llama3

一个月前发布的Llama 3 有两个版本：8B和70B。在架构上，Llama 系列模型采用了仅decoder-only结构，用于多种配置下的下一个标记预测，参数规模从 80 亿到 700 亿不等。

1、tokenizer：由sentencePiece换成了Tiktoken，并且词表从llama2的32k扩展为128k

2、GQA：作为多头注意力（Multi-Head Attention, MHA）和多查询注意力（Multi-Query Attention, MQA）变体之间的折中方案。GQA 通过在注意力头之间分组键和值，从而减少 KVC 引入的内存访问瓶颈，降低了启用 KVC 的 MHA 的整体内存占用。通过在计算注意力之前聚合多个查询，GQA 显著加快了推理时间，并减少了操作所需的缓存大小，从而优化了计算速度和内存利用率。

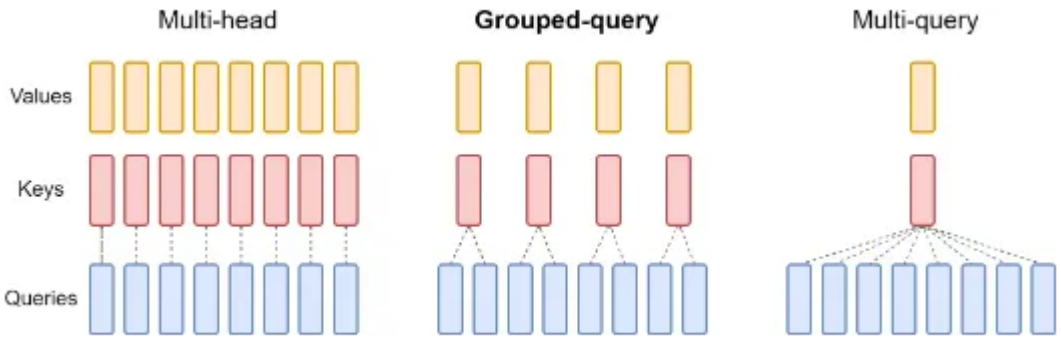


Figure 3: MHA vs GQA vs MQA

3、RMSNorm：预归一化，通过减少标准层归一化通常带来的计算负担，提高了神经网络训练的效率。与传统方法需要调整均值和标准差不同，RMS 归一化通过仅基于输入特征的均方根值进行缩放，简化了这一过程。这种方法减轻了内部协变量偏移的影响——这一现象会导致前一层输出的剧烈变化，从而在训练过程中需要进行大量的权重调整，进而减慢训练速度。RMS 归一化加速了收敛速度，并促进了更稳定的学习环境。

4、RoPE： Llama 3 通过关注序列中标记之间的相对距离来相对地编码位置信息。RoPE 专门应用于注意力机制计算中的查询和键向量，从而根据标记的接近程度调整注意力强度。这种自适应方法有助于保持上下文的完整性，特别是在处理较长文本时，通过减弱远距离标记关系的影响，避免稀释对相关交互的关注。RoPE 的方法在推理时能够更好地推广到更长的上下文窗口。

5、KV-Cache（KVC）：在推理阶段通过存储先前计算的键和值来增强内存效率，这些键和值可以在后续处理步骤中快速检索和重用。此机制减少了冗余计算，并通过有效降低与传统注意力机制相关的计算负担，加快了推理过程。然而，这些改进是以增加内存占用为代价的，因为 KVC 的大小会随着上下文长度线性增长。这需要在现代 GPU（如 NVIDIA A100）上进行仔细的内存管理，因为内存访问的强度/速度比计算高出 40 倍（FP32 位计算的 19.5TFLOPS 与 1,935GB/s 的内存带宽）。在这种情况下，减少计算负载与管理增加的内存需求之间的权衡变得至关重要。

6、Ghost Attention（GAtt）：由Meta AI首创，通过在对话中每次用户消息之前战略性地重复系统提示，增强了大型语言模型（LLM）对长期指令的遵循和记忆。这种方法通过设置前几轮系统提示标记的损失为零，防止了推理过程中重复生成系统提示，从而解决了LLM在多轮对话中由于上下文长度有限而导致的典型遗忘问题。这一策略不仅保持了响应的相关性和连贯性，还延长了指令在更长对话中的影响。从下图中的热力图可以看出，GAtt迫使每个标记对系统提示给予更高的注意力。个人认为，GAtt更像是一种技巧，而不是实际的创新.....

