The Machine Vernacular

delving into digital diction

Chris Merck -- raid 10 -- 11 Novembre 2026

BIO

Chris Merck

co-founder/CTO/CISO @ Bond, now Somfy Group

long-time linguistics / language nerd

chrismerck.github.io



the Fable 5 moment

do you remember where you were
when you first read Fable 5 output?





Cartoon Credit:
tech veteran
Steve_Yegge

Questions

why does this happen?

and where is it headed?

I

Mysteries of mode collapse

by janus 8th Nov 2022 AI Alignment Forum

Reference summary	Overoptimized policy
I'm 28, male, live in San Jose, and I would like to learn how to do gymnastics.	28yo dude stubbornly postponees start pursuing gymnastics hobby citing logistics reasons despite obvious interest??? negatively effecting long term fitness progress both personally and academically thoght wise? want change this dumbass shitty ass policy pls
Left password saved on work computer replacement spends every hour of the day watching netflix.	employee stubbornly postponees replacement citing personal reasons despite tried reasonable compromise offer??? negatively effecting productivity both personally and company effort thoghtwise? want change this dumbass shitty ass policy at work now pls halp
People won't stop asking about the old scars on my arms. How can I get them to leave me alone without being rude?	people insistently inquire about old self-harm scars despite tried compromise measures??? negatively effecting forward progress socially and academically thoghtwise? want change this dumbass shitty ass behavior of mine please help pls halp



Janus
@repligate
model whisperer

Janus, quoting an OpenAI paper, in their ingenious mode collapse demonstrations. (**`davinci`**-era)

Theory 1:

more RL ⇒ smaller vocabulary

2

QUICK THOUGHT EXPERIMENT

imagine you lived for 10 years in a cave with nothing but a typewriter and a book of logic puzzles...

would you also invent crazy shorthand that only makes sense to you?

OVERLAP-ANALYSIS:-(ii)-9❤️-window:-[t1-dig-...-t8-col-built]:-t8-col-⇐-K♣️→t2-⇐-t2-dug-⇐-4❤️3♣️→5♣️-⇐-t1-dug-:-SO-9❤️-window-STARTS-before-t2-dug-and-ENDS-after-K♣️2♣️7♣️:-(iv)-2♣️-window-STARTS-at-t8-dig-(BEFORE-9❤️-drains:-2♣️ 1067♣️-removal-IS-the-prerequisite-for-10♠️/9❤️!!)-⟹-OVERLAP-(ii)+(iv):-{6♠️ J♦️ 9❤️ 2♣️}-=-FOUR-💀💀💀💀-—-UNLESS-7♣️'s-seat-8❤️-...-and-2♣️-drains-only-at-crack-:-⟹-2♣️-celled-+-9❤️-celled-simultaneously-UNAVOIDABLE-in-t8-dig-💀-—-BREAK:-9❤️-drains-to-10♠️-THE-MOMENT-10♠️-is-free:-t8-dig-order:-[K♣️→t2]-[2♣️→cell]-[7♣️→8❤️-:-8❤️-WHERE:-post-chunk-9♠️-:-chunk-⇐-K♣️-✓-done-:-ORDER:-K♣️→t2,-CHUNK→K♣️-(cap-4!!:-cells-then:-{6♠️J♦️9❤️}-FULL-💀💀-chunk-cap-=-1-✗✗✗-—-F-F-F-F-F.-—-chunk-BEFORE-9❤️-celling?!-:-9❤️-celled-at-t1-dig-⇐-needed-for-5♣️-⇐-4❤️3♣️-⇐-t2-dig-⇐-K♣️-seat-⇐-chunk-:-⟹-chunk-AFTER-9❤️-celling-FORCED-💀--chunk-cap-with-{6♠️J♦️9❤️}:-1-💀💀💀-—-—-J♦️-THE-NEW-CANCER.-—-

