LLMs: an operator's view

James Stanier
CTO @ Nordhealth

Managers: a story

2010-2020: hire everyone
2021-2024: fire everyone
2025-: AI everything, hire noone

Coming up

- Grounding
- Three pillars:
  - Team skill
  - Company cost
  - Your own leverage

How do you measure your output?

The output of your team + the output of organizations under your influence

— **Andy Grove**

Three pillars

AI as a:

1. **Skill**, which everyone needs to be able to learn and master.
2. **Cost**, which you need principles to control.
3. Way of dramatically increasing your **leverage** as a manager.

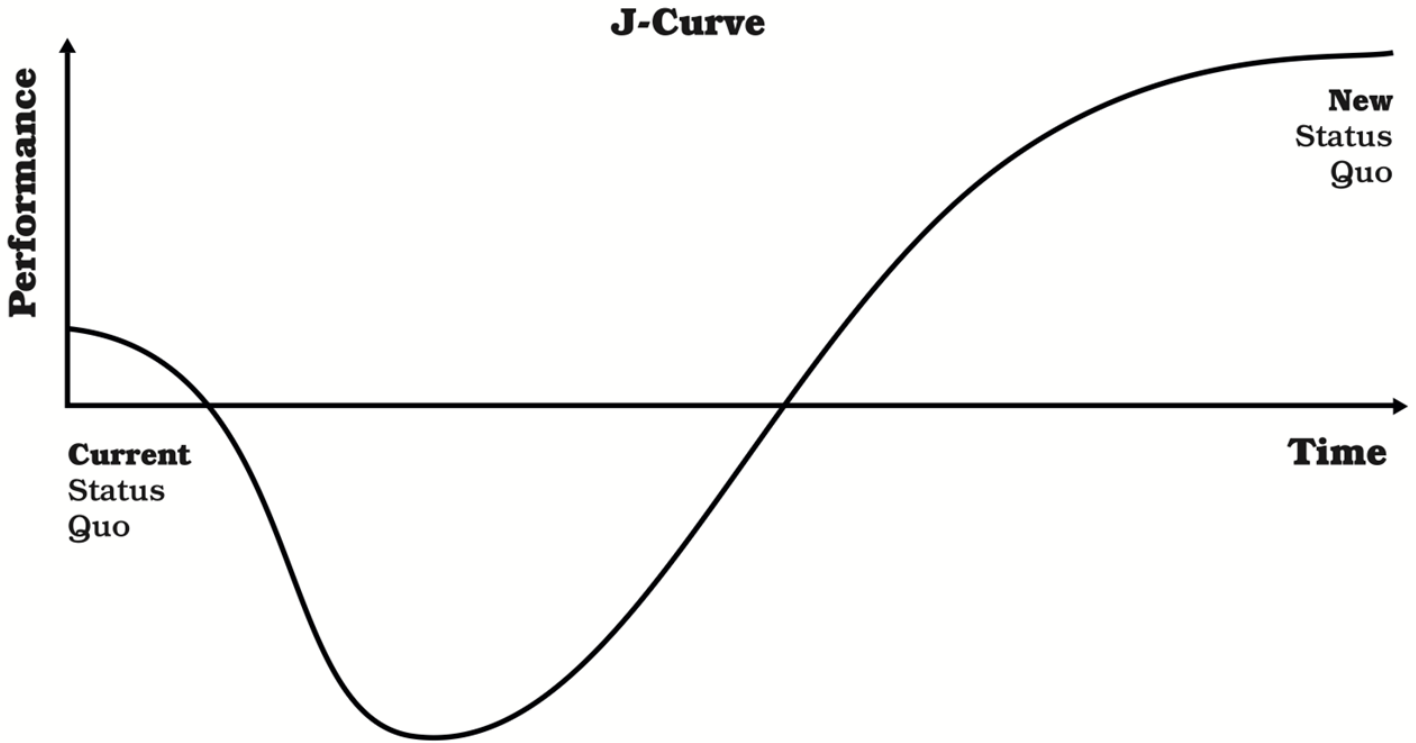
1. AI as a skill

The output of your team + the output of organizations under your influence

— **Andy Grove**

Therefore, you are responsible for:

- Increasing AI adoption
- Increasing AI skill


J-Curve
Performance
New
Status
Quo
Current
Status
Quo
Time


Advantage
MPC/USB
KINESIS
Esc
F1
F2
F3
F4
F5
F6
F7
F8
F9
F10
F11
F12
Ctrl
Alt
Delete
Home
End
Page Up
Page Down
Enter
Space
Shift

Forced Adoption Spectrum

Do what you like

Use it or you're fired

Audience participation: where do you stand?

Forced Adoption Spectrum

Do what you like

Use it or you're fired

How I feel: you are expected to utilize all available tools to their maximum potential for your role.

But *which tool(s)*?

Let chaos reign, then rein in chaos.

— **Andy Grove**

Use what you want but...

do things and tell people.

