Subconscious

Inference Designed for Long-Horizon Agents

How context compression allows developers to run more agents, on more complex tasks, for longer periods of time, all at a lower cost.

Subconscious

Connect with me!



I’m Jack O’Brien

- Co-Founder & CEO of **Subconscious**
- HQ in Boston, MA. Based on MIT research.
- Prev. 2x exited founder, SWE at Google in NYC.

Subconscious

Companies need to cut back on AI spend, but engineers want *more*.

Context compression enables both.

Agents can run for hundreds of steps and millions of tokens across long periods of time. This is a unique challenge because:

Subconscious

Agents can run for hundreds of steps and millions of tokens across long periods of time. This is a unique challenge because:

As the context window fills up

Agents can run for hundreds of steps and millions of tokens across long periods of time. This is a unique challenge because:

As the context window fills up → They get dumber

Agents can run for hundreds of steps and millions of tokens across long periods of time. This is a unique challenge because:

As the context window fills up → They get dumber

As the context window fills up

Subconscious

Agents can run for hundreds of steps and millions of tokens across long periods of time. This is a unique challenge because:

As the context window fills up → They get dumber

As the context window fills up → They get slower

Agents can run for hundreds of steps and millions of tokens across long periods of time. This is a unique challenge because:

As the context window fills up → They get dumber

As the context window fills up → They get slower

Task length is limited by

Agents can run for hundreds of steps and millions of tokens across long periods of time. This is a unique challenge because:

As the context window fills up	→	They get dumber
As the context window fills up	→	They get slower
Task length is limited by	→	The context window size

Agents can run for hundreds of steps and millions of tokens across long periods of time. This is a unique challenge because:

As the context window fills up	→	They get dumber
As the context window fills up	→	They get slower
Task length is limited by	→	The context window size
They are extremely expensive		

Subconscious

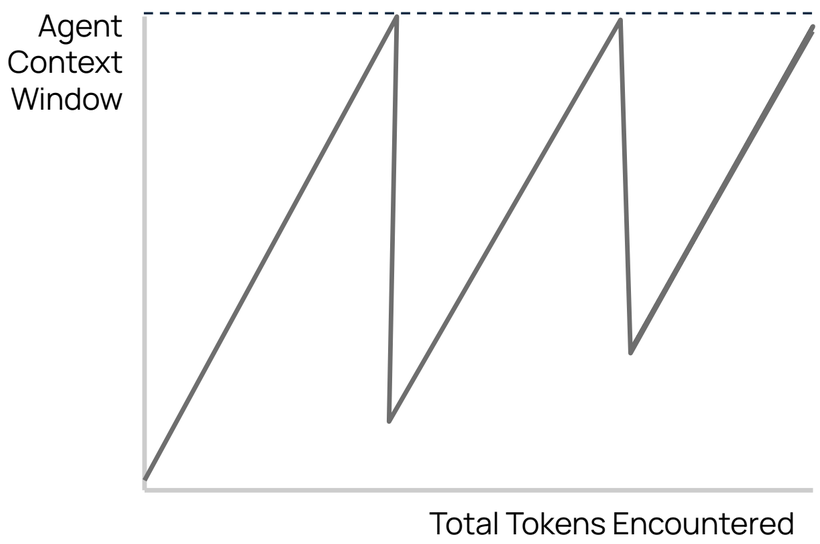
Agents can run for hundreds of steps and millions of tokens across long periods of time. This is a unique challenge because:

As the context window fills up	→	They get dumber
As the context window fills up	→	They get slower
Task length is limited by	→	The context window size
They are extremely expensive	→	*When the context window fills up*

Subconscious

Let's look at a standard coding agent reasoning trace like Claude Code.

Over the course of the run the agent accumulates context.
A large amount of that information becomes ***garbage***.



