Llama 4 模型系列

2025年4月3日　8 分钟阅读

#AI #Llama #LLM #技术

> *本文介绍了Llama 4 模型系列详细解读。*

到目前为止还没有明确的证据表明有单独的技术报告或论文。可能的细节将在2025年4月29日的LlamaCon上分享。

Llama 4 最大的亮点就是它从"文本选手"华丽转身为"多模态全能王"，这意味着 Llama 4 **原生支持理解和处理文本以及图像输入，并能生成文本和代码输出。**这对于我们构建更智能、更具互动性的应用来说，无疑打开了全新的大门。

那么，Llama 4 究竟有哪些核心技术特点，让它如此引人注目呢?

预训练阶段的创新技术

技术名称	特点	优势
混合专家架构 (MoE)	单个 token 激活模型中一部分参数，而非全部参数。	提高计算效率，降低推理成本，支持大规模模型在单个 GPU 上运行。
原生多模态支持	使用早期融合技术 (**early fusion**)，将文本和视觉 token 无缝集成到统一架构中。	增强模型的视觉理解能力，支持多模态数据（文本、图像、视频）的联合训练。
MetaP 超参数优化技术	可靠设置关键模型超参数（如层学习率和初始化比例）。	跨规模迁移能力强，确保模型扩展时保持高质量和稳定性。
FP8 精度训练	使用 FP8 精度进行训练，同时保持高 FLOPs 利用率。	提高训练效率，支持大规模数据处理（30 万亿 token），显著提升模型性能。
长上下文扩展	使用专门数据集进行"中期训练"，延长模型上下文长度。	支持行业领先的 10M token 上下文长度，适用于多文档摘要和大型代码库推理。
iRoPE 架构	采用交错注意力层，移除传统位置嵌入，动态调整注意力温度。	提升上下文长度泛化能力，实现"无限"上下文目标。

目录

文章信息

字数

阅读时间

发布时间

更新时间

标签

#AI #Llama #LLM

后训练阶段的创新技术

技术名称	特点	优势
轻量级监督微调 (SFT)	对模型进行轻量级监督微调，专注于困难数据集。	避免过度约束模型，提高推理、编码和数学领域的准确性。
在线强化学习 (RL)	采用连续在线强化学习策略，动态过滤和保留中等到困难提示。	提高计算效率，构建递增难度课程，优化模型推理能力。
直接偏好优化 (DPO)	处理模型响应质量的边缘情况，轻量级优化模型偏好。	平衡模型的智能与对话能力，确保多模态任务和对话任务间的最佳表现。
动态蒸馏损失函数	在蒸馏过程中动态调整软目标和硬目标的权重。	提高蒸馏效率，降低资源密集型前向传播的计算成本。
强化学习基础设施优化	为两万亿参数模型重新设计 RL 基础设施，采用异步在线 RL 框架。	提高训练效率（10 倍提升），优化 MoE 并行化速度。

以上表格清晰地展示了 Llama 4 系列在训练过程中的技术创新，

混合专家架构 (MoE)：Llama 4 模型系列**首次采用了混合专家 (MoE) 架构。**你可以把它想象成一个拥有众多“专家”的大脑，但对于每一个输入，只有一小部分最相关的“专家”会被激活来处理。这种架构的优势在于：

更高的计算效率：相比于传统的稠密模型，MoE 模型在训练和推理时**只需要激活模型总参数的一小部分**，从而大大降低了计算成本和延迟。

更高的模型质量：在相同的计算资源下，MoE 架构通常能够**提供比稠密模型更高的性能。**

具体实现：Llama 4 系列中包含两款高效模型：

Llama 4 Scout：拥有 **170 亿的激活参数和 1090 亿的总参数，以及 16 个专家。**值得一提的是，它**可以在单个 NVIDIA H100 GPU 上运行。**

Llama 4 Maverick：拥有 **170 亿的激活参数和 4000 亿的总参数，以及 128 个路由专家和一个共享专家。**它的推理过程使用了**交替的稠密层和混合专家层**，每个 token 会被发送到共享专家以及 128 个路由专家中的一个，从而在保证性能的同时提高了推理效率。Llama 4 Maverick **可以在单个 H100 DGX 主机上运行。**

惊人的上下文窗口长度：Llama 4 在处理长文本方面也展现出了强大的能力.

