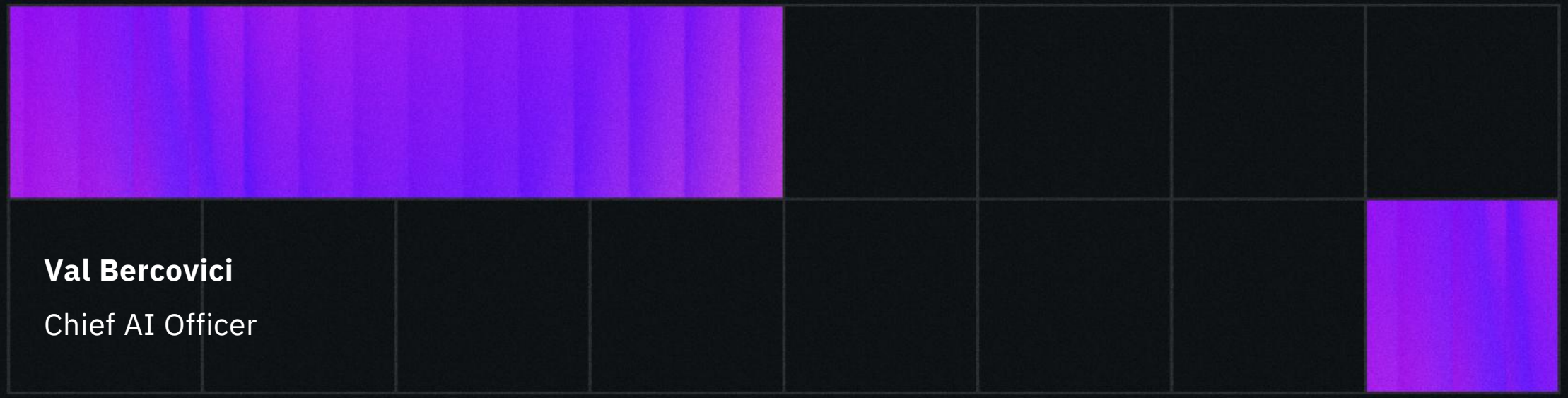
WEKA®

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



**Val Bercovici**

Chief AI Officer

# 2025

## The Rise of Agent Swarms



WEKA®


2025
4M
developers
800M+
weekly ChatGPT users
6B
tokens per minute
on our API

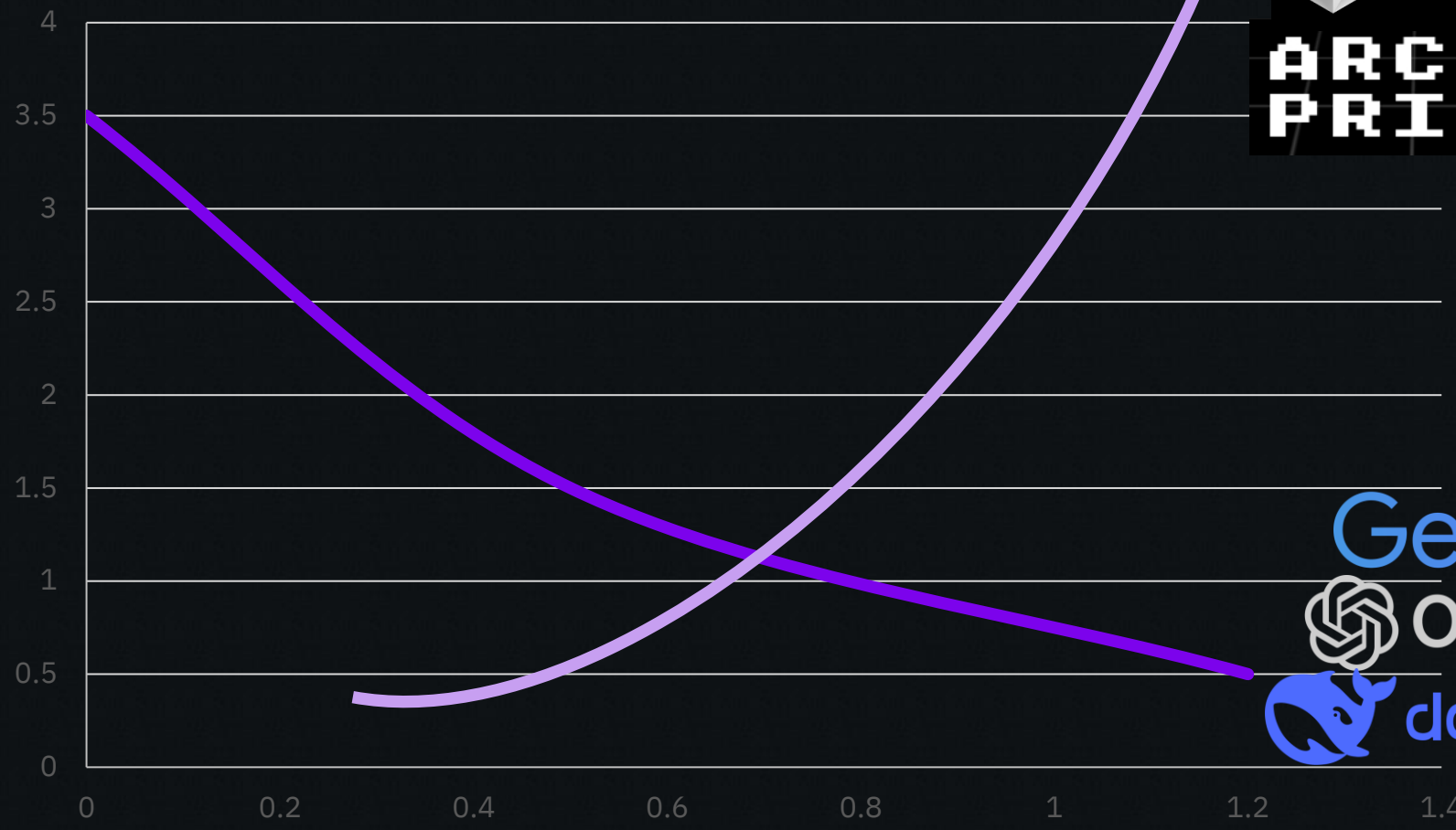

Apps inside ChatGPT
Building agents
Writing [code]

# Paradox:
# Upside-Down Tokenomics



WEKA®

# Token Cost vs Token Volume




CLAUDE CODE

🔥 $3B → $4B (1-QRR)

CURSOR

>$550M ARR