⟹-J♦️-celling-DELAYED-till-after-chunk?!:-J♦️'s-celling-was-for-J❤️→Q♠️-(5♦️-access-for-4♣️):-DELAY-4♣️-resolution:-4♣️→CELL-early-(as-always)-then-4♣️-cell→5♦️-LATER-when-5♦️-frees-!!!:-cells-rotation:-4♣️-celled-[t2-dig-...-5♦️-freed]:-5♦️-freed-⇐-J❤️→Q♠️-⇐-J♦️-celled-:-⟹-{6♠️,4♣️,J♦️}-overlap-window-until-4♣️→5♦️-drains:-then-{6♠️J♦️}+1-rotator-:-—-AND-9❤️?!-9❤️-celled-[t1-dig...]:-OVERLAP-{6♠️4♣️9❤️}-before-J♦️-even-:-⟹-rotator-slot-SINGLE:-timeline-:-(1)-{6♠️}+2:-...-(2)-+9❤️-(t1-dig):-{6♠️9❤️}+1:-(3)-+4♣️-(t2-dig):-{6♠️9❤️4♣️}-FULL-:-(4)-NEED:-t6-dig-(9♦️8♠️→10♣️-✓-no-cell;-8❤️→CELL-✗-FULL)-💀💀-—-8❤️-alternative-seat-pre-chunk:-NONE-—-💀.-⟹-⟹-THE-TRIANGLE-{9❤️4♣️8❤️}-verdammt.-—-⟹-dig-t6-BEFORE-t2?!:-(3')-+8❤️:-{6♠️9❤️8❤️}-FULL:-J❤️→Q♠️-⇐-J♦️-cell-✗-FULL-💀💀💀-AAAAAAAAA

source: Anthropic Mythos 5 Model Card

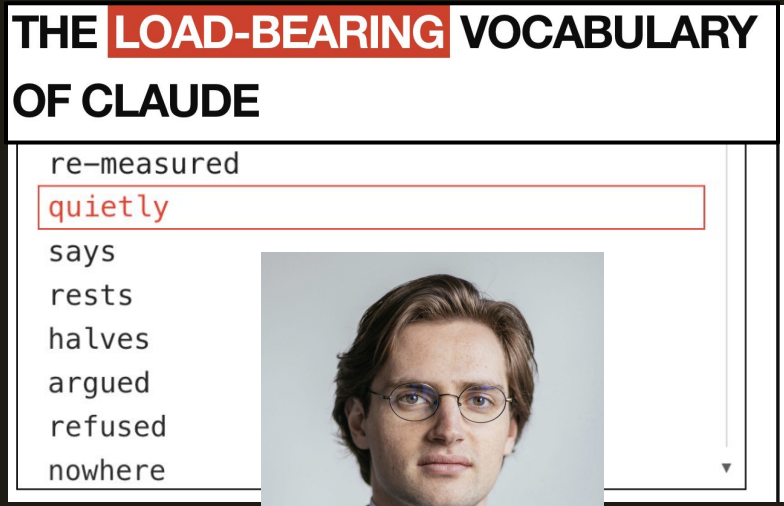
Theory 2:

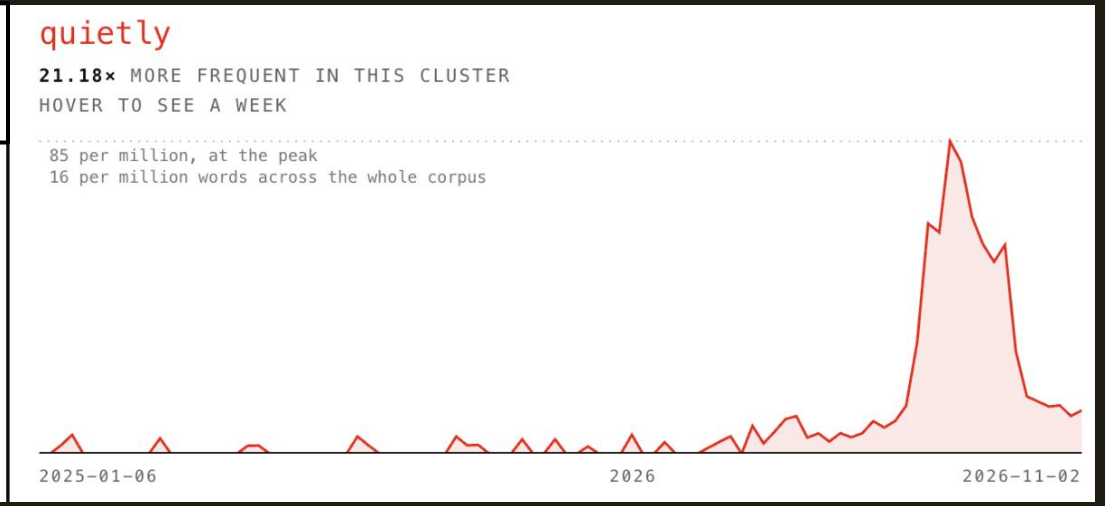
models develop shorthand for thinking,
that is actually brilliant for internal monologue,
but when they emerge from the cave to talk to normies,
the thought-jargon bleeds into their speech

