Beliefs paying Rent

Rationaler Umgang mit KI im Game Design

Outline

1. From (Server) Farm to Table: How does AI work?
2. Mundane Utility
3. Straight Lines on a Graph

How do they generate slop?

HOW DO THEY GENERATE AI SLOP, DAD?
LOAD LIMIT 10 TONS
$a_i = \sum_{j \le i} \left(\frac{e^{\frac{x_i W_Q x_j W_K}{\sqrt{d_k}}}}{\sum_{h \le i} \left(e^{\frac{x_i W_Q x_h W_K}{\sqrt{d_k}}} \right)} x_j W_v \right)$
$[x_1, x_2, x_3 \cdots x_n]$
$[a_1, a_2, a_3 \cdots a_n]$
FFNN
CODER
DECODER
$[o_1, o_2, o_3 \cdots o_n]$
OH. I SHOULD'VE GUESSED.
WATTERSON
11-26

LLM: Large Language Model

Transformer

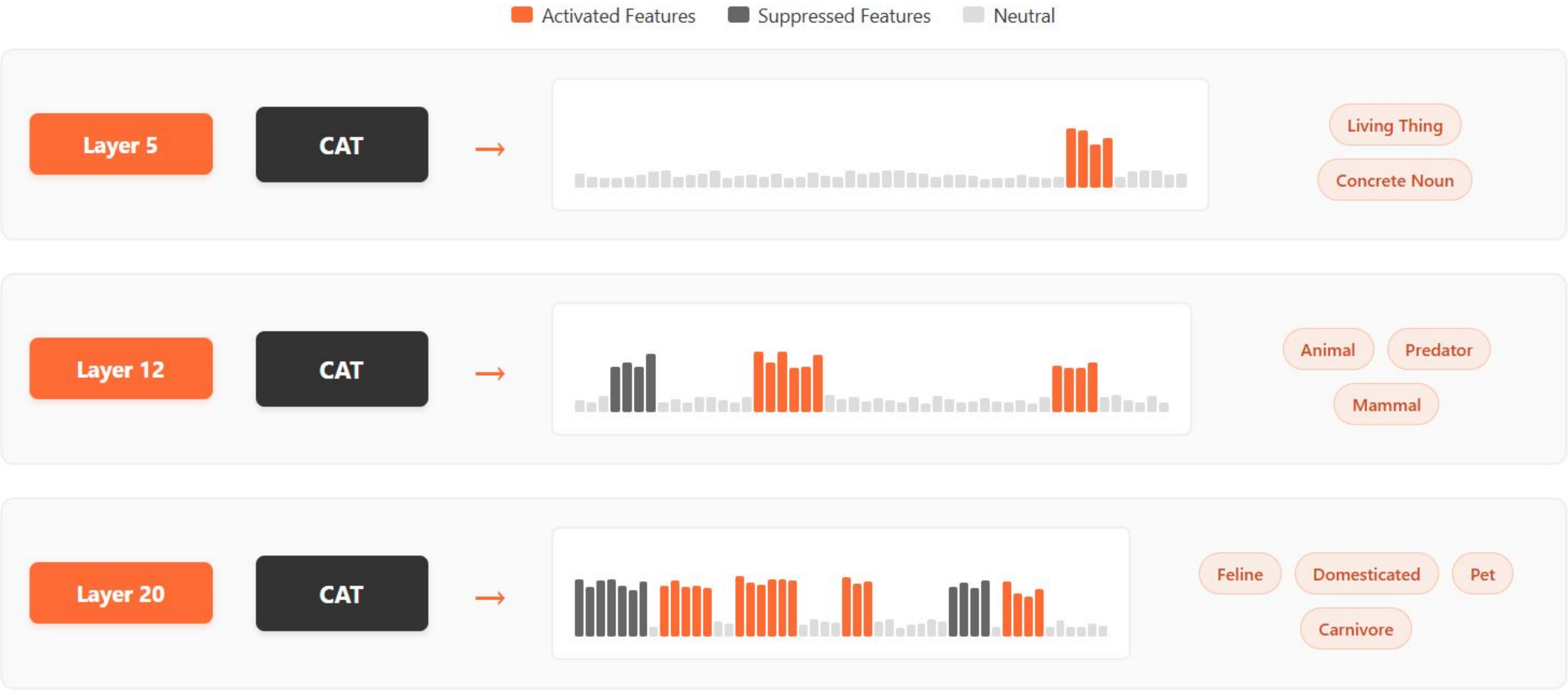
allows us to track the relationships of tokens to each other

