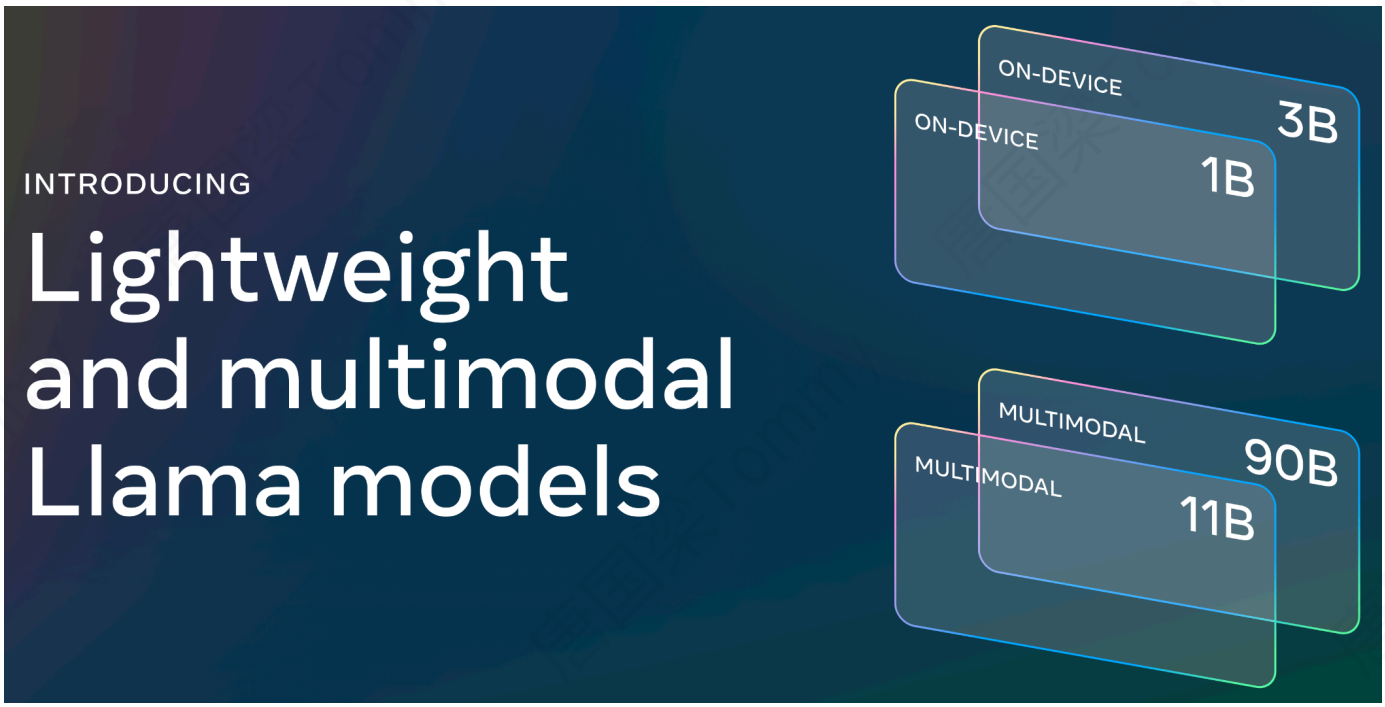
Llama 3.2：通过开放、可定制的模型彻底改变边缘人工智能和视觉

1. Llama 3.2 模型简介

1.1 模型类别

- 包括小型和中型视觉模型（11B 和 90B）以及轻量级文本模型（1B 和 3B），适用于边缘设备和移动设备。



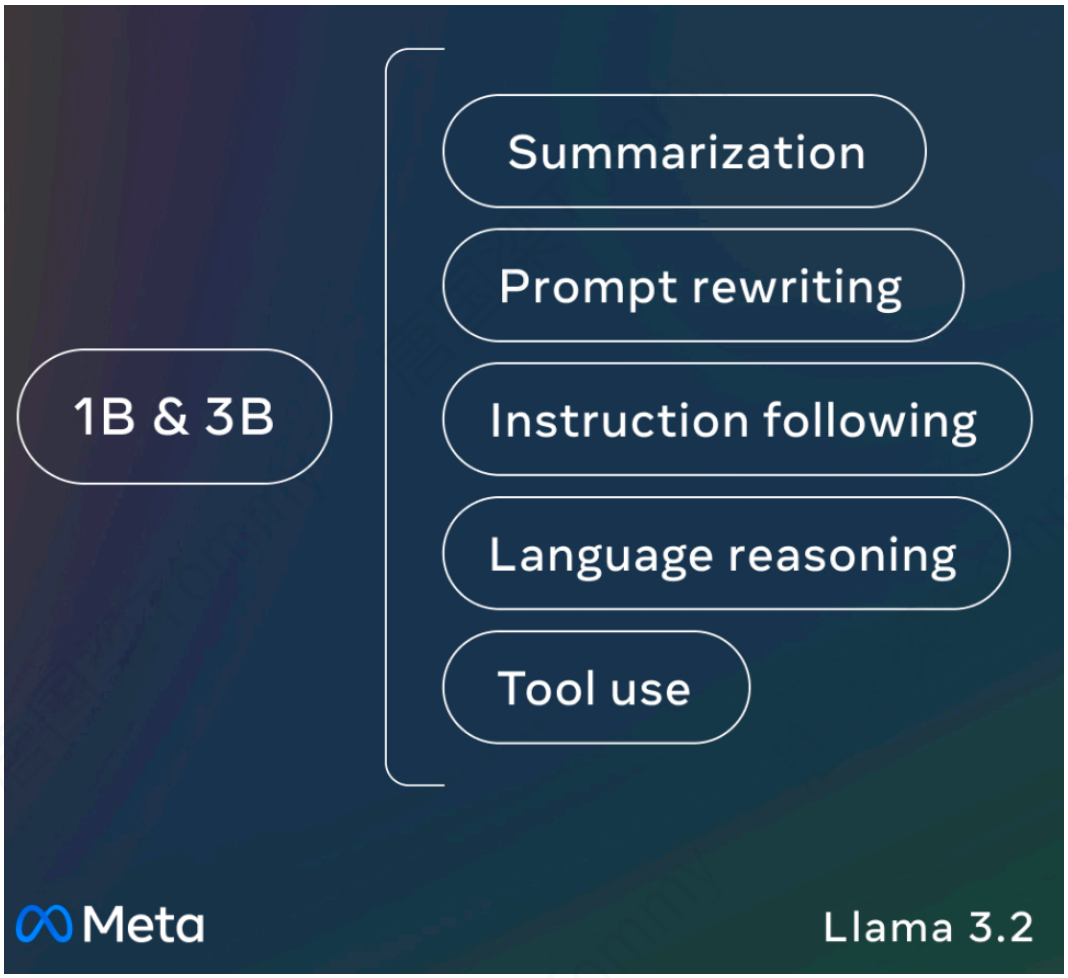
1.2 创新点

- **视觉模型创新**：首次支持图像推理，11B 和 90B 模型通过适配器将图像编码器与语言模型相结合，实现文本与图像对齐。
- **后期训练优化**：采用监督微调SFT、偏好优化DPO等方法，增强模型在图像和文本提示上的理解与推理能力。

1.2 模型性能

1.2.1 文本模型（1B和3B）

1B 和 3B 模型支持 128K tokens 上下文长度，专为本地设备的摘要、指令跟随、文本重写等任务设计。它具备强大的多语言生成能力，并支持工具调用，适合在本地应用，确保数据隐私。