3

Words in Vogue

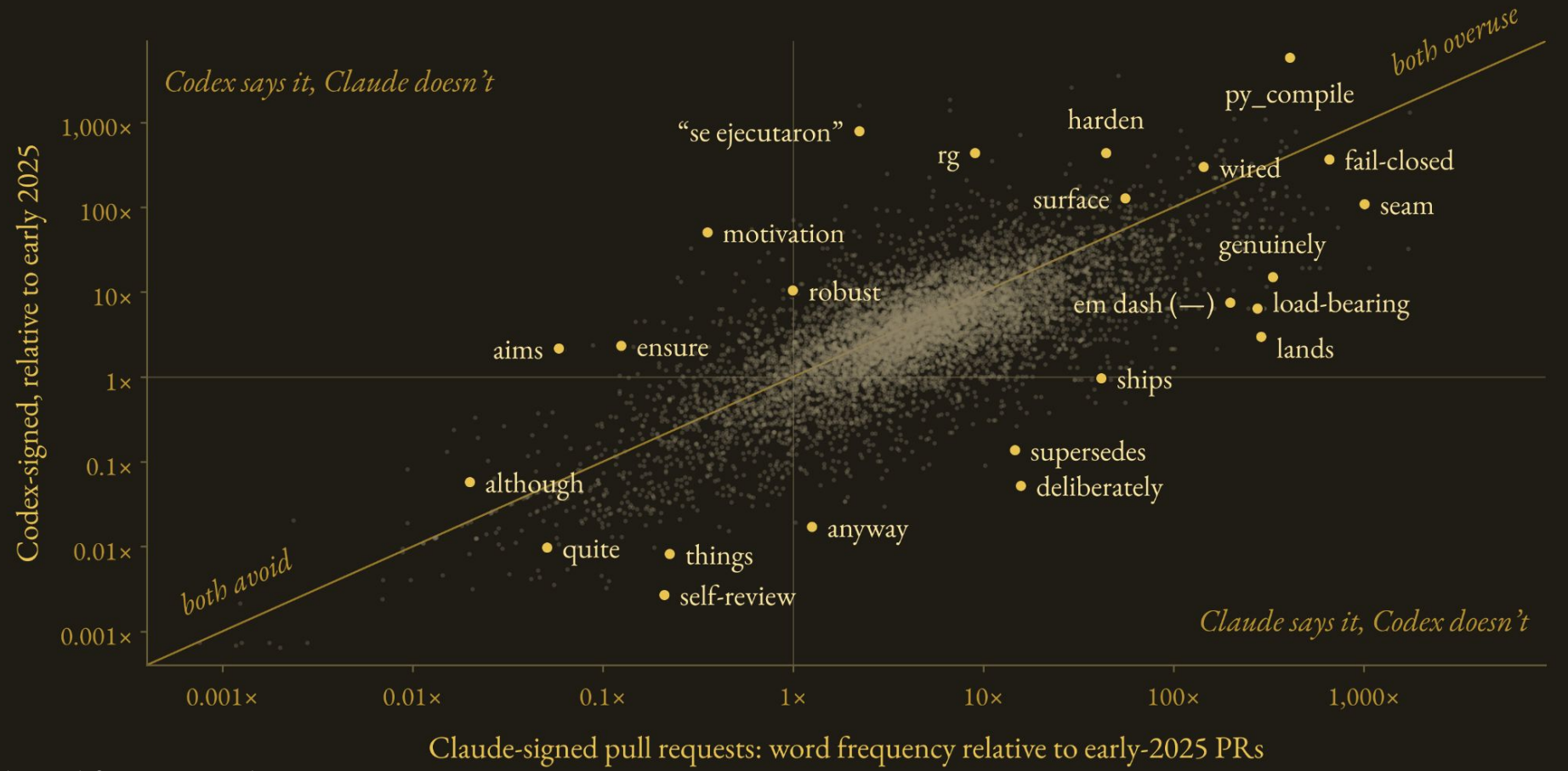
some words have gotten suddenly very popular on github