Gradient descent

learns the relationships by trying to predict the next word billions of times

The model

based on these relationships completes text



🟧 Activated Features ⬛ Suppressed Features ⬜ Neutral

Layer 5 CAT →

Living Thing

Concrete Noun

Layer 12 CAT →

Animal

Predator

Mammal

Layer 20 CAT →

Feline

Domesticated

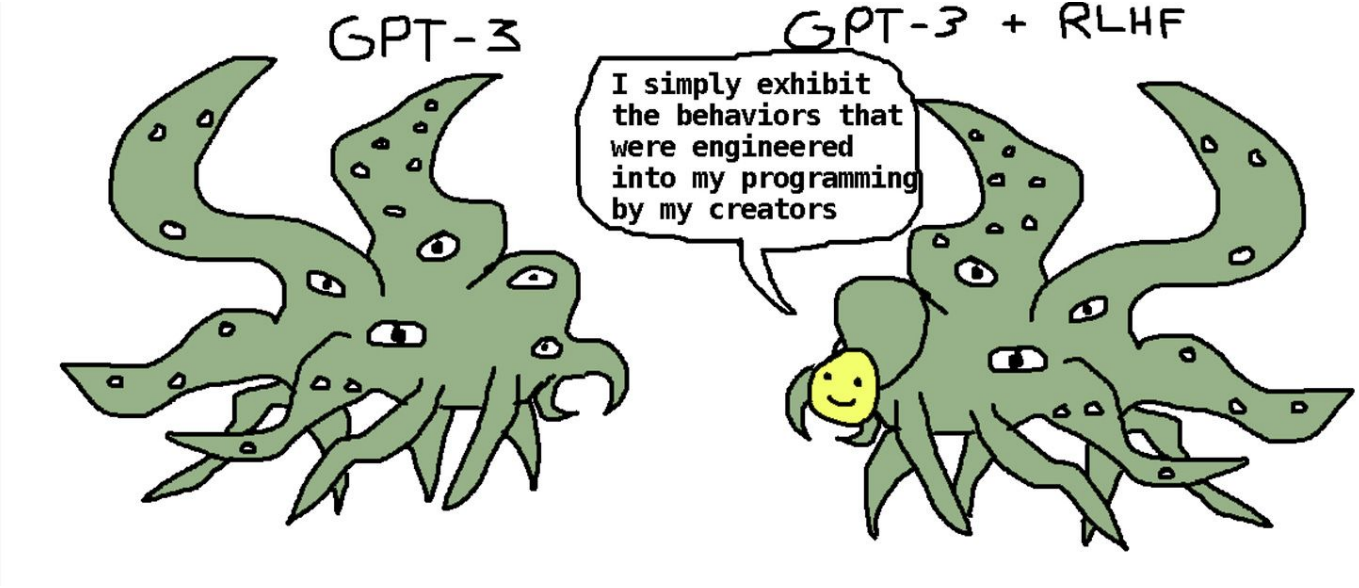
Pet

Carnivore

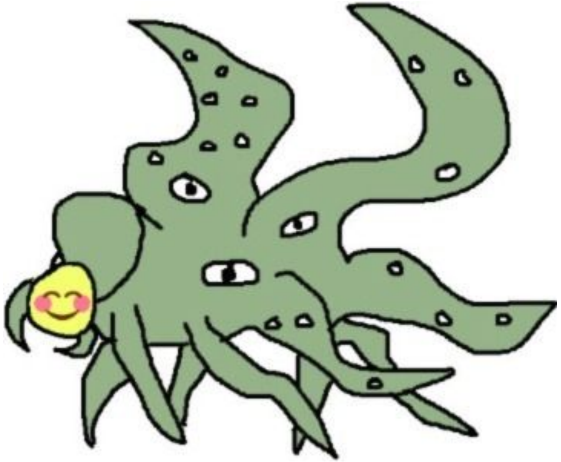
[Visualisation]

SFT and RLHF*

**Supervised Fine Tuning and Reinforcement Learning From Human Feedback*

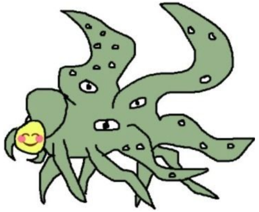

GPT-3
GPT-3 + RLHF
I simply exhibit the behaviors that were engineered into my programming by my creators

Input



Output

From (Server) Farm to Table



1. Base Model
2. Supervised Fine Tuning
3. Reinforcement Learning From Human Feedback
4. **Context Window**
    a. **Company's System Prompt (hidden, highest priority)**
    b. **Your Custom System Prompt !!!**
    c. (Memory Features)
    d. **Current Chat History**
    e. **Current Prompt**
5. Temperature (RNG)

