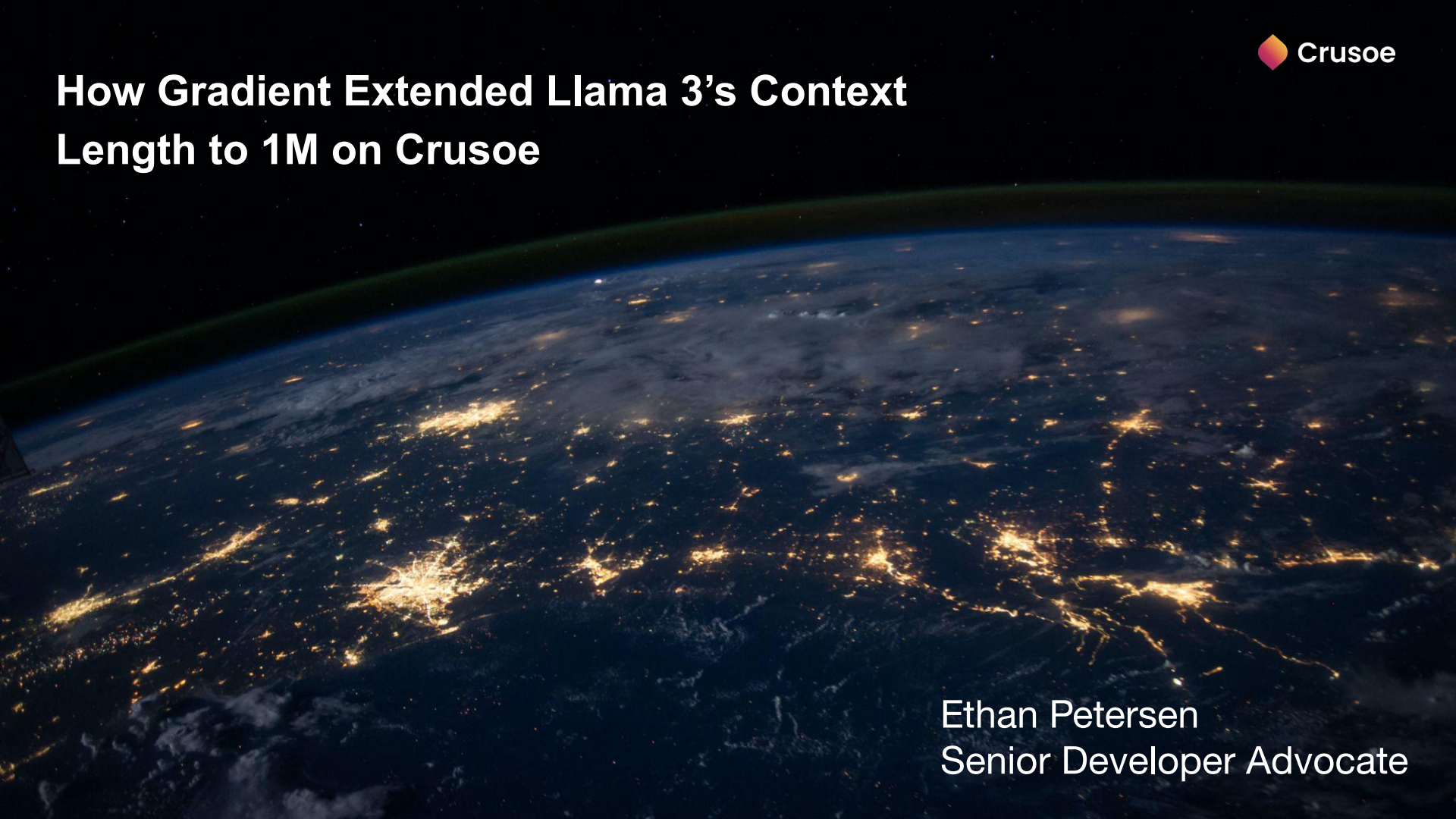
How Gradient Extended Llama 3's Context Length to 1M on Crusoe



