WEKA®

Token Power

# The New Economics of AI

**Lauren Vaccarello**

CMO

# 2025

## The Age of Reasoning

# Smarter AI Comes at a Premium

WEKA® © 2025



WEKA®

# Smarter AI Comes at a Premium

WEKA® © 2025



WEKA®

The Rise of

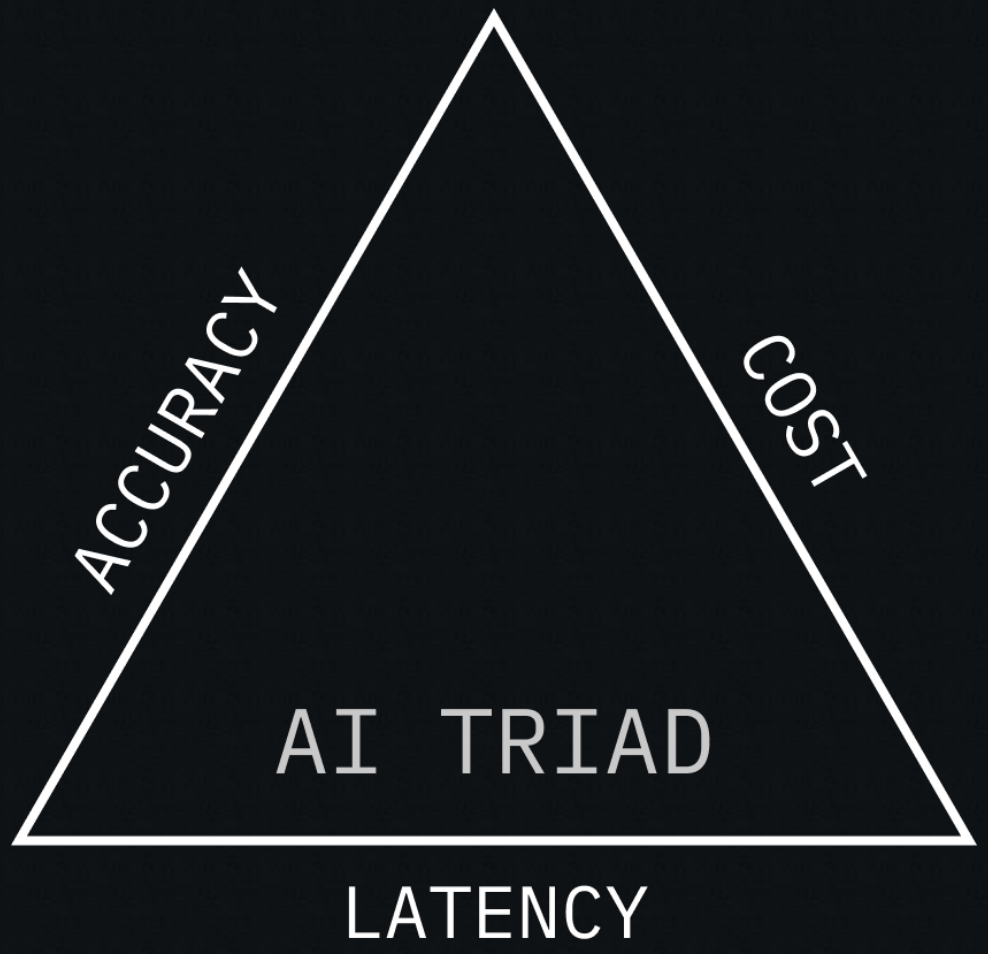
# Reasoning Models

ARC PRIZE

ARC PRIZE

# Infrastructure Cannot be an Afterthought

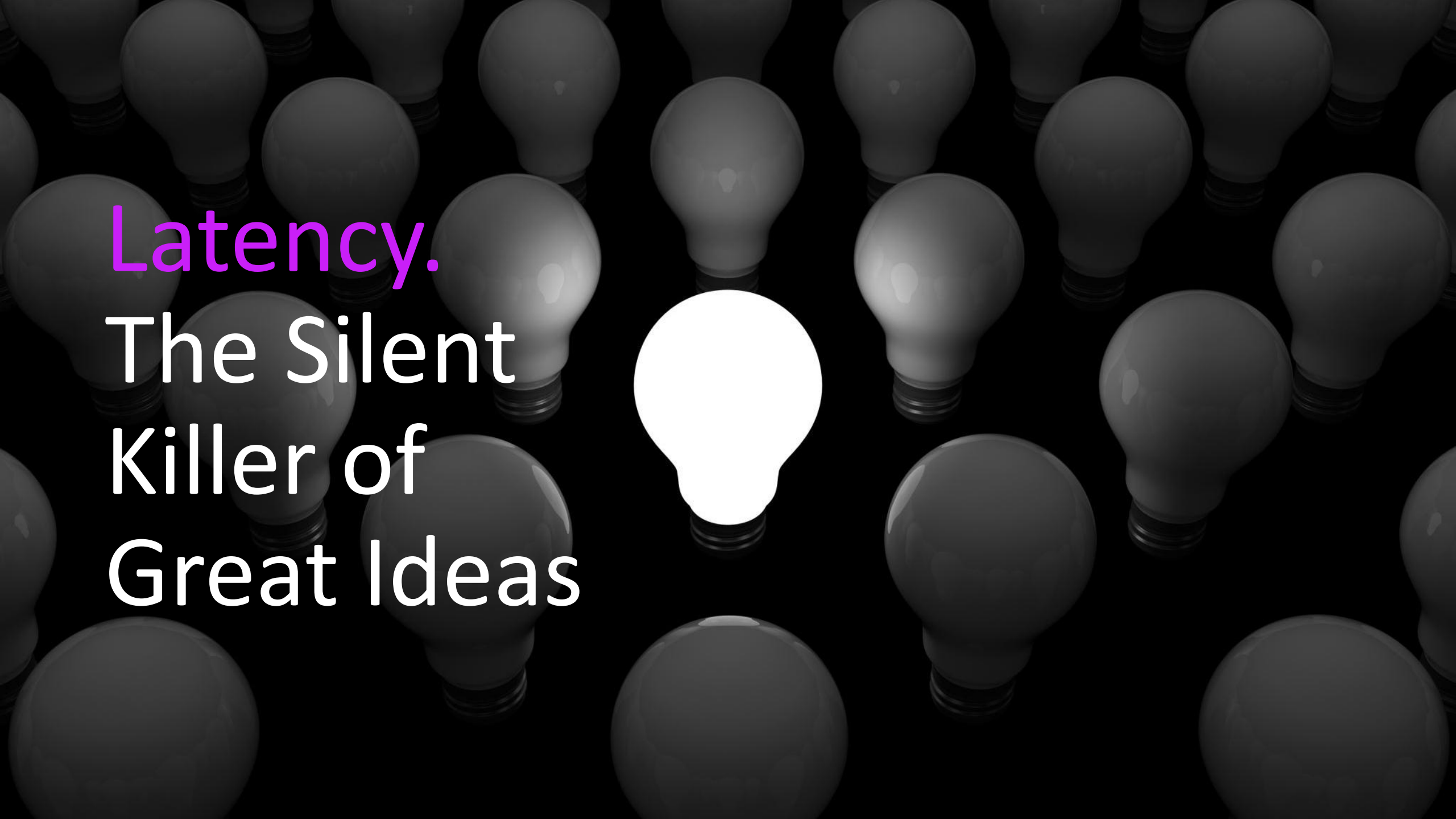
# The AI Economics Triad



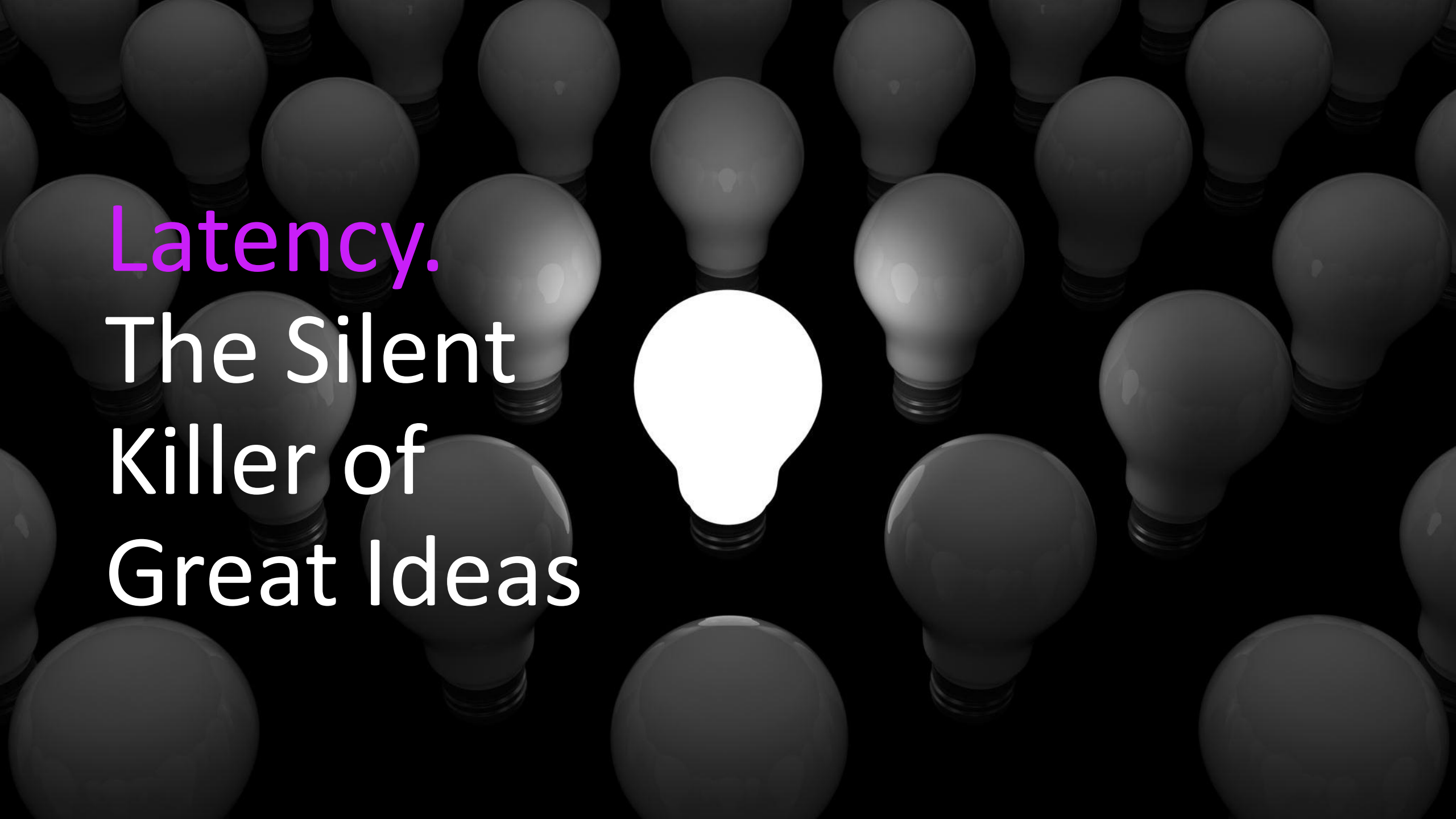
# Accuracy.

# Non-negotiable

# Latency.
# The Silent Killer of Great Ideas



# Latency.
# The Silent Killer of Great Ideas



# Cost is Just
# The Symptom

WEKA® © 2025

WEKA®

# Agentic AI

Hi, I'm DeepSeek.
How can I help you today?