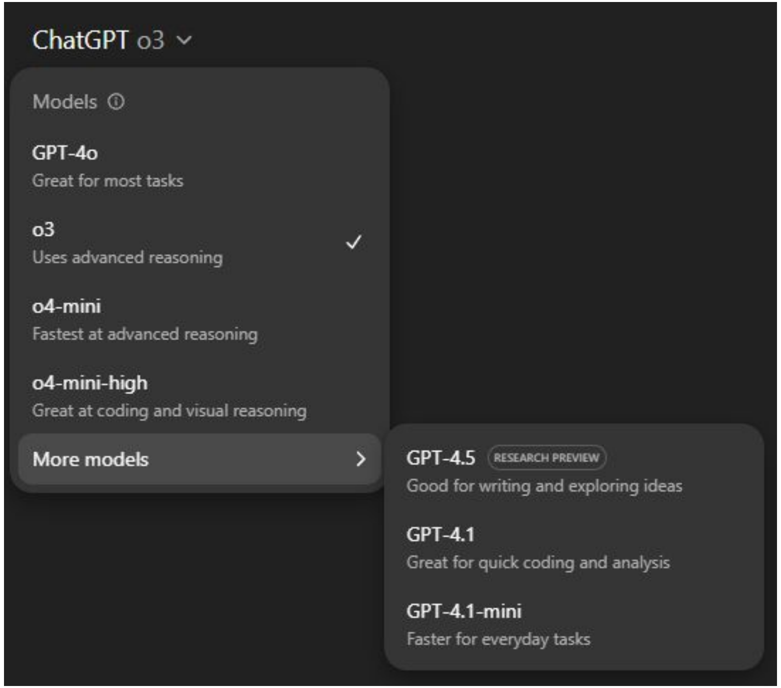
Core Takeaways

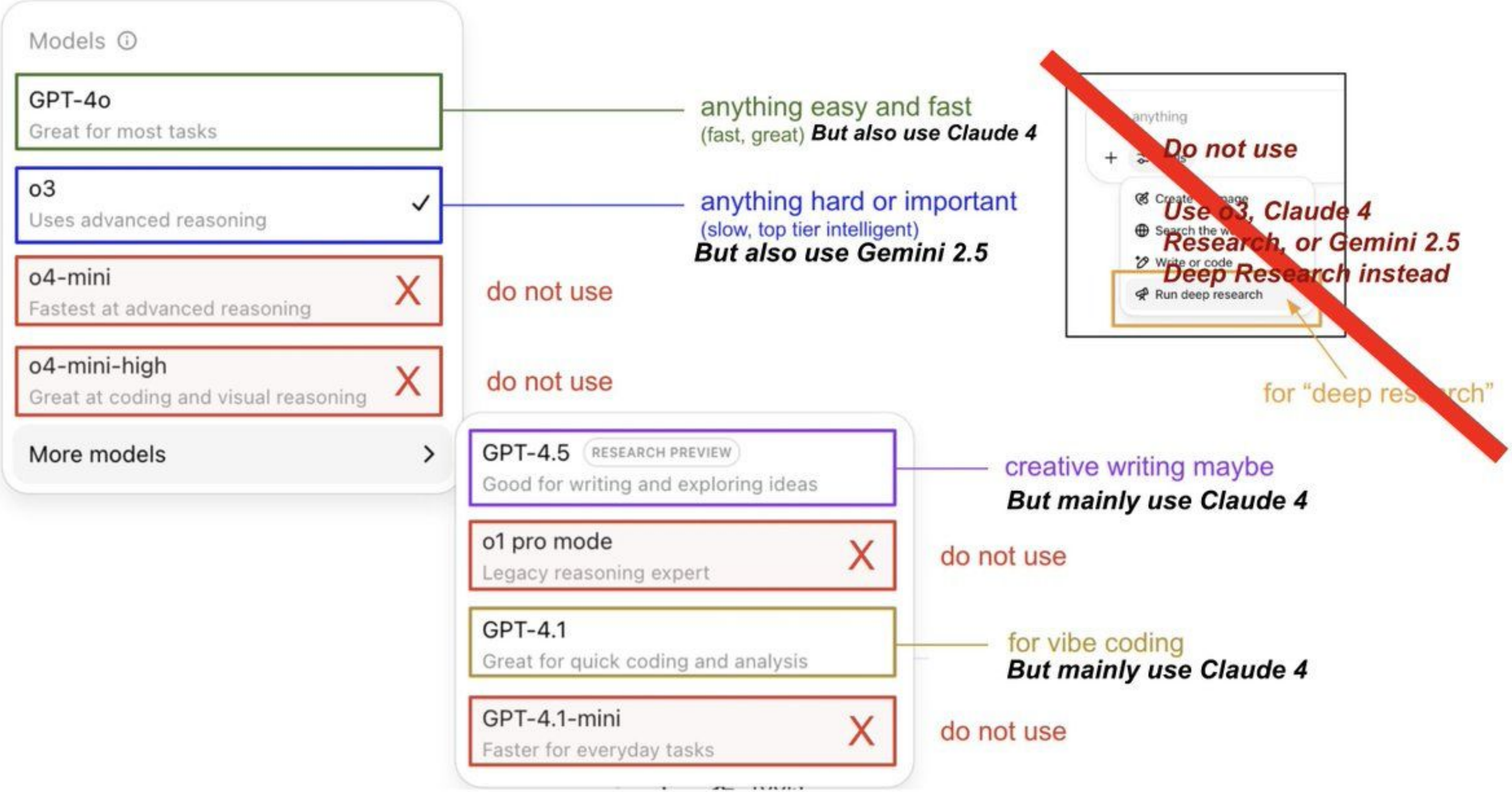
- Prediction Engine
- Scale changes everything.
- Excels when close to it’s training data.
  Struggles out of distribution.

Paying Rent

General Principles for Utilizing AI

- Doing lots of boring work. Fast.
- Context, Context, Context.
- Manage a Smart Intern
- Write a System Prompt
- Choose your Fighter


ChatGPT o3
Models
GPT-4o
Great for most tasks
o3
Uses advanced reasoning
✓
o4-mini
Fastest at advanced reasoning
o4-mini-high
Great at coding and visual reasoning
More models
GPT-4.5
RESEARCH PREVIEW
Good for writing and exploring ideas
GPT-4.1
Great for quick coding and analysis
GPT-4.1-mini
Faster for everyday tasks



GPT-4o
Great for most tasks

anything easy and fast
(fast, great) ***But also use Claude 4***

o3
Uses advanced reasoning

anything hard or important
(slow, top tier intelligent)
But also use Gemini 2.5

o4-mini
Fastest at advanced reasoning

do not use

o4-mini-high
Great at coding and visual reasoning

do not use

Do not use

Use o3, Claude 4 Research, or Gemini 2.5 Deep Research instead

~~for “deep research”~~

GPT-4.5
Good for writing and exploring ideas

creative writing maybe
But mainly use Claude 4

o1 pro mode
Legacy reasoning expert

do not use

GPT-4.1
Great for quick coding and analysis

for vibe coding
But mainly use Claude 4

GPT-4.1-mini
Faster for everyday tasks

do not use

@peterwildeford on Twitter

Please, pay for one these.
And use “reasoning”.

