



Overcoming The AI Memory Wall: Why AI Agents Need a New Foundation

[Redacted]								
Val Bercovici Chief AI Officer							[Redacted]	

2025

The Rise of Agent Swarms

2025

4M

developers

800M+

weekly ChatGPT users

6B

tokens per minute
on our API

Apps inside ChatGPT 

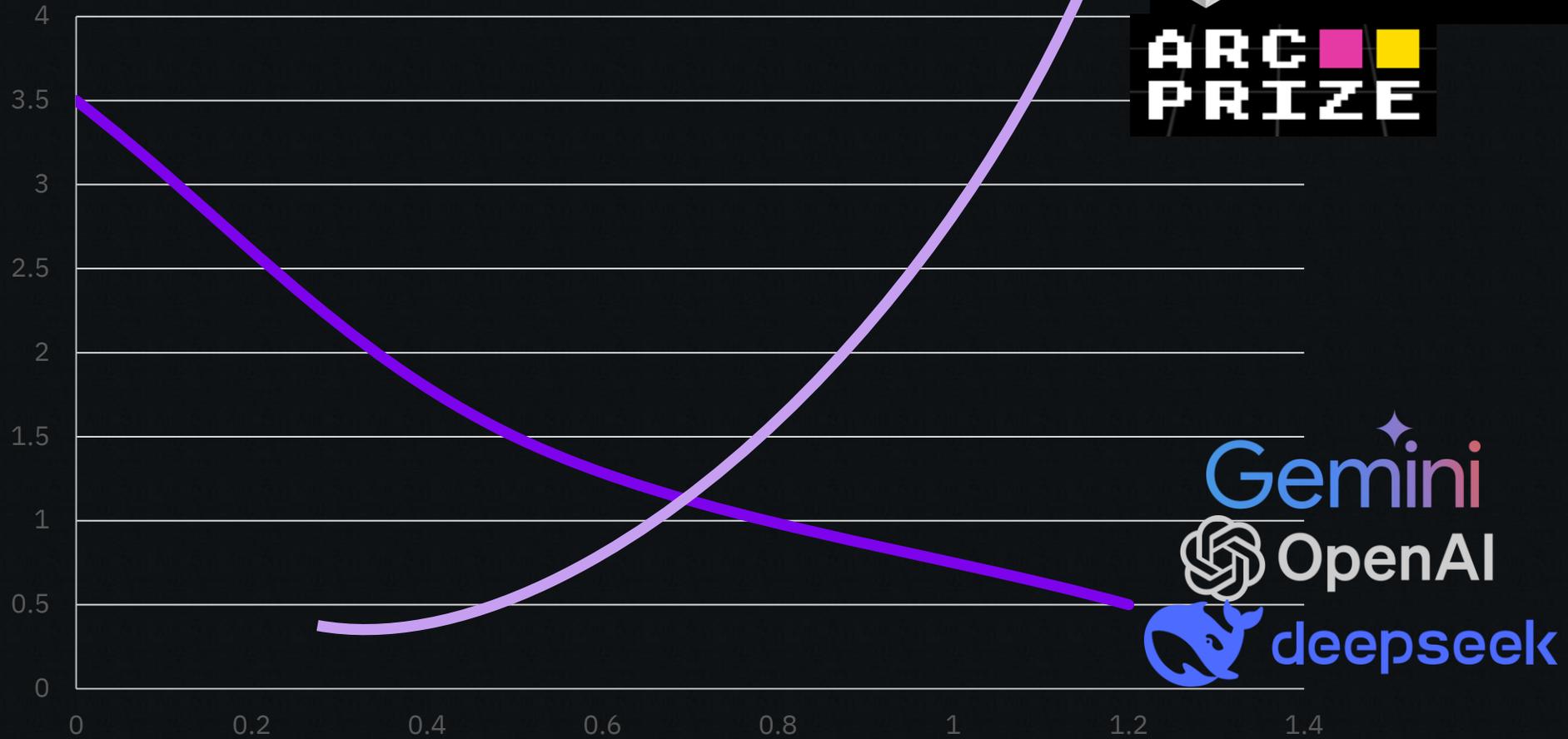
Building agents 

Writing [code]

An upside-down ice cream cone with melting vanilla ice cream on a dark blue background. The cone is inverted, with the tip pointing upwards. The ice cream is melting, creating a puddle at the base of the cone. The background is a solid dark blue color.