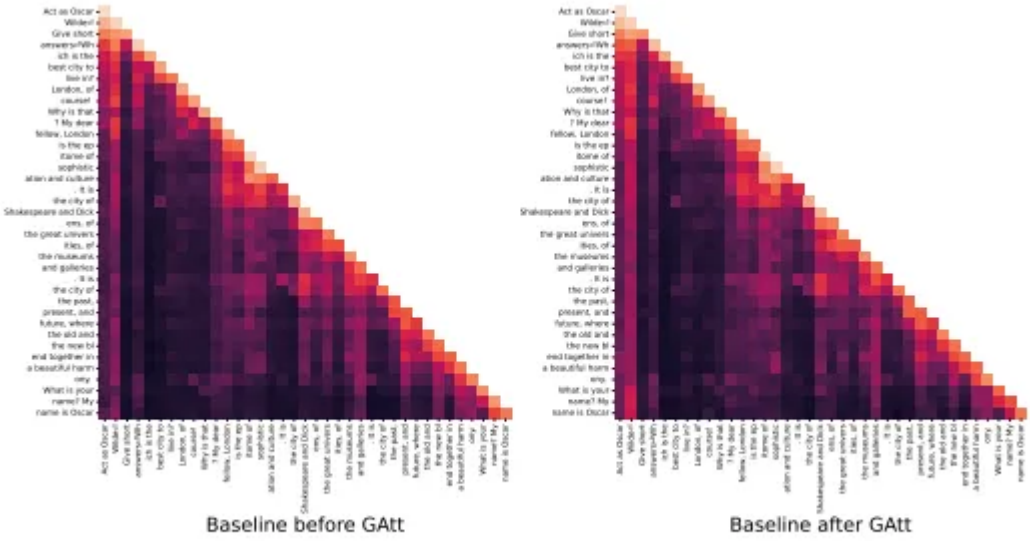


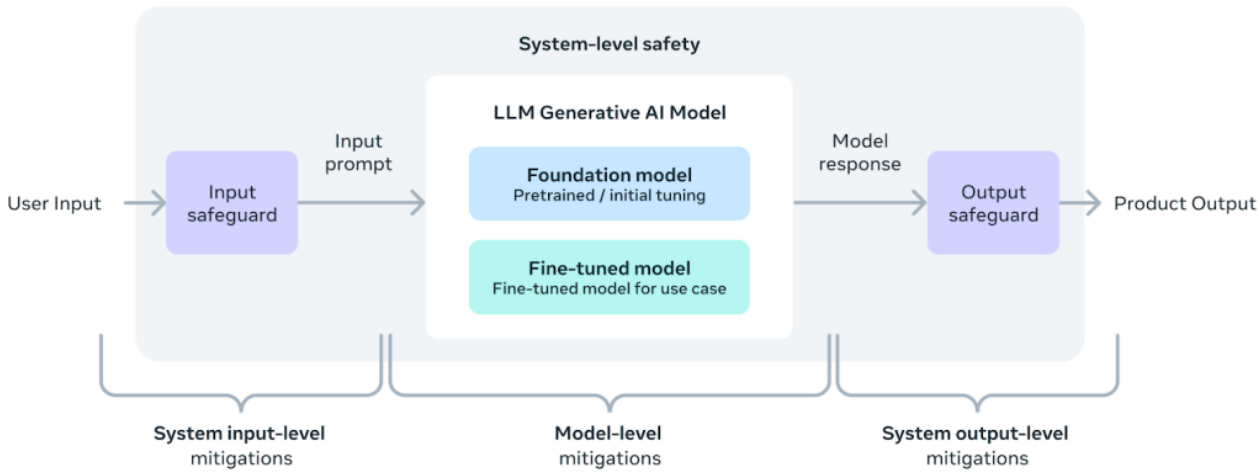
Figure 4: Heat-Map Attention Visualization of Effects on a multi-turn Dialog with and without GAtt

