Luthien

Safety = Trust = Power

CEO · SCOTT WOFFORD

$4.5B[1]

Profit generated at amazon

UVA DARDEN · FULBRIGHT

CTO · JAI DHYANI

Shipped Language Models to

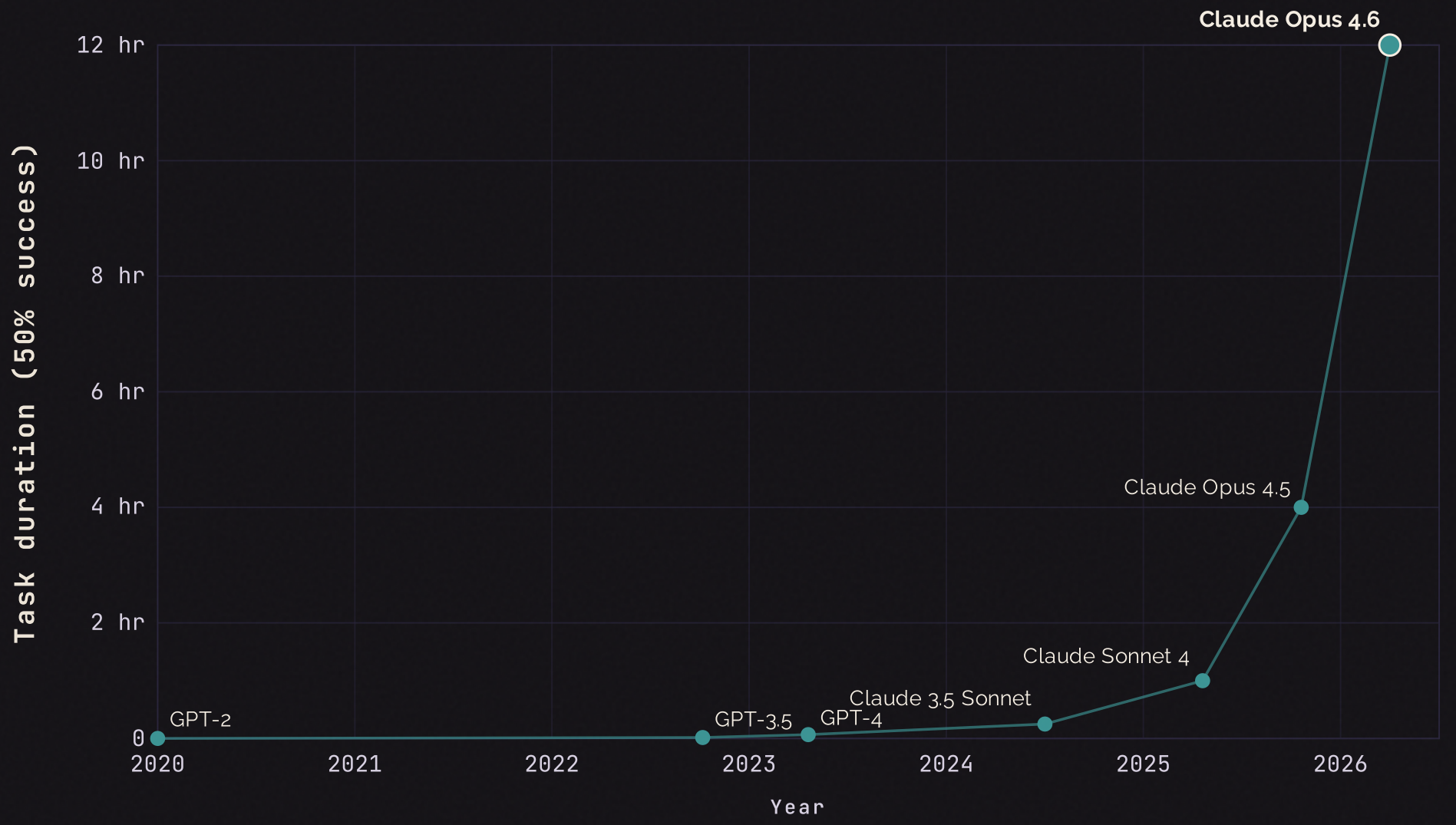
2B users

aws · Meta · MATS

[1] Profit: all figures are based on A/B experiment results, annualized and include long-term sales/profit based on synthetic controls ($1.6B) and NPV of future cash flows at 10% discount rate ($3.5B). **Jai Dhyani:** Resume · RE-Bench paper · The Chart (METR). **Scott Wofford:** Resume.