Llama 4 Scout 提供了**行业领先的 1000 万 (10M) token 的上下文窗口。**这为处理海量文档、分析用户行为和理解大型代码库带来了前所未有的潜力. 为了实现如此长的上下文，Llama 4 Scout 在预训练和后训练阶段都使用了 **256K 的上下文长度**，并采用了**交错注意力层 (interleaved attention layers) 而没有位置嵌入 (positional embeddings)**，以及一种称为 **iRoPE 架构** 的推理时注意力温度缩放技术，旨在支持“无限”上下文长度。

Llama 4 Maverick 也拥有 **100 万 (1M) token 的上下文窗口。**

强大的“老师”——Llama 4 Behemoth：Llama 4 Scout 和 Llama 4 Maverick 的卓越性能离不开它们强大的“老师”—— **Llama 4 Behemoth。**这是一款拥有 **2880 亿激活参数和近两万亿总参数，以及 16 个专家**的超大规模多模态混合专家模型。尽管 Llama 4 Behemoth 仍在训练中，但它已经展现出了**在多个**

STEM 基准测试中超越 GPT-4.5、Claude Sonnet 3.7 和 Gemini 2.0 Pro 的潜力。通过**知识蒸馏 (distillation)** 的方法，Llama 4 Behemoth 将其强大的知识和能力传递给了更小的 Llama 4 模型。

高效的训练过程：为了训练如此强大的模型，Meta 采取了多种创新方法来提高效率.

FP8 精度训练：在保证模型质量的前提下，使用了 **FP8 精度**进行模型训练，提高了计算效率. 这个和DeepSeek V3类似。

海量训练数据：Llama 4 的预训练数据超过 **30 万亿 token**，是 Llama 3 的两倍多，并且包含了多样化的文本、图像和视频数据.

MetaP 训练技术：开发了一种新的训练技术 **MetaP**，可以可靠地设置关键的模型超参数.

长上下文扩展：通过在"mid-training"阶段使用专门的数据集进行长上下文扩展训练，提升了模型质量并实现了 10M 的超长上下文窗口.

优化的强化学习 (RL)：针对拥有两万亿参数的 Llama 4 Behemoth，对 RL 基础设施进行了全面革新，**优化了 MoE 的并行化**，并开发了**完全异步的在线 RL 训练框架**，实现了约 **10 倍的训练效率提升**.

广泛的语言支持：Llama 4 在预训练阶段使用了 **200 种语言**的数据，并且**明确支持 12 种语言的输出**：阿拉伯语、英语、法语、德语、印地语、印度尼西亚语、意大利语、葡萄牙语、西班牙语、塔加路语、泰语和越南语.

原生多模态能力的技术细节：Llama 4 通过 **早期融合 (early fusion)** 的方式，将文本和视觉 tokens 无缝集成到统一的模型骨干中。这意味着模型在预训练阶段就可以同时学习文本、图像和视频数据之间的关联。Llama 4 还**改进了视觉编码器**，该编码器基于 MetaCLIP，但与一个冻结的 Llama 模型联合训练，以更好地适应 LLM。模型在预训练阶段最多可以处理 **48 张图像**，并在后训练中测试了最多 **8 张图像**，效果良好. 此外，**Llama 4 Scout 在图像 grounding 方面表现出色**，能够将用户提示与相关的视觉概念对齐，并将模型响应锚定到图像的特定区域，从而实现更精确的视觉问答.

强大的后训练流程：Llama 4 采用了**轻量级监督微调 (SFT) > 在线强化学习 (RL) > 轻量级直接偏好优化 (DPO)** 的后训练流程。为了提高性能，特别是在推理、编码和数学领域，Meta **移除了超过 50% 的被 Llama 模型判断为简单的 SFT 数据**，并专注于更难的数据集进行轻量级 SFT。在多模态在线 RL 阶段，通过精心选择更难的 prompts，实现了性能的显著提升，并采用了**持续在线 RL 策略和自适应数据过滤**. 最后，使用轻量级 DPO 来处理模型响应质量的边界情况. 针对 Llama 4 Behemoth 这样的大模型，后训练流程进行了大幅调整，例如**修剪了 95% 的 SFT 数据**，并侧重于大规模强化学习以提升推理和编码能力.