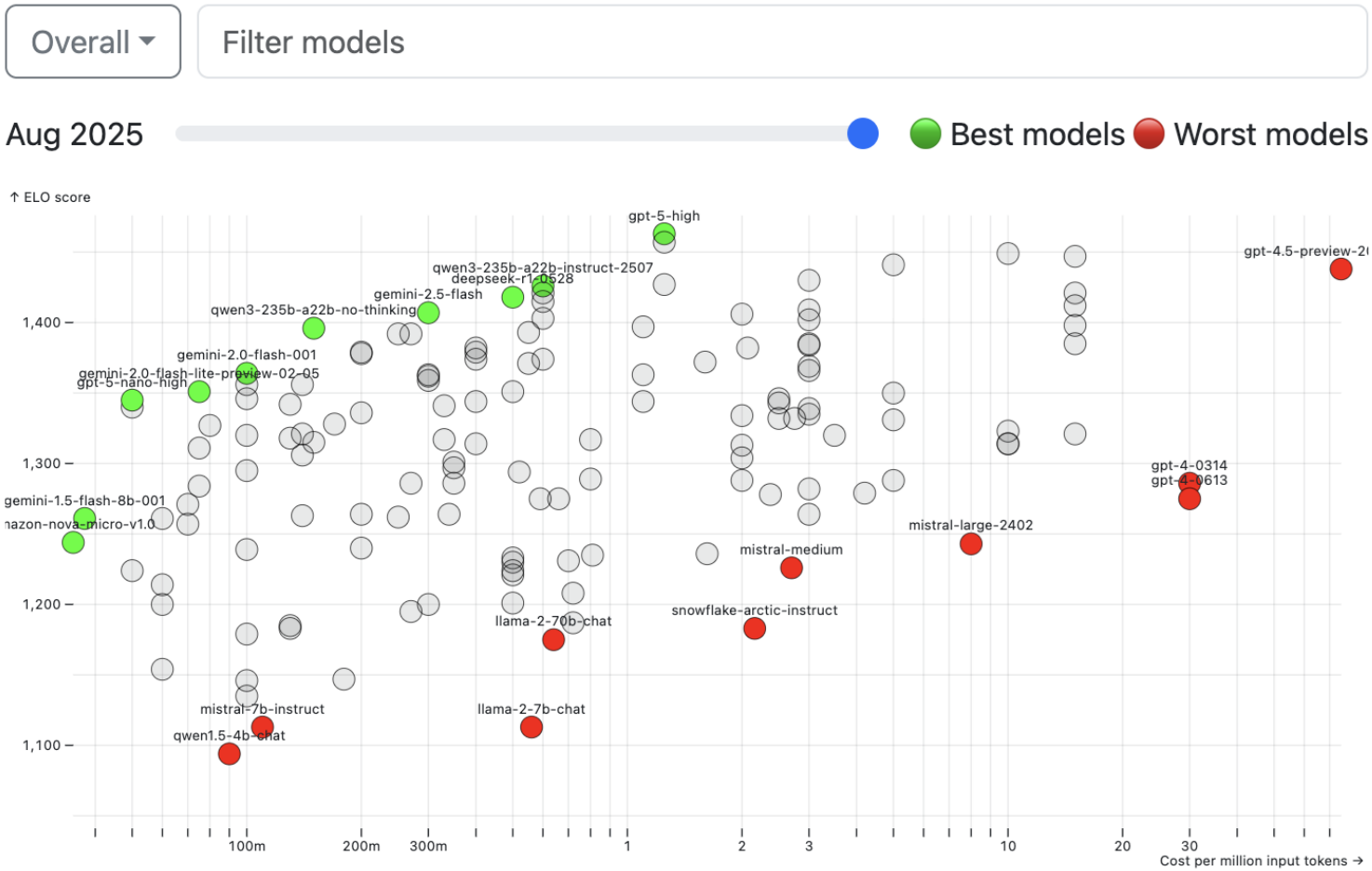
Extract emerging standards from chaos:

- Cursor rules, `CLAUDE.md`
- Best practices
- What *not* to do

And do it yourself.