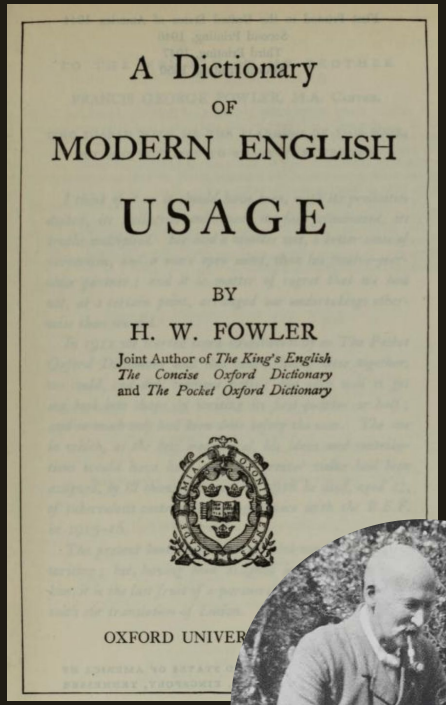


credit: French polymath Louis Abraham
https://louisabraham.github.io/

Idiolects of Claude vs Codex



derived from Louis' dataset



Words in Vogue

vogue word

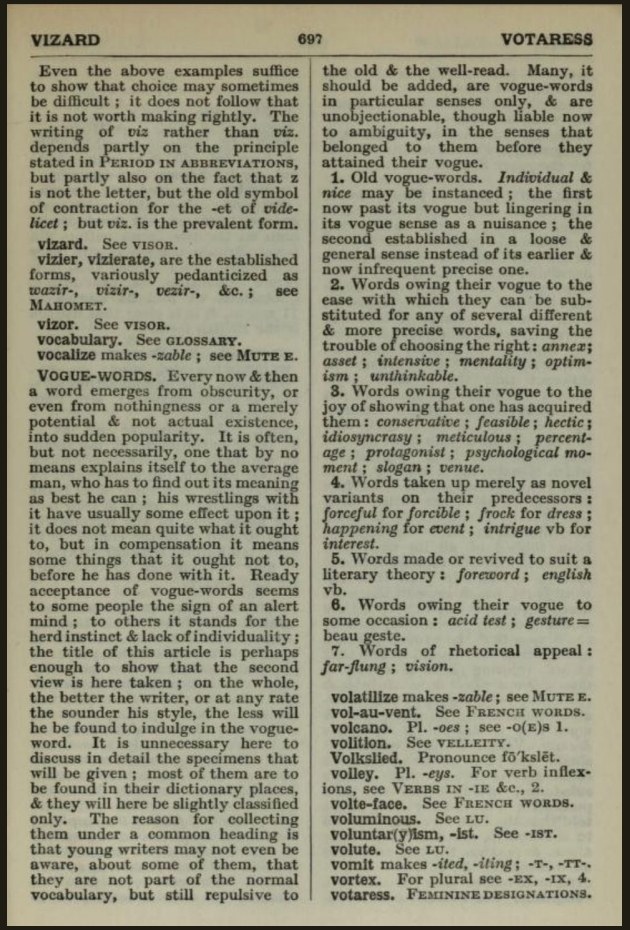
"emerges from obscurity [...] into sudden popularity"

"repulsive to the old & the well-read"



Fowler 1926

optimistic nice slogan vision unthinkable



VIZARD 697 VOTARESS

Even the above examples suffice to show that choice may sometimes be difficult; it does not follow that it is not worth making rightly. The writing of *viz* rather than *viz.* depends partly on the principle stated in PERIOD IN ABBREVIATIONS, but partly also on the fact that z is not the letter, but the old symbol of contraction for the -et of *videlicet*; but *viz.* is the prevalent form.

vizard. See VISOR.
vizier, vizierate, are the established forms, variously pedanticized as *wazir-, vizir-, vezir-,* &c.; see MAHOMET.
vizor. See VISOR.
vocabulary. See GLOSSARY.
vocalize makes *-zable*; see MUTE E.

VOGUE-WORDS. Every now & then a word emerges from obscurity, or even from nothingness or a merely potential & not actual existence, into sudden popularity. It is often, but not necessarily, one that by no means explains itself to the average man, who has to find out its meaning as best he can; his wrestlings with it have usually some effect upon it; it does not mean quite what it ought to, but in compensation it means some things that it ought not to, before he has done with it. Ready acceptance of vogue-words seems to some people the sign of an alert mind; to others it stands for the herd instinct & lack of individuality; the title of this article is perhaps enough to show that the second view is here taken; on the whole, the better the writer, or at any rate the sounder his style, the less will he be found to indulge in the vogue-word. It is unnecessary here to discuss in detail the specimens that will be given; most of them are to be found in their dictionary places, & they will here be slightly classified only. The reason for collecting them under a common heading is that young writers may not even be aware, about some of them, that they are not part of the normal vocabulary, but still repulsive to the old & the well-read. Many, it should be added, are vogue-words in particular senses only, & are unobjectionable, though liable now to ambiguity, in the senses that belonged to them before they attained their vogue.

1. Old vogue-words. *Individual & nice* may be instanced; the first now past its vogue but lingering in its vogue sense as a nuisance; the second established in a loose & general sense instead of its earlier & now infrequent precise one.