Subconscious

Let's look at a
reasoning trac

Over the course of the
A large amount of tha

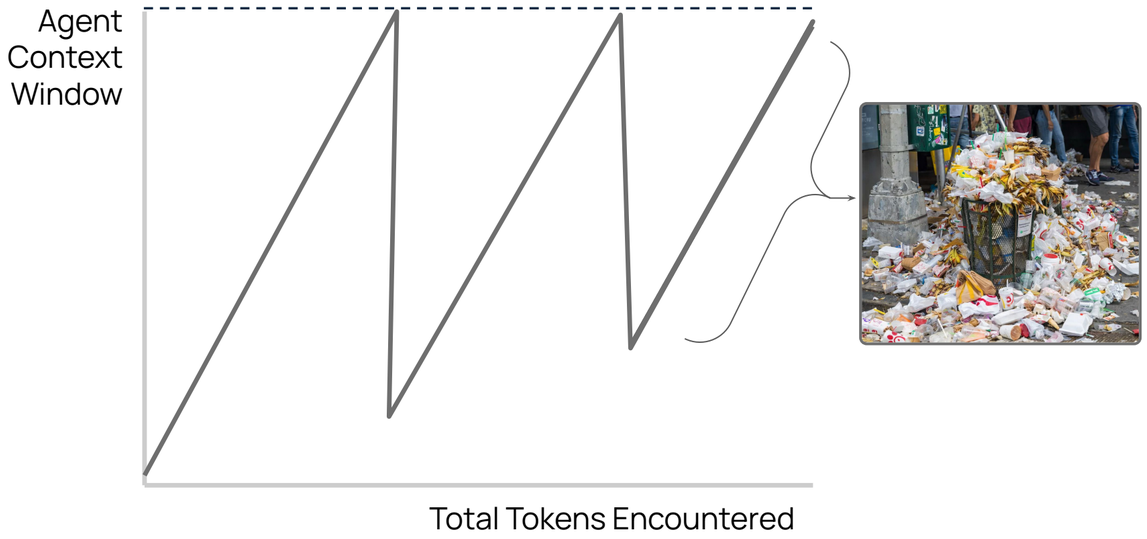
Agent
Context
Window



Subconscious

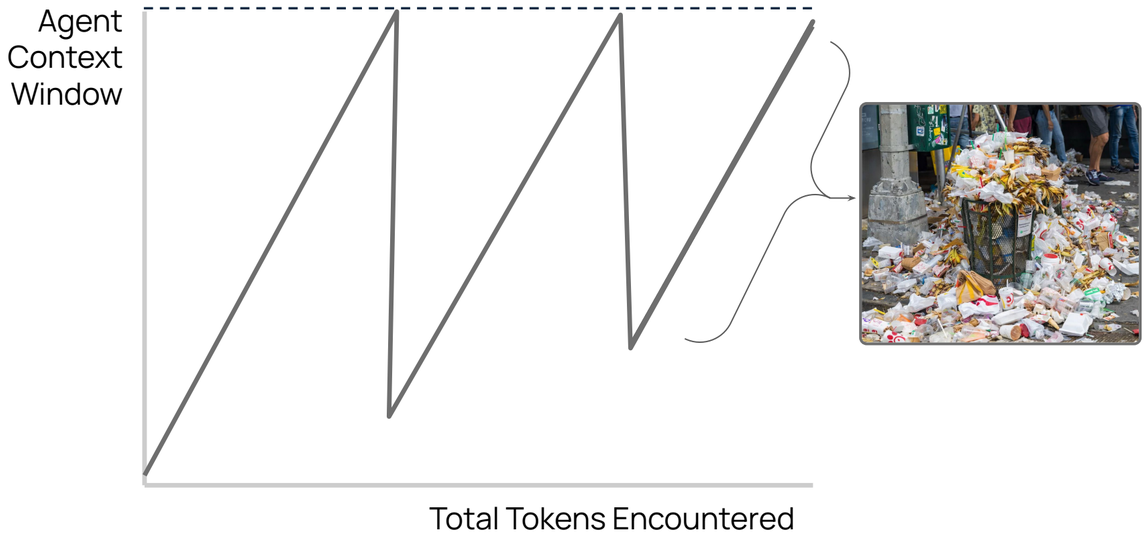
Let's look at a standard coding agent reasoning trace like Claude Code.