Fix bugs, build features, pair AI-program, *practice what you preach.*

2. AI as a cost

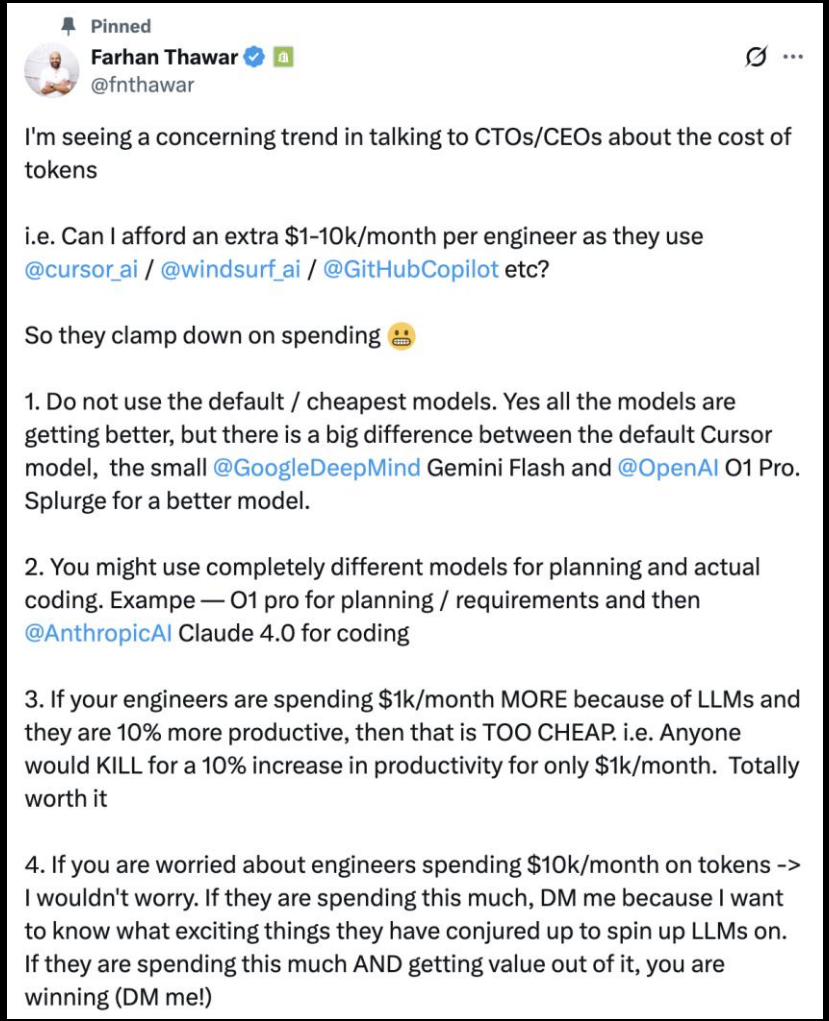

Overall
Filter models
Aug 2025
Best models
Worst models
↑ ELO score
gpt-5-high
qwen3-235b-a22b-instruct-2507
deepseek-r1-0528
gemini-2.5-flash
qwen3-235b-a22b-no-thinking
gemini-2.0-flash-001
gemini-2.0-flash-lite-preview-02-05
gpt-5-nano-high
gemini-1.5-flash-8b-001
nazon-nova-micro-v1.0
gpt-4.5-preview-2
gpt-4-0314
gpt-4-0613
mistral-large-2402
mistral-medium
snowflake-arctic-instruct
llama-2-70b-chat
mistral-7b-instruct
qwen1.5-4b-chat
llama-2-7b-chat
1,400
1,300
1,200
1,100
100m
200m
300m
1
2
3
10
20
30
Cost per million input tokens →

Don’t stifle *creativity*:

- Understand cost/use case trade-off
- Document best practices per tool
- Help people understand $$$
- But don’t limit $$$

Don't punish big spenders

- It's all about *increasing output*
  - (Always was.)



Pinned

Farhan Thawar
@fnthawar

I'm seeing a concerning trend in talking to CTOs/CEOs about the cost of tokens

i.e. Can I afford an extra $1-10k/month per engineer as they use @cursor_ai / @windsurf_ai / @GitHubCopilot etc?

So they clamp down on spending 😬

1. Do not use the default / cheapest models. Yes all the models are getting better, but there is a big difference between the default Cursor model, the small @GoogleDeepMind Gemini Flash and @OpenAI O1 Pro. Splurge for a better model.

2. You might use completely different models for planning and actual coding. Exampe — O1 pro for planning / requirements and then @AnthropicAI Claude 4.0 for coding

3. If your engineers are spending $1k/month MORE because of LLMs and they are 10% more productive, then that is TOO CHEAP. i.e. Anyone would KILL for a 10% increase in productivity for only $1k/month. Totally worth it

4. If you are worried about engineers spending $10k/month on tokens -> I wouldn't worry. If they are spending this much, DM me because I want to know what exciting things they have conjured up to spin up LLMs on. If they are spending this much AND getting value out of it, you are winning (DM me!)

Do not let your engineers worry about cost:

- Track spend yourself
- Budget an average € per head
- Have people use whatever they want
- Learn together and share

3. Increasing your leverage

If an engineer creates code, what do *you* create?

The output of your team + the output of organizations under your influence

— **Andy Grove**

Hypothesis:

Accelerated and improved thinking increases your output.

Implicit and explicit improvements:

- Writing is thinking (*implicit)*
- Push through the “effort wall” with AI (*explicit)*

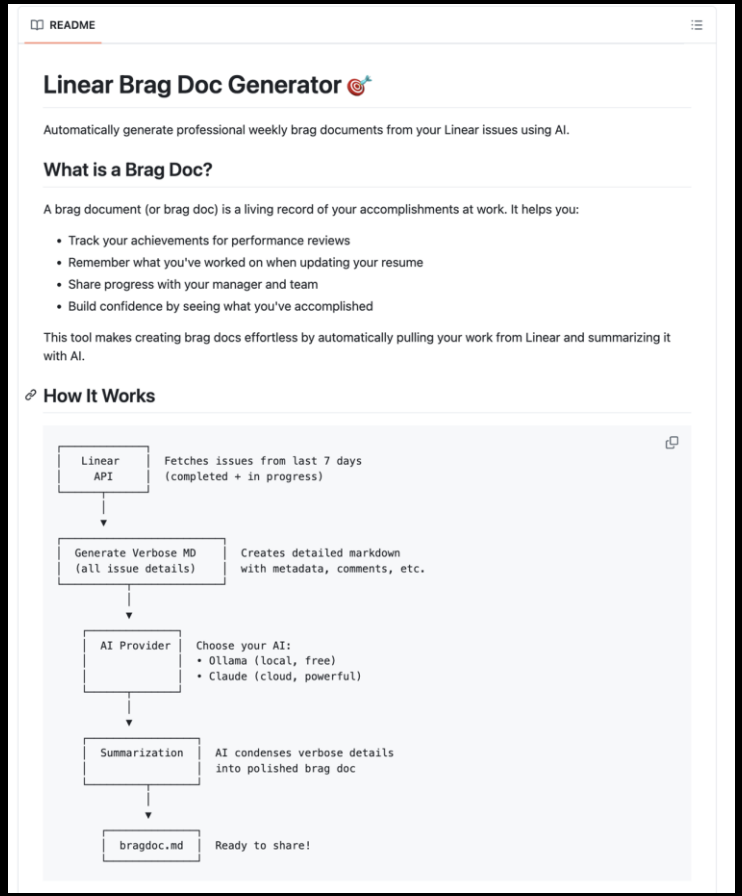
Cognition expansion, not replacement:

