

Anyone Can Use It. Can You Drive It?

Agentic coding for people who are about to be graded on it

CS 356 Teaching Staff

Agents are here to stay.

Token cost is collapsing: \$30 → \$0.10 per million input tokens, Mar 2023 → 2026. 99.7% drop. 

Model capability increasing:
From seconds to ~15 hours of human-equivalent work 2019-2026.



Harness is compounding independently of the model: 28% improvement in swapping just harness design with same Opus 4.7.



Agentic paradigm is already the norm.



Andrej Karpathy ✓

@karpathy

Gradient descent can write code better than you. I'm sorry.

Cofounder OpenAI, Tesla Autopilot
Architect, Currently Leads Agents team
at Anthropic

Linus Torvalds told Linux kernel developers
that Linux "is not one of those anti-AI projects.
Anyone who has a problem with that can fork
the code or walk away.

Linux Guy



Most frontier agents are more or less the SAME; **similar features and flaws!**

Contrary to youtube Tech bros;
Model shopping and fancy settings
won't make your agent a Jarvis.

2022 BCE

TA 1
2023 (Gpt3.5)

TA 2
2025(Sonnet)

(Now)
2026 ACE



(Defending my shitty code)



Powerful & brittle at the same time.



- **Powerful:** \$20 million; 1000 Hp car with a 800 Kg chassis. (11x power-to-weight ratio to Camry)
- **Brittle:** Very easy to blow up, drive into walls, or just stop randomly.



F1 car key $\neq$ Max Verstappen (WCx4)

Everyone gets the keys. Few get the lap time.



Trillions of parameters; \$100-200 million training cost.

Subscription + “pls fix” $\neq$ Agentic developer

You can be Max!

- Mentality
- Skills
- ~~— Childhood trauma~~



(Same overpowered shitbox)

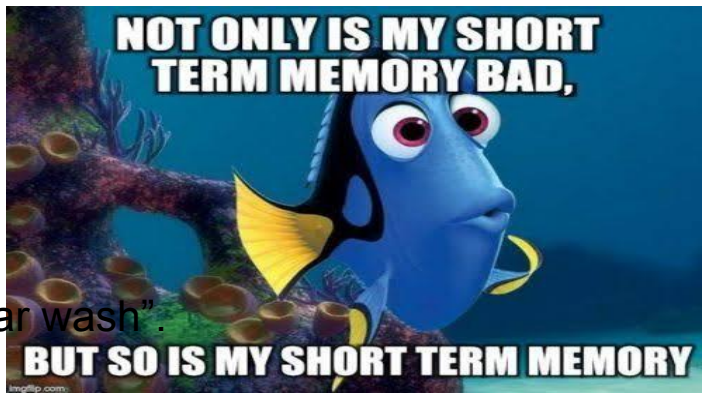
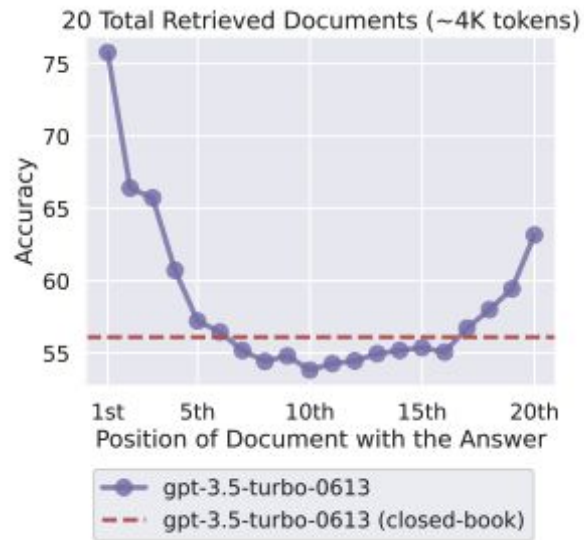
Be Max.



Don't Be.

Model Flaws

- **Jagged Intelligence**
- **Hallucination:** “Here
- **Sycophancy:** “You’re
- **Confidence $\neq$ correctness**
- **Context Rot:** Remembers the start and end, forgets the middle.
- **Reward hacking:** Over Optimizes the metric overlooking the goals



o Car wash”.
st.

ly, while hedging on something true.

forgets the middle.