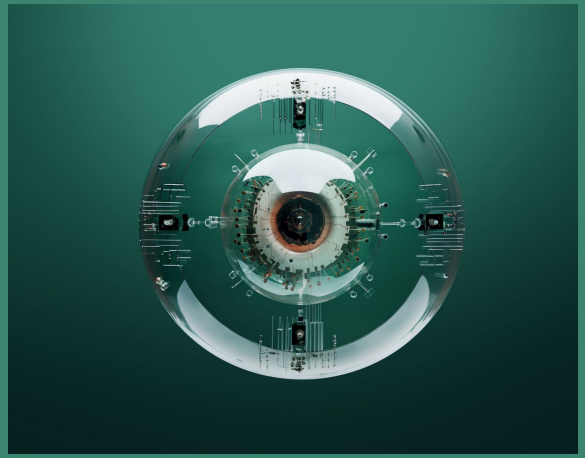
Ethan Petersen
Senior Developer Advocate

Crusoe Cloud

Aligning the future of computing with the future of the climate

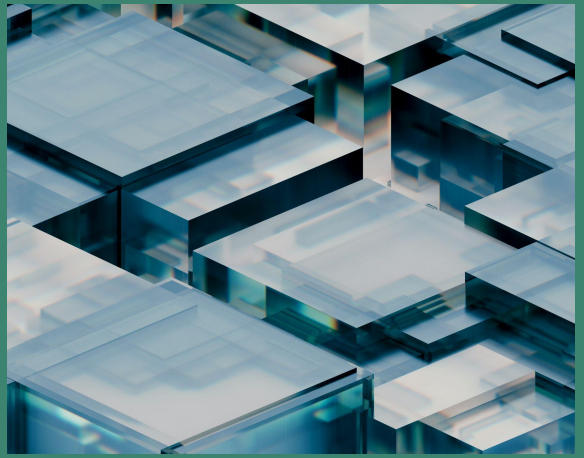
High-performance, AI-first

Compute infrastructure for AI training and Inference



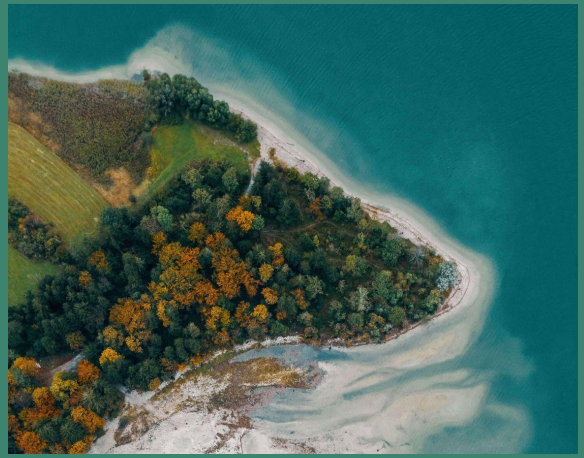
Easy to Use

Simple user interface enabling developers to get started quickly and seamlessly manage their compute environment