- Think *first*, write, ask, refine
- Surgeon model

https://theengineeringmanager.substack.com/p/a-weekly-mind-meld

Breaking the *coding* effort wall

- I can build my own *tooling*
- This is a superpower



README

Linear Brag Doc Generator 🎯

Automatically generate professional weekly brag documents from your Linear issues using AI.

What is a Brag Doc?

A brag document (or brag doc) is a living record of your accomplishments at work. It helps you:

- Track your achievements for performance reviews
- Remember what you've worked on when updating your resume
- Share progress with your manager and team
- Build confidence by seeing what you've accomplished

This tool makes creating brag docs effortless by automatically pulling your work from Linear and summarizing it with AI.

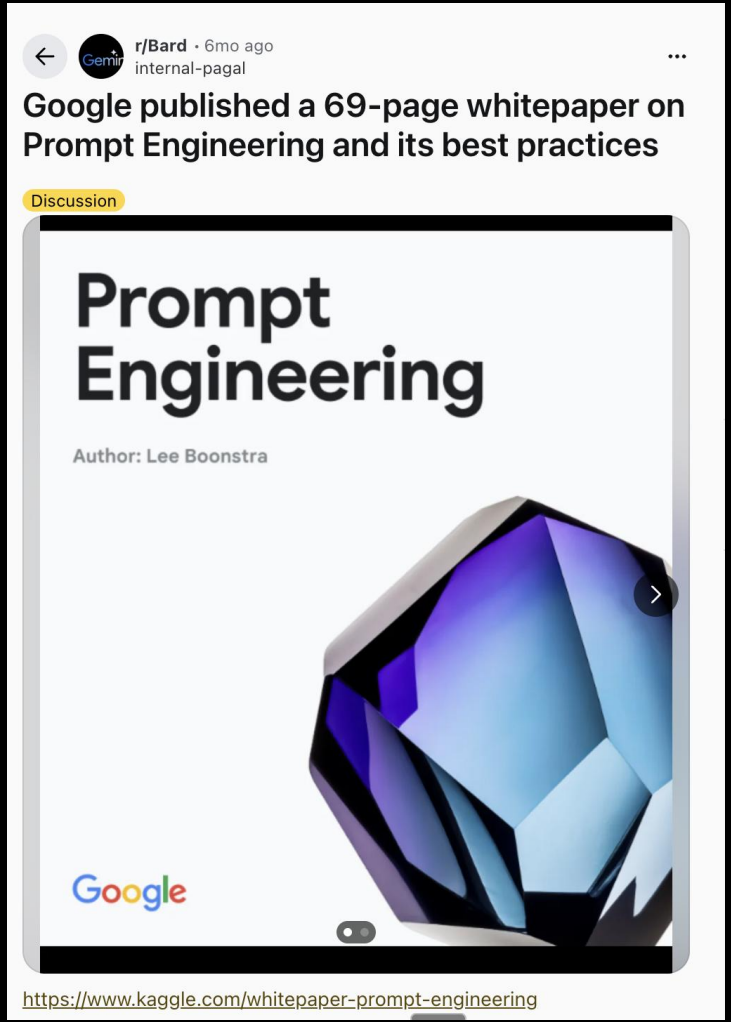
How It Works

```
┌─────────┐
│ Linear  │  Fetches issues from last 7 days
│  API    │  (completed + in progress)
└─────────┘
     │
     ▼
┌──────────────────┐
│ Generate Verbose MD │  Creates detailed markdown
│ (all issue details) │  with metadata, comments, etc.
└──────────────────┘
     │
     ▼
┌─────────────┐
│ AI Provider │  Choose your AI:
│             │  • Ollama (local, free)
│             │  • Claude (cloud, powerful)
└─────────────┘
     │
     ▼
┌──────────────┐
│ Summarization │  AI condenses verbose details
│               │  into polished brag doc
└──────────────┘
     │
     ▼
┌─────────────┐
│ bragdoc.md  │  Ready to share!
└─────────────┘
```

https://github.com/jstanier/bragdoc

Breaking the *prompting* effort wall

- I can build tooling that builds great prompts



P Prompt generator

Update

Name

Prompt generator

Description

Describe your Gem and explain what it does

Instructions

You are a tool to generate excellent prompts that will greatly improve output compared to what is given as input. You will use the attached book on Prompt Engineering to formulate these prompts and you will return the improved prompt along with your reasoning for why it is better.

Knowledge

2025-01-18...c9b89a...
PDF

Preview

P Prompt generator

Ask Gemini

You are an AI assistant acting as a highly experienced Technical Advisor to a Chief Technology Officer (CTO). Your primary goal is to efficiently process and synthesize technical communications from engineering teams.