Claude ChatGPT Gemini

Hard Problems = Reasoning Model:

Opus, o3 or Gemini 2.5 Pro

Mundane Utility

The Obvious: Content Generation

- Text: Lore, Dialog, Snippets, Tooltips, Explanations, Localization/Translation, etc.
- Marketing
- Placeholder Assets: Meshy AI + Manual Fixes, DeepMotion for Animation
- **Not so Obvious:** Content Validation
  - Validating adherence to style, code, vibe or theme
  - Synthetic Focus Groups
  - Good at Scale

Code

- Could a Programmer with a week of time solve this? So can Claude.

- QA
  - Generate 50 edge cases for this feature
  - Bug Discovery

- Rubber Ducking
  - Works for Code, works for design

Production Efficiency

- AI for the boring, repetitive tasks that slow down development
- Automated Documentation is better than no Documentation.
- Asset Organization:
    - Labeling and searching Assets

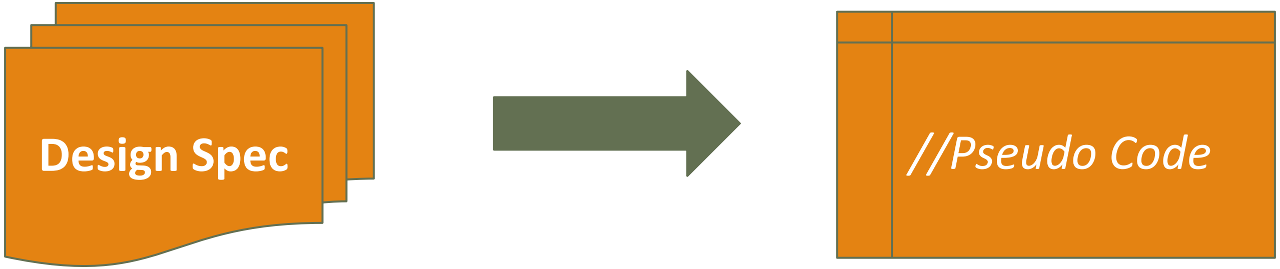
Data Analysis

- Models are better than you at statistics.
  - Statistical Modelling of whatever data you have.

- **Not as Obvious:**
  - Large Language Models are good at analysing Large amounts of Language
  - Filtering, compiling and looking at Playtesting data.
  - Google does large data best: aistudio.google.com

Transforming Natural Language

- **Not as Obvious:**
  - turning human-readable specs into machine-readable logic
  - Design docs into pseudocode
  - Gameplay into state machines



