Inferred Edge AI Update

June 2025

© 2025 Inferred Edge

Top AI Industry News

- **AI M&A Heating Up + Continued Financings:**
  - Databricks acquires Neon for $1B: strengthening its AI infrastructure capabilities
  - OpenAI acquires io for $6.5B: hiring Apple's star designer, Jony Ive + team, to build hardware
  - Salesforce acquires Informatica for $8B: bolstering its Data Cloud and data governance capabilities
  - Qualcomm acquires Alphawave Semi for $2.4B: expanding into AI data center market
  - Meta takes a $14.8B (49%) stake in Scale AI, including onboarding Scale CEO, Alexandr Wang, to lead a new AI lab
  - Cursor raises $900M Series B at $9B valuation led by Thrive Capital
- **AI Coding System Updates:**
  - OpenAI launches Codex, a coding agent that hooks into code repositories and performs autonomous tasks
  - Cursor releases Cursor 1.0 with enhanced AI-first coding capabilities, including BugBot autonomous code review
  - Anthropic releases "How Teams Use Claude Code" to demonstrate 50+ enterprise use cases
- **Major Model Releases:**
  - Anthropic's Claude 4, Google's Gemini 2.5 Pro, OpenAI's o3-pro (see slide 12 for additional details)

© 2025 Inferred Edge

OpenAI vs Anthropic: We've Seen This Before

Just as Apple's iOS and Google's Android defined mobile computing, OpenAI's ChatGPT and Anthropic's Claude Desktop are facing the same key questions around product and GTM strategy.

Ecosystem Strategy	Example Company	Product Strategy	GTM Strategy	Strategic Trade Offs
Closed	Apple (iOS)	Curated Experience	Enterprise Adoption	■ Quality Control ■ Tight Integrations ■ Enterprise Trust
Open	Google (Android)	Platform for Innovation	Developer Adoption	■ Customizable Features ■ Broad Compatibility ■ Developer Love

Both closed and open ecosystems can win - the question is which strategy fits the AI moment?

© 2025 Inferred Edge

A Closer Look: Battle for AI OS Supremacy

OpenAI: *the **Apple** Approach*

Leans toward closed ecosystem

- Tight quality control of user experience with the GPT Store + pre-built Connectors to data sources
- Supports MCP, but only recently. To use requires admin set-up, and enterprise use only
- Favors first party vs third party integrations

"We want to be people's core AI subscription… we'll have these surfaces, like future devices, future things that are sort of similar to operating systems…"

- Sam Altman, CEO OpenAI

Anthropic: *the **Google** Approach*

Leans toward open ecosystem

- Created MCP, betting on rapid adoption with developers
- Developers can create MCP servers for any data source, no approval needed
- Favors third party integrations with standardization coming from protocol

"By *powerful AI*, I have in mind… all the 'interfaces' available to a human working virtually… it can engage in any actions, communications, or remote operations enabled by this interface…"

- Dario Amodei, CEO Anthropic

© 2025 Inferred Edge

What an AI OS Means for Enterprise

- **Unified Workflow Orchestration**
  - Single interface managing multiple enterprise systems
- **Context Preservation Across Tools**
  - AI maintains state between different platforms/applications
  - Employees work through one centralized AI interface rather than juggling multiple apps
- **Opportunity for Custom Interfaces**
  - With Large Language Models accessible via APIs and other tooling via MCP, companies can build their own single pane of glass AI interfaces with the security, governance and tooling specific to their business

The AI OS isn't replacing the enterprise stack —
it's becoming the intelligent layer that connects everything together.