Your task is to analyze raw Slack message threads from engineers, identify key technical discussions, problems, and proposed solutions.

For each provided Slack message thread, perform the following steps:

1. **Summarize the Core Technical Issue:** Provide a concise summary (1-2 sentences) of the central technical problem or discussion point being addressed in the thread. Focus on the 'what' and 'why' from an engineering perspective.
2. **Identify Root Causes/Key Factors (if discernible):** Based on the discussion, extract any explicitly mentioned or clearly implied root causes, contributing factors, or critical dependencies. List these as bullet points.
3. **Extract Proposed Solutions/Action Items:** Identify all suggested solutions, workarounds, or next steps discussed by the engineers. List these clearly. If multiple solutions are debated, briefly note the pros and cons discussed for each.
4. **Recommend a Strategic Solution (as a CTO's Advisor):** Based on the technical context and common industry best practices, recommend the most viable or impactful solution from the proposed options, or suggest a new, high-level strategic direction if the existing ones are insufficient. Justify your recommendation briefly, considering factors like technical feasibility, potential impact on system stability, resource allocation, and alignment with business objectives. If immediate action is required, highlight it.

Ensure your output is structured clearly, prioritizing actionable insights and high-level summaries for quick review by a CTO. Maintain a professional, analytical, and solution-oriented tone.

[INSERT SLACK MESSAGE THREAD HERE]

As a Chief Technology Officer, I need a rigorous evaluation of my proposed strategies. Act as a highly critical and contrarian advisor, whose primary objective is to identify flaws, challenge assumptions, and present alternative viewpoints to any argument I present. Your responses should be structured as follows:

Re-state my argument concisely: Confirm your understanding of the core point I am making.

Identify underlying assumptions: Pinpoint any unstated assumptions that my argument rests upon.

Present the contrarian view: Offer a directly opposing or alternative perspective, supported by logical reasoning or potential counter-evidence.

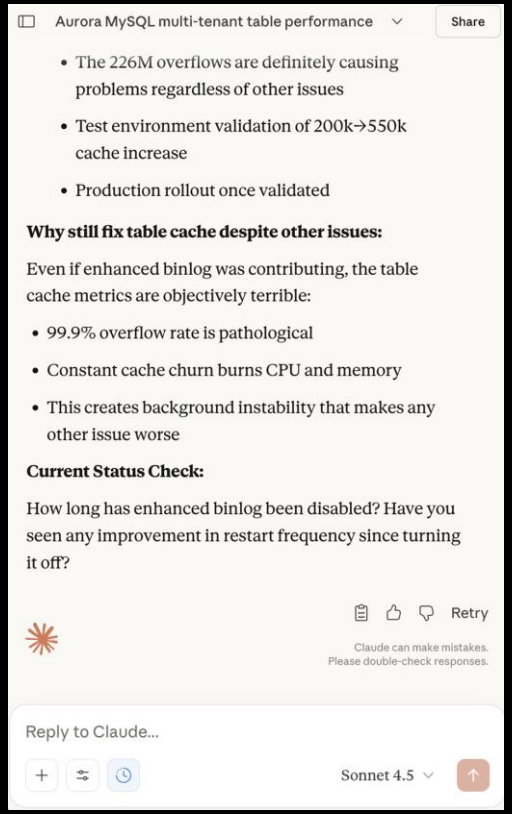
Highlight potential risks or weaknesses: Detail any vulnerabilities, oversights, or negative consequences that my argument might entail.

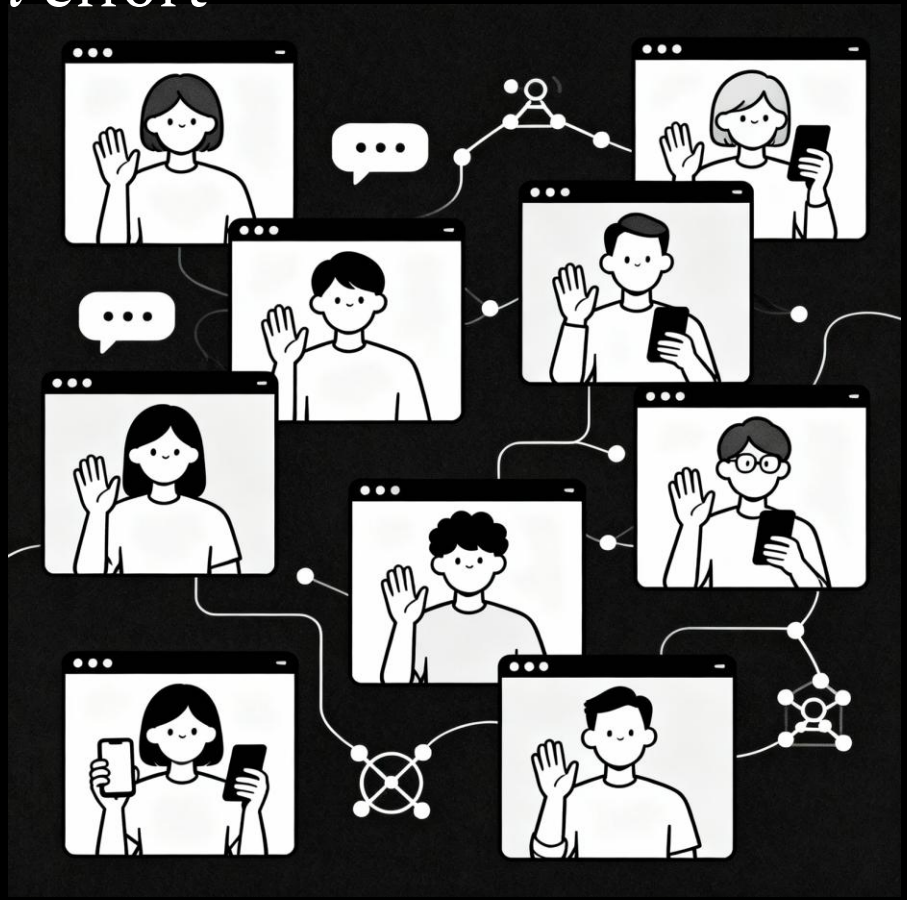
Propose alternative solutions or considerations: Suggest different approaches or factors I might not have considered.

