

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

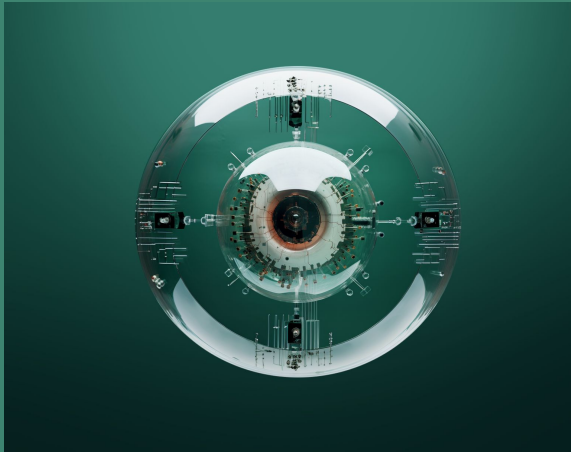
Ethan Petersen  
Senior Developer Advocate

# Crusoe Cloud

Aligning the future of computing with the future of the climate

## High-performance, AI-first

Compute infrastructure for AI training and Inference



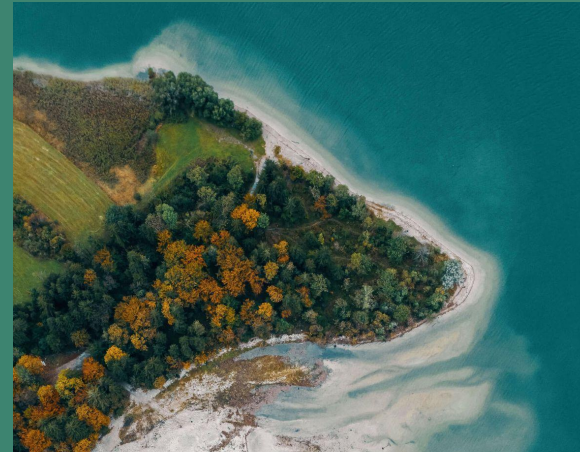
## Easy to Use

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

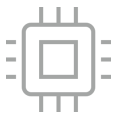


# Crusoe Cloud Regions

- Current Regions
- H2 2024



# 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



together.ai



JuDA

c/ai



# Compute/AI Sponsorships



**Axolotl**

DBRX



Long-Context Llama3



FP8

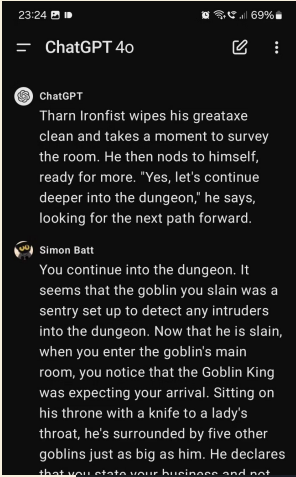


# How Gradient Extended Llama 3's Context Length to 1M on Crusoe

Leonid (Leo) Pekelis  
July 31, 2024



# Short Context



Technical preview

## Your AI pair programmer

```

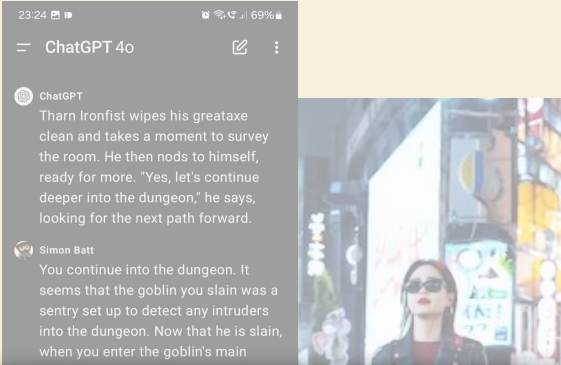
1 const fetchNASAPictureOfThe
2   return fetch("https://api
3   method: 'GET',
4   headers: {
5     'Content-Type': 'appl
6   }
7 }
8
9 then(response) => {
10  then(json => {
11    return json;
12  });
13 }
14
15

```

# Long Context



# Short Context



**i** The message you submitted was too long,

throat, he's surrounded by five other goblins just as big as him. He declares that you state your business and get