安全与防护：Meta 在 Llama 4 的开发过程中高度重视安全问题，从预训练到后训练再到系统层面都采取了多项缓解措施. 这包括**数据过滤**、**安全微调**以及开源的**系统级防护工具**，如 **Llama Guard**（用于检测输入/输出是否违反策略）、**Prompt Guard**（用于检测恶意 prompt 和 prompt 注入）和 **CyberSecEval**（用于评估和降低网络安全风险）. Meta 还进行了广泛的**评估和红队测试**，使用了 **Generative Offensive Agent Testing (GOAT)** 等自动化工具来模拟对抗性攻击，并特别关注儿童安全、网络攻击和CBRNE等关键风险领域. 此外，Llama 4 在**减少偏见**方面也取得了显著进展，例如在有争议的政治和社会话题上的拒绝率更低，并且对不同观点的响应更加平衡.

开源与易于获取：秉承开放的理念，**Llama 4 Scout 和 Llama 4 Maverick 模型已在 llama.com 和 Hugging Face 上开放下载**. 这使得开发者和研究人员可以轻松地使用和构建基于这些先进模型的新应用.

总而言之，Llama 4 模型系列通过其**原生多模态能力**、**创新的混合专家架构**、**超长的上下文窗口**以及对**效率、性能和安全性**的全面提升，无疑站在了当前AI技术的最前沿。

以下是Llama 4与其他知名大模型（如GPT-4o、Gemini 2.0、DeepSeek v3.1等）的横向对比表：

Llama 4 模型系列

版本	发布日期	多模态能力	参数规模	训练成本 (petaFLOP-day)	上下文长度 (tokens)	语料规模 (tokens)	架构类型	MoE使用	备注
Llama 1 骆驼1	2023-02	否，文本-only	6.7B, 13B, 32.5B, 65.2B	6,300	2048	1~1.4T	变压器 (decoder-only)	否	自动回归语言模型，仅文本输入输出
Llama 2 骆驼2	2023-07	否，文本-only	6.7B, 13B, 69B	21,000	4096	2T	变压器 (decoder-only), GQA	否	扩展了语言支持，但无图像处理
Code Llama	2023-08-24	否，文本-only	13B, 33.7B, 69B	-	-	-	-	-	专注代码生成任务
Llama 3 骆驼3	2024-04	否，文本-only	8B, 70.6B	100,000	8192	15T	变压器 (decoder-only), GQA	否	输入输出均为文本
Llama 3.1 骆驼3.1	2024-07-23	否，文本-only	70.6B, 405B	440,000	128,000	-	-	-	专注于对话任务
Llama 3.2 骆驼3.2	2024-09-25	是，支持文本和图像	1B, 3B, 11B, 90B	-	128,000	-	变压器 + 视觉适配器	否	支持视觉任务
Llama 3.3 骆驼3.3	2024-12-07	否，文本-only	70B	-	128,000	-	变压器 (decoder-only)	否	专注于多语言对话，无图像处理
Llama 4 骆驼4	2025-04-05	是，支持文本和图像	109B, 400B, 2T	71,000, 34,000, ?	10M, 1M, ?	40T, 22T, ?	混合专家 (MoE) + 变压器	是	原生多模态，采用混合专家架构

Llama 4与其他模型横向比较

模型对比表格

指标	Llama 4 Scout	Llama 4 Maverick	Llama 4 Behemoth	GPT-4o	GPT-4.5	Gemini 2.0	Gemini 2.5	DeepSeek v3.1
参数规模	17B 活跃参数 / 109B 总参数	17B 活跃参数 / 400B 总参数	288B 活跃参数 / 2T 总参数	~~220B (MoE, 8x~~27B) [非官方]	未公开 (推测~5T 总参数)	未公开	未公开	37B 活跃参数 / 671B 总参数
专家数	16(MoE + Dense)	128(MoE + Dense)	16(MoE + Dense)	8 (MoE) [非官方]	MoE (数量未公开)	无专家架构	未公开	MoE 架构
总参数量	109B	400B	2T	~220B [非官方]	未公开	未公开	未公开	**671B**
上下文长度	10M tokens	1M tokens	未明确	128K tokens	128K tokens	256K tokens (基于 1.5 Pro 推测)	1M tokens	128K tokens
多模态能力	文本+图像输入	文本+图像输入	文本+图像+视频输入	文本+图像+音频输入	文本+图像输入	文本+图像+音频输入	未公开	文本+图像输入