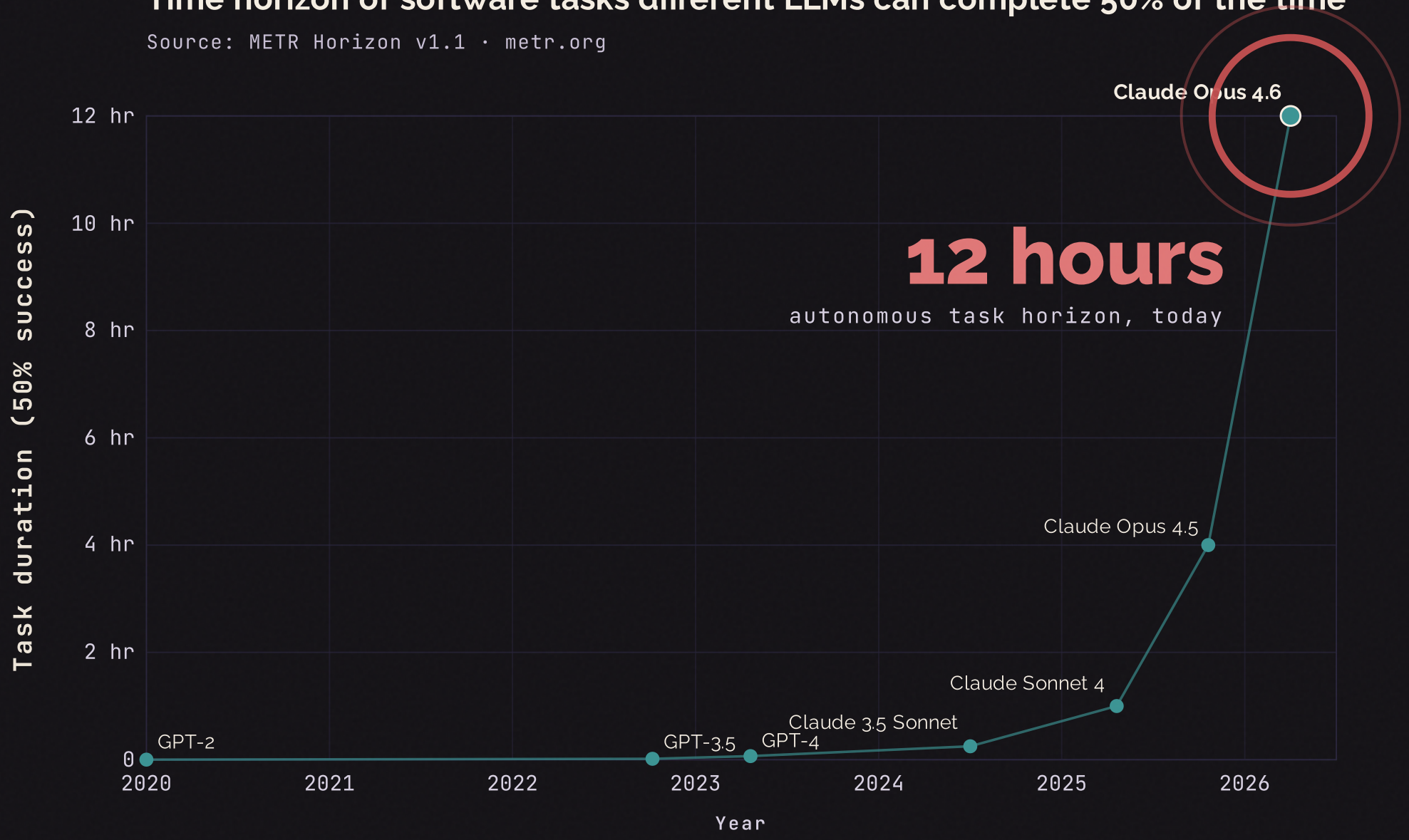
Time horizon of software tasks different LLMs can complete 50% of the time