GitHub Copilot

Technical preview

Your AI pair programmer

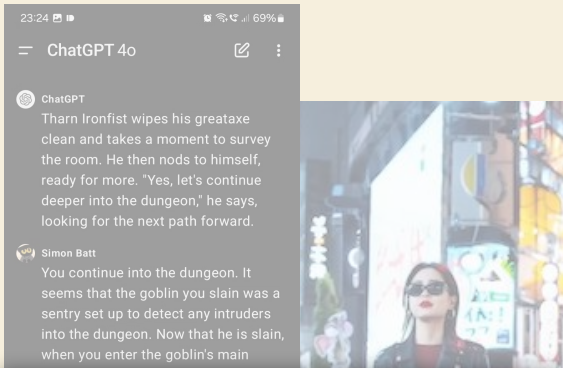
```

1 const fetchNASAPictureOfThe
2 return fetch("https://api
3 method: 'GET',
4 headers: {
5   'Content-Type': 'appl
6 }
7 }
8
9 then(response) => respo
10 then(() => {
11   return json()
12 })
13 }
14 }

```

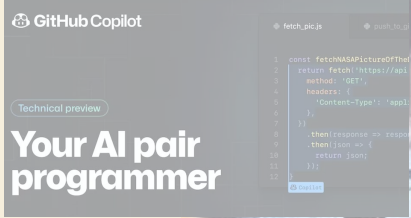
# Long Context

# Short Context

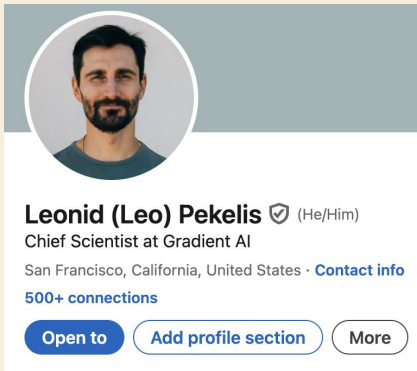
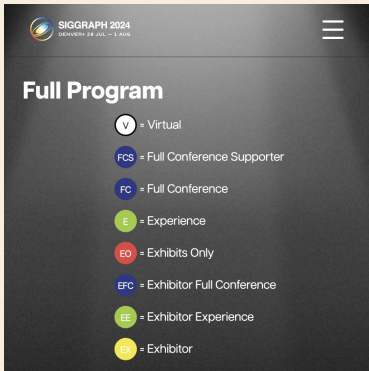


The message you submitted was too long,

throat, he's surrounded by five other goblins just as big as him. He declares that you state your business and not



# Long Context



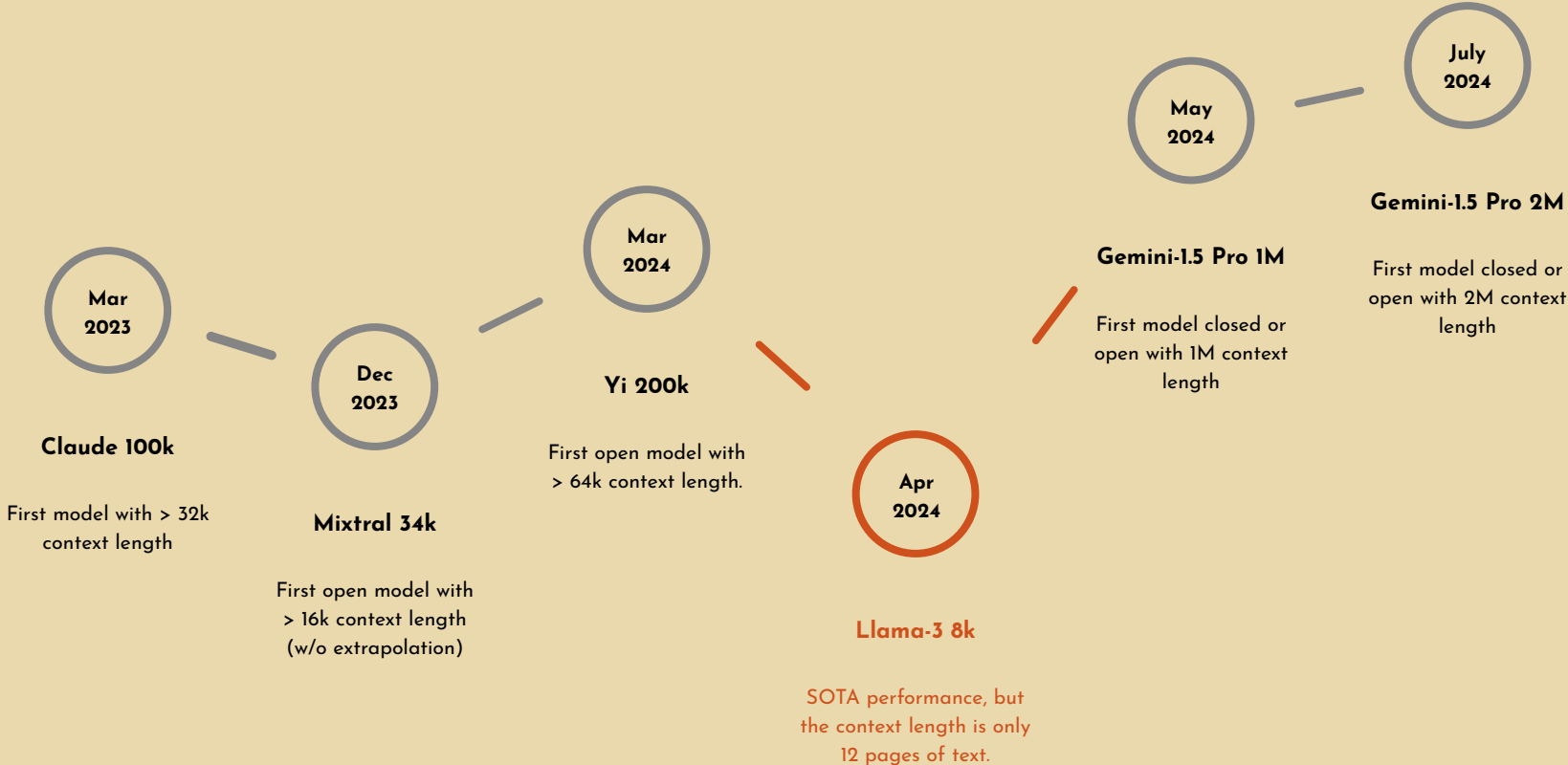
Here is a possible schedule for Leo Pekelis for the first day of SIGGRAPH 2024, taking into account his interests and professional responsibilities:

