nemotron-4	15B	**58.7**	64.2

	Size	GSM8K	HumanEval	MBPP
LlaMA-2	13B	28.7	18.3	30.6
	34B	42.2	22.6	33.0
Baichuan-2	13B	52.8	17.1	30.2
QWEN	14B	**60.1**	32.2	40.8
Mistral	7B	35.4*	30.5	40.2*
Gemma	7B	46.4	**32.3**	**44.4**
Nemotron-4	15B	46.0	31.6	40.6

小结与思考

洞察与思考：Scaling Law

- 过去研究主要针对模型大小进行缩放（13B to 175B），最近发表 LLM 研究受到 Chinchilla 模型「缩放定律 Scaling Law」的启发，需要在固定算力、数据和模型规模大小协同优化。

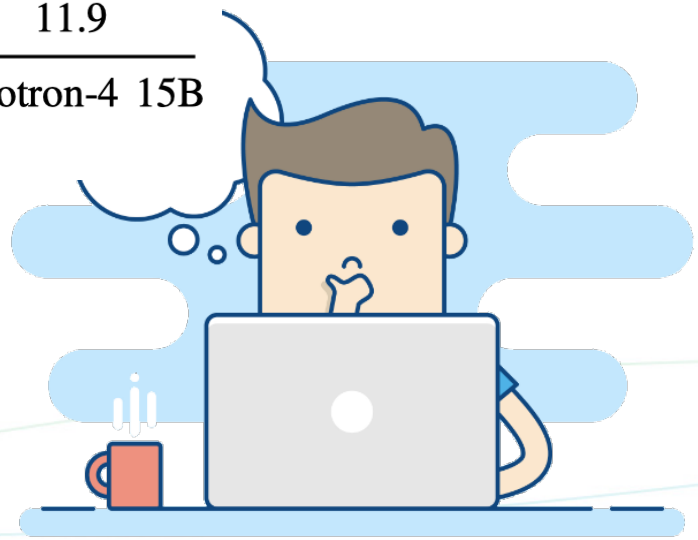


洞察与思考：分布式并行

1. Batch Size 3 阶段，包括每次迭代时间和模型 FLOP/s 利用率（MFU），MFU量化了GPU在模型训练中的利用效率，训练大约在13天内完成。

Data-parallel size	GPUs	Iteration time (secs)	MFU (%)	Batch size	Tokens (B)	Time (days)
96	768	0.57	34.3	384	200	0.8
192	1,536	0.58	33.3	768	200	0.4
288	2,304	0.64	30.5	1,152	7,600	11.9

Table 2: Batch size rampup schedule, along with time and efficiency metrics for the Nemotron-4 15B parameter model.



洞察与思考：大模型规模发展

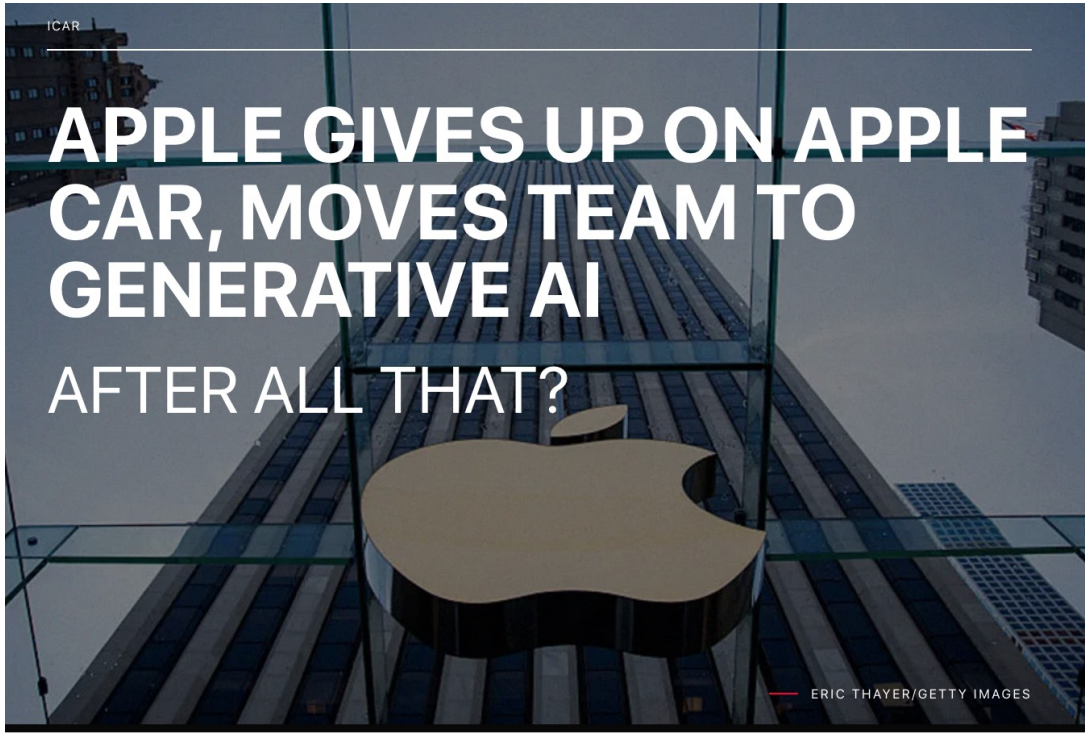
1. Google GEMMA：Gemma 模型是可以在台式机 and/or 笔记本电脑上本地运行的大模型。
2. NVDIA Nemotron-4：成为能在单个 A100 or H100 GPU上运行的最佳「通用大模型」。