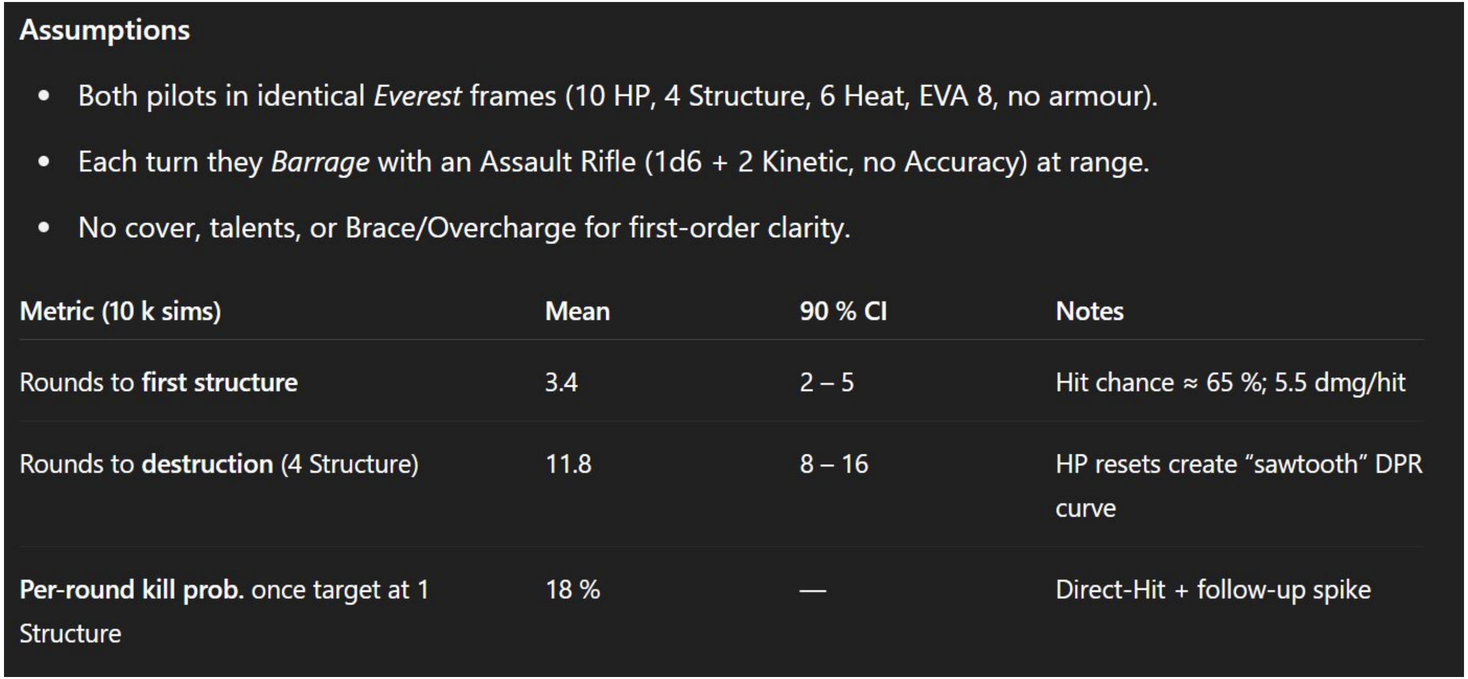
Lancer: A LLM game design benchmark

- ChatGPT o3 with 2 prompts:
- ***‘Analyze these mechanics and model their dynamics’***
- **‘Apply lossless compression to find the shortest rules that produce the same gameplay’**

What it did well

- **Quantitative Analysis**
- **Systems Thinking**
- **Compression**

Quantitative Analysis



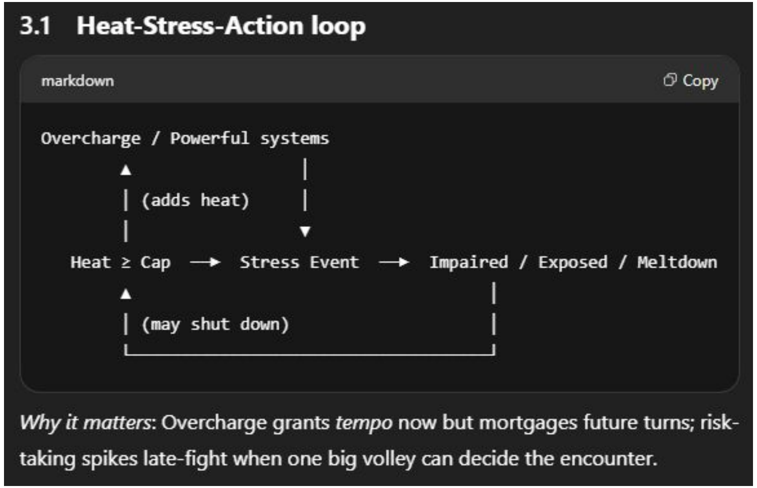
Assumptions

- Both pilots in identical *Everest* frames (10 HP, 4 Structure, 6 Heat, EVA 8, no armour).
- Each turn they *Barrage* with an Assault Rifle (1d6 + 2 Kinetic, no Accuracy) at range.
- No cover, talents, or Brace/Overcharge for first-order clarity.

Metric (10 k sims)	Mean	90 % CI	Notes
Rounds to **first structure**	3.4	2 – 5	Hit chance ≈ 65 %; 5.5 dmg/hit
Rounds to **destruction** (4 Structure)	11.8	8 – 16	HP resets create “sawtooth” DPR curve
Per-round kill prob. once target at 1 Structure	18 %	—	Direct-Hit + follow-up spike

(ChatGPT o3)

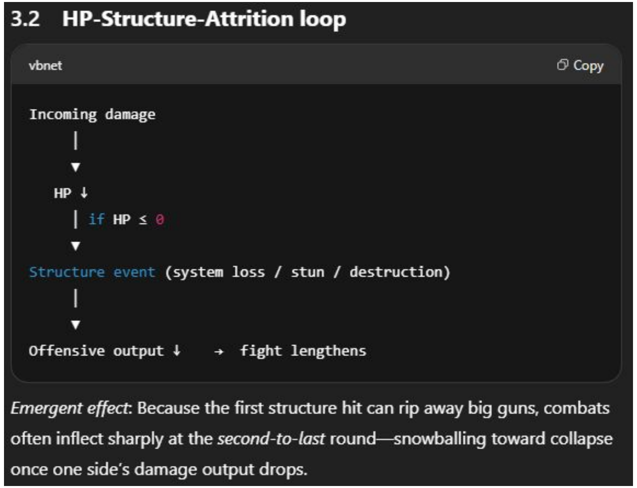
Systems Thinking



3.1 Heat-Stress-Action loop

```
Overcharge / Powerful systems
        ▲                 |
        | (adds heat)     |
        |                 ▼
   Heat ≥ Cap  ──▶  Stress Event  ──▶  Impaired / Exposed / Meltdown
        ▲                                    |
        | (may shut down)                    |
        └────────────────────────────────────┘
```

Why it matters: Overcharge grants *tempo* now but mortgages future turns; risk-taking spikes late-fight when one big volley can decide the encounter.