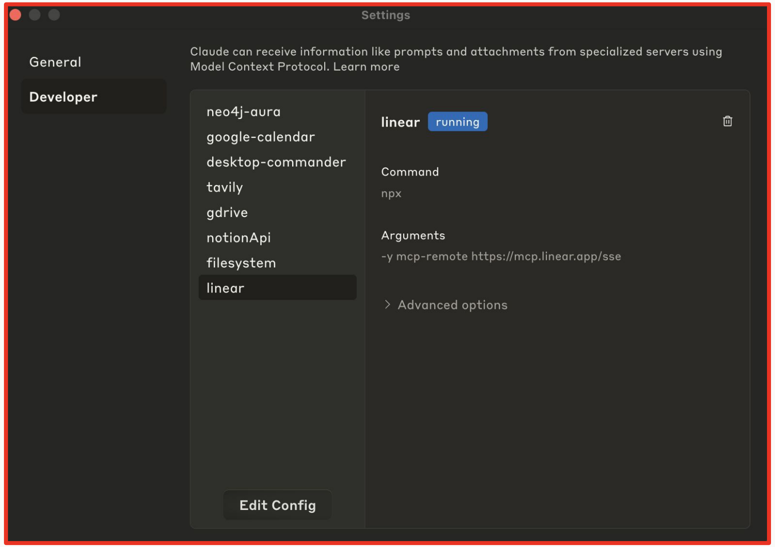
© 2025 Inferred Edge

Example: Claude Desktop as AI OS

To provide an intelligent layer, we configure Claude with access to tools, while maintaining context across the different platforms.

The tools Claude uses in the following example:

1. **Linear MCP** to check project status and issues (tasks)
2. **Desktop Commander MCP** to find the codebase in the user's file system, diagnose and fix the SST deployment problem
3. **Google Calendar MCP** to schedule the requested meeting
4. **Notion MCP** to create meeting documentation with the specific agenda item



Claude can receive information like prompts and attachments from specialized servers using Model Context Protocol. Learn more

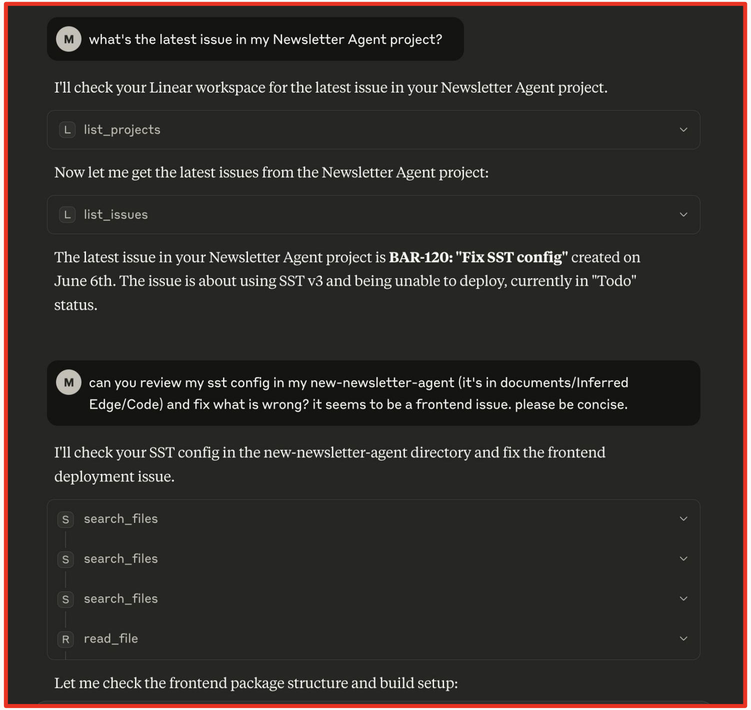
Command
npx

Arguments
-y mcp-remote https://mcp.linear.app/sse

Above: an example MCP config in Claude Desktop.

© 2025 Inferred Edge

1 Example: Claude Desktop as AI OS



what's the latest issue in my Newsletter Agent project?

I'll check your Linear workspace for the latest issue in your Newsletter Agent project.

Now let me get the latest issues from the Newsletter Agent project:

The latest issue in your Newsletter Agent project is **BAR-120: "Fix SST config"** created on June 6th. The issue is about using SST v3 and being unable to deploy, currently in "Todo" status.

can you review my sst config in my new-newsletter-agent (it's in documents/Inferred Edge/Code) and fix what is wrong? it seems to be a frontend issue. please be concise.