Over the course of the run the agent accumulates context.
A large amount of that information becomes ***garbage***.



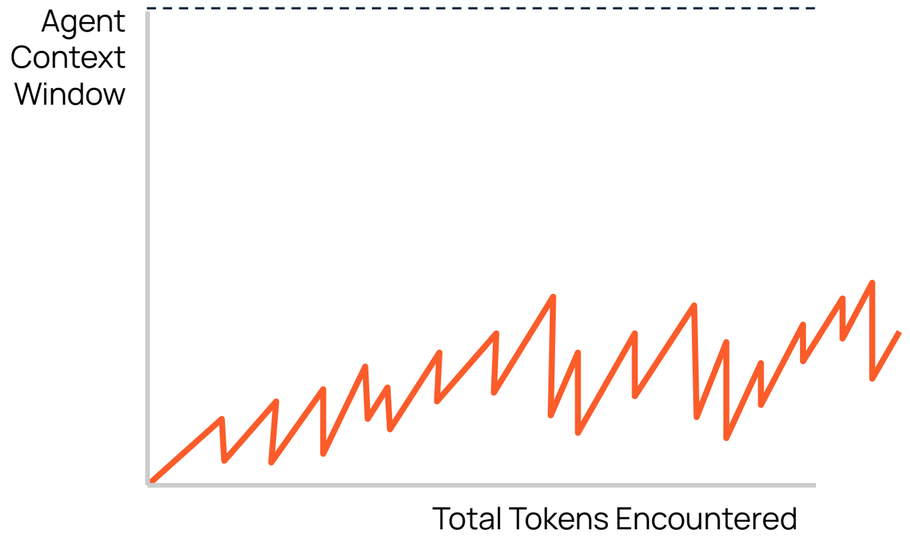
Let's look at a standard coding agent reasoning trace like Claude Code.

Over the course of the run the agent accumulates context.
A large amount of that information becomes ***garbage***.



An inference runtime designed for long-horizon agent workloads.

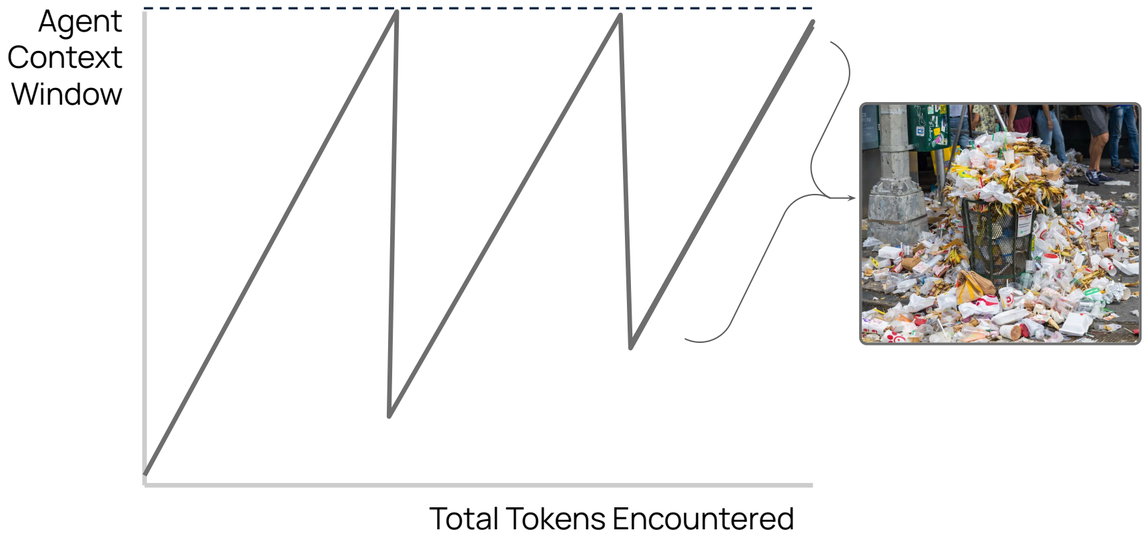
Using **context compression** and **highly efficient caching**, massively increase efficiency and accuracy of running agents.