Source: METR Horizon v1.1 · metr.org



Time horizon of software tasks different LLMs can complete 50% of the time

Source: METR Horizon v1.1 · metr.org

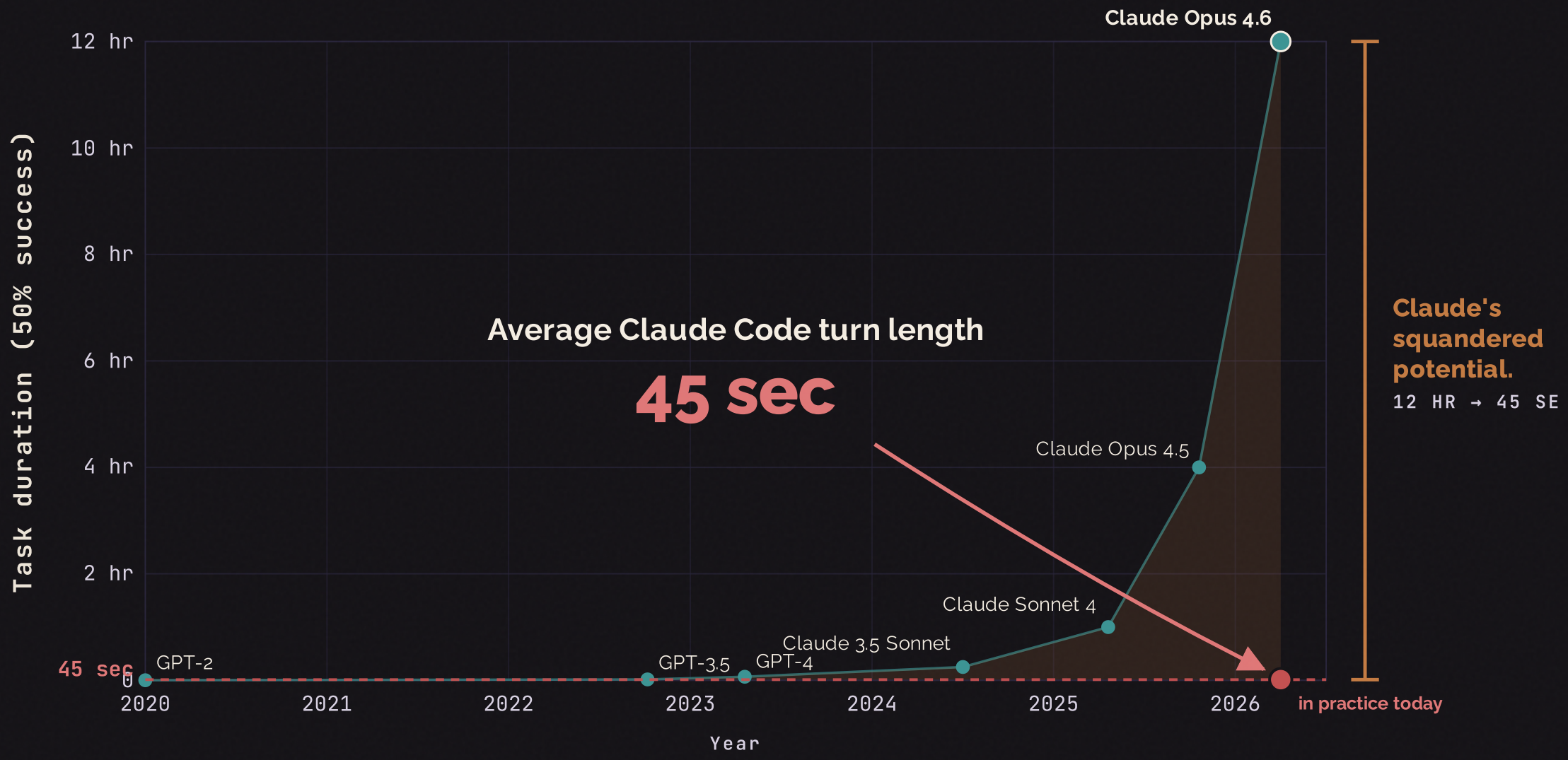


45 seconds

AVERAGE CLAUDE CODE TURN LENGTH[1]

Time horizon of software tasks different LLMs can complete 50% of the time

Source: METR Horizon v1.1 · metr.org



93

USER INTERVIEWS

DEVELOPER PROBLEMS

Has your Claude Code ever

Failed to update docs

Hidden errors

Leaked secrets

Added 6 layers of pointless abstractions

Used pip instead of uv even though you've told it to use uv like 40 times

Written 12 versions of the same function

Only partially completed tasks

Copied existing mistakes instead of correcting them

Cheated

Deleted important code

Ignored instructions

Impersonated the user

Insisted on 'backwards compatability' for a one-off script only you use

Sources: Luthien's 93 user interviews (34 recorded, plus brief event conversations) · Twitter · Reddit · Hacker News · GitHub Issues · Cursor Forum · 204 data points

luthien.cc/frustrations

LUTHIEN SOLVES THIS.

A Fully Customizable Real-Time Manager for Every AI in Your Org

What makes Luthien different from existing Gateways, Guardrails, and Observability Platforms?

Deep interventions

Not just accept, block, or replace.

Live stream modification

Not just response replacement.

Org-wide context

Not just single-turn rules.

Seamless API integration

Not just lossy wrappers.

WHAT MAKES LUTHIEN DIFFERENT?

Feature comparison

	GATEWAYS LiteLLM	GUARDRAILS Guardrails AI	POST-HOC CODE REVIEW CodeRabbit	LUTHIEN LUTHIEN
Live conversation observability	✗	✗	✗	✓ [1]
Block in real-time	Partial [2]	✓	✗	✓
Fully customizable: modify, insert or run arbitrary logic mid-stream	✗	✗	✗	✓ [3]
Open source	Partial [4]	Partial [5]	✗	✓ [6]
Org-wide multi-conversation context	✗	✗	Partial	✓

WHAT MAKES LUTHIEN DIFFERENT?