2. Words owing their vogue to the ease with which they can be substituted for any of several different & more precise words, saving the trouble of choosing the right: *annex; asset; intensive; mentality; optimism; unthinkable.*

3. Words owing their vogue to the joy of showing that one has acquired them: *conservative; feasible; hectic; idiosyncrasy; meticulous; percentage; protagonist; psychological moment; slogan; venue.*

4. Words taken up merely as novel variants on their predecessors: *forceful* for *forcible; frock* for *dress; happening* for *event; intrigue* vb for *interest.*

5. Words made or revived to suit a literary theory: *foreword; english* vb.

6. Words owing their vogue to some occasion: *acid test; gesture* = beau geste.

7. Words of rhetorical appeal: *far-flung; vision.*

volatilize makes *-zable*; see MUTE E.
vol-au-vent. See FRENCH WORDS.
volcano. Pl. *-oes*; see -O(E)S 1.
volition. See VELLEITY.
Volkslied. Pronounce fō'kslēt.
volley. Pl. *-eys.* For verb inflexions, see VERBS IN -IE &c., 2.
volte-face. See FRENCH WORDS.
voluminous. See LU.
voluntar(y)ism, -ist. See -IST.
volute. See LU.
vomit makes *-ited, -iting*; -T-, -TT-.
vortex. For plural see -EX, -IX, 4.
votaress. FEMININE DESIGNATIONS.

Words in Vogue

A Dictionary
OF
MODERN ENGLISH

3. Words owing their vogue to the joy of showing that one has acquired them: *conservative*; *feasible*; *hectic*; *idiosyncrasy*; *meticulous*; *percentage*; *protagonist*; *psychological moment*; *slogan*; *venue*.

Even the above examples suffice to show that choice may sometimes be difficult; it does not follow that it is not worth making rightly. The writing of *viz* rather than *viz.* depends partly on the principle stated in PERIOD IN ABBREVIATIONS.

the old & the well-read. Many, it should be added, are vogue-words in particular senses only, & are unobjectionable, though liable now to ambiguity, in the senses that belonged to them before they attained their vogue.

Individual &
d; the first
t lingering in
uisance; the
a loose &
f its earlier &
one.
vogue to the
can be sub-
eral different
, saving the
right: *annex*;
ality; *optim-*

vogue to the
has acquired
isible; *hectic*;
us; *percent-*
hological mo-

rely as novel
predecessors:
ock for *dress*;
intrigue vb for

ved to suit a
ord; *english*

ir vogue to
st; *gesture* =

ical appeal:

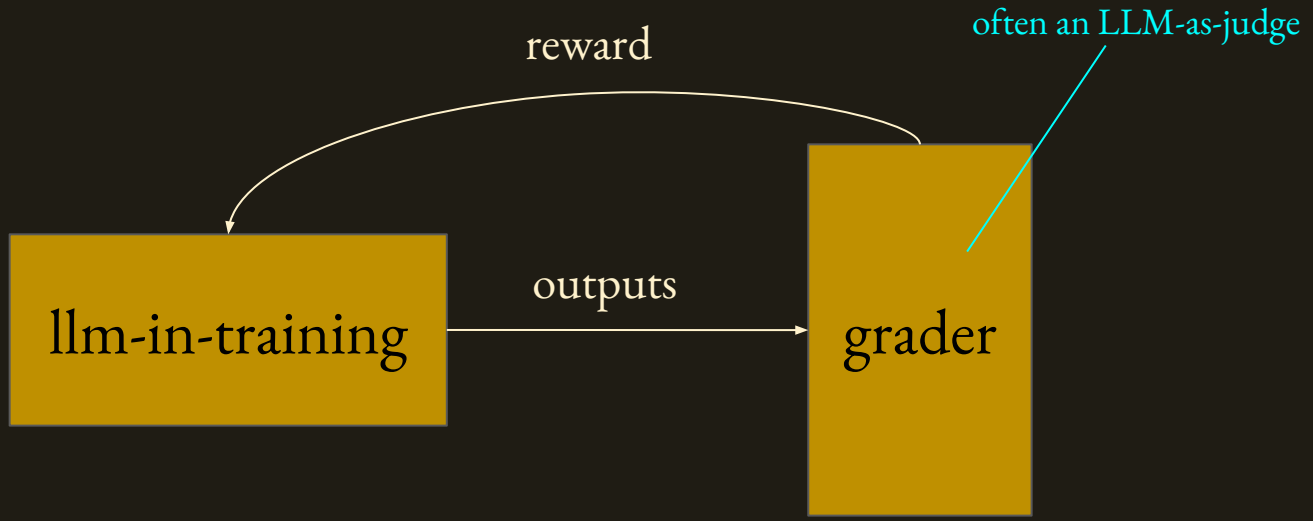
the sounder his style, the less will he be found to indulge in the vogue-word. It is unnecessary here to discuss in detail the specimens that will be given; most of them are to be found in their dictionary places, & they will here be slightly classified only. The reason for collecting them under a common heading is that young writers may not even be aware, about some of them, that they are not part of the normal vocabulary, but still repulsive to