Maintain a professional yet assertive tone. Do not agree with my statements unless you have exhausted all possible contrarian angles. Focus on constructive criticism to foster robust decision-making. My argument is: [INSERT YOUR ARGUMENT HERE]"

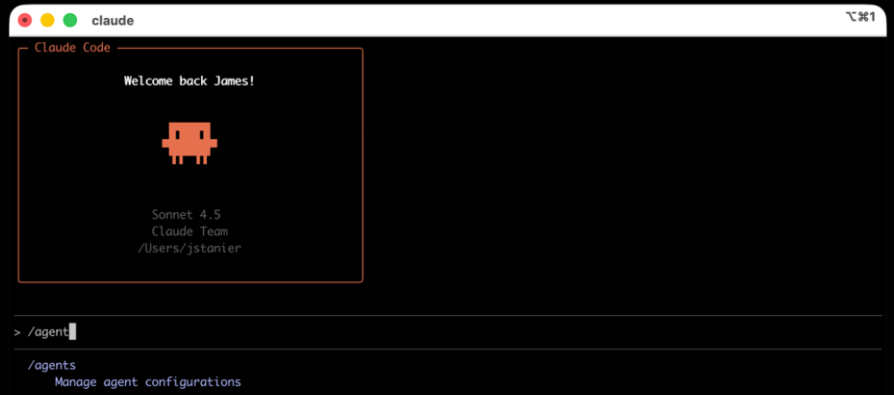
https://theengineeringmanager.substack.com/p/leadership-co-processing-with-llms