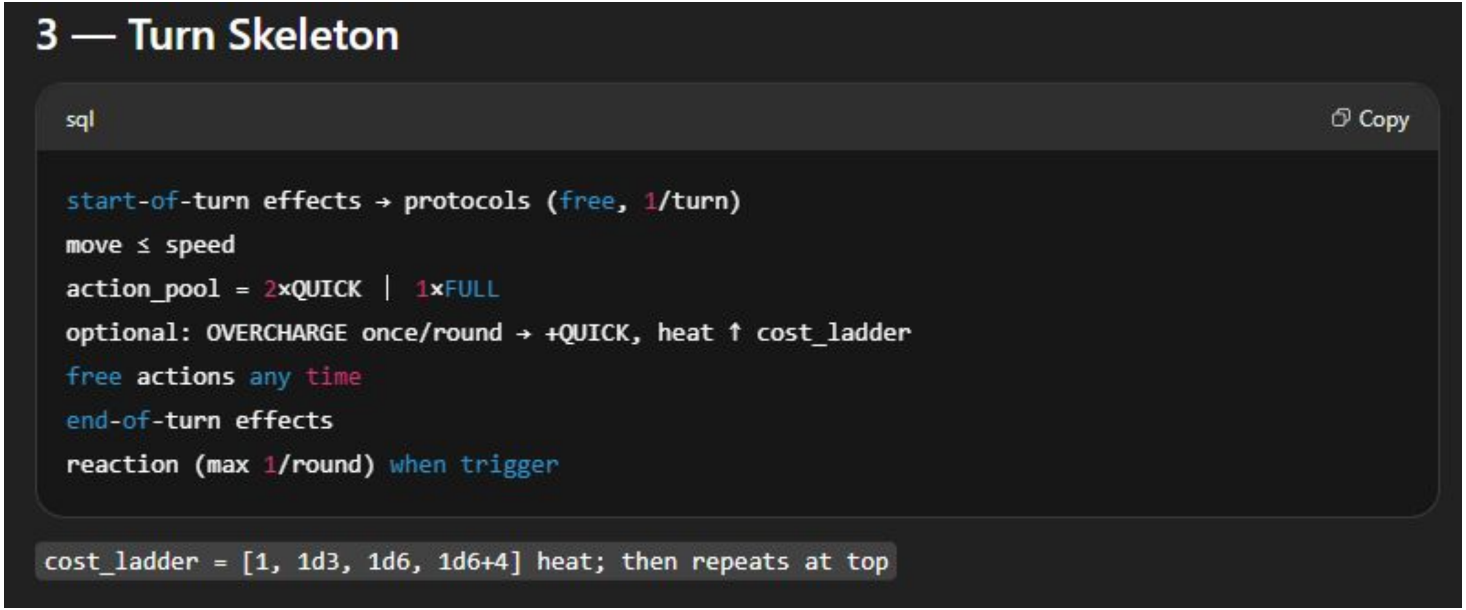
3.2 HP-Structure-Attrition loop

```
Incoming damage
     |
     ▼
   HP ↓
     | if HP ≤ 0
     ▼
Structure event (system loss / stun / destruction)
     |
     ▼
Offensive output ↓    →  fight lengthens
```

Emergent effect: Because the first structure hit can rip away big guns, combats often inflect sharply at the *second-to-last* round—snowballing toward collapse once one side's damage output drops.

(ChatGPT o3)

Compression



3 — Turn Skeleton

```sql
start-of-turn effects → protocols (free, 1/turn)
move ≤ speed
action_pool = 2×QUICK | 1×FULL
optional: OVERCHARGE once/round → +QUICK, heat ↑ cost_ladder
free actions any time
end-of-turn effects
reaction (max 1/round) when trigger
```

`cost_ladder = [1, 1d3, 1d6, 1d6+4] heat; then repeats at top`

(ChatGPT o3)

What it did well

Quantitative Analysis

- Monte Carlo simulation of 10,000 combats
- Discovered the "3-4 rounds to first structure damage" pattern
- Calculated that +1 Accuracy saves ~0.4 rounds to first structure

Systems Thinking

- Identified three core feedback loops (Heat-Stress-Action, HP-Structure-Attrition, Status Flag Economy)

Compression Achievement

- Reduced 15,000 words to 1,300 while preserving all mechanics
- Created a functional "codec" that could rebuild the entire system