Subconscious

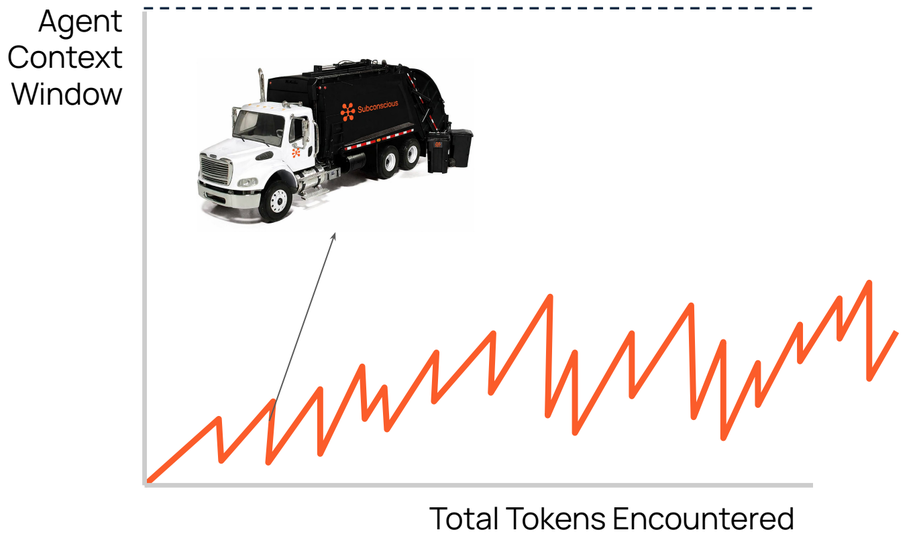
Let's look at a standard coding agent reasoning trace like Claude Code.

Over the course of the run the agent accumulates context.
A large amount of that information becomes ***garbage***.



An inference runtime designed for long-horizon agent workloads.

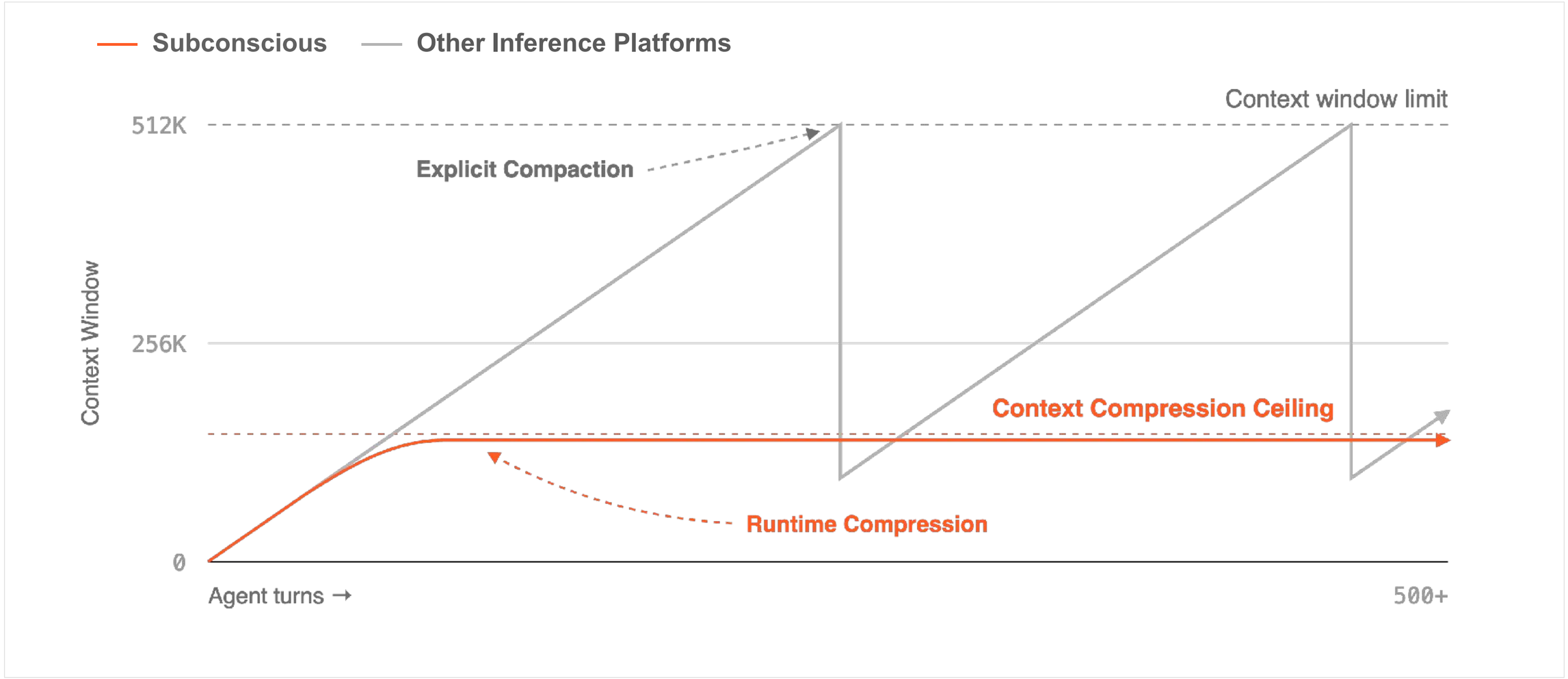
Using **context compression** and **highly efficient caching**, massively increase efficiency and accuracy of running agents.