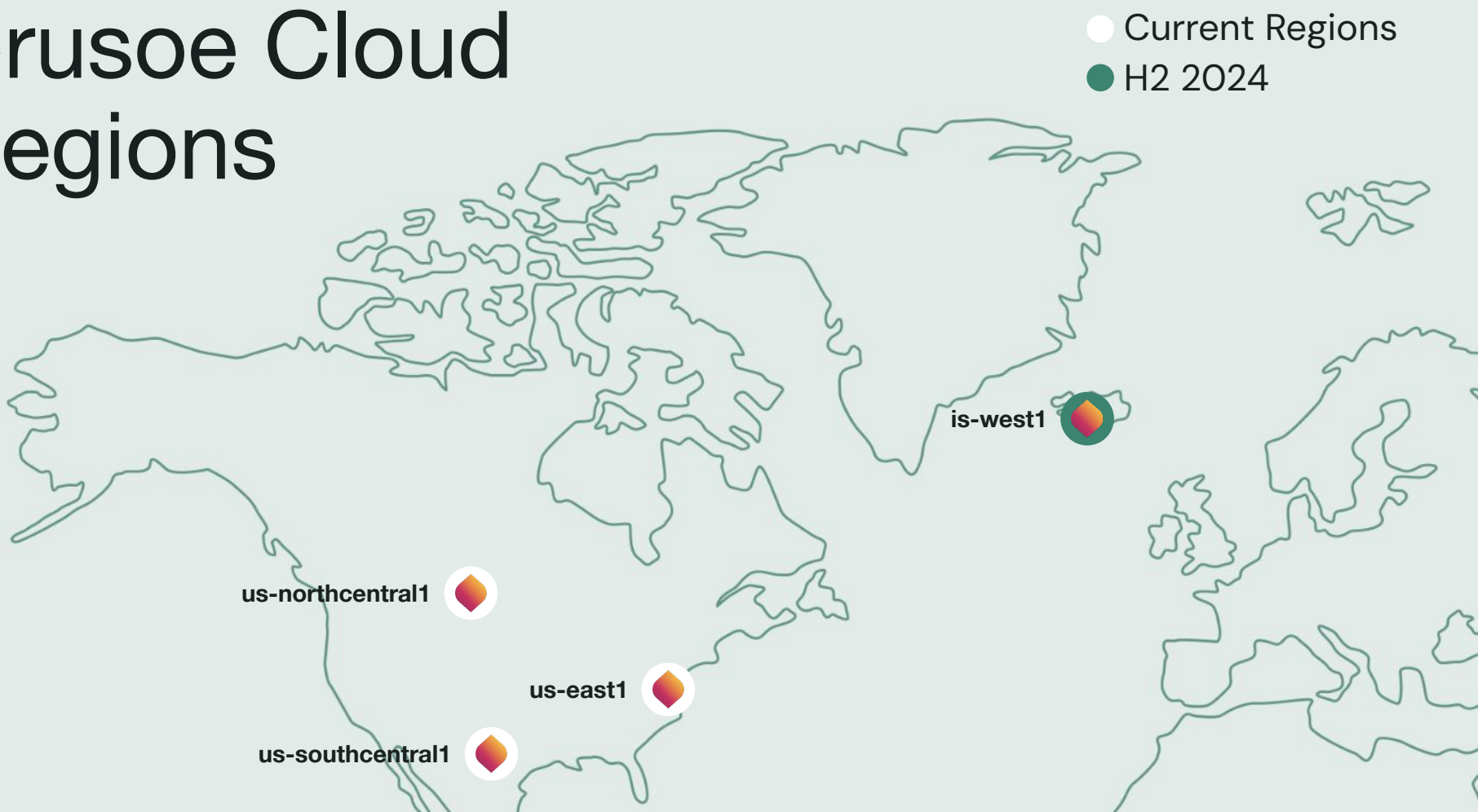


Climate-Aligned

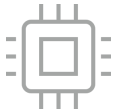
Data centers co-located with sources of clean energy