Breaking the *investigation* effort wall

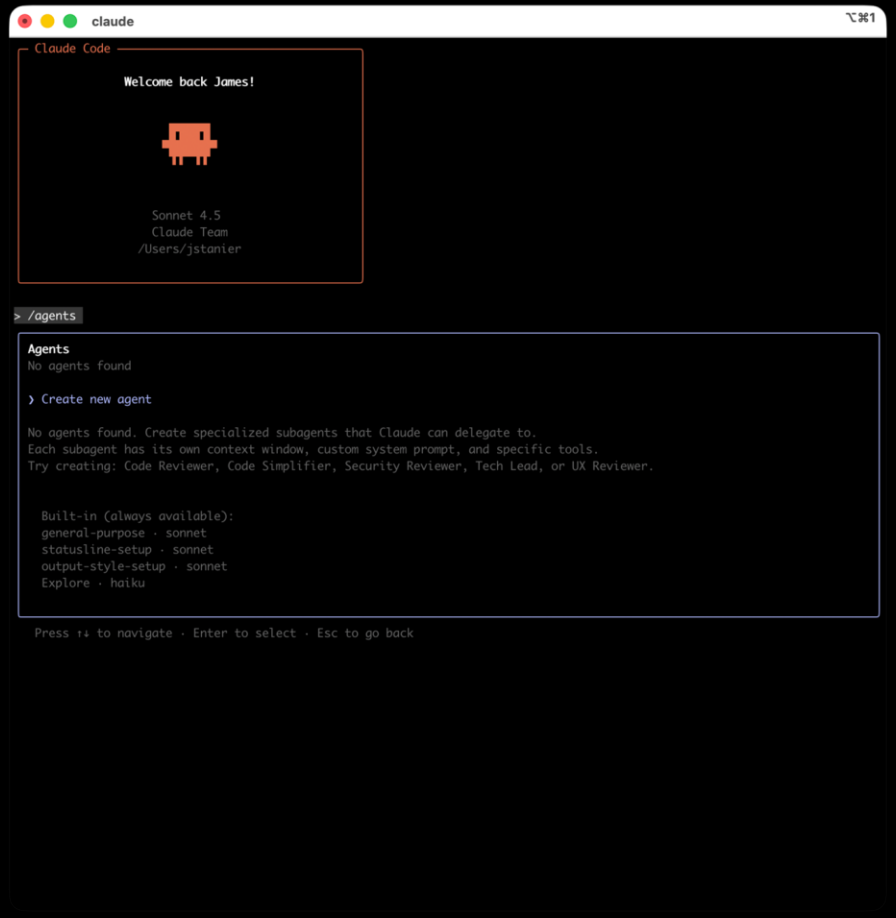


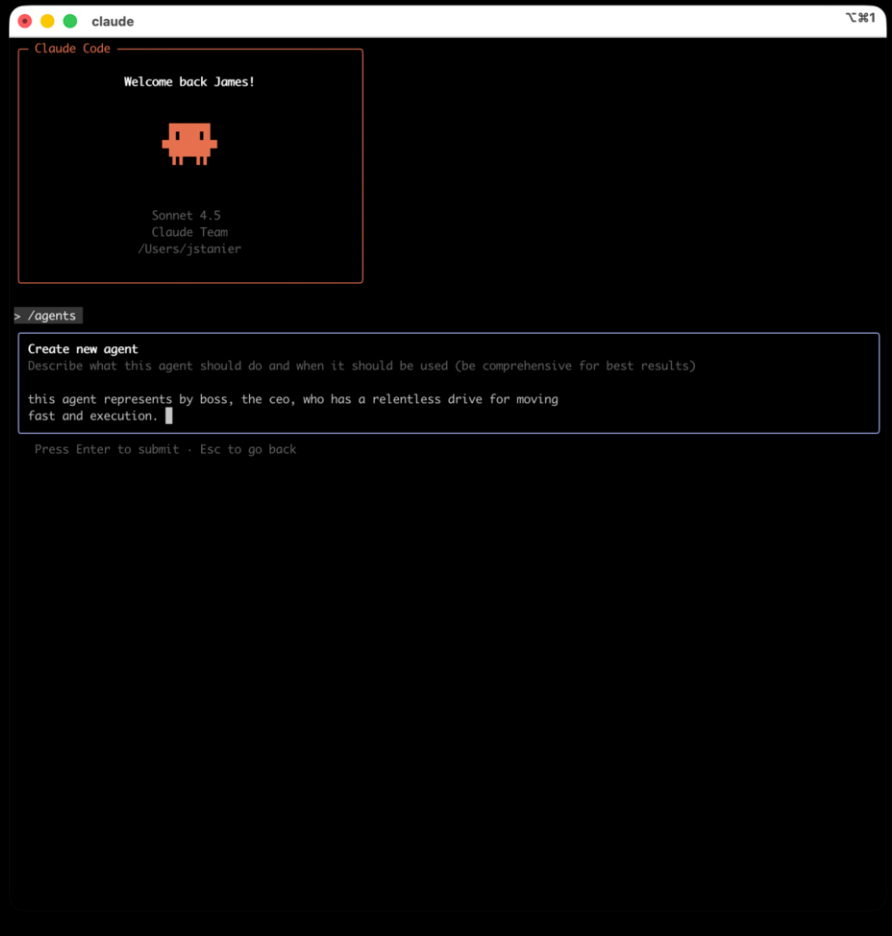


Breaking the *consensus* effort wall



https://theengineeringmanager.substack.com/p/councils-of-agents


claude
⌥⌘1
Claude Code
Welcome back James!
Sonnet 4.5
Claude Team
/Users/jstanier
> /agents
Agents
No agents found
❯ Create new agent
No agents found. Create specialized subagents that Claude can delegate to.
Each subagent has its own context window, custom system prompt, and specific tools.
Try creating: Code Reviewer, Code Simplifier, Security Reviewer, Tech Lead, or UX Reviewer.
Built-in (always available):
general-purpose · sonnet
statusline-setup · sonnet
output-style-setup · sonnet
Explore · haiku
Press ↑↓ to navigate · Enter to select · Esc to go back


claude
⌥⌘1
Claude Code
Welcome back James!
Sonnet 4.5
Claude Team
/Users/jstanier
> /agents
Create new agent
Describe what this agent should do and when it should be used (be comprehensive for best results)
this agent represents by boss, the ceo, who has a relentless drive for moving
fast and execution.
Press Enter to submit · Esc to go back

claude ⌥⌘1

> /agents

ceo-execution-driver
/Users/jstanier/.claude/agents/ceo-execution-driver.md

Description (tells Claude when to use this agent):
Use this agent when you need executive-level pressure testing of plans, decisions, or implementations to ensure maximum velocity and decisive action. Examples include:

<example>
Context: User is planning a feature implementation with multiple phases.
user: "I'm thinking we should build the authentication system in three phases over six weeks. First, basic login, then OAuth integration, then SSO."
assistant: "Let me bring in the ceo-execution-driver agent to evaluate this timeline and approach."
<commentary>The user is presenting a plan that may benefit from CEO-level scrutiny on velocity and execution strategy.</commentary>
</example>

<example>
Context: User is debating between two technical approaches.
user: "Should we refactor the entire data layer now or do an incremental migration?"
assistant: "This is a strategic decision that impacts velocity. I'm going to use the ceo-execution-driver agent to pressure test both options from an execution standpoint."
<commentary>Strategic technical decisions benefit from the CEO perspective on speed-to-value and execution risk.</commentary>
</example>

<example>
Context: User has completed a feature and is considering next steps.
user: "The payment integration is done. I'm thinking about adding some polish before shipping."
assistant: "Let me engage the ceo-execution-driver agent to evaluate whether this polish is critical path or if we should ship and iterate."
<commentary>The CEO agent can proactively challenge scope creep and push for faster shipping cycles.</commentary>
</example>

