[illegible]

Llama 3

- 128K token [illegible]
- [illegible] GQA[illegible]
- [illegible] 8,192 [illegible] Token [illegible]
- [illegible]
- [illegible] NSFW [illegible] Llama 2 [illegible] Llama 3 [illegible]
- [illegible] 16K [illegible] GPU [illegible] GPU [illegible] 400 TFLOPS [illegible] 24K GPU [illegible]
- [illegible]
- [illegible] SFT[illegible] PPO[illegible] DPO[illegible] SFT [illegible] PPO [illegible] DPO [illegible]

Infini-Transformer

Infini-attention[illegible] compressive memory [illegible] Transformer [illegible]

1. [illegible] Infini-attention [illegible]
2. [illegible] Infini-attention[illegible] Transformer[illegible] masked local [illegible] long-term linear [illegible]

3. 高效的参数重用：在 Infini-attention中，标准注意力计算中的键（ key）、值（ value）和查询（ query ）状态被重用于长期记忆的整合和检索。这种重用策略不仅提高了参数效率，还加速了模型的训练和推理过程。
4. 流式处理能力： Infini-attention 支持对输入序列进行流式处理，这意味着模型可以逐步处理和推理新输入的数据，而不需要一次性处理整个序列。这对于处理实时数据流或非常长的文本序列尤其有用。
5. 持续预训练和任务适应： Infini-attention 设计上支持持续的预训练和针对长上下文的适应，使得模型可以通过持续学习来提高对长序列的处理能力，并且可以通过任务特定的微调来适应不同的应用场景。

总之， Infini-attention 的核心思想是通过压缩记忆和注意力机制的结合，实现对长序列数据的有效处理，同时保持计算和内存效率，这对于推动大型语言模型在各种长文本处理任务中的应用具有重要意义。

LNNs具备两个核心特点

动态架构和持续学习与适应性。

动态架构让 LNNs的神经元比传统神经网络更具表现力，从而提高了模型的可解释性。

而持续学习与适应性则使 LNNs能够在训练后继续适应变化的数据，这一点更接近生物大脑的工作机制。

相较于传统神经网络， LNNs 展现出多重优势。它们不需要大量的标记训练数据就能产生准确的结果，而且模型规模较小，计算需求较低，这使得它们在企业级应用中具有很好的可扩展性。

此外， LNNs对输入信号中的噪声和干扰也表现出更强的鲁棒性。

总的来说，液体神经网络作为一种有前景的神经网络方法，能够出色地处理复杂的实时数据处理任务，并具备良好的适应性，因此在多种应用中都能成为有价值的工具。

Revision #1
Created 11 January 2025 09:44:04 by Colin
Updated 11 January 2025 09:44:04 by Colin