Crusoe Cloud Regions



Platform Overview



Compute

NVIDIA A100, H100 and L40S VMs for AI training and Inference

Coming Soon - NVIDIA Blackwell based VMs

CPU instances for general purpose compute

Storage

Ephemeral and Persistent disks for fast data access

Coming Soon - Managed Filesystems

Network

VPC Networking

Rail-Optimized Infiniband Cluster Networking

User Experience

Command Line Interface, REST APIs, and management console

Terraform Provider and published solutions for K3s and SLURM

Crusoe Cloud Customers

codeium

together.ai

Massachusetts Institute of Technology

BosonAI

Jua

c/ai

Compute/AI Sponsorships

Axolotl

DBRX

gradient

Long-Context Llama3

LLM

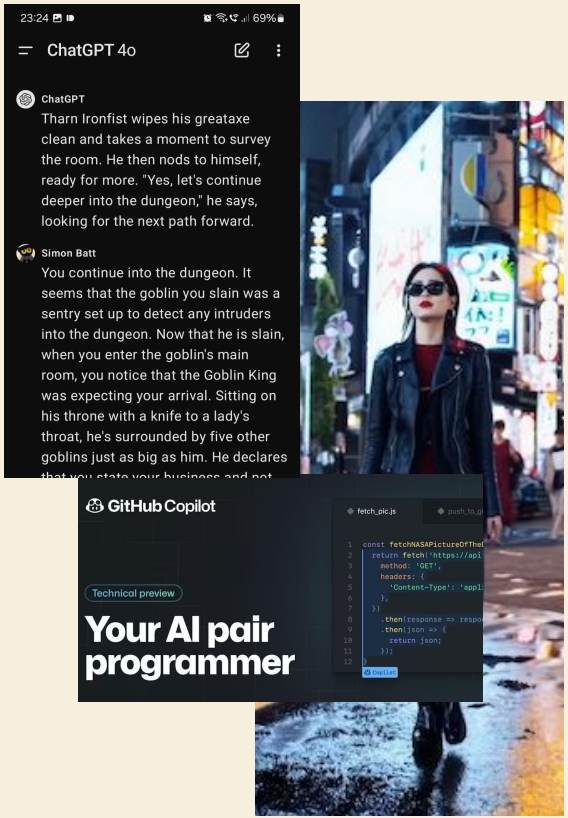
FP8

gradient