Subconscious

An inference system designed for agents.

A system optimized for agent workloads with context compression and highly efficient caching.



Subconscious

Runtime compression and caching

Kicks in at 50+ messages (user input, reasoning, tool calls, tool responses, output)

01 Identify

Score and mark messages with low relevance to agent performance.



02 Compress

Marked spans are compressed in place. The thread keeps its order.



03 Cache both sides

Cache tokens before *and* after each compressed span, prefix and suffix.



Subconscious

Context compression unlocks agent performance

5m+

Extended context window

3.5×

Sustained faster token throughput

40%

Fewer tokens required

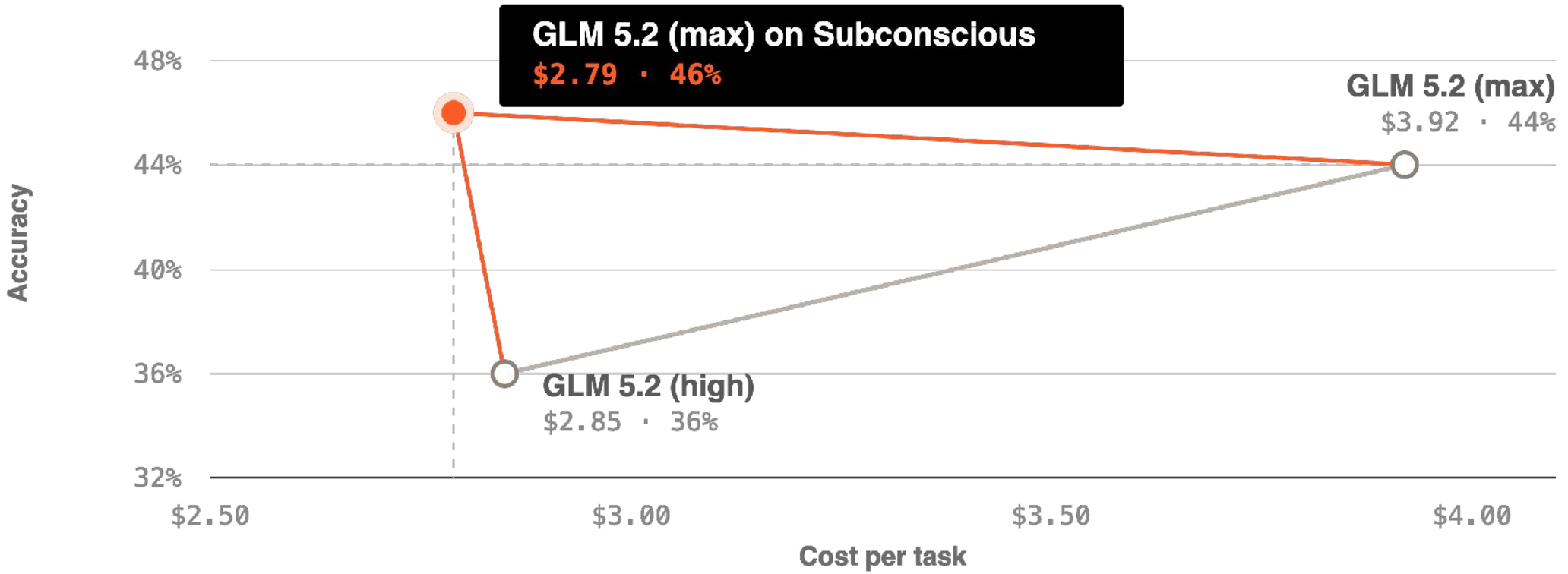
10%

performance *increase* on key benchmarks like DeepSWE

Run faster agents, on harder tasks, for longer periods of time, for a lot less.