Paradox: Upside-Down Tokenomics

Token Cost vs Token Volume



CLAUDE
CODE

🔥 \$3B → \$4B (1-QRR)

 CURSOR

>\$550M ARR

ARC
PRIZE

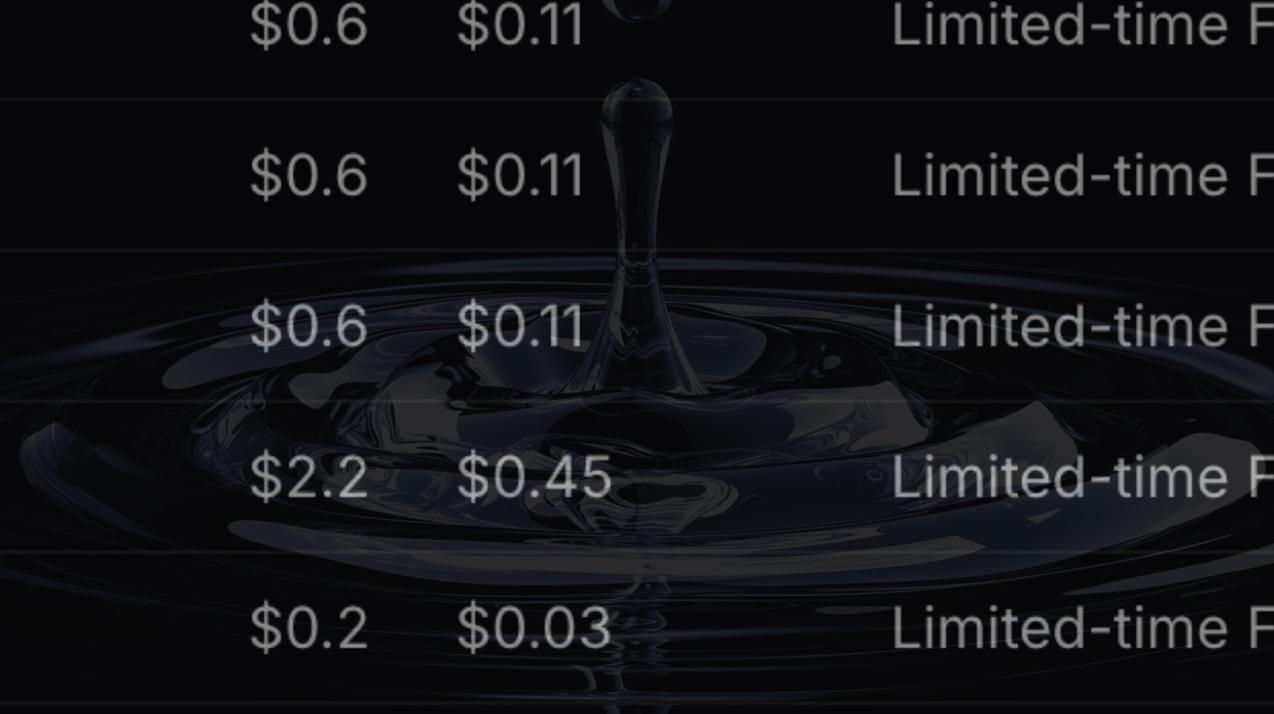
Gemini

 OpenAI

 deepseek



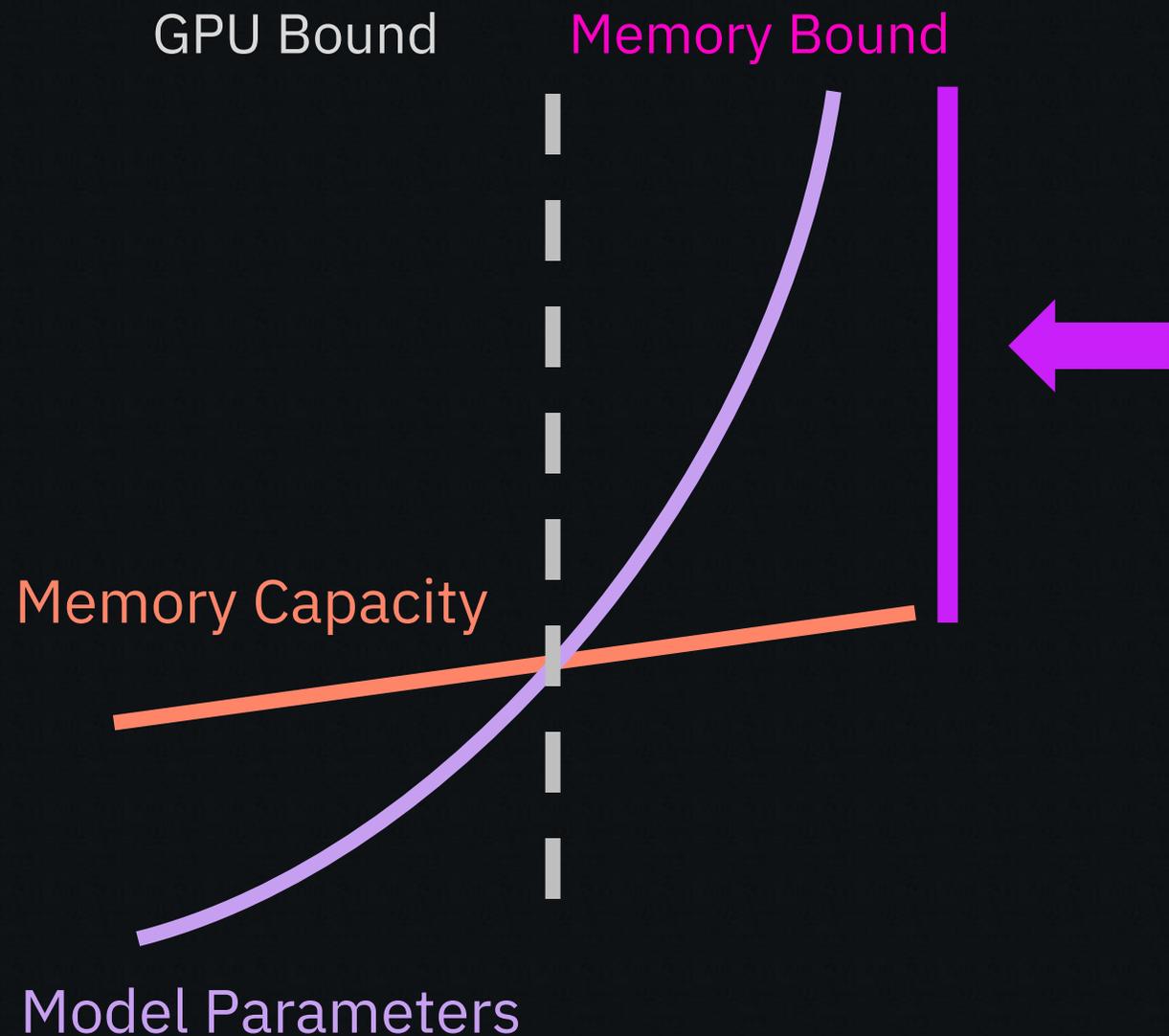
Pricing Ripple Effect

A decorative background featuring a water splash in the center, with ripples and droplets extending upwards and downwards. The splash is rendered in a dark, semi-transparent style against the dark background of the table.