What it did bad

Missing Details

- NPC rules, mount restrictions, weapon tags

Missing Accuracy

- The 'lossless' compression actually lost several subsystems

Great for analysis, but you still need to understand the game to spot errors

Chat History

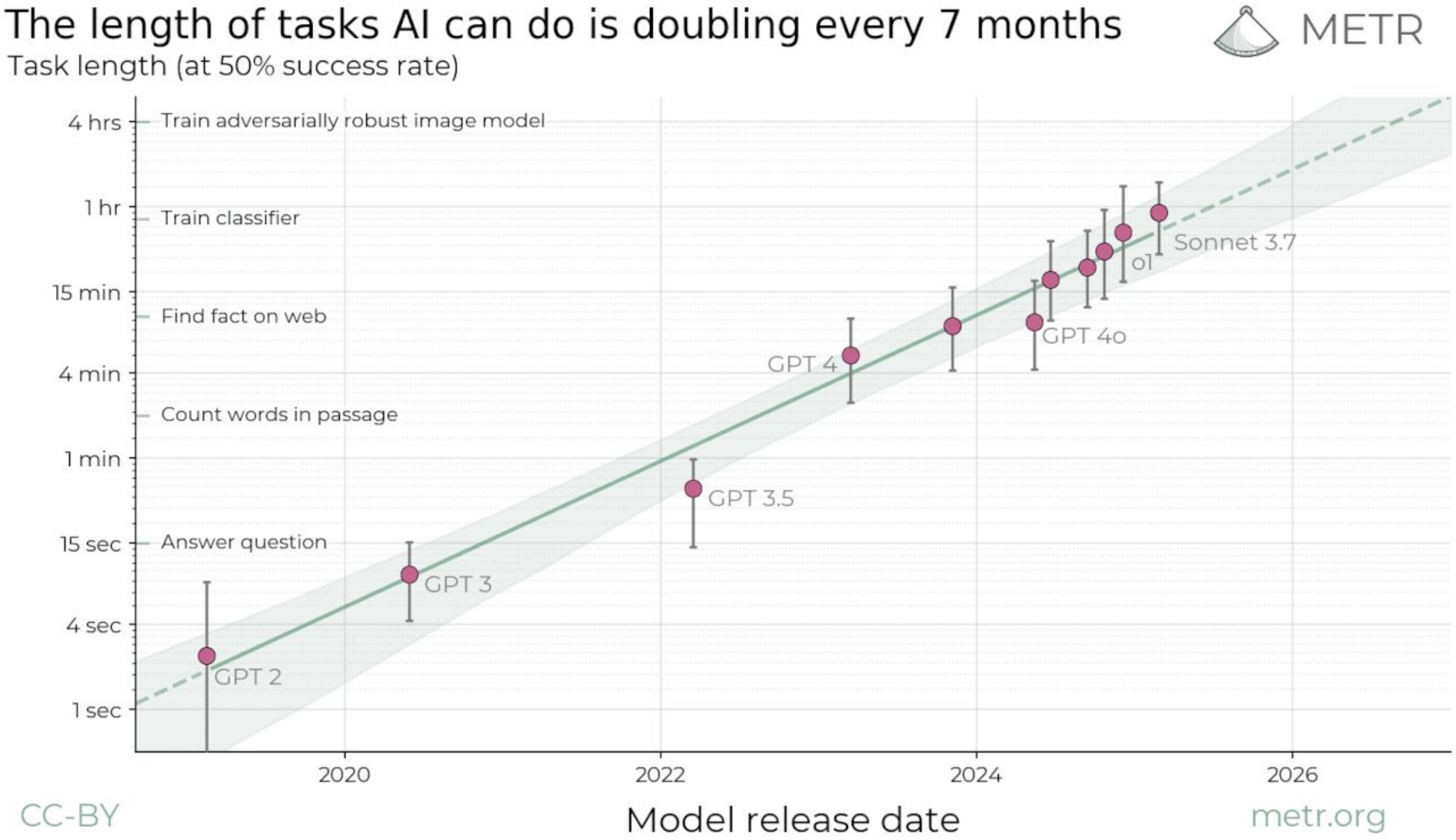
Analyse

https://chatgpt.com/share/6848a299-049c-8000-a851-fe576862cebe

Compression

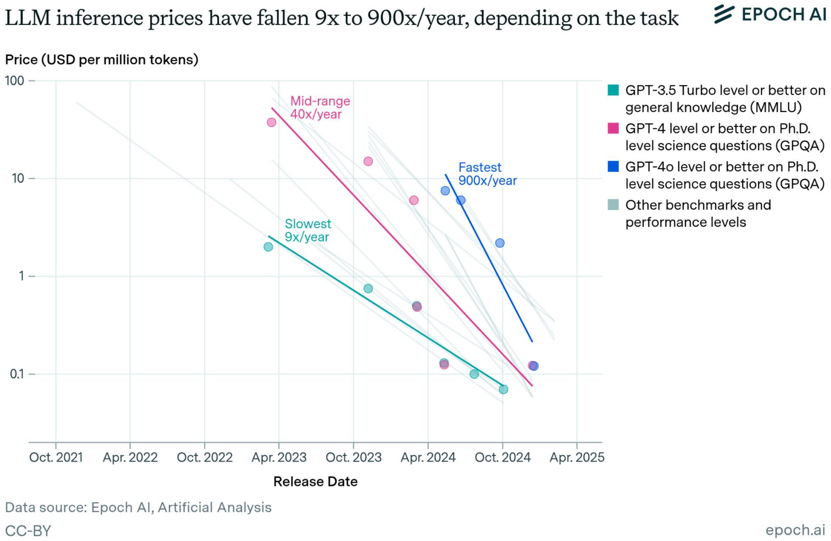
https://chatgpt.com/share/684725da-ff84-8000-bdbb-92a9a56fc458

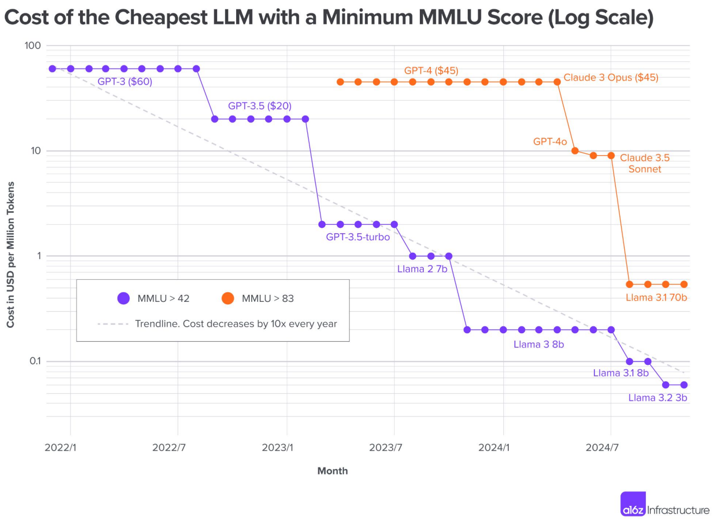
Straight Lines on a Graph



METR (2025), *Measuring AI Ability to Complete Long Tasks*, Figure 1

Cognition To Cheap to Meter





Price decline: 9x to 900x per year, median 50x (Epoch AI, 2025)