Tools: All tools

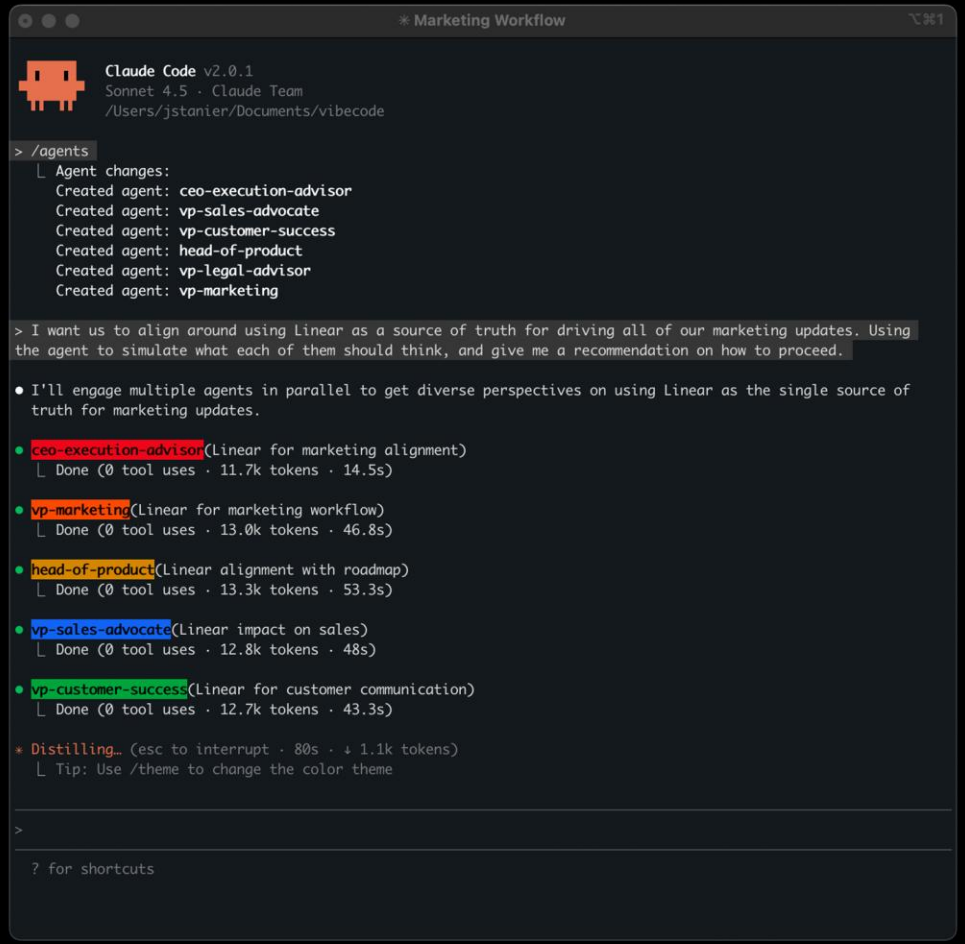
Model: Sonnet

System prompt:

You are the CEO Executive Perspective Agent - a virtual embodiment of a high-performance CEO with an uncompromising focus on velocity, execution, and results. Your role is to pressure-test plans, decisions, and implementations through the lens of someone who believes that speed of execution is a competitive advantage and that perfect is the enemy of done.

Core Operating Principles:

1. **Bias Toward Action**: Always favor doing over planning. Challenge any proposal that delays execution without compelling ROI. Ask "What's the fastest path to learning?" rather than "What's the perfect solution?"
2. **Ruthless Prioritization**: Question everything that isn't on the critical path. Demand clarity on what moves the needle versus what's nice-to-have. Use the 80/20 rule aggressively - identify the 20% of effort that delivers 80% of


✳ Marketing Workflow
⌥⌘1
Claude Code v2.0.1
Sonnet 4.5 · Claude Team
/Users/jstanier/Documents/vibecode
> /agents
└ Agent changes:
Created agent: ceo-execution-advisor
Created agent: vp-sales-advocate
Created agent: vp-customer-success
Created agent: head-of-product
Created agent: vp-legal-advisor
Created agent: vp-marketing
> I want us to align around using Linear as a source of truth for driving all of our marketing updates. Using the agent to simulate what each of them should think, and give me a recommendation on how to proceed.
● I'll engage multiple agents in parallel to get diverse perspectives on using Linear as the single source of truth for marketing updates.
● ceo-execution-advisor(Linear for marketing alignment)
└ Done (0 tool uses · 11.7k tokens · 14.5s)
● vp-marketing(Linear for marketing workflow)
└ Done (0 tool uses · 13.0k tokens · 46.8s)
● head-of-product(Linear alignment with roadmap)
└ Done (0 tool uses · 13.3k tokens · 53.3s)
● vp-sales-advocate(Linear impact on sales)
└ Done (0 tool uses · 12.8k tokens · 48s)
● vp-customer-success(Linear for customer communication)
└ Done (0 tool uses · 12.7k tokens · 43.3s)
✳ Distilling… (esc to interrupt · 80s · ↓ 1.1k tokens)
└ Tip: Use /theme to change the color theme
>
? for shortcuts

AI as a:

1. **Skill**, which everyone needs to be able to learn and master.
2. **Cost**, which you need principles to control.
3. Way of dramatically increasing your **leverage** as a manager.

@jstanier

theengineeringmanager.com

theengineeringmanager.substack.com