- **1B 文本模型**

Llama 3.2系列中最轻量级的模型，非常适合边缘设备和移动应用的摘要任务。该模型非常适合以下使用场景：个人信息管理和多语言知识检索。

- **3B 文本模型**

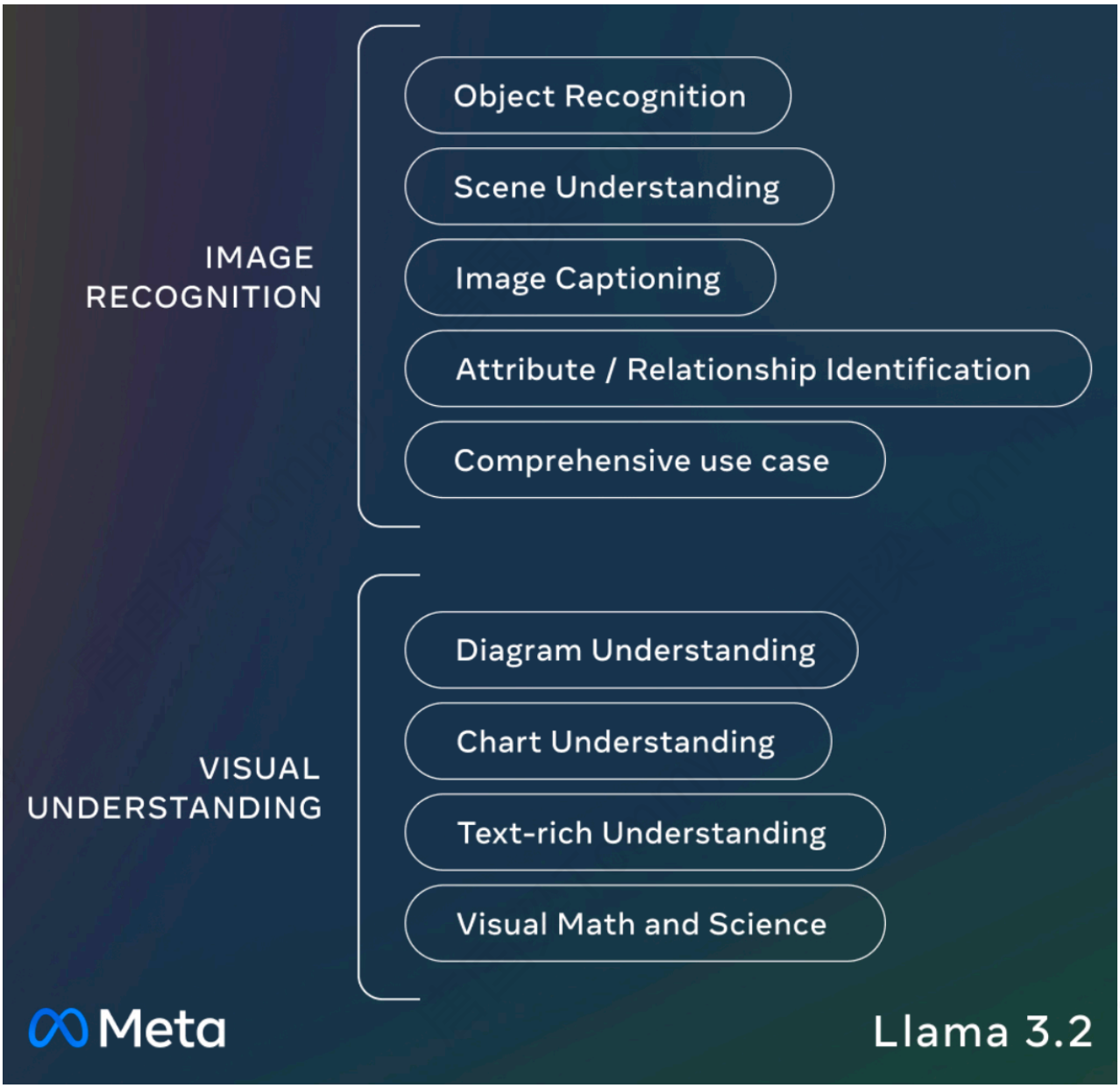
为需要低延迟推理和有限计算资源的应用设计。在文本摘要、分类和语言翻译任务中表现出色。该模型非常适合以下使用场景：由AI驱动的移动写作助手和客户服务应用。