指标	Llama 4 Scout	Llama 4 Maverick	Llama 4 Behemoth	GPT-4o	GPT-4.5	Gemini 2.0	Gemini 2.5	DeepSeek v3.1
推理能力	强，支持代码推理	接近 DeepSeek v3.1	STEM 领域表现卓越	强	强，综合推理能力领先	强 (基于 Gemini 1.5 Pro)	未公开	**顶尖数学与代码能力**
性能表现	- 超越 Gemini 2.0 Flash-Lite - 长上下文任务第一	- 超越 GPT-4o - 强大的多模态能力	- STEM 超越 GPT-4.5 和 Claude 3.7	综合性能强，但长上下文受限	多项基准超越 GPT-4o	多模态领先，推理弱于 Llama 4	未公开	**推理与代码生成顶尖**
成本效率	高 (单 H100 可运行)	高 (单 H100 主机)	未明确	中等	成本高 (API: $75 输入 +$150输出/M tokens)	中等 (对比 GPT-4o)	未公开	高 (¥2输入 +¥8输出/M tokens)
训练架构	混合专家架构（MoE）	混合专家架构（MoE）	混合专家架构（MoE）	混合专家架构（MoE）[非官方]	MoE + SFT/RLHF	密集架构	未公开	混合专家架构（MoE）
安全性	Llama Guard、Prompt Guard	同 Scout	同 Scout	安全过滤和监控	安全对齐机制	安全过滤和监控	未公开	未明确

性能比较

以下数据来自https://artificialanalysis.ai/ （2025-04-06）

Metrics	Llama 4 Maverick	Llama 4 Scout	DeepSeek V3 (Mar' 25)
AA Intelligence Index	49	36	**53**
Output Speed, Tokens/s	**127**	104	27
Price, Blended USD/1M Tokens	$0.40	**$0.30**	$0.50
Coding Index	35	23	**38**
Math Index	64	56	**73**
MMLU-Pro (%)	80	58	**82**
GPQA Diamond (%)	60	34	**66**
Humanity's Last Exam (%)	4.8	4.3	**5.2**
LiveCodeBench (%)	38	30	**41**
SciCode (%)	33	17	**36**
HumanEval (%)	88	83	**92**
MATH-500 (%)	89	84	**94**
AIME 2024 (%)	39	28	**52**

技术细节

这部分在4月29日后，在得到更具体Tech Report后可能要更新更多技术细节。

如何实现原生多模态

lama 4 通过 早期融合 (early fusion) 的方式，将文本和视觉 tokens 无缝集成到统一的模型骨干中。 好的，我们来详细展开"Llama 4 通过 **早期融合 (early fusion)** 的方式，将文本和视觉 tokens 无缝集成到统一的模型骨干中。"这句话。

根据来源，**Llama 4 模型系列是首个原生支持多模态 (natively multimodal)** 的 Llama 模型，这意味着它们从一开始就被设计成能够理解和处理文本以及图像输入，并生成文本和代码输出. 实现这一能力的关键技术之一就是**早期融合 (early fusion)**.

