



Can LLMs Play Pokèmon?

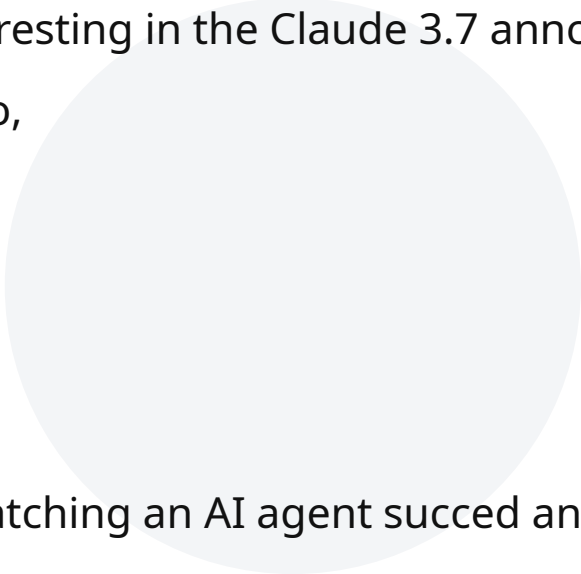
Today's AI agent effectiveness
and the time-horizon view

Hazel Shanks
Christchurch AI Meetup
Monday 7th of July 2025



Who am I?

Follow LLM news obsessively



Spotted something interesting in the Claude 3.7 announcement blogpost:
It's playing pokemon. So,

- what is that,
- why do it,
- and: does it work?

We can learn a lot by watching an AI agent succeed and fail, livestreamed



Claude Sonnet 3.7

Anthropic announced new (late Feb '25) frontier model (now obsolete)

- And it has a simulated reasoning mode ('Extended Thinking')
- Trained & eval'd on agent tasks:

“...an improved capability that allows it to iteratively call functions, respond to environmental changes, and continue until an open-ended task is complete.”



<https://www.anthropic.com/news/visible-extended-thinking>



Agent tasks?

A new AI interaction paradigm

Contrast the old paradigm:

- Chatbot: constant back-and-forth with a chat window (or voice call)
- Or: integration of AI 'calls' to extend existing tools (super-autocomplete in IDE).
- Or: 1-shot 'generative' creation of art assets (image/audio models)

Notice the user experience: micro-managing the AI.



Agent tasks?

A new AI interaction paradigm

Agents paradigm:

- Give an open-ended task, come back in days/weeks.

AI runs continuously. Takes action, gets results, takes new action, repeat.

Goal: Handing off *whole tasks* to an AI, to complete *independently*.

Totally different UI. Asynchronous.

It's the future. But is it our *present* (mid-2025)?

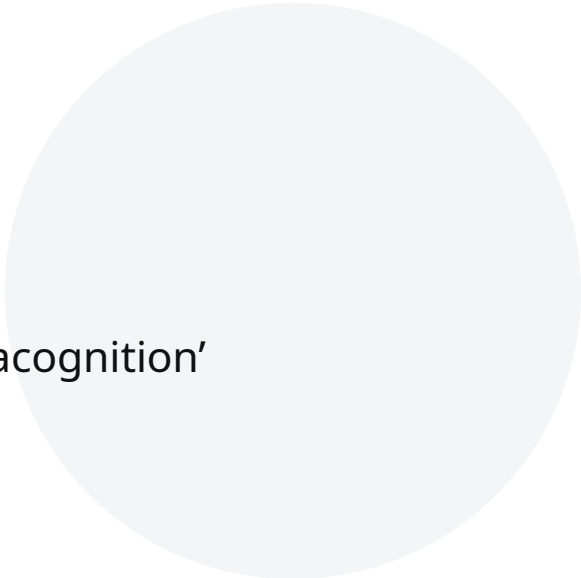
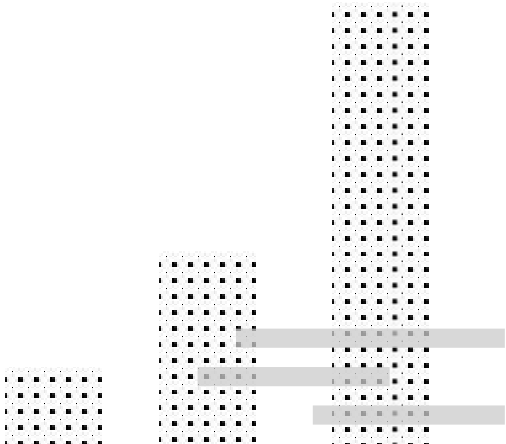




Agent tasks?

A new AI interaction paradigm

Agents need:

- Reliability
 - Tool use
 - Long context
 - Self-evaluation / 'metacognition'
 - ... and a simple loop.
- 
- 



Agent tasks?

A new AI interaction paradigm



Neat!

Tell the AI to do your job, then go for coffee. Sounds great.

But: does it **work**, today?





Agent evals

Figuring out if it works



Can be tricky to eval agents!

(Lots of good work happening here, eg, IBM doing a recent survey.)

This talk: a look at two agent eval techniques from earlier in the year.

First up: why not just **give the AI a real task**, and see how it does?

For example....



Can it play Pokémon?