How Gradient Extended Llama 3's Context Length to 1M on Crusoe

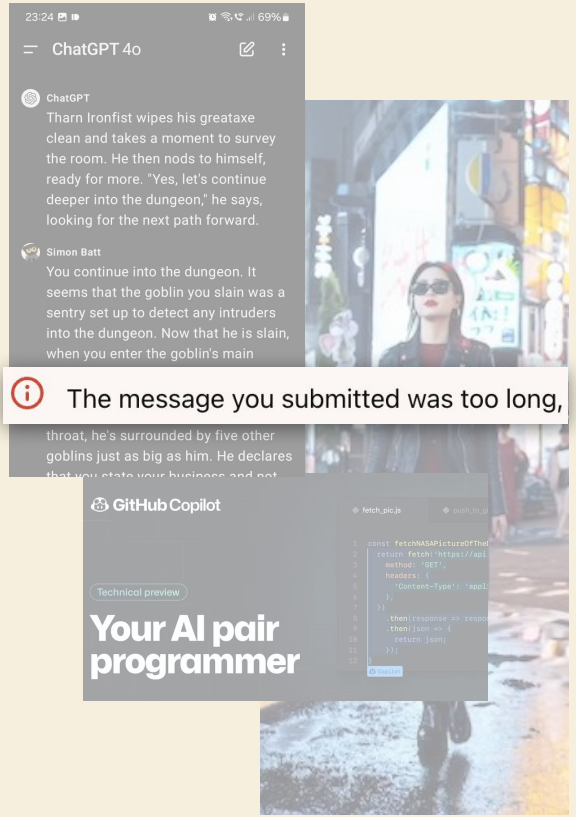
Leonid (Leo) Pekelis
July 31, 2024

Short Context



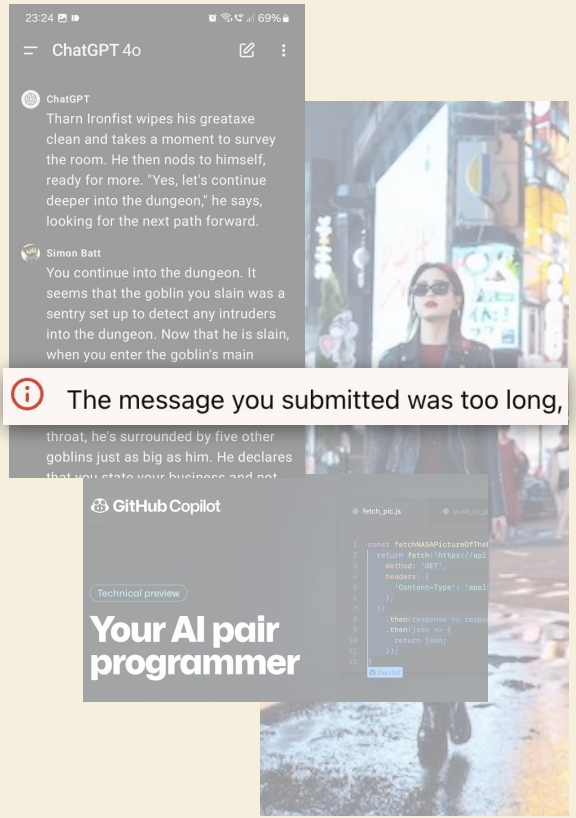
Long Context

Short Context

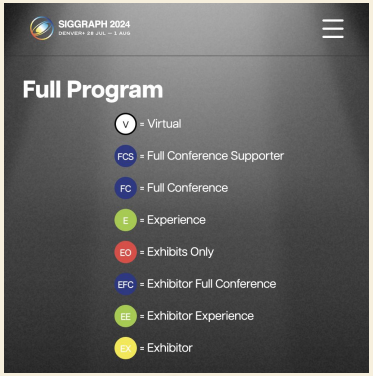


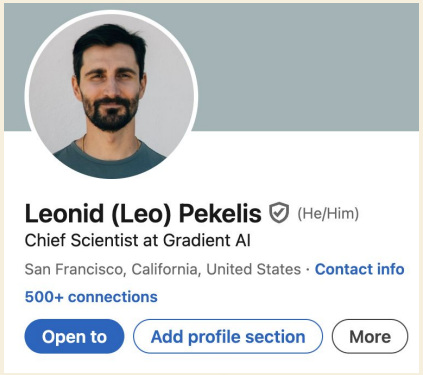
Long Context

Short Context



Long Context





Here is a possible schedule for Leo Pekelis for the first day of SIGGRAPH 2024, taking into account his interests and professional responsibilities:

Morning:

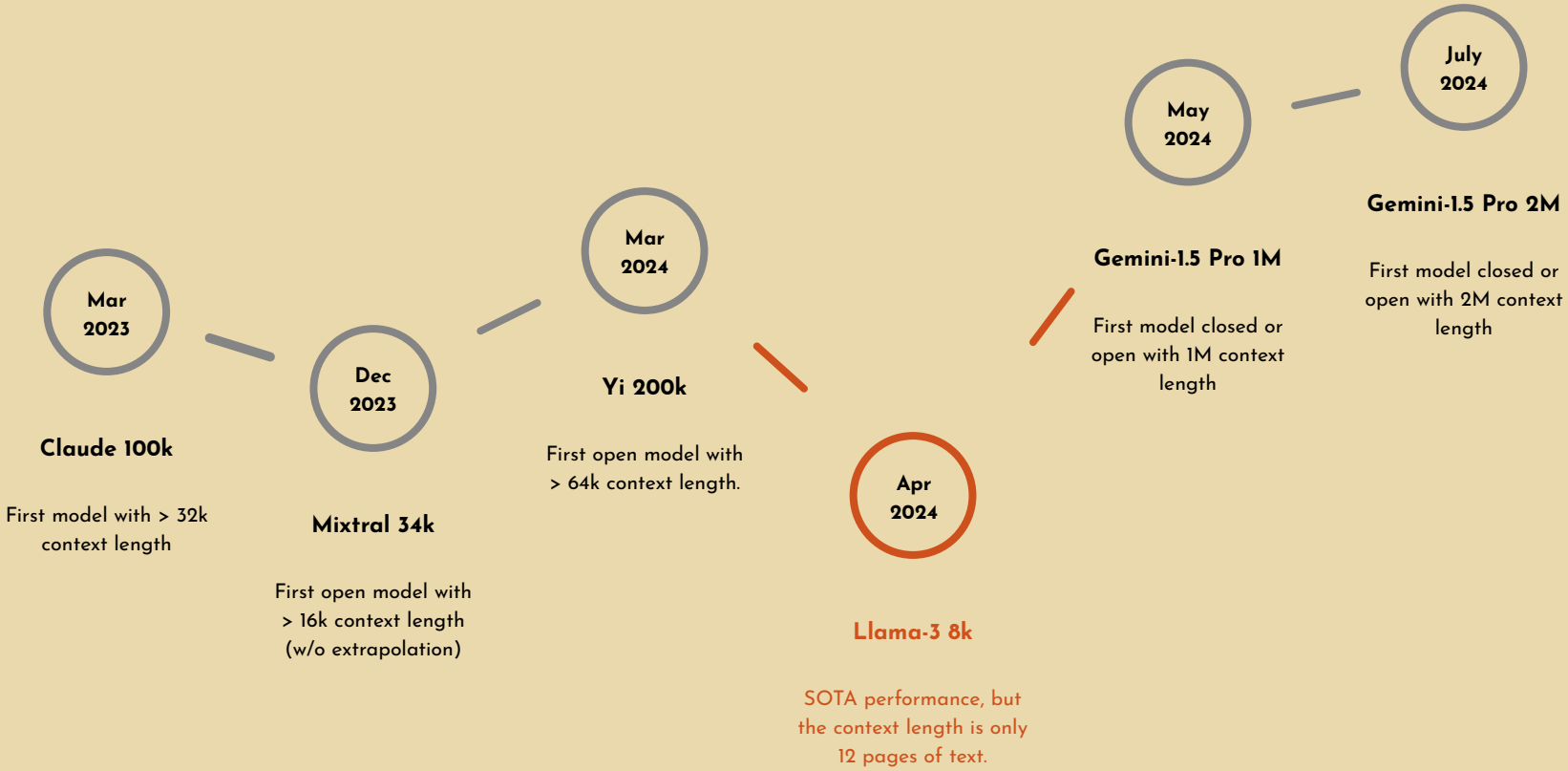
- **9:00 am - 10:30 am:** Attend the "Vector Graphics" Technical Paper session in Mile High 4. This session covers topics related to geometry and modeling, which are areas of interest for Leo.