早期融合指的是在模型的**早期阶段**，就将来自不同模态（例如文本和视觉）的信息进行整合. 具体来说，对于 Llama 4 而言，这意味着：

> **无缝集成文本和视觉 tokens**: Llama 4 的架构能够**将文本和图像的 tokens (tokens)** 在**统一的模型骨干 (unified model backbone)** 中进行无缝集成. 这与一些后期融合 (late fusion) 的方法不同，后者可能先独立处理不同模态的信息，然后在模型的较后阶段才进行融合。
>
> **联合预训练 (Joint Pre-training)**: 早期融合是实现**联合预训练 (joint pre-training)** 的重要一步. 通过在**大量的未标注文本、图像和视频数据**上进行联合预训练，Llama 4 能够学习不同模态数据之间的**关联性 (associations)**. 这使得模型能够更好地理解文本描述与对应图像之间的关系，以及视频中包含的视觉和文本信息。

改进的视觉编码器: 为了更好地实现早期融合，Llama 4 **改进了其视觉编码器 (vision encoder)**. 这个视觉编码器是**基于 MetaCLIP** 构建的，但是它是**与一个冻结 (frozen) 的 Llama 模型联合训练**的. 这样的联合训练使得视觉编码器能够更好地**适应 LLM (Large Language Model)** 的特性和需求，从而更有效地将视觉信息转化为模型可以理解的 tokens。

处理多张图像: Llama 4 在预训练阶段**最多可以处理 48 张图像**. 并且在后训练 (post-training) 阶段测试了**最多 8 张图像**，并取得了良好的效果. 这表明早期融合的架构能够有效地处理多个视觉输入。

图像 Grounding 能力: Llama 4 Scout 在**图像 grounding (image grounding)** 方面表现出色. 这意味着它能够将用户的文本提示与图像中**相关的视觉概念对齐**，并将模型的回复**锚定到图像的特定区域**. 这种能力也得益于早期融合带来的更深层次的文本和视觉信息理解。

总而言之，Llama 4 通过早期融合的技术，在模型架构的早期就将文本和视觉信息融合在一起进行处理和学习。这种方法使得模型能够更自然、更深入地理解多模态输入，为构建更强大的多模态 AI 应用奠定了坚实的基础.

如何采用MoE架构的

以下是 Llama 4 实现 MoE 的具体细节：

MoE 架构的核心思想：在 MoE 模型中，对于每一个输入的 **token (tokens)**，只有模型总参数的一个**子集 (fraction)** 会被激活。这种机制使得 MoE 架构在相同的计算资源下，能够拥有比密集模型**更高的模型质量 (higher quality)**。同时，这种稀疏激活也提高了**训练和推理的计算效率 (compute efficient for training and inference)**。

Llama 4 模型中的 MoE 层：Llama 4 的模型架构中使用了 **交替的密集层 (dense layers) 和混合专家层 (mixture-of-experts layers)**，以提高**推理效率 (inference efficiency)**。

Llama 4 Maverick 的 MoE 实现：

Llama 4 Maverick 模型拥有 **170 亿的激活参数 (17B active parameters)** 和 **4000 亿的总参数 (400B total parameters)**。

在 MoE 层中，模型使用了 **128 个路由专家 (128 routed experts)** 和 **1 个共享专家 (a shared expert)**。

当一个 token 输入到 MoE 层时，该 token 会被发送到 **共享专家** 以及 **128 个路由专家中的一个**。

这意味着，虽然模型的所有参数都存储在内存中，但在服务 (serving) 这些模型时，**只有总参数的一个子集被激活**。

这种设计通过**降低模型服务成本和延迟 (lowering model serving costs and latency)** 来提高推理效率。因此，Llama 4 Maverick 可以在**单个 NVIDIA H100 DGX 主机 (a single NVIDIA H100 DGX host)** 上轻松部署，或者通过**分布式推理 (distributed inference)** 来实现更高的效率。

Llama 4 Scout 的 MoE 实现：

Llama 4 Scout 模型同样拥有 **170 亿的激活参数 (17 billion active parameter model)**，但它使用了 **16 个专家 (16 experts)**。其总参数为 **1090 亿 (109B total)**。

Llama 4 Scout (Int4 量化后) 可以容纳在**单个 NVIDIA H100 GPU (a single H100 GPU)** 上运行。

Llama 4 Behemoth 的 MoE 实现：

作为 Llama 4 Scout 和 Maverick 的教师模型，Llama 4 Behemoth 也是一个 **多模态混合专家模型 (multimodal mixture-of-experts model)**，拥有 **2880 亿的激活参数 (288B active parameters)** 和 **近两万亿的总参数 (nearly two trillion total parameters)**，使用了 **16 个专家 (16 experts)**。