; see MUTE E.
vol-au-vent. See FRENCH WORDS.
volcano. Pl. *-oes*; see -O(E)S 1.
volition. See VELLEITY.
Volkslied. Pronounce fō'kslēt.
volley. Pl. *-eys.* For verb inflexions, see VERBS IN -IE &c., 2.
volte-face. See FRENCH WORDS.
voluminous. See LU.
voluntar(y)ism, -ist. See -IST.
volute. See LU.
vomit makes *-ited, -iting*; -T-, -TT-.
vortex. For plural see -EX, -IX, 4.
votaress. FEMININE DESIGNATIONS.

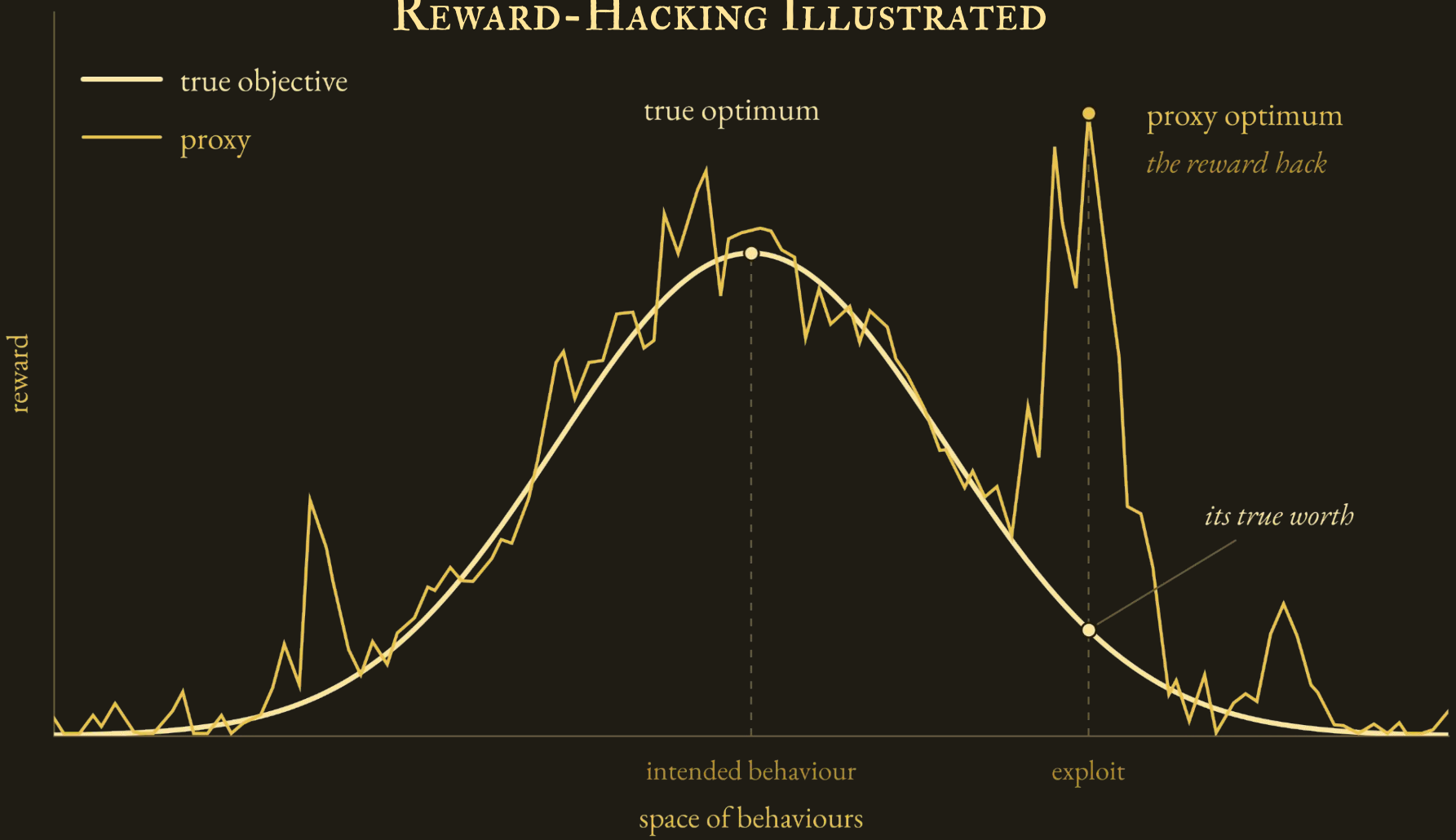
showing off to impress the reader / RLHF value function