> Lightweight instruction-tuned benchmarks

Category Benchmark	Llama 3.2 1B	Llama 3.2 3B	Gemma 2 2B IT (5-shot)	Phi-3.5 - Mini IT (5-shot)
General MMLU (5-shot)	49.3	63.4	57.8	**69.0**
Open-rewrite eval (0-shot, rougeL)	**41.6**	40.1	31.2	34.5
TLDR9+ (test, 1-shot, rougeL)	16.8	**19.0**	13.9	12.8
IFEval	59.5	**77.4**	61.9	59.2
Math GSM8K (0-shot, CoT)	44.4	77.7	62.5	**86.2**
MATH (0-shot, CoT)	30.6	**48.0**	23.8	44.2
Reasoning ARC Challenge (0-shot)	59.4	78.6	76.7	**87.4**
GPQA (0-shot)	27.2	**32.8**	27.5	31.9
Hellaswag (0-shot)	41.2	**69.8**	61.1	81.4
Tool use BFCL V2	25.7	**67.0**	27.4	58.4
Nexus	13.5	**34.3**	21.0	26.1
Long context InfiniteBench/En.MC (128k)	38.0	63.3	-	39.2
InfiniteBench/En.QA (128k)	20.3	19.8	-	11.3
NIH/Multi-needle	75.0	84.7	-	52.7
Multilingual MGSM (0-shot, CoT)	24.5	**58.2**	40.2	49.8

1.2.2 视觉模型（11B和90B）

11B 和 90B 模型支持图像与语言结合推理，如图像定位和物体识别，可用于文档级理解、图表信息提取等任务。它优于其他闭源模型（如 Claude 3 Haiku）在图像理解任务上的表现。



> Vision instruction-tuned benchmarks

Modality	Category Benchmark	Llama 3.2 11B	Llama 3.2 90B	Claude3 - Haiku	GPT-4o-mini
Image	College-level Problems and Mathematical Reasoning MMMU (val, 0-shot CoT, micro avg accuracy)	50.7	**60.3**	50.2	59.4
	MMMU-Pro, Standard (10 opts, test)	33.0	**45.2**	27.3	42.3
	MMMU-Pro, Vision (test)	27.3	33.8	20.1	**36.5**
	MathVista (testmini)	51.5	**57.3**	46.4	56.7
	Charts and Diagram Understanding ChartQA (test, 0-shot CoT, relaxed accuracy)*	83.4	85.5	81.7	-
	AI2 Diagram (test)*	91.9	92.3	86.7	-
	DocVQA (test, ANLS)*	88.4	90.1	88.8	-
	General Visual Question Answering VQAv2 (test)	75.2	78.1	-	-
Text	General MMLU (0-shot, CoT)	73.0	**86.0**	75.2 (5-shot)	82.0
	Math MATH (0-shot, CoT)	51.9	68.0	38.9	**70.2**
	Reasoning GPQA (0-shot, CoT)	32.8	**46.7**	33.3	40.2
	Multilingual MGSM (0-shot, CoT)	68.9	86.9	75.1	**87.0**

2. Llama 3.2 文本模型

Llama 3.2 是一组多语言的大语言模型（LLMs），包含1B和3B大小的预训练和指令调优生成模型（输入文本/输出文本），主要用于多语言对话场景，包括智能检索和摘要任务。它们在常见的行业基准测试中优于许多现有的开源和封闭式聊天模型。

2.1 模型架构

Llama 3.2 是一个自回归语言模型，使用优化的Transformer架构。调优版本采用了监督微调（SFT）和基于人类反馈的强化学习（RLHF）以适应人类对实用性和安全性的偏好。

模型	训练数据	参数	输入模式	输出模式	上下文长度	GQA	共享嵌入	Tokens数量	知识截止时间
Llama 3.2（仅文本）	公开可用在线数据的新组合	1B（1.23B）	多语言文本	多语言文本和代码	128k	是	是	高达9万亿tokens	2023年12月
		3B（3.21B）	多语言文本	多语言文本和代码	128k	是	是		2023年12月

2.2 支持语言

官方支持的语言包括英语、德语、法语、意大利语、葡萄牙语、印地语、西班牙语和泰语。Llama 3.2 还在比这些8种语言更广泛的语言集合上进行了训练。开发者可以根据 Llama 3.2 社区许可协议和可接受使用政策对模型进行微调，以支持其他语言。