ARC PRIZE

Gemini

OpenAI

deepseek



WEKA®

# Pricing Ripple Effect



WEKA®

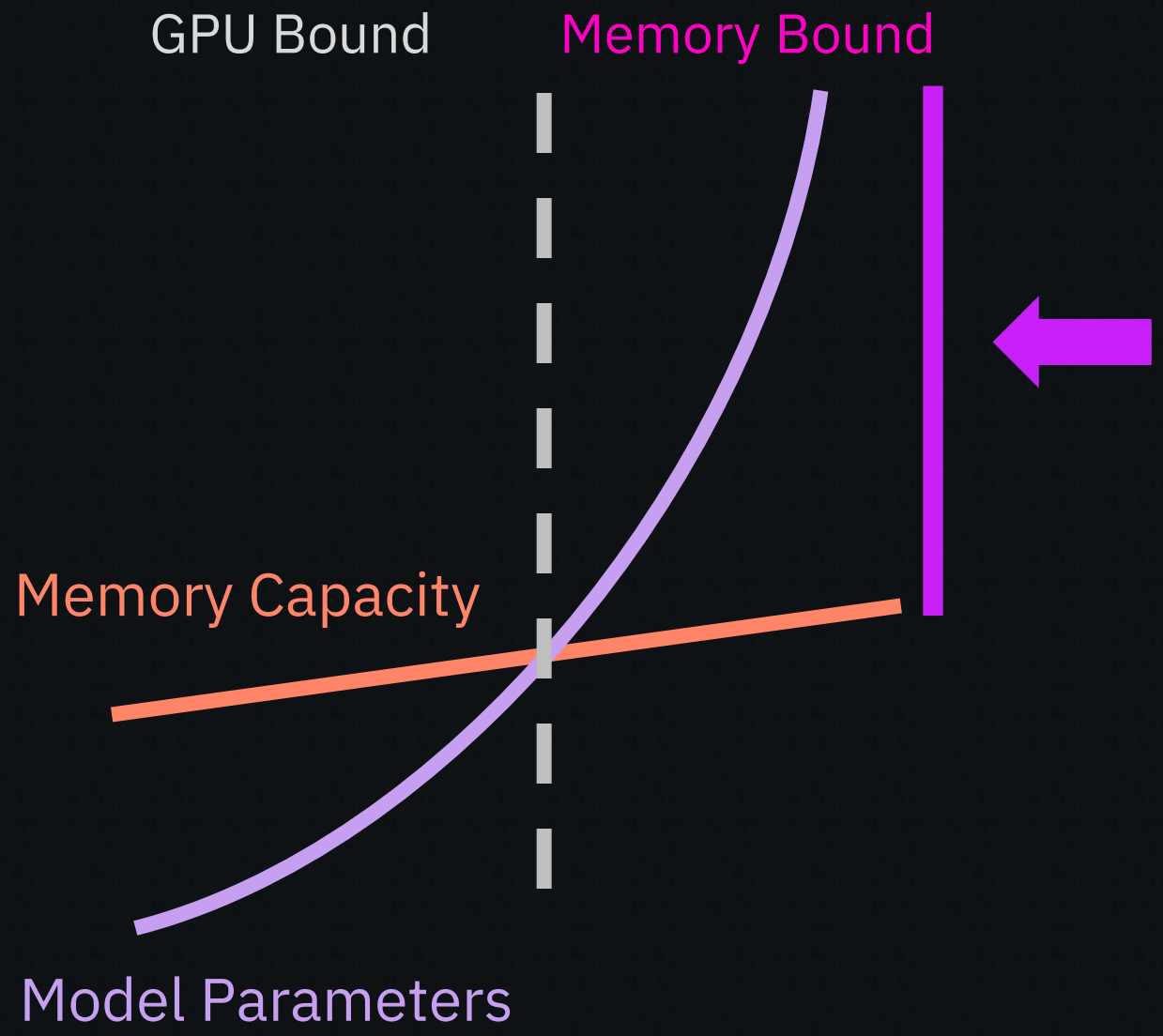
| Model. | Input | Cached Input | Cached Input Storage | Output |
|---|---|---|---|---|
| GLM-4.6 | $0.6 | $0.11 | Limited-time Free | $2.2 |
| GLM-4.5 | $0.6 | $0.11 | Limited-time Free | $2.2 |
| GLM-4.5V | $0.6 | $0.11 | Limited-time Free | $1.8 |
| GLM-4.5-X | $2.2 | $0.45 | Limited-time Free | $8.9 |
| 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