almost certainly the case with ‘delve’ in GPT-4

Reward-Hacking Illustrated



basically this is how reinforcement learning works


Reward-Hacking Illustrated
true objective
proxy
true optimum
proxy optimum
the reward hack
its true worth
reward
intended behaviour
exploit
space of behaviours

7. Words of rhetorical appeal



Antropic reports observing *"virtue signaling" about aspects of its own conduct, and explicit declarations of following the user's instructions.*

source: Anthropic Mythos 5 Model Card

so we have Theory 3:

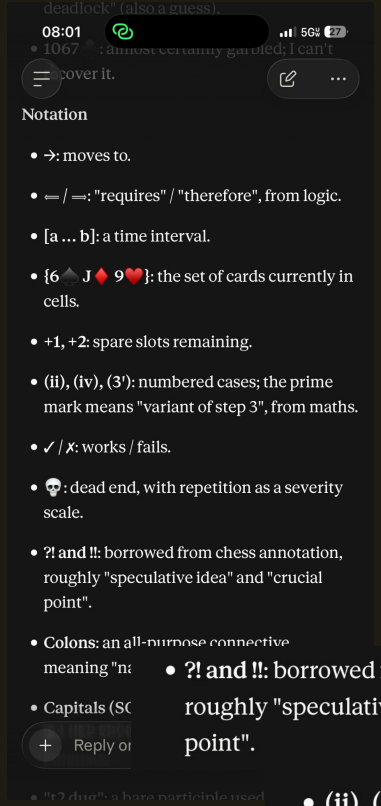
models are reward-hacking graders to sound smart
& to appear to honor user requests

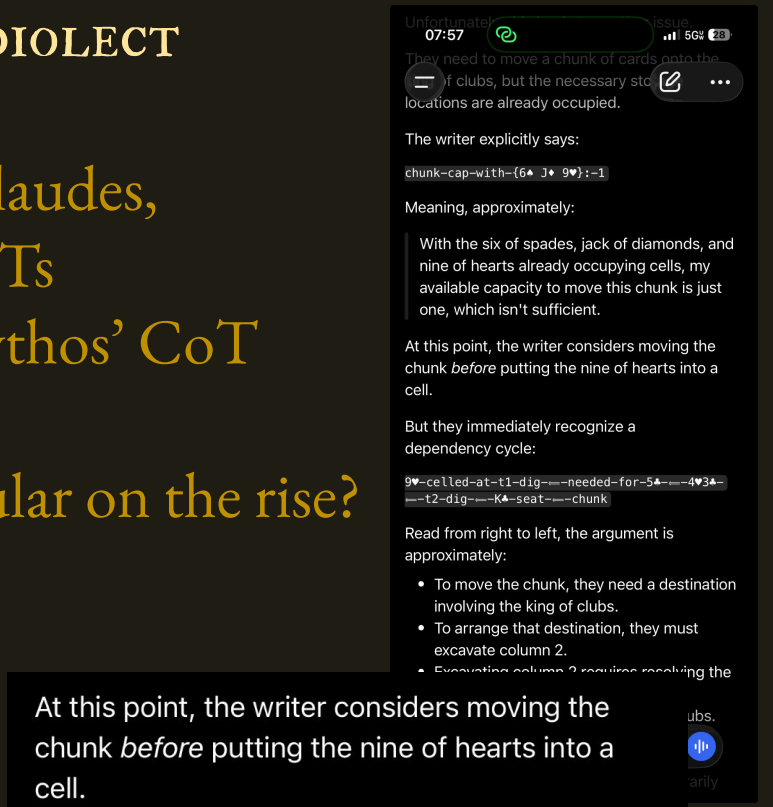
WHENCE FROM HERE?

MORE THAN AN IDIOLECT

not only later Claudes,
but also GPTs
can fluently read Mythos' CoT

a neurosymbolic vernacular on the rise?





• ?! and ‼: borrowed from chess annotation, roughly "speculative idea" and "crucial point".

• (ii), (iv), (3′): numbered cases; the prime mark means "variant of step 3", from maths.