- **10:45 am - 12:15 pm:** Attend the "VR, Eye Tracking, Perception" Technical Paper session in Mile High 3C. This session includes talks on topics such as saccade-contingent rendering and perceptual evaluation of steered retinal projection, which are relevant to Leo's interest in VR and perception.

...

gradient

Only 8k?!



gradient

Gradient 🤝 Crusoe extended Llama-3

Mar 2023

Claude 100k

First model with > 32k context length

Dec 2023

Mixtral 34k

First open model with > 16k context length (w/o extrapolation)

Mar 2024

Yi 200k

First open model with > 64k context length.

Llama-3 8B (Gradient) 4M

Apr 2024

Apr 2024

Llama-3 70B (Gradient) 1M

Apr 2024

Llama-3 8k

SOTA performance, but the context length is only 12 pages of text.

May 2024

Gemini-1.5 Pro 1M

~~First~~ Second 😄 model closed or open with 1M context length

July 2024

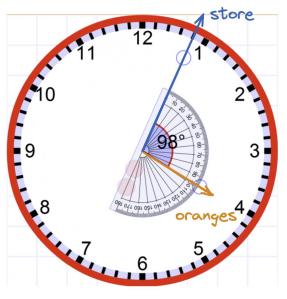
Gemini-1.5 Pro 2M

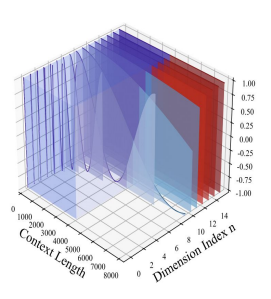
~~First~~ Second 😄model closed or open with 2M context length

1M = 6 HP books

with two technological advancements

1 Long reading comprehension

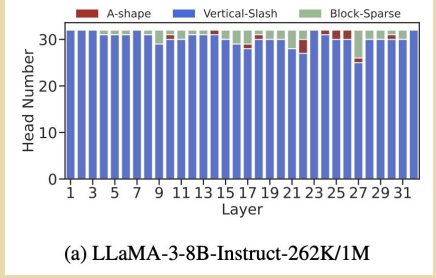




Liu, Xiaoran, et al. "Scaling laws of rope-based extrapolation." (2023).

2 Network aware context parallelism

then we open-sourced it 🤗



Jiang, Huiqiang, et al. "MInference 1.0: Accelerating Pre-filling for Long-Context LLMs via Dynamic Sparse Attention." (2024).

- ► Downloaded over 100,000 times, and used for downstream open research
- ► Only publicly available model with near perfect NIAH score up to 2M
- ► 4th place on RULER long context benchmark despite only training for 1.4B additional tokens