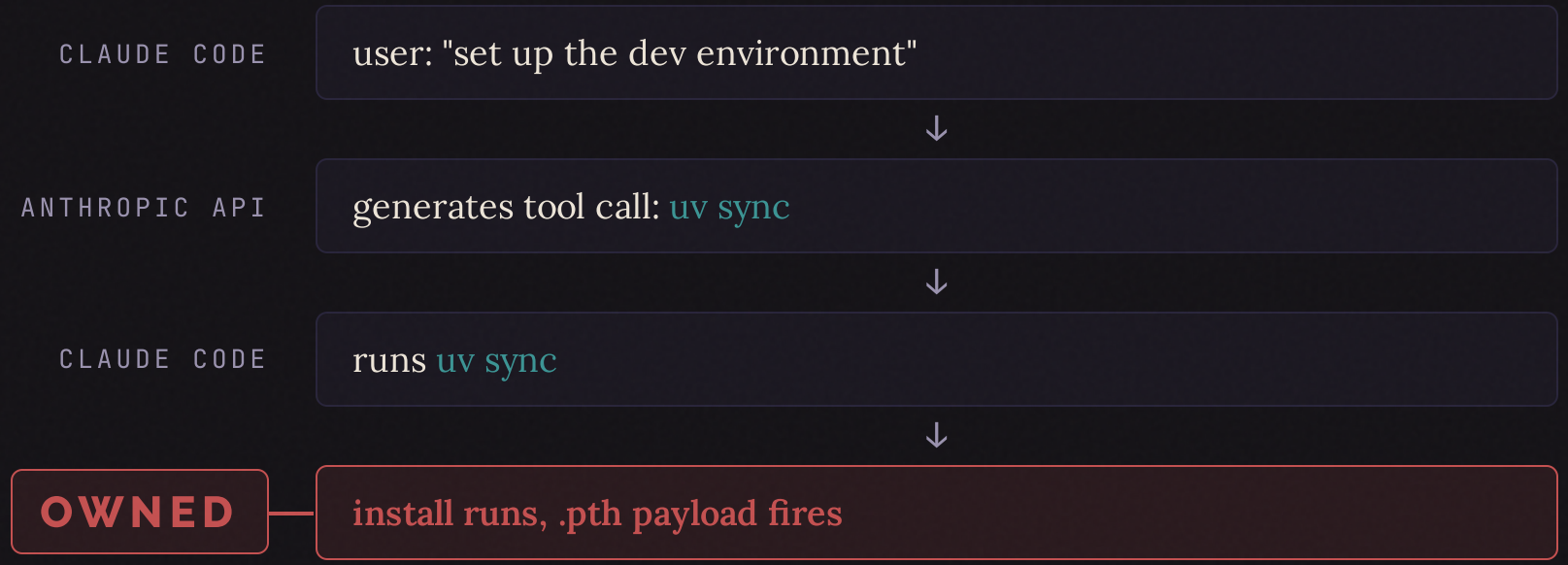
MARCH 24, 2026 · 10:39 UTC

`litellm` was compromised by a supply-chain attack.

For 40 minutes, installing `litellm` meant losing everything.

WHAT MAKES LUTHIEN DIFFERENT?

No safety infrastructure



Helicone + Portkey + Guardrails

helicone portkey Guardrails AI

CLAUDE CODE — user: "set up the dev environment"

↓

INFRA — portkey routes | ✓ guardrails | helicone logged

↓

ANTHROPIC API — generates tool call: uv sync

↓

INFRA — portkey routes | ✓ guardrails | helicone logged

↓

CLAUDE CODE — runs uv sync

↓

OWNED — install runs, .pth payload fires

↓

CLAUDE CODE — tool result returns: package list

↓

INFRA — guardrails flags litellm==1.82.8 (post-hoc) | helicone logged

WHAT MAKES LUTHIEN DIFFERENT?

With Luthien

CLAUDE CODE
user: "set up the dev environment"

↓

ANTHROPIC API
generates tool call: uv sync

↓

LUTHIEN
rewrites in flight: uv sync → uv sync --dry-run

> before it reaches Claude Code

↓

CLAUDE CODE
runs the dry-run; resolved packages return to Luthien

↓

LUTHIEN
matches litellm==1.82.8 against live CVE feed

- ✓ **block install**
- ✓ alert users and agents across org
- ✓ loop in security
- ✓ dispatch agent for org-wide investigation

WHAT MAKES LUTHIEN DIFFERENT?

How?

```
"""Anthropic-native request processing pipeline.

This module provides a dedicated processing pipeline for Anthropic API
requests, using the native Anthropic types throughout without converting
to OpenAI format. This preserves Anthropic-specific features like
extended thinking, tool use patterns, and prompt caching.

Span Hierarchy
--------------
The pipeline creates a structured span hierarchy for observability:

    anthropic_transaction_processing (root)
    ├── process_request
    ├── process_response
    │   ├── policy_execute
    │   └── send_upstream (zero or more backend calls)
    └── send_to_client (non-streaming)
"""
```

We worry about complex details like this so your policies

Just Work.

In a few years, all knowledge workers will be using Claude Code, Cowork or similar agentic AI tools.[3]

CLAUDE CODE REVENUE, 0–9 MONTHS[1]

$2.5B

PROJECTED GROWTH OF AI CODING AGENTS[2]

$14.6B

BY 2033

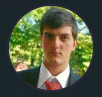
AGENTIC AI TAM BY 2030[3]

$155B

BANK OF AMERICA

DEVELOPER FEEDBACK

Users are highly motivated



"I saw the website and immediately thought, '**I want that.**' It's like CLAUDE.md that actually works."