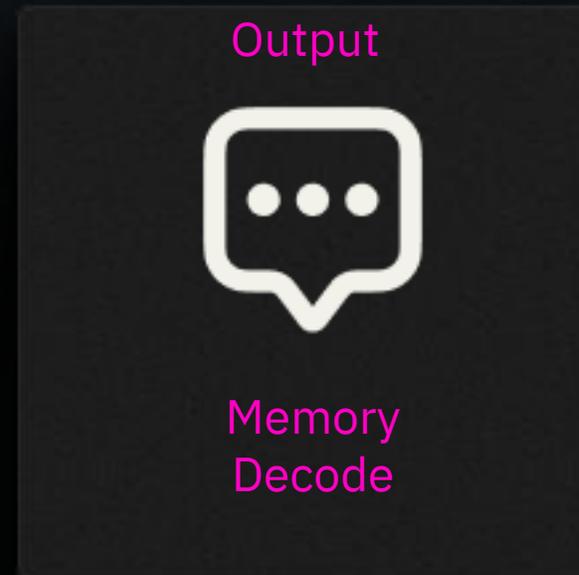
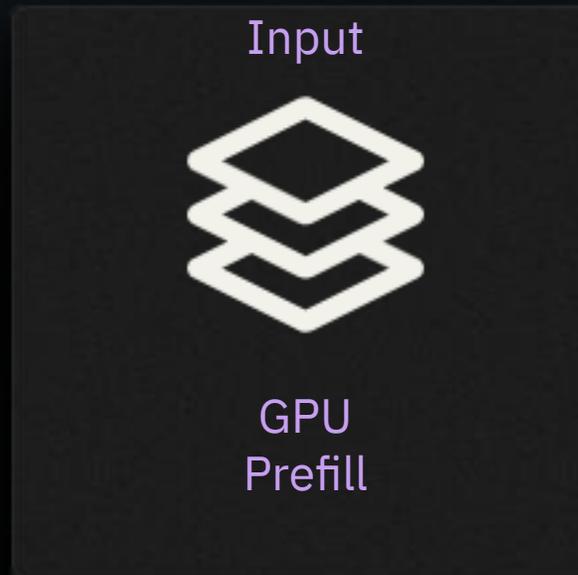
Model.	Input	Cached Input	Cached Input Storage	Output
GLM-4.6	\$0.6	\$0.11	Limited-time Free	\$2.2
GLM-4.5	\$0.6	\$0.11	Limited-time Free	\$2.2
GLM-4.5V	\$0.6	\$0.11	Limited-time Free	\$1.8
GLM-4.5-X	\$2.2	\$0.45	Limited-time Free	\$8.9
GLM-4.5-Air	\$0.2	\$0.03	Limited-time Free	\$1.1
GLM-4.5-AirX	\$1.1	\$0.22	Limited-time Free	\$4.5