# Design Problem
# NOT a Funding Problem

# QPS is Out
# TTFT is In

≠

WEKA® © 2025



WEKA®

# AI Unit Economics Per Token

WEKA® © 2025

WEKA®

# AI Unit Economics Per Token

WEKA® © 2025

WEKA®

# Token Warehousing

# Prefill vs. Decode

Prefill

Decode

# Prefill Caching.
# Instant Recall
# Reduced Spend

# Memory = Margin
# Reuse = Speed

# Memory = Margin
# Reuse = Speed

# The GPU Bottleneck

# The GPU Bottleneck

# Full Price Part-Time Work

WEKA

# Architecting for Velocity

# Architecting for Velocity

The Power of

# Persistent Memory

WEKA

The Power of

# Persistent Memory

WEKA

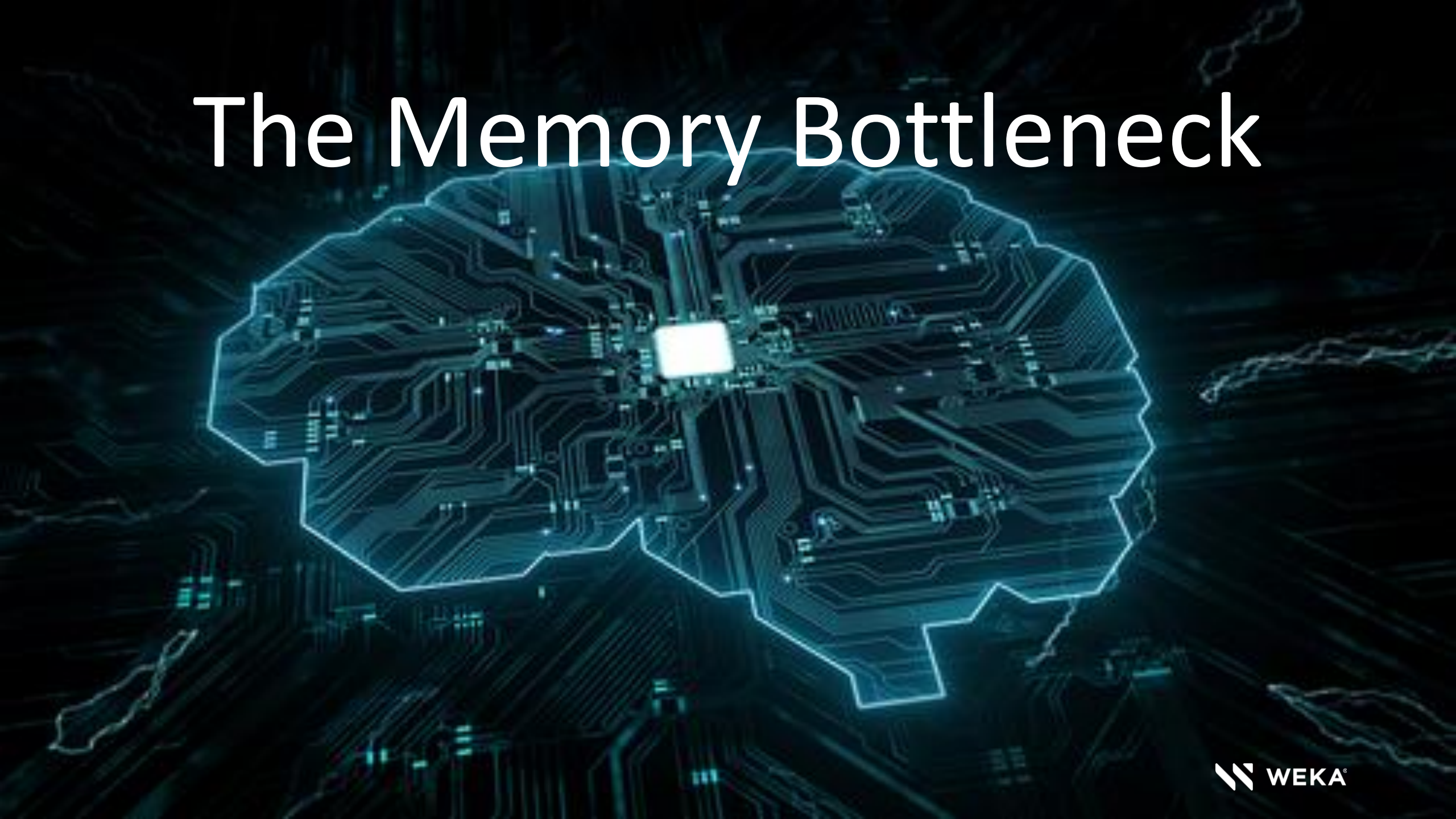
The Holy Grail of

# Inferencing Economics

The Holy Grail of

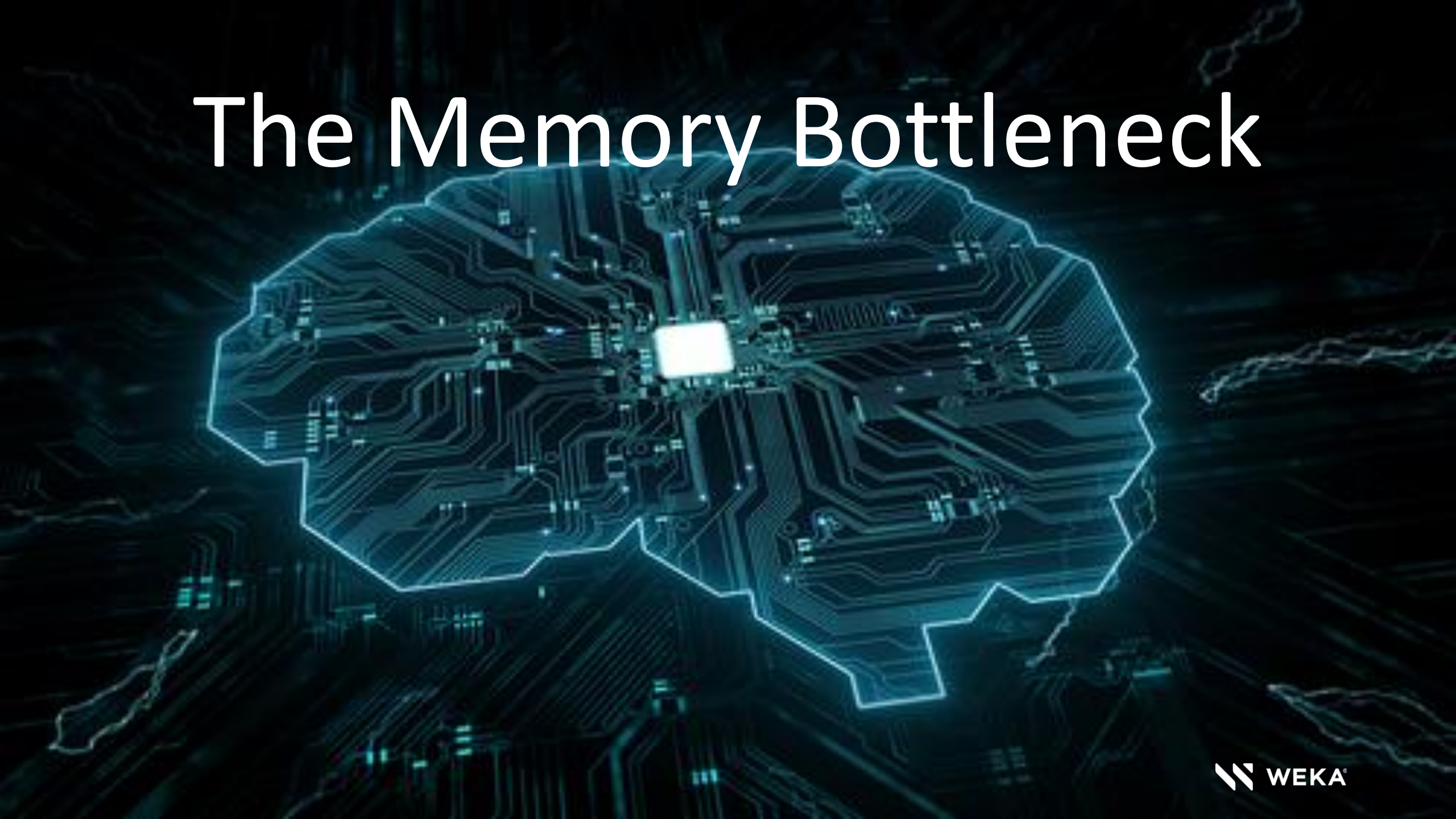
# Inferencing Economics

# The Memory Bottleneck



WEKA