为了支持如此大规模模型的强化学习训练，Meta **优化了其 MoE 并行化的设计以提高速度 (optimized the design of our MoE parallelization for speed)**。

总而言之，Llama 4 通过 MoE 架构，特别是通过在推理时只激活部分参数，实现了在保持甚至提升模型性能的同时，提高了计算效率和降低了部署成本。不同型号的 Llama 4 模型（如 Scout 和 Maverick）在激活参数数量和专家数量上有所不同，以适应不同的使用场景和性能需求。

MetaP - 新的训练技术

Meta 在 Llama 4 模型的预训练过程中开发了一种新的训练技术，称为 **MetaP**。这项技术的主要作用是能够**可靠地设置关键的模型超参数**，例如**每层的学习率 (per-layer learning rates)** 和**初始化尺度 (initialization scales)**。

Meta 发现通过 MetaP 选择的超参数在不同的**批量大小 (batch size)**、**模型宽度 (model width)**、**模型深度 (depth)** 以及**训练 tokens 数量**上都能够很好地迁移。

总而言之，MetaP 是一种用于**自动化和优化模型超参数设置**的训练技术，它能够提高 Llama 4 模型在不同训练配置下的鲁棒性和性能。

长上下文扩展

Llama 4 模型在预训练之后，进行了一个称为 **"mid-training"** 的阶段，目的是通过新的训练方法和**专门的数据集**来提升模型的核心能力，其中包括**长上下文扩展**。这个阶段对于提升模型质量，并最终使 Llama 4 Scout 实现了 **10M 的超长输入上下文窗口**起到了关键作用。

以下是这个阶段的技术细节展开：

目标：提升核心能力和扩展上下文长度。"Mid-training" 的主要目标是通过专注于特定能力（例如处理更长的文本序列）来进一步提高模型的性能。

使用专门的数据集进行长上下文扩展。在这个阶段，Meta 使用了**特殊设计的数据集**，这些数据集可能包含非常长的文本序列，旨在训练模型处理和理解更广泛的上下文信息。

Llama 4 Scout 的 256K 上下文长度基础。值得注意的是，Llama 4 Scout 在**预训练和后训练阶段都使用了 256K 的上下文长度。**这为模型提供了先进的长度泛化能力，为后续的 10M 上下文窗口奠定了基础。

iRoPE 架构的关键创新。Llama 4 架构中的一个关键创新是使用了 **交错的注意力层 (interleaved attention layers) 而不使用位置嵌入 (positional embeddings)**。这种架构被称为 **iRoPE (interleaved RoPE)**，其中 "i" 代表

"interleaved" 注意力层，强调支持 "无限" 上下文长度的长期目标，而 "RoPE" 指的是大多数层中使用的旋转位置嵌入 (rotary position embeddings)。

推理时注意力温度缩放 (Inference time temperature scaling of attention)。 为了进一步增强长度泛化能力，Llama 4 Scout 在推理时采用了**注意力温度缩放 (inference time temperature scaling of attention)** 技术。

总而言之，"mid-training" 阶段通过引入专门的长上下文数据集，并结合 Llama 4 架构本身在处理长序列方面的创新（如 iRoPE 架构和推理时注意力温度缩放），使得 Llama 4 Scout 能够显著扩展其上下文窗口至 10M tokens，同时也提升了模型的整体质量。在预训练和后训练阶段就具备的 256K 上下文长度也为这一突破奠定了坚实的基础。

优化的强化学习 (RL)

根据来源，Llama 4 模型系列在**后训练 (post-training)** 阶段采用了**优化的强化学习 (RL)** 技术，以提升模型的性能，尤其是在对话、推理和编码等方面。以下是详细的技术细节：

Llama 4 Maverick 的后训练流程: Llama 4 Maverick 的后训练流程包括了**轻量级的监督微调 (SFT)**、**在线强化学习 (RL)** 和**轻量级的直接偏好优化 (DPO)**。

解决 SFT 和 DPO 的过度约束: Meta 发现 SFT 和 DPO 可能会**过度约束模型**，限制在线 RL 阶段的探索，导致在推理、编码和数学等领域表现次优。