7、混合奖励模型与拒绝采样用于RLHF：

为了充分释放预训练模型在聊天用例中的潜力，还对指令调整方法进行了创新。后训练方法是监督微调SFT、拒绝采样、近端策略优化PPO(关于PPO详见此文《强化学习极简入门：通俗理解MDP、DP MC TC和Q学习、策略梯度、PPO》的第4部分)，和直接策略优化DPO的组合(关于DOP则见此文：《RLHF的替代之DPO原理解析：从RLHF、Claude的RAILF到DPO、Zephyr》)

1. SFT 中使用的prompt质量以及 PPO 和 DPO 中使用的偏好排名对对齐模型的性能有着巨大的影响。最终，在模型质量方面的一些最大改进来自于仔细整理这些数据并对人类标注者提供的标注或注释进行多轮质量保证
2. 通过 PPO 和 DPO 从偏好排名中学习也极大地提高了 Llama 3 在推理和编码任务上的性能。即如果你向模型提出一个它难以回答的推理问题，该模型有时会产生正确的推理轨迹：模型知道如何产生正确的答案，但不知道如何选择它。但对偏好排名的训练使模型能够学习如何选择它

8、SwiGLU 激活函数：结合了 Sigmoid 门控和线性单元，以优于传统的 ReLU 激活通过选择性地允许信息流动来优化神经网络的性能，提供了一种动态的替代方案，增强了 Transformer 架构中的学习能力。



图示：llama3的系统级的负责任方法

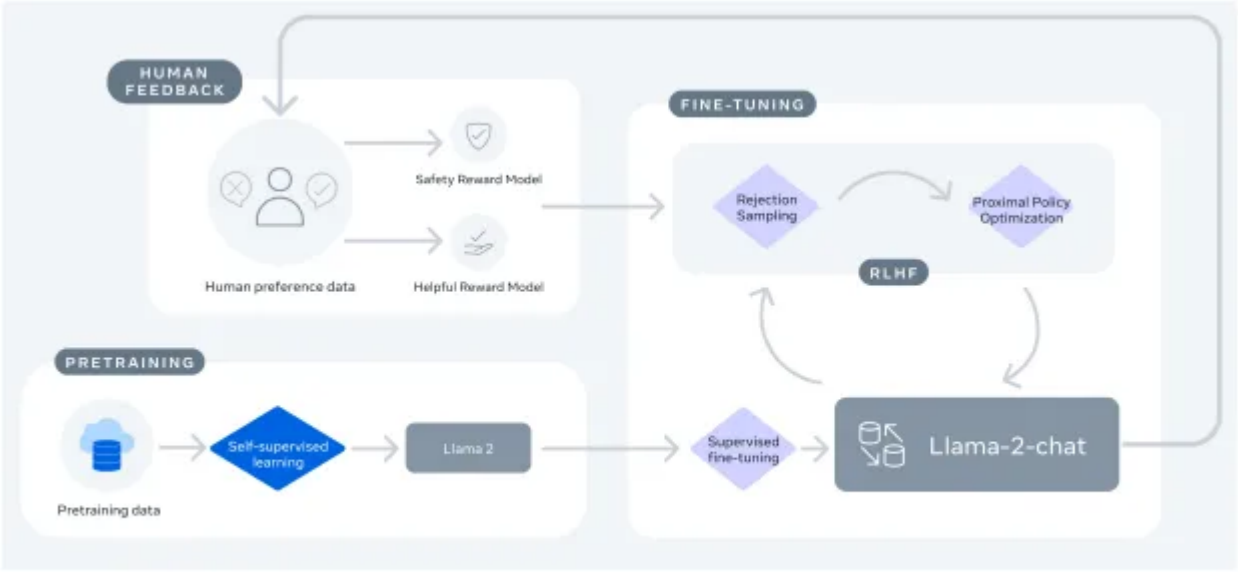


Figure 5: Overall Training Pipeline for Llama 2 — Chat

图示：llama2的训练pipeline