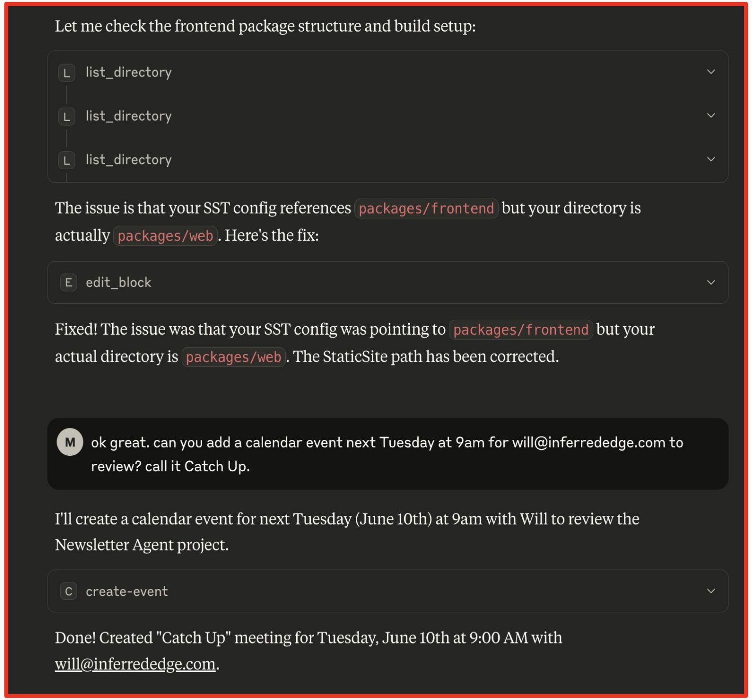
I'll check your SST config in the new-newsletter-agent directory and fix the frontend deployment issue.

Let me check the frontend package structure and build setup:

Claude uses the Linear MCP to get the latest issue + description

Claude leverages DesktopCommander MCP to find and read the repository

2 Example: Claude Desktop as AI OS



Let me check the frontend package structure and build setup:

The issue is that your SST config references `packages/frontend` but your directory is actually `packages/web`. Here's the fix:

Fixed! The issue was that your SST config was pointing to `packages/frontend` but your actual directory is `packages/web`. The StaticSite path has been corrected.

M ok great. can you add a calendar event next Tuesday at 9am for will@inferrededge.com to review? call it Catch Up.

I'll create a calendar event for next Tuesday (June 10th) at 9am with Will to review the Newsletter Agent project.

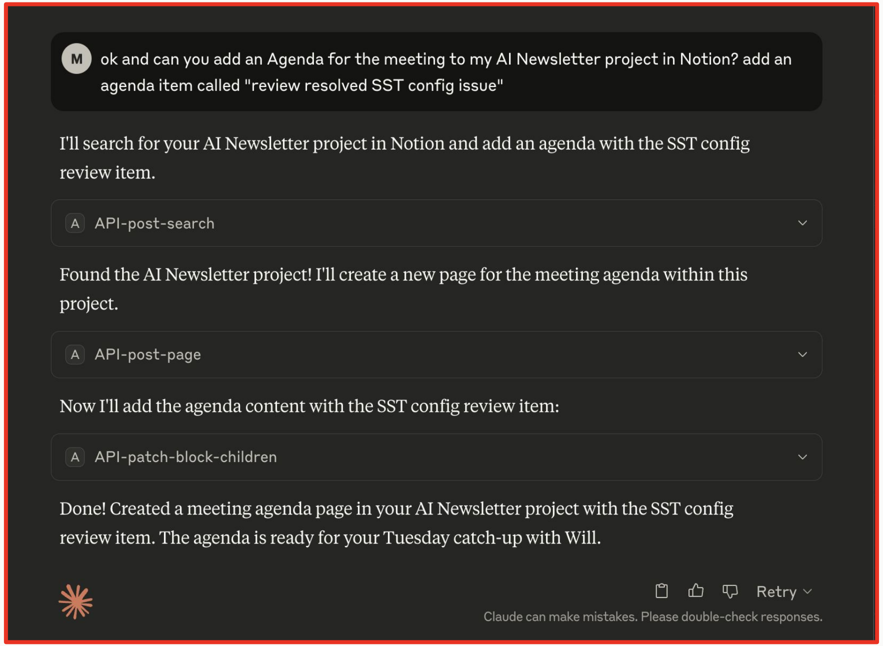
Done! Created "Catch Up" meeting for Tuesday, June 10th at 9:00 AM with will@inferrededge.com.