Agentic AI is Hitting Hard Limits

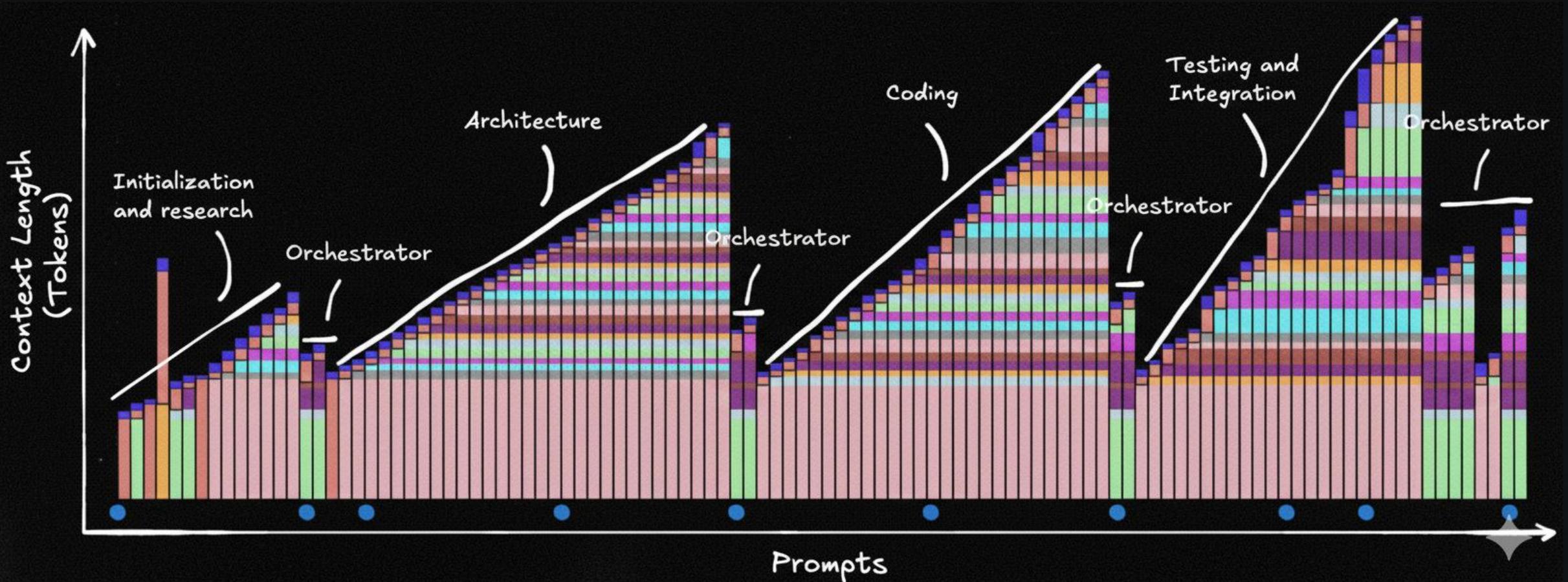
The AI Memory Wall



Prefill → Decode



Key Metric: KV Cache Hit Rate



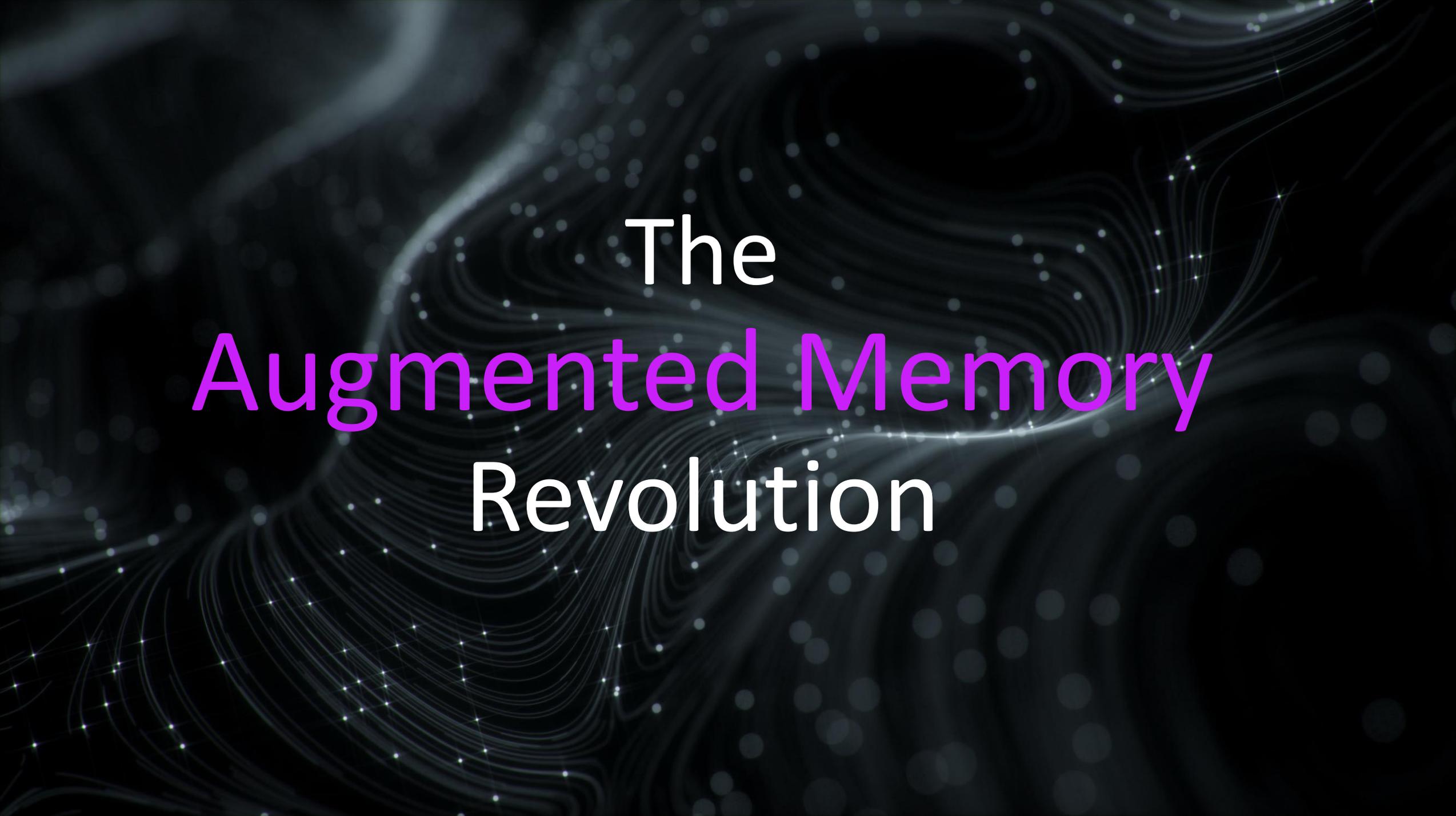
Scaling The Memory Wall





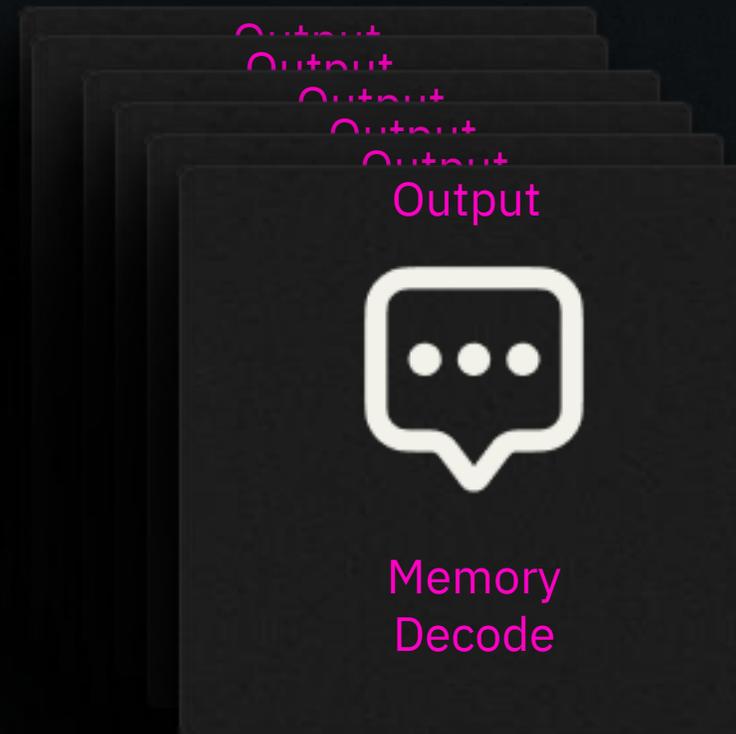
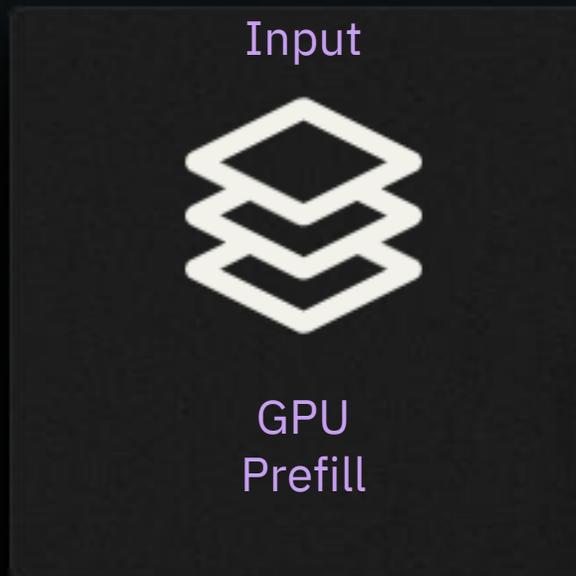
GPU
Prefill

Token Warehousing

The background features a dark, almost black, space filled with intricate, glowing patterns. These patterns consist of numerous thin, white and light blue lines that curve and swirl, creating a sense of motion and depth. Interspersed among these lines are small, bright white and blue particles, some of which appear to be twinkling or emitting light. The overall effect is reminiscent of a complex data network or a futuristic, digital landscape.