自适应数据过滤的连续在线 RL 策略: 为了解决上述问题，Meta 对其后训练流程进行了改进，采用了**连续在线 RL 策略**，并结合了**自适应数据过滤**。

难度分级数据: 他们使用 Llama 模型作为裁判，**移除了超过 50% 被标记为容易的数据**，并对剩余的难度较高的数据进行了轻量级 SFT。

选择困难提示: 在随后的多模态在线 RL 阶段，通过**精心选择更具挑战性的提示 (harder prompts)**，实现了性能的显著提升。

持续在线 RL 和动态过滤: 他们实施了一种**持续在线 RL 策略**，模型在训练的同时被用于**不断过滤和保留中等到高难度的提示**。这种策略在计算成本和准确性之间取得了很好的平衡。

Llama 4 Behemoth 的大规模 RL: 对于拥有两万亿参数的 Llama 4 Behemoth 模型，其 RL 过程也进行了大规模的优化。

剪枝 SFT 数据: 为了最大化性能，Behemoth 模型**剪枝了 95% 的 SFT 数据**，以专注于质量和效率。

大规模 RL 的重要性: 轻量级 SFT 之后进行**大规模强化学习 (RL)**，显著提升了模型的推理和编码能力。

基于策略模型的 pass@k 分析: RL 的重点是通过对策略模型进行 **pass@k 分析**来**采样困难提示**，并构建一个难度逐渐增加的训练课程。

动态过滤零优势提示和混合提示批次: 在训练过程中**动态过滤掉零优势的提示**，并使用来自多种能力的**混合提示构建训练批次**，这对于提升数学、推理和编码方面的性能至关重要。

多样化的系统指令采样: 采样各种**系统指令**对于确保模型保留其在推理和编码方面的指令遵循能力，并在各种任务中表现良好至关重要。

RL 基础设施的革新: 为了支持两万亿参数模型的 RL 训练，Meta **彻底改造了其底层的 RL 基础设施**。

优化 MoE 并行化: 他们优化了 **MoE 并行化的设计以提高速度**，从而加快了迭代。

全异步在线 RL 训练框架: 开发了一个**全异步在线 RL 训练框架**，增强了灵活性。

灵活的 GPU 资源分配: 相较于将所有模型堆叠在内存中而牺牲计算内存的现有分布式训练框架，新的基础设施能够**将不同的模型灵活地分配到独立的 GPU 上**，基于计算速度平衡跨多个模型的资源。这使得训练效率比以前提升了约 10 倍。

总而言之，Llama 4 通过在后训练阶段采用精细化的 RL 策略，包括针对中高难度数据的持续在线 RL、自适应数据过滤以及对大规模模型 RL 基础设施的革新，显著提升了模型的智能和对话能力。这些优化使得 Llama 4 模型在多个基准测试中都取得了优异的成绩.

参考文献

截至2025年4月7日，有证据表明，GitHub上的模型卡是Llama 4的主要技术文档，提供了详细的规格和性能指标。目前还没有迹象表明会有单独的技术报告或论文发布，但Meta提到LlamaCon表明有可能会有额外的文档，可能包括一份正式的论文，将于4月晚些时候发布。这与历史模式一致，因为以前的版本，如Llama 3有模型卡和arXiv文件，但考虑到最近的Llama 4的发布，目前的文档可能已经足够了。

Llama 4 模型卡

Llama 4 博客

Llama 维基百科

Llama 4 吃瓜

根据2025年4月第三方评测机构及开发者社区的公开测试结果，Llama 4系列模型的实际表现引发广泛争议，尤其在代码生成、多模态能力和长上下文支持等核心领域存在显著落差。以下是关键测试结论的综合分析：

一、代码生成能力：实测表现远低于预期

Aider Polyglot编码基准测试

Llama 4-Maverick（402B参数） 在六种主流编程语言（Python、Java等）测试中仅得16%得分，与32B参数的Qwen-QWQ-32B相当，显著落后于DeepSeek V3（45.8%）和GPT-4o（40.3%）。

Llama 4-Scout（109B参数） 表现接近Grok-2（14.1%）和ERNIE 4.5（15.6%），处于行业尾部。

LiveCodeBench动态测试

官方宣称的43.4分（Maverick）与第三方实测结果差异显著。例如，在复杂算法实现任务中，Maverick的错误率高达40%，远超DeepSeek V3的22%。