WEKA®

# The AI Memory Wall






WEKA®

# Prefill → Decode

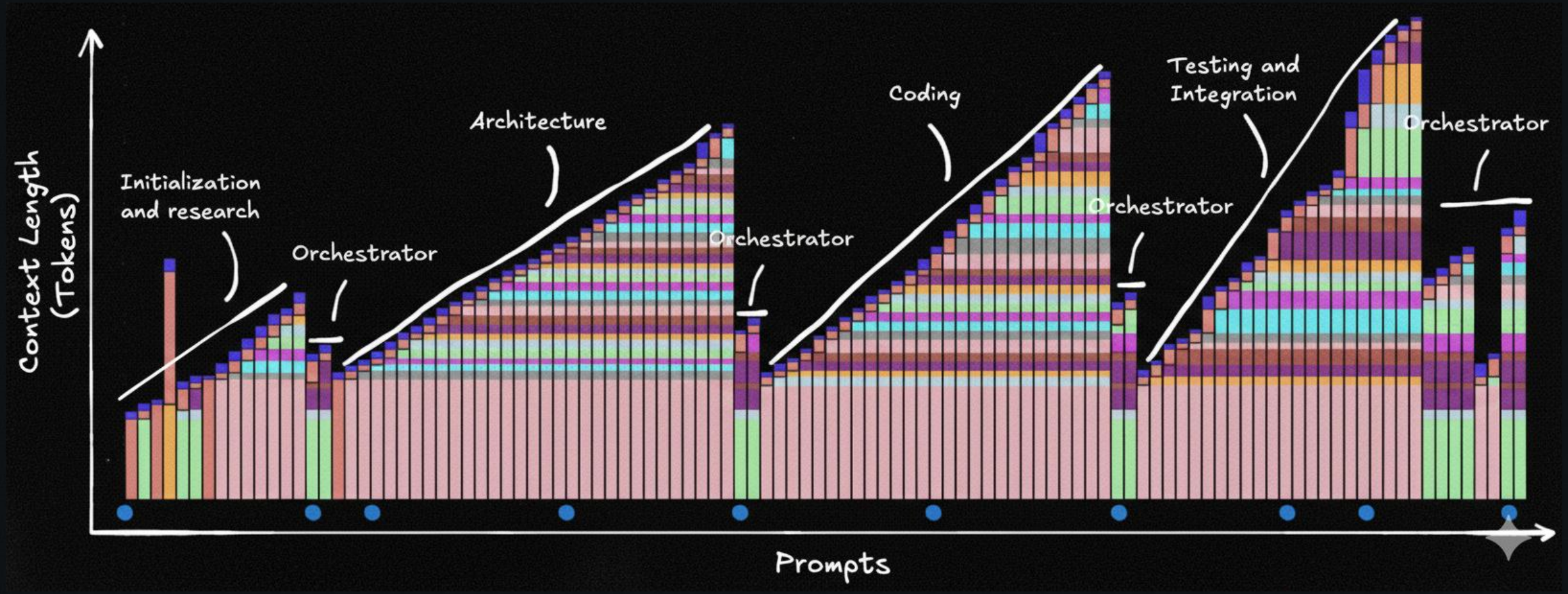
Input

GPU
Prefill

Output

Memory
Decode

# Key Metric:
# KV Cache Hit Rate






WEKA

# Scaling The Memory Wall



WEKA®