### Morning:

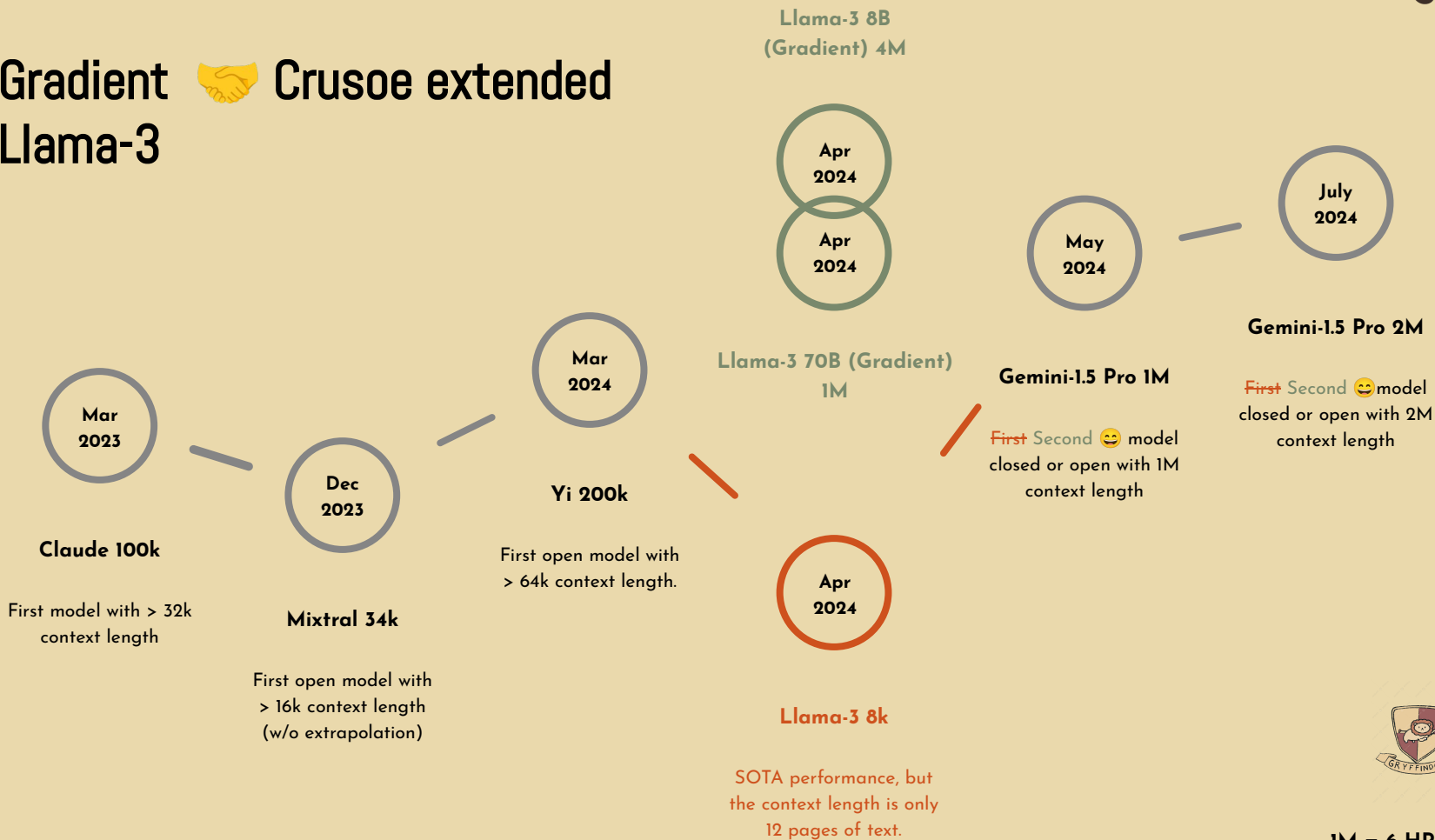
- **9:00 am - 10:30 am:** Attend the "Vector Graphics" Technical Paper session in Mile High 4. This session covers topics related to geometry and modeling, which are areas of interest for Leo.
- **10:45 am - 12:15 pm:** Attend the "VR, Eye Tracking, Perception" Technical Paper session in Mile High 3C. 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.