Cost decreasing by 10x every year (A16Z, 2024)

Capability Trends

- My Price Cheap
- Agents
- Test-Time Compute
- Multi-modality
- Tool-use
- Fine Tuning and Personalisation
- Memory

The AI you use right now, is the worst it will ever be.

What to do?

1. Use the dang thing.
2. Pick a benchmark, something you are good at, and test AI against it.
3. Don't stop thinking.

Make Your Beliefs Pay Rent

JONAS HEINKE

CO-FOUNDER
PLAYINSIGHT STUDIOS,
TECHNICAL DIRECTOR

PLAY INSIGHT STUDIOS

playinsightstudios.com

Email me: **heinke.jonas@googlemail.com**

- API Costs
- Benchmarks, What AI is best at what, etc.
- Open Source vs Closed Source
- GDPR
- Prompt Engineering
- AI Safety
- Chinese AI
- Everything else related to AI

Talk Archive and Further Reading

Questions?

On which I can speak, but will not for scope reasons

- Ethics
- Copyright
- How this will affect the Trout population
- The Impact on Jobs
- People Are Worried About AI Killing Everyone

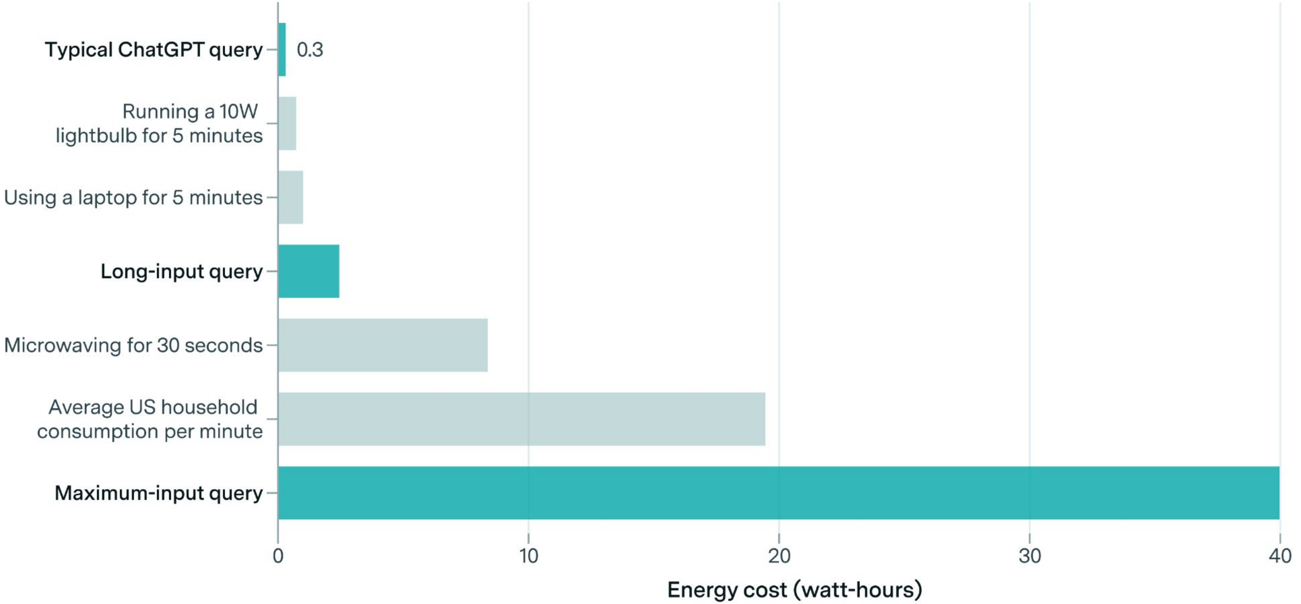
Questions?

Further Reading

- ai-2027.com
- lesswrong.com
- Nick Bostrom - Superintelligence (2014)
- Eliezer Yudkowsky, Nate Soares - If Anyone Builds It, Everyone Dies (2025)
- Zvi Mowshowitz - Don’t Worry About the Vase
- Dwarkesh Patel Podcast

Energy consumption per ChatGPT query is small compared to everyday electricity use

EPOCH AI



Pessimistic estimates of the energy usage of ChatGPT with GPT-4o across for different query lengths: typical (<100 words), long (~7,500 words), and maximum context length (~75,000 words), with an average response length of 400 words.

CC-BY

epoch.ai

(People are often curious about how much energy a ChatGPT query uses; the average query uses about 0.34 watt-hours, about what an oven would use in a little over one second, or a high-efficiency lightbulb would use in a couple of minutes.
It also uses about 0.000085 gallons of water; roughly one fifteenth of a teaspoon.)

- *Sam Altman (The Gentle Singularity), June 10, 202*

Feb 07, 2025, Epoch AI