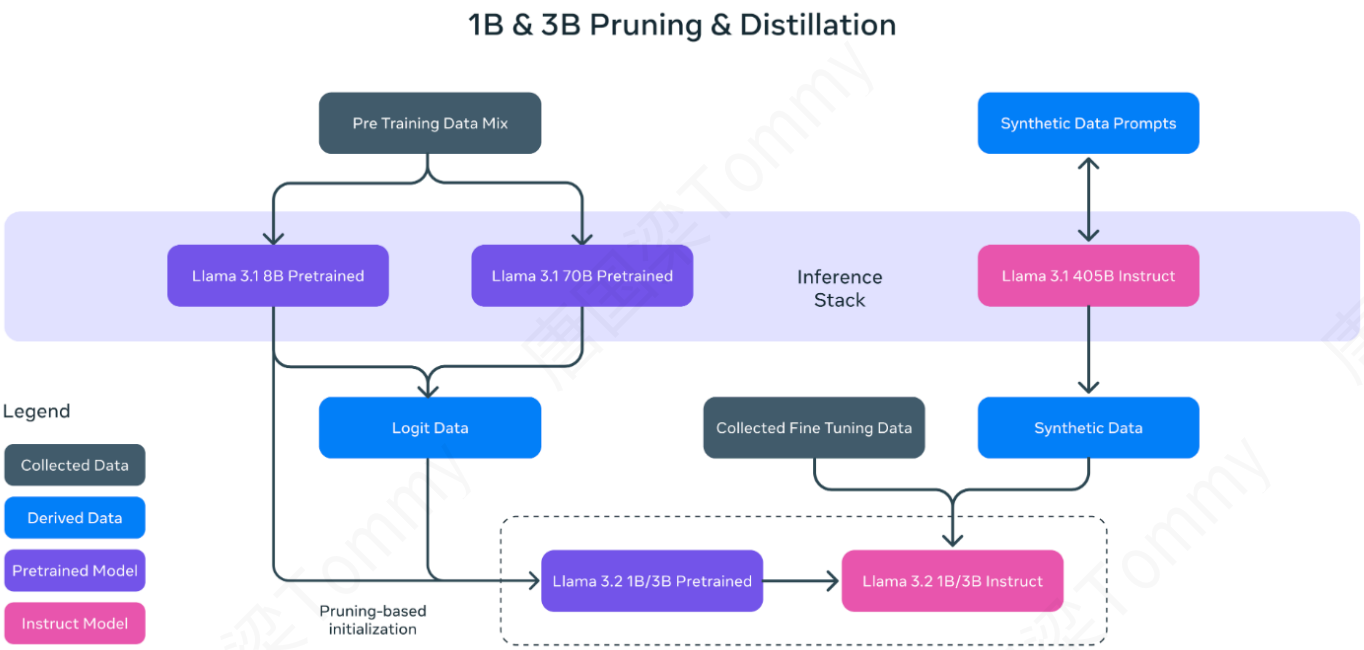
2.3 训练耗时

训练使用了累计91.6万小时的GPU计算时间，硬件为H100-80GB（TDP为700W），表中给出的训练时间为每个GPU设备的总GPU训练时间，经过功率使用效率调整后的数值。

模型	训练时间（GPU 小时）	Logit 生成时间（GPU 小时）	训练功率消耗（W）	基于地点的训练温室气体排放量（吨CO2当量）	基于市场的训练温室气体排放量（吨CO2当量）
Llama 3.2 1B	370k	-	700	107	0
Llama 3.2 3B	460k	-	700	133	0
总计	830k	86k		240	0

2.4 训练数据

- **概览**: Llama 3.2 在多达9万亿个来自公开可用资源的标记数据上进行了预训练。对于1B和3B的Llama 3.2模型，我们将来自Llama 3.1 8B和70B模型的logits数据合并到模型开发的预训练阶段，这些较大模型的输出作为token级目标。修剪后使用知识蒸馏恢复性能。在微调后，我们使用与Llama 3.1相似的配方，通过多轮对齐优化生成最终的聊天模型。每轮包括监督微调（SFT）、拒绝采样（RS）和直接偏好优化（DPO）。