开发者社区反馈

Reddit用户指出，Maverick生成的代码在"20个弹跳球"物理模拟测试中直接穿透墙壁，暴露逻辑推理缺陷。

知乎开发者@deedydas评价其代码生成水平"像实习生"，尤其在多步推理任务中频繁"断片"。

二、综合能力评测：与官方宣传严重不符

Artificial Analysis Intelligence Index

Llama 4-Maverick 综合得分49分，落后于DeepSeek V3（66分）、GPT-4o（64分）和Claude 3.7 Sonnet（62分），仅略高于Gemma 3（36分）。

Llama 4-Scout 得分36分，与GPT-4o Mini持平，但落后于Mistral Small（41分）和Qwen-7B（39分）。

STEM领域表现

数学推理（MathVista）：Maverick得分73.7，低于DeepSeek V3的82.1和GPT-4o的79.5。

科学推理（GPQA Diamond）：Maverick得分57.2，仅为DeepSeek V3（68.0）的84%。

三、多模态与长上下文：实际效果未达预期

多模态理解

图像理解能力落后于GPT-4o和Claude 3.5。在MMMU（多模态多任务理解）测试中，Maverick的图文匹配准确率仅69.4%，而GPT-4o达到82.1%。

长上下文支持

官方宣称的1000万token上下文窗口存在严重衰减：
1K长度时召回率跌破60%，16K时仅剩22%，远低于Gemini 2.5 Pro的92%。

实际应用中，用户反馈长文本生成内容重复率高，公式化表达明显。

四、争议焦点：数据污染与评测作弊质疑

训练数据泄露

匿名Meta前员工爆料，模型训练后期混入多个基准测试的测试集数据，类似"高考前提前拿到试卷"，导致过拟合。

开发者发现竞技场（LM Arena）榜单版本与公开下载版本行为差异显著，例如大量使用表情符号和冗长回答。

榜单刷分争议

TechCrunch指出，Maverick在LM Arena的1417分依赖"针对对话优化的特殊版本"，与实际模型能力脱节。

知名社区Kcores LLM Arena管理员称，Llama 4的排名可能通过"定制化数据清洗"人为提升。

五、行业影响与Meta的应对

开源生态信任危机

开发者转向闭源方案，称"开源模型的承诺正在褪色"。

Hugging Face论坛出现《Llama 4：开源理想主义的终结？》热帖，质疑Meta滥用开源名义。

Meta内部动荡

Llama 4发布前，AI研究主管Joelle Pineau突然离职，匿名员工称其与管理层在技术路线上存在分歧。

核心团队流失导致技术路线调整，例如原计划的MoE架构优化被搁置。

总结：理想与现实的巨大落差

Llama 4的争议暴露了AI竞赛中的深层矛盾：

参数竞赛失效：2万亿参数的"Behemoth"版本未开放下载，实际开放的Scout/Maverick性能未达预期。

评测体系缺陷：现有榜单易被针对性优化攻破，需动态场景化评估标准。

开源商业模式困境：Meta试图通过低价策略（Maverick定价为同类1/10）抢占市场，但技术短板导致用户流失。

建议开发者谨慎选择：若需代码生成，优先考虑DeepSeek V3或Qwen-QWQ；若需多模态能力，GPT-4o和Claude 3.5仍是更可靠的选择。

还是让子弹再飞一会吧。

参考文献

Llama 4 吃瓜

Llama 3.2 从零开始

LLM

minGPT

zero-to-hero

Hacker News

Hacker News

Hacker News

arxiv

Hacker News

Build Your Own Llama 3 Architecture from Scratch Using PyTorch

GPT in 60 Lines of NumPy

分享这篇文章

相关文章推荐

Chain of Draft 论文解读

本文介绍了 Chain of Draft（CoD）论文，并对其技术...

DeepSeek FlashMLA 代码解读

本文介绍了深度求索（DeepSeek）公司FlashMLA...

Test-Time Scaling 相关论文解读

本文介绍了 Test-Time Scaling（测试时扩展）的概念，并对...