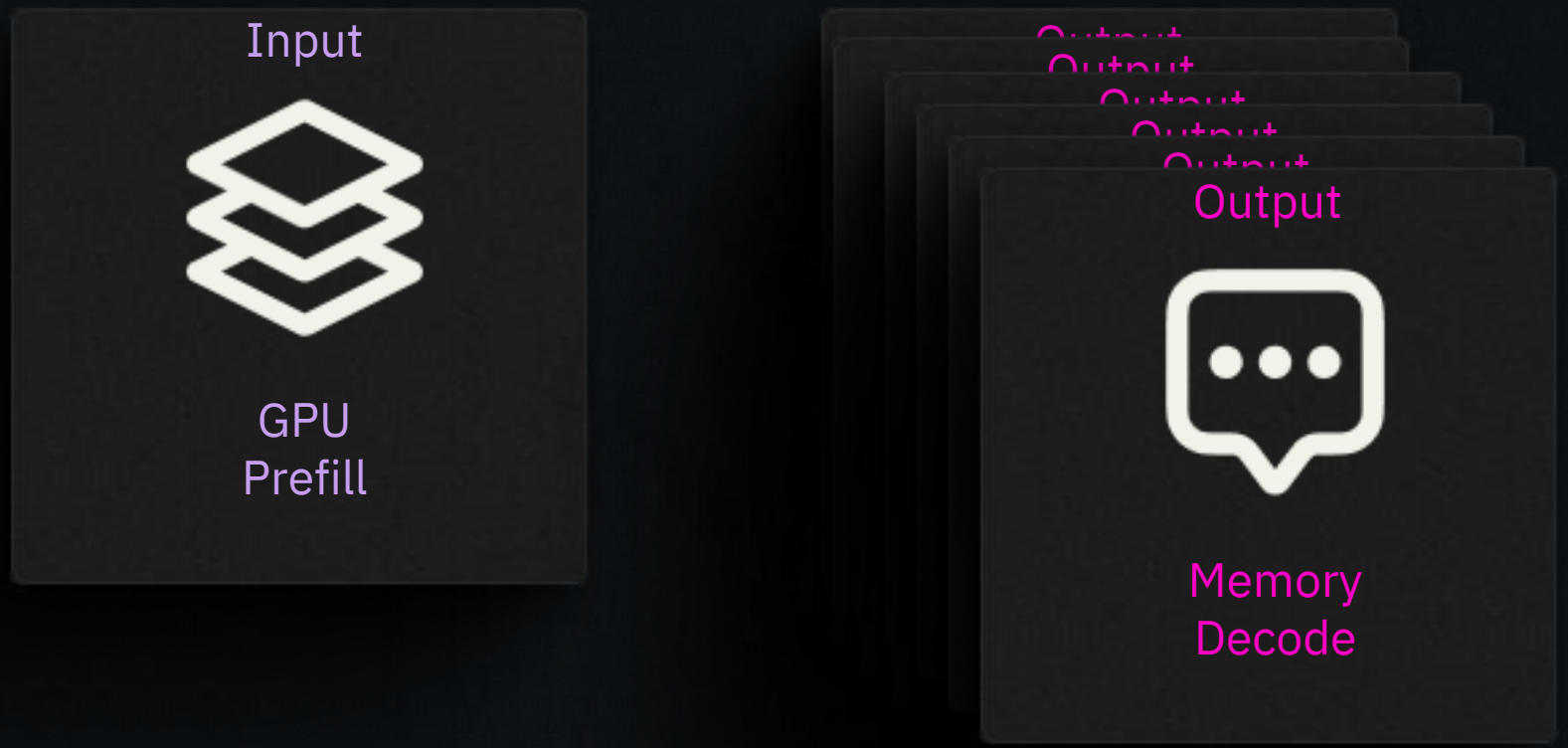

# Token Warehousing



WEKA®

# The Augmented Memory Revolution

# 1 Prefill ∞ Decode



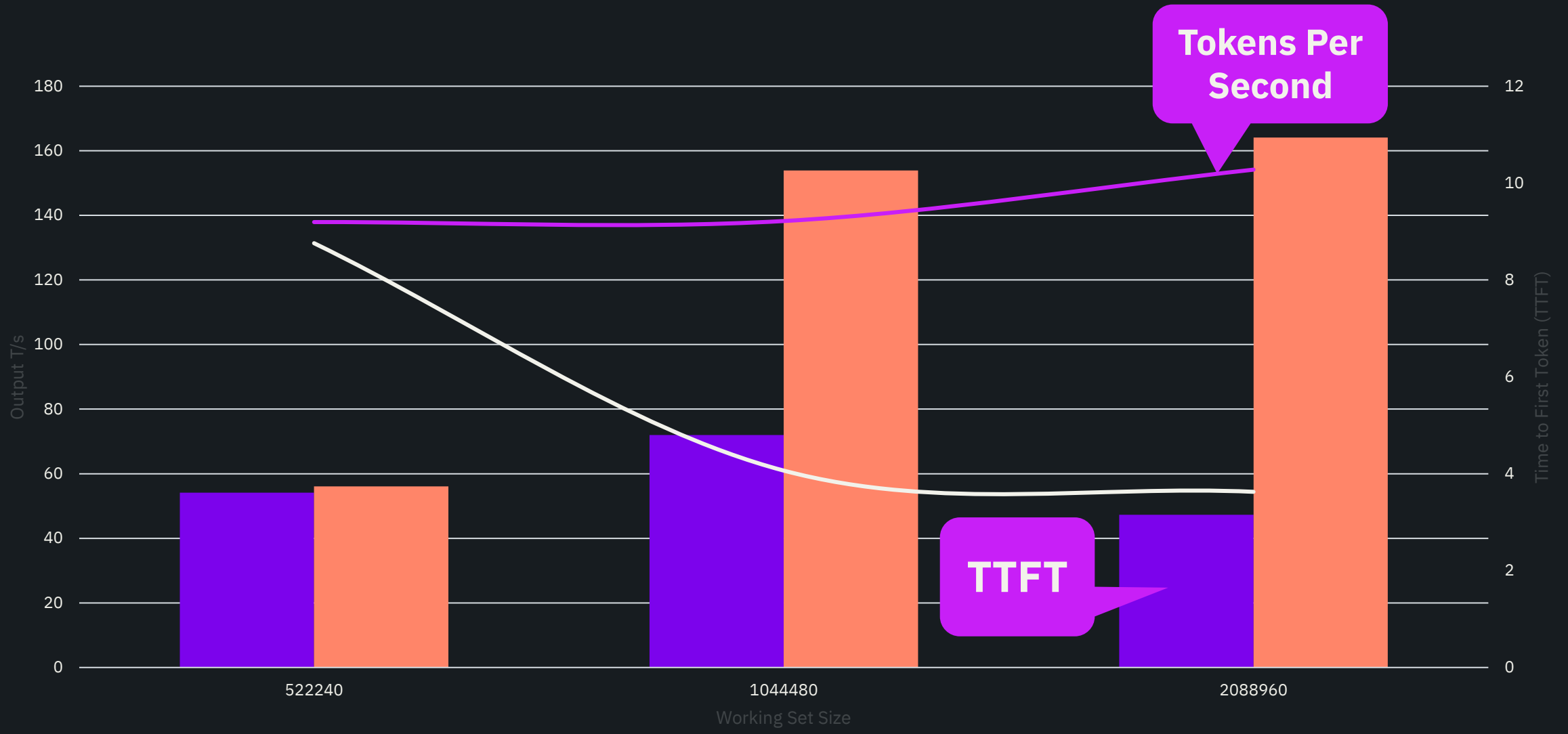



WEKA

# Insights from Our Labs

**4X** More Users
& Agent Sessions
Per GPU

# Real-World Agent Inference Performance




WEKA®

# 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



WEKA

THANKS FOR YOUR TIME

# Learn How to Maximize Your AI Token Production



WEKA®


WEKA®