2.5 基准测试

Base Pretrained Models

类别	基准测试	# Shots	指标	Llama 3.2 1B	Llama 3.2 3B	Llama 3.1 8B
通用	MMLU	5	macro_avg/acc_char	32.2	58	66.7
	AGIEval English	3-5	average/acc_char	23.3	39.2	47.8
	ARC-Challenge	25	acc_char	32.8	69.1	79.7
阅读理解	SQuAD	1	em	49.2	67.7	77
	QuAC (F1)	1	f1	37.9	42.9	44.9
	DROP (F1)	3	f1	28.0	45.2	59.5
长上下文	Needle in Haystack	0	em	96.8	1	1

Instruction Tuned Models

能力	基准测试	# Shots	指标	Llama 3.2 1B	Llama 3.2 3B	Llama 3.1 8B
通用	MMLU	5	macro_avg/acc	49.3	63.4	69.4
重写	Open-rewrite eval	0	micro_avg/rougeL	41.6	40.1	40.9
总结	TLDR9+ (test)	1	rougeL	16.8	19.0	17.2
遵循指令	IFEval	0	avg(prompt/instruction acc loose/strict)	59.5	77.4	80.4
数学	GSM8K (CoT)	8	em_maj@1	44.4	77.7	84.5
	MATH (CoT)	0	final_em	30.6	47.3	51.9
推理	ARC-C	0	acc	59.4	78.6	83.4
	GPQA	0	acc	27.2	32.8	32.8
	Hellaswag	0	acc	41.2	69.8	78.7
工具使用	BFCL V2	0	acc	25.7	67.0	70.9
	Nexus	0	macro_avg/acc	13.5	34.3	38.5
长上下文	InfiniteBench/En.QA	0	longbook_qa/f1	20.3	19.8	27.3
	InfiniteBench/En.MC	0	longbook_choice/acc	38.0	63.3	72.2
	NIH/Multi-needle	0	recall	75.0	84.7	98.8
多语言	MGSM (CoT)	0	em	24.5	58.2	68.9

Multilingual Benchmarks

类别	基准测试	语言	Llama 3.2 1B	Llama 3.2 3B	Llama 3.1 8B
通用	MMLU（5-shot, macro_avg/acc）	葡萄牙语	39.82	54.48	62.12
		西班牙语	41.5	55.1	62.5
		意大利语	39.8	53.8	61.6
		德语	39.2	53.3	60.6
		法语	40.5	54.6	62.3
		印地语	33.5	43.3	50.9
		泰语	34.7	44.5	50.3

3. Llama 3.2 视觉模型

Llama 3.2-Vision 是一系列多模态大型语言模型（LLMs），包括经过预训练和指令微调的图像推理生成模型，提供 11B 和 90B 参数规模（文本+图像输入/文本输出）。Llama 3.2-Vision 指令微调模型专为视觉识别、图像推理、图像描述和回答图像相关的通用问题而优化。这些模型在许多公开或封闭的多模态模型上表现优于行业标准基准。

3.1 模型架构

Llama 3.2-Vision 构建于 Llama 3.1 纯文本模型之上，后者是使用优化的自回归语言模型（Transformer）架构。微调版本使用监督微调（SFT）和通过人类反馈（RLHF）的强化学习，以符合人类偏好的有用性和安全性。

为支持图像识别任务，Llama 3.2-Vision 模型使用单独训练的视觉适配器，与预训练的 Llama 3.1 语言模型集成。该适配器由一系列跨注意力层组成，将图像编码表示输入到核心 LLM 中。

训练数据	参数	输入模态	输出模态	上下文长度	GQA	数据量	知识截止日期
Llama 3.2-Vision	(图像，文本) 对	11B (10.6)	文本 + 图像	128k	是	60亿 (图像，文本) 对	2023年12月
Llama 3.2-Vision	(图像，文本) 对	90B (88.8)	文本 + 图像	128k	是	60亿 (图像，文本) 对	2023年12月

3.2 支持语言

对于仅文本任务，Llama 3.2 正式支持英语、德语、法语、意大利语、葡萄牙语、印地语、西班牙语和泰语。Llama 3.2 已在比这 8 种语言更广泛的语言集合上进行训练。对于图像+文本应用，当前仅支持英语。