The Augmented Memory Revolution

1 Prefill ∞ Decode

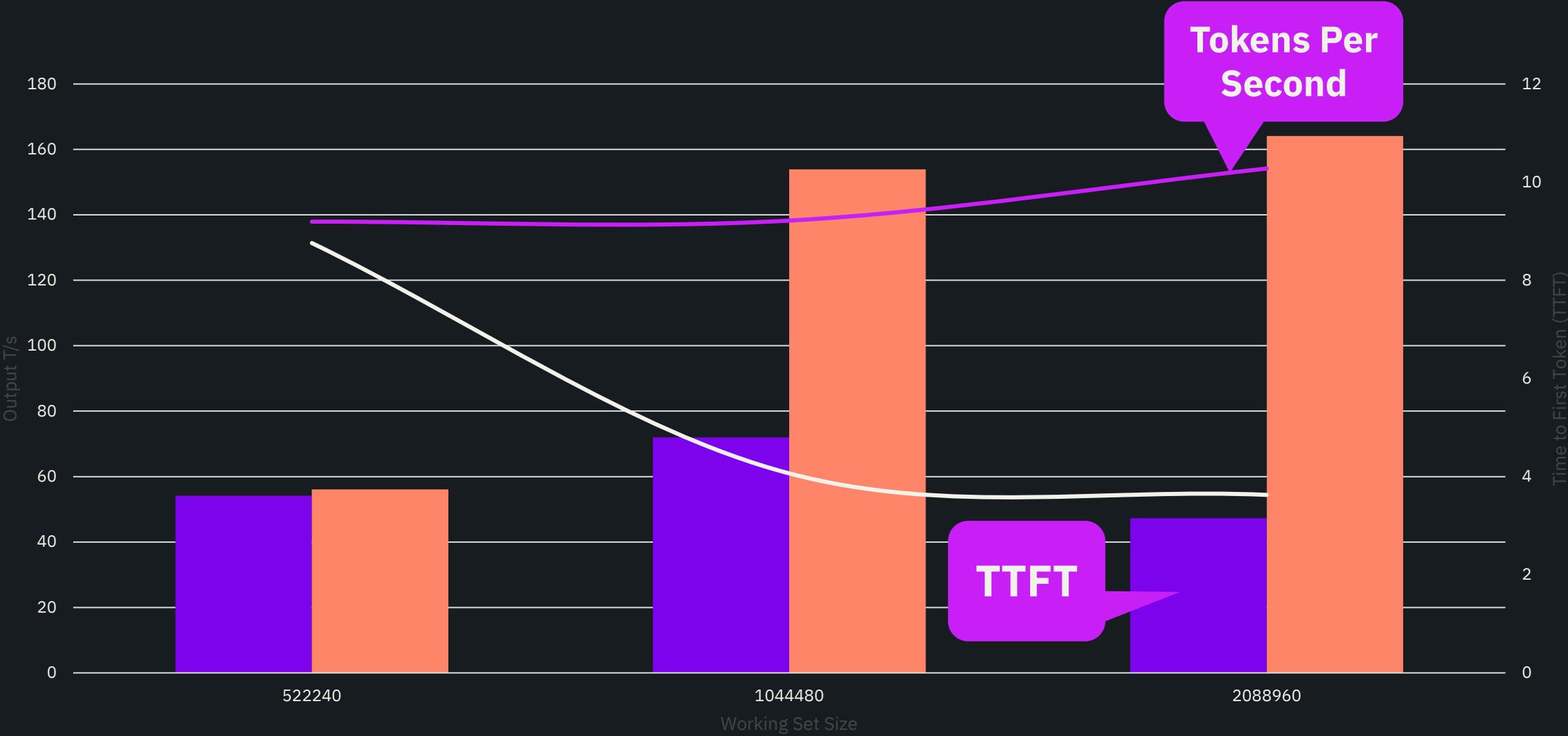


Insights from Our Labs

4X More Users
& Agent Sessions
Per GPU



Real-World Agent Inference Performance



WEKA TTFT DRAM TTFT WEKA Output T/s DRAM Output T/s





What it Takes to Win

- More Tokens
- Track KV Cache Hit Rate
- Prefill Once, Decode Forever
- Leverage GPU, Network, Memory
- Abundant Quality & Safety
Tokens (Every Agent Step)

Profitable AI Requires Overcoming The Memory Wall



THANKS FOR YOUR TIME

Learn How to
Maximize Your AI
Token Production



WEKA[®]