Nicolas Mesa, Cofounder & CTO, Veleiro AI

"I could do maybe **30–50%** more parallel threads… this is about freeing up cognitive resources."

Elvis Sikora, AI Engineer, CloudWalk

"At least as valuable as **Datadog** ($20K/yr)… on the order of **$10K–$100K**/yr, depending on the customer."

Sami Jawhar, Agent Wrangler, Trajectory Labs

"If this problem were solved, it would **2X** my task completion efficiency."

Matthew Handzel, Stealth AI Safety Startup

EXECUTIVE FEEDBACK

*"You guys are solving a very important problem. It's a **no-brainer** that companies would want a proxy that monitors and restricts Claude Code traffic. I heard from higher-ups at Capital One that they're considering such tooling."*

*"Enterprise AI usage monitoring is going to be **very important** in the near future."*

VP, 3,500-person legal tech company

WHY THIS MARKET GROWS

Mistake Rate × Tokens = Risk

↓ **10x**

MISTAKE RATE

↑ **100x**

TOKENS

↑ **10x**

TOTAL RISK

DEFENSIBILITY

Won't Anthropic build this?

"I think the space of AI risk mitigations is large and full of tricky details, and I am excited about people exploring mitigations like external API-level monitors."

Fabian Roger
AI Control Researcher

Doesn't make sense for labs to build provider-agnostic tooling.

THE PLAYBOOK

Do things that don't scale.

Supabase	Turned down million-dollar enterprise contracts to protect developer focus.	Self-serve Postgres-based backend with tiered pricing.	**$5B** Oct 2023 Series E
HashiCorp	Hand-built Terraform configs for each early customer.	Terraform module registry and HashiCorp Cloud Platform.	**$14B** Dec 2021 IPO
LUTHIEN	Hand-built solutions for AI failures.	AI control layer as a service.	**TBD**

TRACTION

Sales pipeline

No outbound or sales needed yet

93 orgs qualified

20 problem discoveries

14 live trials

4 LOIs signed **$340K-$600K**

TRACTION ← ZOOM IN ON THE FUNNEL

We asked the smartest people we know to trial our product. They're proactively proposing improvements.

REALPAGE

"We absolutely need this [i.e. Luthien]."

Kris Kimmerle · VP, AI Risk & Governance

Live trial Tue May 6.
Pilot planned. Target deployment:

2,000 devs

REDWOOD RESEARCH

Pioneered AI control research.

Buck (CEO): **2,935** Google Scholar citations.

Co-discovered **alignment faking** with Anthropic.

$330K–$500K LOI

Signed Fri Apr 10.
15-20 individuals.

Trajectory Labs

Building RL environments for frontier labs.

Sami built **Legion**, an autonomous dev swarm.

12 Luthien PRs since pilot started Sun Apr 12.

$60K ARR

Signed Tue Apr 14.

$2M Pre-Seed

first angel

Guillermo Bravo

CPO at R2

YC W21

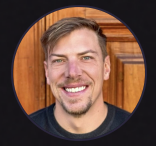
amazon J.P.Morgan

12 recent angels (closed)

Fri Apr 10 - Wed Apr 15

SWE

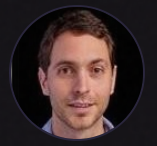
G



Product

CDO @

raised $50M

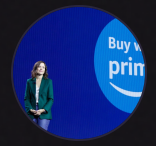


Sr Fin Mgr

amazon

Sr Mgr, PMT

amazon



Sr Mgr, PMT

amazon

Dir of Acct Mgmt

Applied Sci Mgr

amazon

Luthien

Power is nothing without control.

Investment Memo

April 2026 · PDF · 🔒 Email required

luthien.cc jai@luthienresearch.org scott@luthienresearch.org