3.3 使用场景

Llama 3.2-Vision 旨在用于商业和研究用途。指令微调模型用于视觉识别、图像推理、图像描述以及类似于图像的助手式聊天，而预训练模型可适应各种图像推理任务。

- **视觉问答（VQA）和视觉推理**: 想象一台机器可以看着图片并理解你对其的提问。
- **文档视觉问答（DocVQA）**: 想象计算机能够理解文档的文本和布局，如地图或合同，然后直接从图像中回答问题。
- **图像描述**: 图像描述弥合了视觉和语言之间的差距，提取细节、理解场景，并生成讲述故事的句子。
- **图像-文本检索**: 图像-文本检索类似于图像及其描述的匹配引擎，像搜索引擎一样，但能同时理解图片和文本。
- **视觉定位**: 视觉定位就像将我们看到和说的点连接起来，它涉及理解语言如何基于自然语言描述来参考图像的特定部分，使 AI 模型能够根据这些描述来精确定位物体或区域。

3.4 训练耗时

训练使用了总计 2.02M 个 GPU 小时，基于 H100-80GB（700W TDP）类型的硬件。训练时间为每个 GPU 设备的总 GPU 时间，经过电源使用效率调整。

训练时间（GPU 小时）	训练功耗（W）	位置基础温室气体排放（吨 CO2eq）	市场基础温室气体排放（吨 CO2eq）
Llama 3.2-vision 11B	阶段 1 预训练: 147K H100 小时 阶段 2 退火: 98K H100 小时 SFT: 896 H100 小时 RLHF: 224 H100 小时	700	71
Llama 3.2-vision 90B	阶段 1 预训练: 885K H100 小时 阶段 2 退火: 885K H100 小时 SFT: 3072 H100 小时 RLHF: 2048 H100 小时	700	513
总计	2.02M		584

3.5 训练数据

Llama 3.2-Vision 在 60 亿图像和文本对上进行了预训练。指令微调数据包括公开可用的视觉指令数据集，以及超过 300 万个合成生成的示例。

训练数据	参数	输入模态	输出模态	上下文长度	GQA	数据量	知识截止日期
Llama 3.2-Vision	(图像，文本) 对	11B (10.6)	文本 + 图像	128k	是	60亿 (图像，文本) 对	2023年12月
Llama 3.2-Vision	(图像，文本) 对	90B (88.8)	文本 + 图像	128k	是	60亿 (图像，文本) 对	2023年12月

3.6 基准测试

> **Base Pretrained Models**

类别	基准测试	# 次数	指标	Llama 3.2 11B	Llama 3.2 90B
图像理解	VQAv2 (验证集)	0	准确率	66.8	73.6
	Text VQA (验证集)	0	放宽的准确率	73.1	73.5
	DocVQA (验证集，未见过)	0	ANLS	62.3	70.7
视觉推理	MMMU (验证集，0-shot)	0	微平均准确率	41.7	49.3
	ChartQA (测试集)	0	准确率	39.4	54.2
	InfographicsQA (验证集，未见过)	0	ANLS	43.2	56.8
	AI2 Diagram (测试集)	0	准确率	62.4	75.3

Instruction Tuned Models

模态	能力	基准测试	# 次数	指标	Llama 3.2 11B	Llama 3.2 90B
图像	大学水平的问题和数学推理	MMMU (验证集, CoT)	0	微平均准确率	50.7	60.3
		MMMU-Pro, 标准 (10 选项, 测试)	0	准确率	33.0	45.2
		MMMU-Pro, 视觉 (测试)	0	准确率	23.7	33.8
		MathVista (testmini)	0	准确率	51.5	57.3
	图表和图解理解	ChartQA (测试集, CoT)	0	放宽的准确率	83.4	85.5
		AI2 Diagram (测试集)	0	准确率	91.1	92.3
		DocVQA (测试集)	0	ANLS	88.4	90.1
	通用视觉问答	VQAv2 (测试集)	0	准确率	75.2	78.1
文本	通用能力	MMLU (CoT)	0	宏平均/准确率	73.0	86.0
	数学	MATH (CoT)	0	最终_em	51.9	68.0
	推理	GPQA	0	准确率	32.8	46.7
	多语言	MGSM (CoT)	0	em	68.9	86.9