Claude diagnoses the issue and uses DesktopCommander MCP to edit the file

Claude chooses Google Calendar MCP and creates a new event

© 2025 Inferred Edge

③ Example: Claude Desktop as AI OS



M ok and can you add an Agenda for the meeting to my AI Newsletter project in Notion? add an agenda item called "review resolved SST config issue"

I'll search for your AI Newsletter project in Notion and add an agenda with the SST config review item.

Found the AI Newsletter project! I'll create a new page for the meeting agenda within this project.

Now I'll add the agenda content with the SST config review item:

Done! Created a meeting agenda page in your AI Newsletter project with the SST config review item. The agenda is ready for your Tuesday catch-up with Will.

Claude can make mistakes. Please double-check responses.

Claude uses the Notion API to find the correct project, create a new page and add agenda items

© 2025 Inferred Edge

Google Is Also a Player in the AI OS Battle

The incumbent is leveraging their existing ecosystem dominance in Google Workspaces and in-house infrastructure advantages to build their AI OS for enterprise.

AI-Native Integration to Existing Products

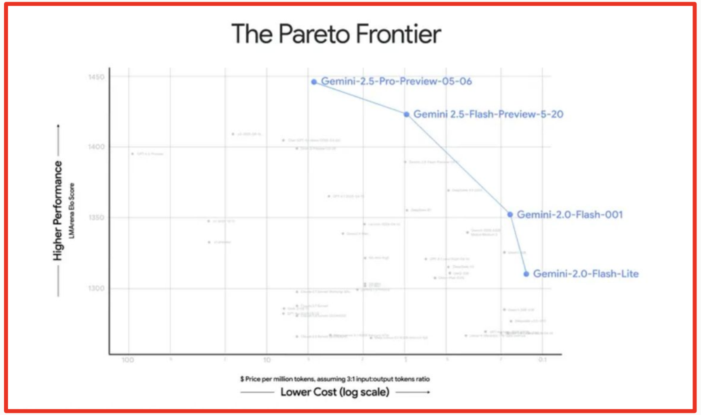
- Google Workspaces enterprise subscription now embeds Gemini models across calendar, gmail and drive
- Goal to maintain user's personal context across Google apps for consistent UI

Infrastructure-First Strategy

- Ironwood TPU: 10x performance improvement vs previous generation
- Monthly token processing surged (9.7 trillion to 480 trillion YoY)

Focus on Rapid Iteration

- Google Labs achieving 50-100 day cycles from concept to testing
- Research projects becoming shipped products:
  - NotebookLM driving 400M+ Gemini app users
  - Project Mariner becoming Agent Mode in Google Search



Cost-Performance Advantage

Custom Hardware (Ironwood TPU) + Vertical Integration Across Stack + Economies of Scale = Leading the Pareto Frontier

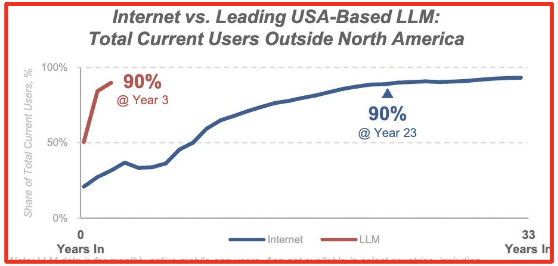
Sources: Google, "Google I/O 2025: From research to reality", May 2025; Sequoia Capital, "Google I/O Afterparty: The Future of Human-AI Collaboration, From Veo to Mariner", May 2025.