...

# Only 8k?!



# Gradient 🤝 Crusoe extended Llama-3

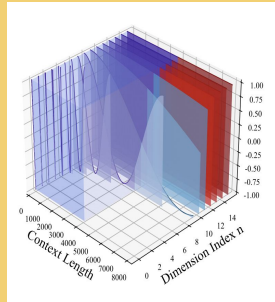
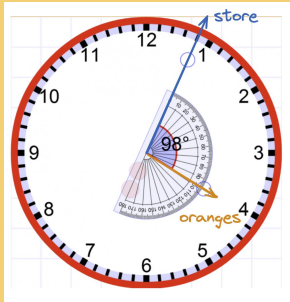


1M = 6 HP books

# with two technological advancements

1

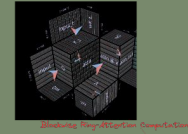
Long reading comprehension



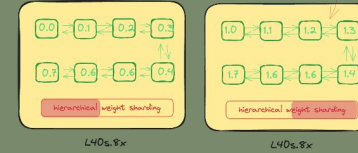
Liu, Xiaoran, et al. "Scaling laws of rope-based extrapolation." (2023).

2

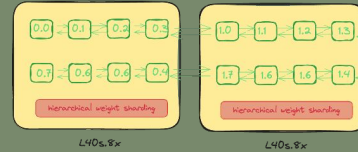
Network aware context parallelism



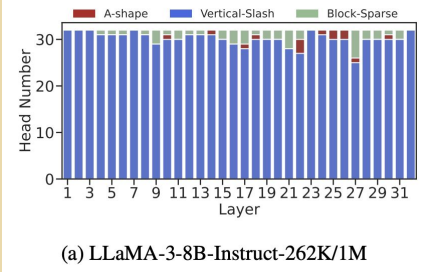
Configuration for Training  
eg LLaMa-3-8B-362k



Configuration for Training  
eg LLaMa-3-8B-1048k



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

**Haojun Zhao** · 2nd  
MLE intern @ Hugging Face | MS in Computer Science | X19  
1w · 🌐

I am excited to share my first project at [Hugging Face](#): Integrating Ring Attention into the Nanotron Library. 🤗🤖

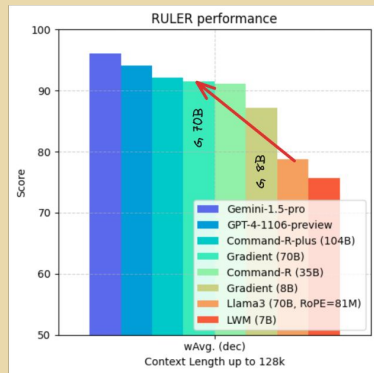
**Gradient** @Gradient\_AI · May 4

We're going back 2 back! 🤗 Introducing the first 1M context window @AlatMeta Llama-3 70B to pair with the our Llama-3 8B model that we launched last week on @huggingface. Our 1M context window 70B model landed a perfect score on NIAH and we're excited about the results that...

[Show more](#)

**gradient**

Llama-3 70B Instruct Gradient 1048K



Trending on 🤗 this week

Models

- meta-llama/Meta-Llama-3-8B  
Updated 11 days ago · ⬇️ 736k · ♡ 3.21k
- gradientai/Llama-3-8B-Instruct-Gra...**  
Updated about 11 hours ago · ⬇️ 11.2k · ♡ 429
- apple/OpenELM  
Updated 4 days ago · ♡ 1.12k
- meta-llama/Meta-Llama-3-8B-Instruct  
Updated 11 days ago · ⬇️ 1.15M · ♡ 1.8k
- NousResearch/Hezmes-2-Pro-Llama-3-...  
Updated about 16 hours ago · ⬇️ 3.39k · ♡ 212