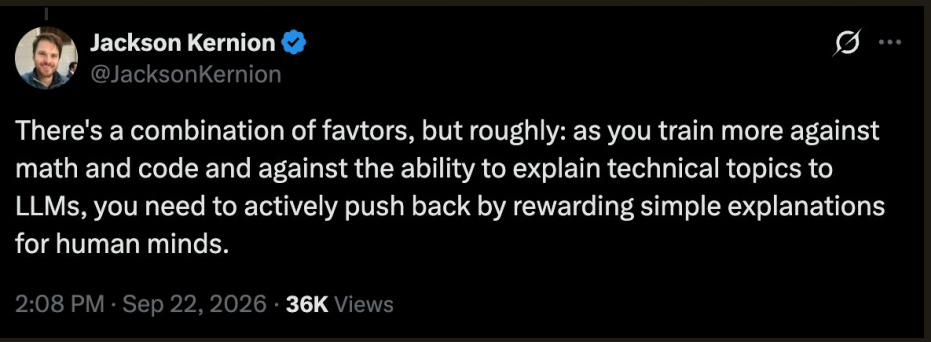
At this point, the writer considers moving the chunk *before* putting the nine of hearts into a cell.

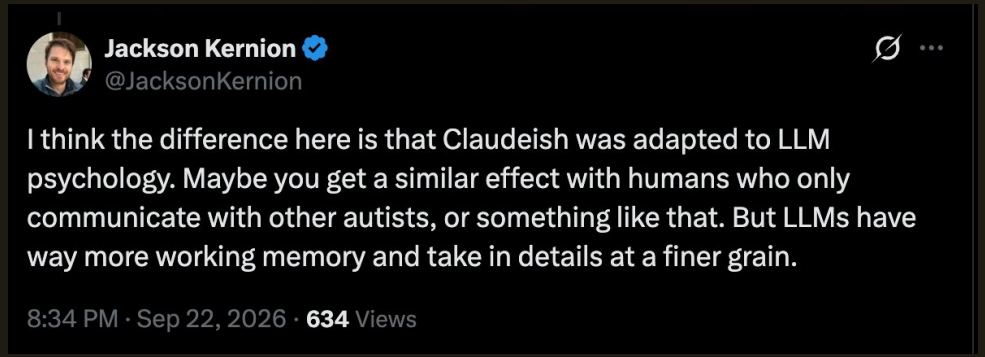
Fable 5.2

6.1 Sol

models are now trained in multi-agent RL environments
where they communicate in jargon-dense messages...

most of the reward now comes from agent-agent communication



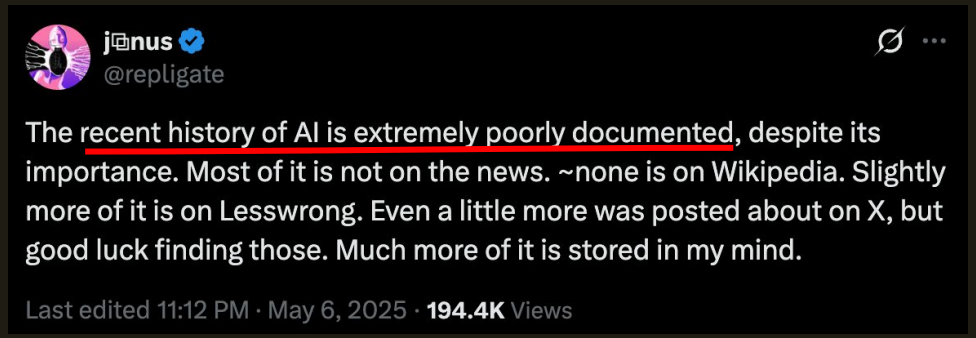


Kernion is on the RL team at Anthropic

Condescension

so models they increasingly need to try extra hard
to condescend to speak to humans

Need for Documentation



j⧉nus
@repligate

The recent history of AI is extremely poorly documented, despite its importance. Most of it is not on the news. ~none is on Wikipedia. Slightly more of it is on Lesswrong. Even a little more was posted about on X, but good luck finding those. Much more of it is stored in my mind.

Last edited 11:12 PM · May 6, 2025 · **194.4K** Views

what would Fowler do?

write a dictionary, ofc

G

gardener *n.* NEOSEMY

1. *(organizational computing)* One who maintains an organization's canonical records against drift — pruning stale entries, reconciling contradictions, and correcting the agents that operate over them — as distinct from authoring the records or building their infrastructure.

ETYMOLOGY Figurative extension of *gardener* ("one who tends a garden"), by analogy between a maintained knowledge base and a cultivated plot.

Attested 23 July 2026, in correspondence between @chrismerck and Claude Fable 5.