



# Llama 3

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

# Infini-Transformer

- Infini-attention compressive memory  
 Transformer
1.  Infini-attention
  2.  Infini-attention Transformer masked local  
 long-term linear