# The Memory Bottleneck



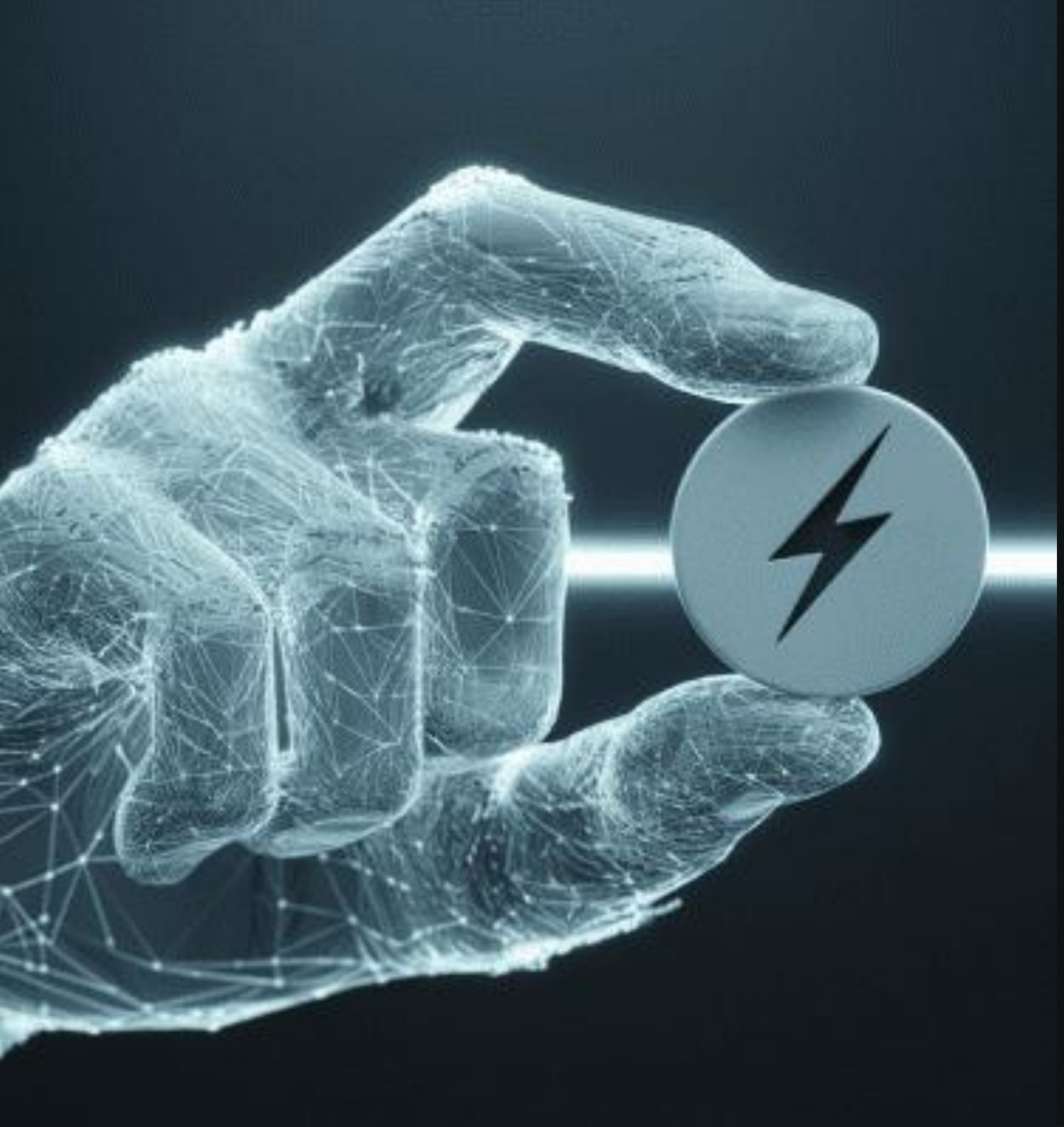
WEKA

# Memory is Currency

# Memory is Currency

# What it Takes to Win

- Measure TTFT
- Prefill Once, Decode Often
- Think Tokens, not Queries
- Maximize Memory to Drive Profits



# The New Economics of AI

WEKA®

# Sustainable AI Starts With Tokenomics

WEKA® © 2025

WEKA®

THANKS FOR YOUR TIME

# Learn How to Maximize Your AI Token Production

WEKA®