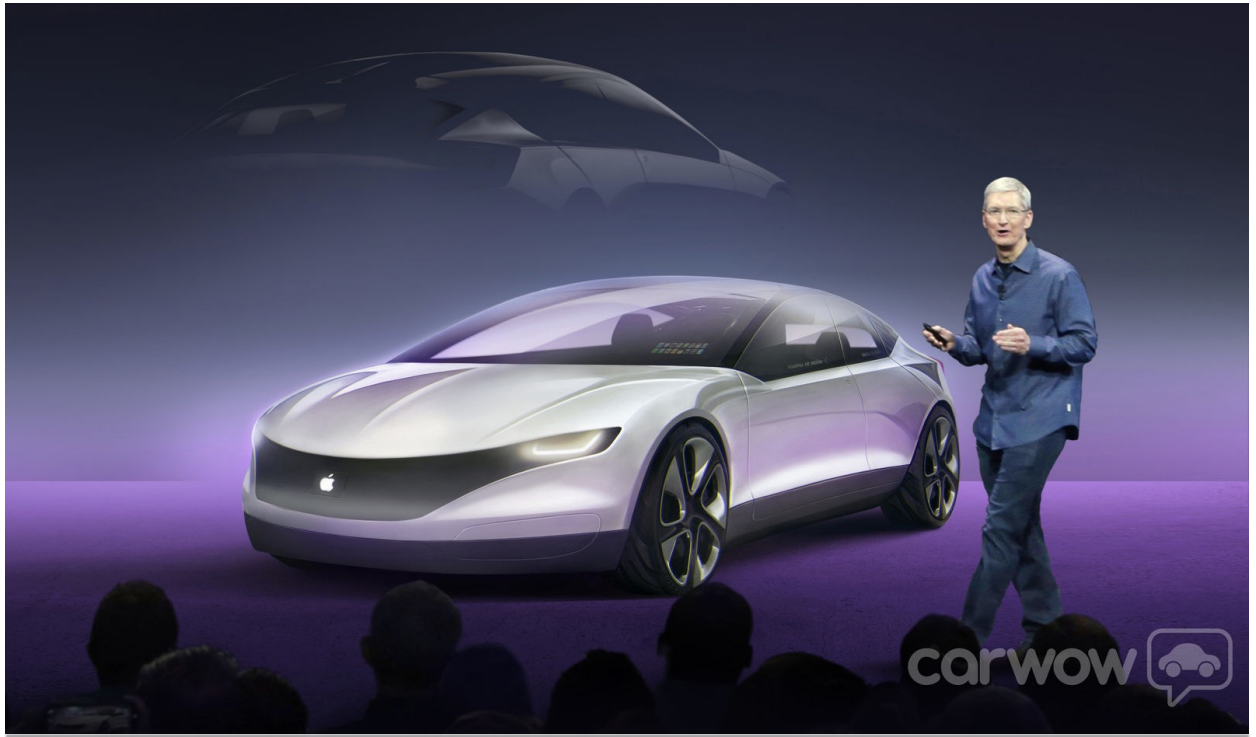
提供小规模的大模型，满足端侧场景

AI PC


ANNOUNCING WAVE OF AI PCs
New GeForce RTX 40 Series Laptops
ACER
Swift X 14
ASUS
ROG Zephyrus G16
DELL
XPS 14
HP
OMEN Transcend 14
LENOVO
Yoga Pro 9i
MSI
Stealth 16 AI Studio
RAZER
Blade 16
SAMSUNG
Galaxy Book4 Ultra

AI 手机





RIP Titan

After around a decade of extremely secretive development, Apple is officially giving up on building an electric car.

AI 手机

Top stories

Futurism

Apple Gives Up on Apple Car, Moves Team to Generative AI

2 days ago

Bloomberg

Apple to Wind Down Electric Car Effort After Decadelong Odyssey

2 days ago

Time Magazine

Apple's Scrapped Car Project Makes AI, Headset Bets More Urgent

1 day ago

Thank you

把AI系统带入每个开发者、每个家庭、每个组织，构建万物互联的智能世界

Bring AI System to every person, home and organization for a fully connected, intelligent world.

ZOMI

Course chenzomi12.github.io

GitHub github.com/chenzomi12/DeepLearningSystem