Subconscious

Compression *improves* accuracy on DeepSWE agentic coding



Runtime improvements are additive to any AI stack

Harness

1000's of AI products and apps

Models


Z.ai



Nemotron


Qwen


Gemma


Kimi

Inference Runtime

Subconscious

Chips



Samsung



Nvidia



AMD



Etched

Additive to

The rest of the AI stack: chips, models, and app-layer frameworks all keep working, untouched.

We replace

We are a drop-in replacement for vLLM, SGLang, llama.cpp, and other inference systems.

Now serving

GLM 5.2 Marathon

Available on

Subconscious

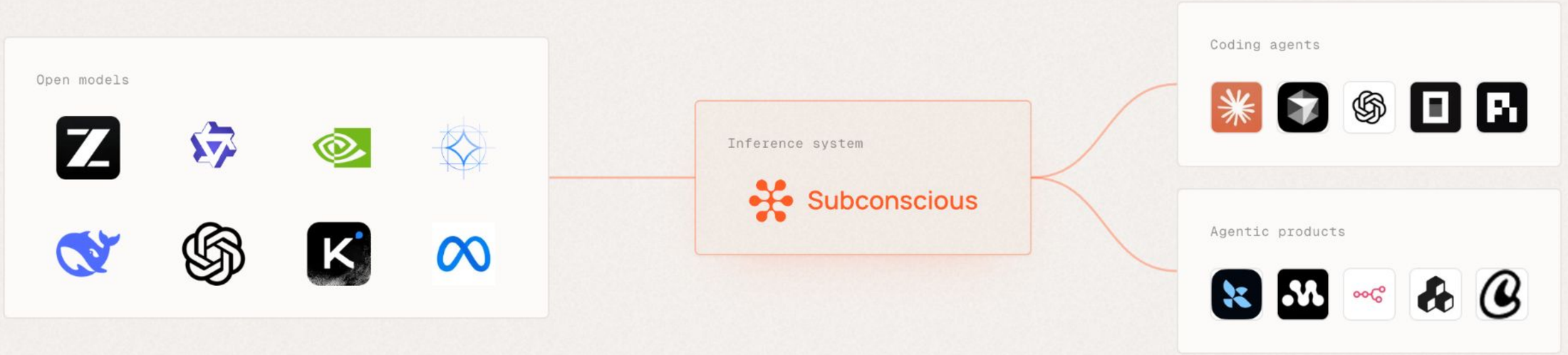
CLOUDFLARE

baseten

MODELS.DEV

Open models in, powerful agents out.

The best open models are compatible with our our inference system. Use our API or host on your own infrastructure, then plug straight into your coding agents and agentic products.