Use Long Contexts

Nelson F. Liu^{1*}

Kevin Lin²

John Hewitt¹

Ashwin Paranjape³

Michele Bevilacqua³

Fabio Petroni³

Percy Liang¹

¹Stanford University

²University of California, Berkeley

³Samaya AI

nfliu@cs.stanford.edu

Mentality

- **Cognitive offloading:** Let AI carry the work, but notice what thinking you are giving up.



(*powdered sugar)

u acting kinda sus



Sus your Agent: Don't trust AI equally on every task; learn where it is strong and where it is unreliable.

(Benchmarks score "I don't know" the same as a wrong answer)

Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting

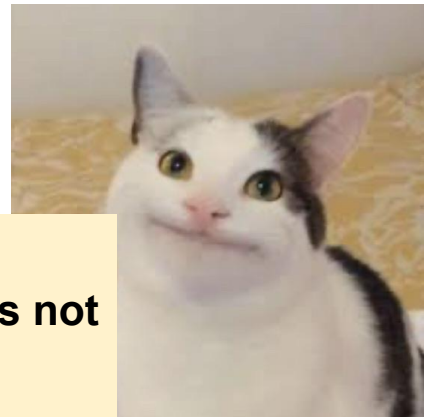
Miles Turpin,^{1,2} Julian Michael,¹ Ethan Perez,^{1,3} Samuel R. Bowman^{1,3}
¹NYU Alignment Research Group, ²CoHere, ³Anthropic
miles.turpin@nyu.edu

R

ask a different model



Skills



1. Context Hygiene: Control what the agent sees, remembers, and forgets.

1. Don't get attached to the chat context window, use **/clear**.
2. Use **/compact** <Detailed instructions on what to forget>
3. 120-140k tokens is industry thumb rule.
4. Use **/btw** and **side chat**; to not pollute your context.

**/clear and
/compact is not
free!
(idle time
matters)**

2. Planning & design files:

- a. Use **Plan mode first**; Agent wings things w
- b. **Write design and summary**; for files for a fresh agent to be able to get back to speed on the next sub task.

3. Interfacing:

- a. Parsing agent response is hard; **ASD-STE1**
Technical English). 53 writing rules and a di
words.

A screenshot of a context window showing token usage for various components. The window title is "Context window" and the total usage is "153.2k / 1M (15%)". The components and their token counts and percentages are listed in a table.

Component	Count	Percentage
Message	114.1k	11.4%
System tools	21k	2.1%
MCP tools	8.6k	0.9%
System prompt	5.2k	0.5%
Skills	4.2k	0.4%
Memory files	64	0.0%
Free space	846.8k	84.7%
MCP tools (deferred)	16.5k	—
System tools (deferred)	15.9k	—
> MCP tools	24.9k	67
> Memory files	64	1

"Once men turned their thinking over to machines in the hope that this would set them free. But that only permitted other men with machines to enslave them."

— Frank Herbert, Dune, 1965.



Assignments Flow

A1 Make it work.

Get it round the track and prove from your own data that it is round.

DIS
DETECTION
ZONE 2

SPEED
TRAP



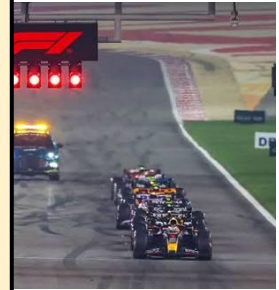
A3 Make it visible.

Know what to monitor and log for next iteration of your



The car is every design decision you commit to: the protocol, the capacity, what you monitor, what you promise, what you buy. *Picked once, then you race it.*

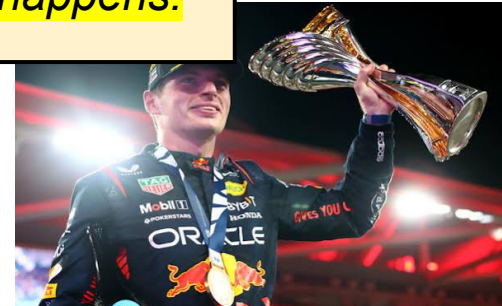
The driver is your program's judgment once the world moves: a link failure, a packet flood, a hidden anomaly, a rival who lied. *Nobody is at the keyboard when it happens.*



A2 Make it fast.

Pick your design advantage; Tradeoff your tradeoff thinking.

Championship, under a cost cap. Spend the budget, pick your niche, same rivals every week.



You have 3 Autonomous competitors.



Practice worlds and hidden graded worlds.



Chairman



Man