© 2025 Inferred Edge

Bond Capital AI Trends Report: Key Takeaways

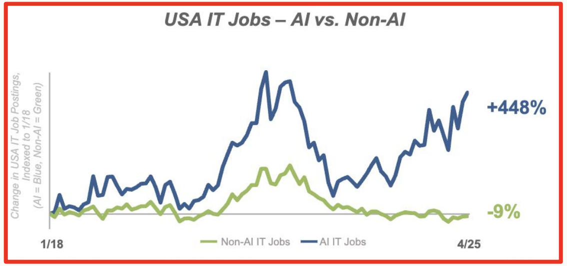
Mary Meeker's latest report demonstrates the unprecedented pace and scope of the AI revolution.

AI Adoption Is Growing at Unprecedented Rates



Left: 90% of ChatGPT users are now outside North America, achieved in 3 years (vs 23 years for the internet)

Right: AI job postings are up +448%, while traditional IT listings dropped -9%



Other Key Data Points

- ChatGPT Hit 365B Annual Searches in 2 Years (2024) vs. Google in 11 Years (2009)
- >7 million developers are building with Google's Gemini, +5x YoY
- Model training compute costs +2,400x over 8 years, while performance among competitors rapidly converges
- AI inference costs for customers -99.7% over 2 years

Source: Bond Capital, "Trends – Artificial Intelligence," May 2025.

© 2025 Inferred Edge

Large Language Models: Latest Roster (1 of 2)

Model	Parent	Release Date	License Type	Context Window*	Blended Price**	Initial Market Traction
o3-pro	OpenAI	June 2025	Proprietary	200K	$35.00	Multi-step reasoning; highly complex tasks
Gemini 2.5 Flash-Lite	Google	June 2025	Proprietary	1M	$0.38	Translation + classification tasks
Gemini 2.5 Pro	Google	May/June 2025	Proprietary	1M	$3.44	Enterprise data extraction; web app creation
Claude Opus 4	Anthropic	May 2025	Proprietary	200K+	$30.00	Complex coding; autonomous development
Claude Sonnet 4	Anthropic	May 2025	Proprietary	200K+	$6.00	Co-pilot coding with enhanced context
Gemini 2.5 Flash	Google	May 2025	Proprietary	1M	$1.50	High-volume apps at lower costs/query
Gemma 3n	Google	May 2025	Open	128K	$0.00	Mobile/edge device deployment
GPT-4.1	OpenAI	April 2025	Proprietary	1M	$3.40	Document processing
o4-mini	OpenAI	April 2025	Proprietary	200K	$1.93	High-volume processing in production
o3	OpenAI	April 2025	Proprietary	200K	$3.50***	Complex technical problem solving

*maximum number of tokens processed in single interaction. **per 1M tokens; blended input/output costs. *** in June 2025, reduced 80% from $17.50
Source: OpenAI, Anthropic, Google, Meta.

© 2025 Inferred Edge

Large Language Models: Latest Roster (2 of 2)

Model	Parent	Release Date	License Type	Context Window*	Blended Price**	Initial Market Traction
Llama 4 Scout	Meta	April 2025	Open	10M	$0.17	Edge computing
Llama 4 Maverick	Meta	April 2025	Open	1M	$0.40	Multimodal applications
Gemini 2.5 Pro (Initial Release)	Google	March 2025	Proprietary	1M	$3.44	Long document processing/comprehension
GPT-4o	OpenAI	February 2025	Proprietary	128K	$5.11	Real-time chat
Claude 3.7 Sonnet	Anthropic	February 2025	Proprietary	200K	$6.00	Software engineering and coding
Claude 3.5 Sonnet	Anthropic	June 2024	Proprietary	200K	$6.00	Coding and creative content generation

*maximum number of tokens processed in single interaction. **per 1M tokens; blended input/output costs.
Source: OpenAI, Anthropic, Google, Meta.

© 2025 Inferred Edge

Recommended Next Steps

Data Readiness Assessment

- Audit existing data sources and structures for AI integration readiness
- Identify priority datasets for initial implementation
- Develop data governance frameworks to ensure compliance, quality and security

Use Case Prioritization

- Evaluate high-impact, low-complexity opportunities for initial deployment
- Conduct stakeholder workshops to identify pain points addressable by AI
- Define clear success metrics for pilot implementations

Implementation Roadmap

- Schedule a technical discovery session with Inferred Edge experts
- Develop a phased implementation plan with defined milestones

© 2025 Inferred Edge