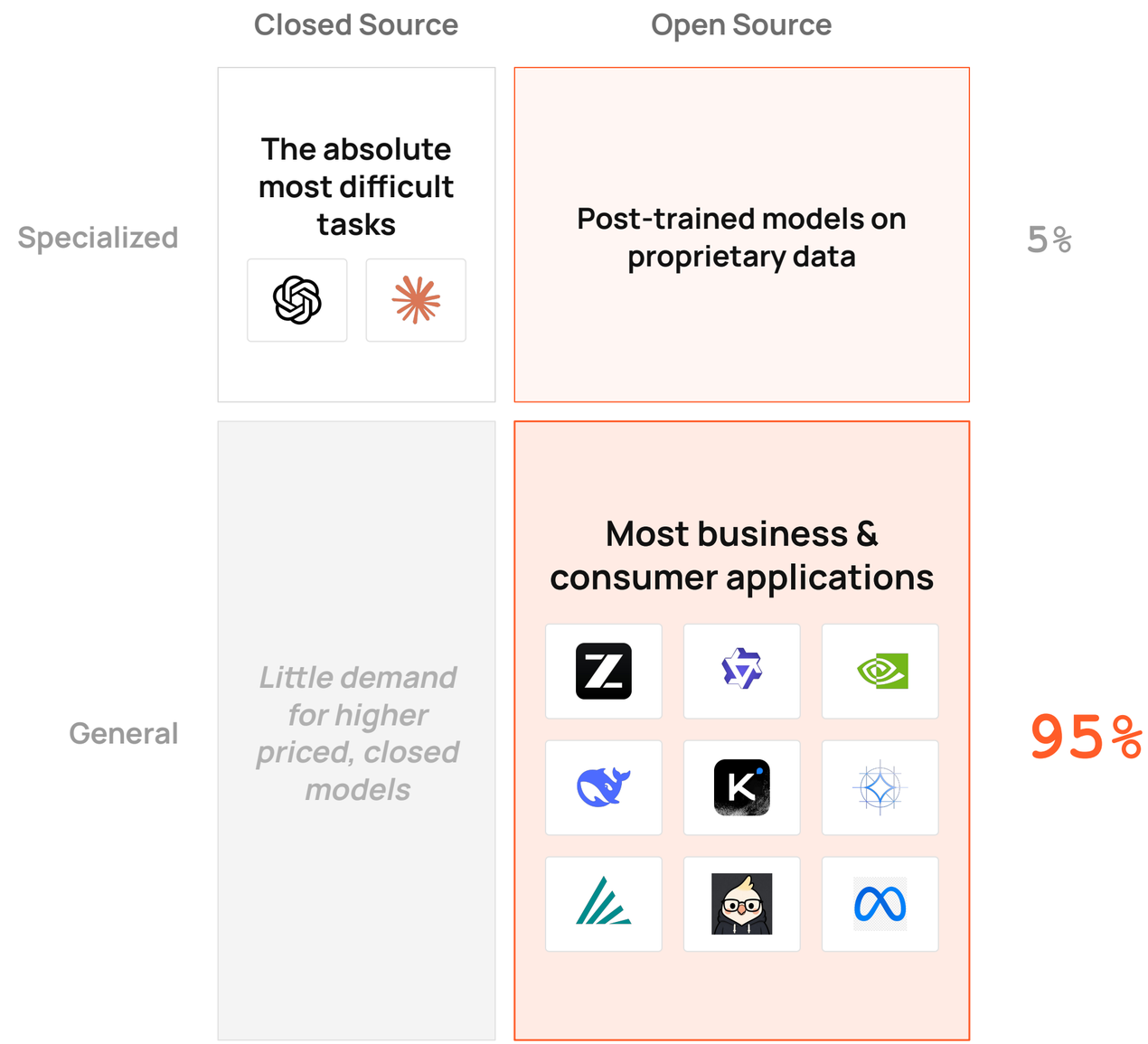


Subconscious

Open models will power most AI workloads.

Cheaper, customizable, and private.

These optimizations would not have been possible without open models.



Product demo

Thank you!

Subconscious

subconscious.dev

Try Subconscious

Connect with me