Anthropic's blog says: yes

Single researcher pet project: Claude playing **Pokémon Red**.
Just Sonnet 3.7 with basic GameBoy emulator interaction tools.
— no training on gaming, pokemon, or the gameboy.



Can it play Pokémon?

Anthropic's blog says: yes ... at least the start of it

Single researcher pet project: Claude playing **Pokémon Red**.

Just Sonnet 3.7 with basic GameBoy emulator interaction tools.

— no training on gaming, pokemon, or the gameboy.

... and **it does work**.

“Claude 3.7 Sonnet’s improved agentic capabilities helped it advance much further, **successfully battling three Pokémon Gym Leaders** (the game’s bosses) and winning their Badges. Claude 3.7 Sonnet is super effective at trying multiple strategies and questioning previous assumptions, which allow it to improve its own capabilities as it progresses.”



(out of 8 total, plus the E4 (+ champ) boss rush at the end.)

From the announcement blog:

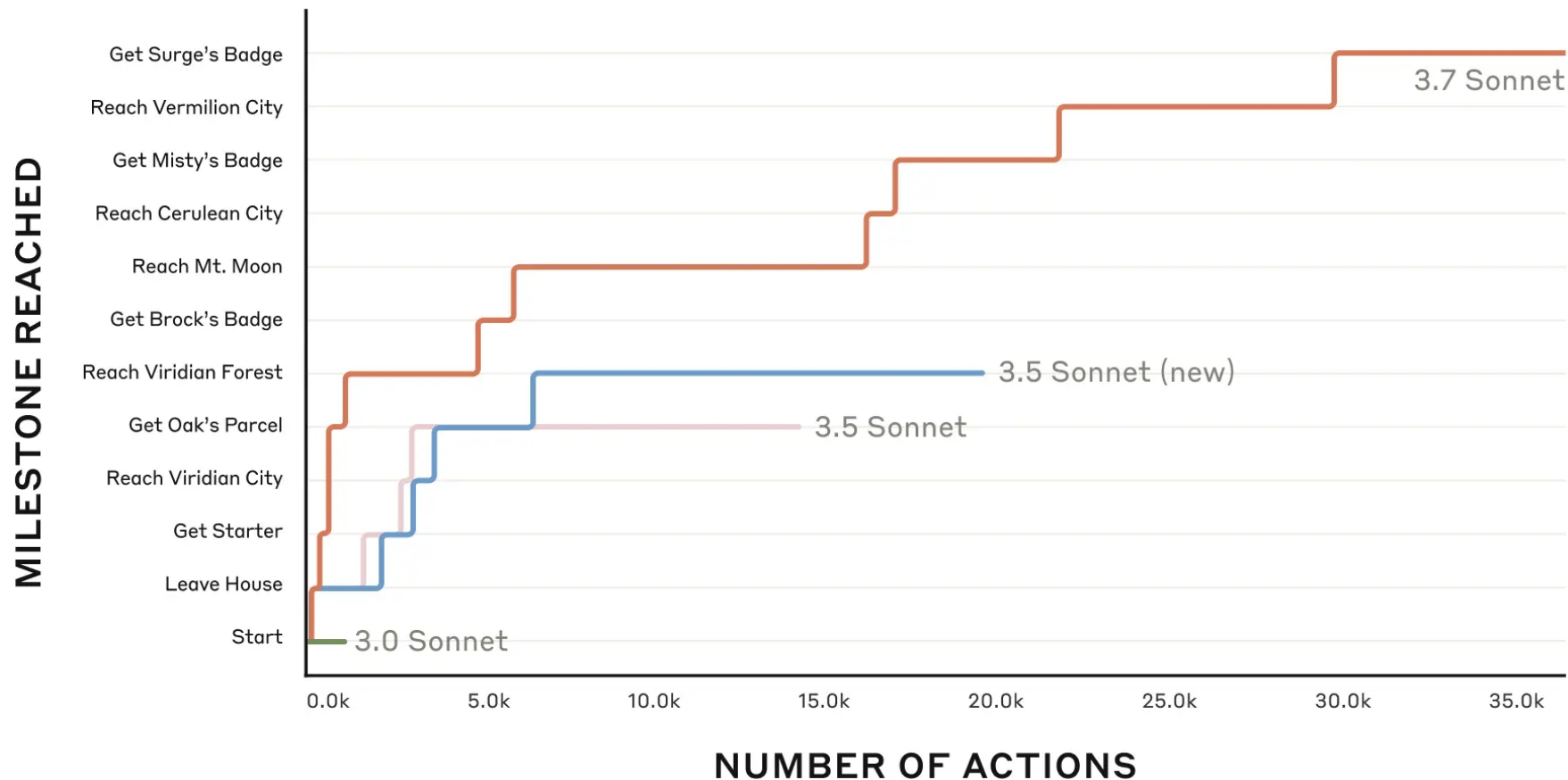
Claude plays Pokémon

Together, Claude's extended thinking and agent training help it do better on many standard evaluations like OSWorld. But they also give it a major boost on some other, perhaps more unexpected, tasks.

Playing Pokémon—specifically, the Game Boy classic *Pokémon Red*—is just such a task. We equipped Claude with basic memory, screen pixel input, and function calls to press buttons and navigate around the screen, allowing it to play Pokémon continuously beyond its usual context limits, sustaining gameplay through tens of thousands of interactions.

Claude models playing Pokémon*

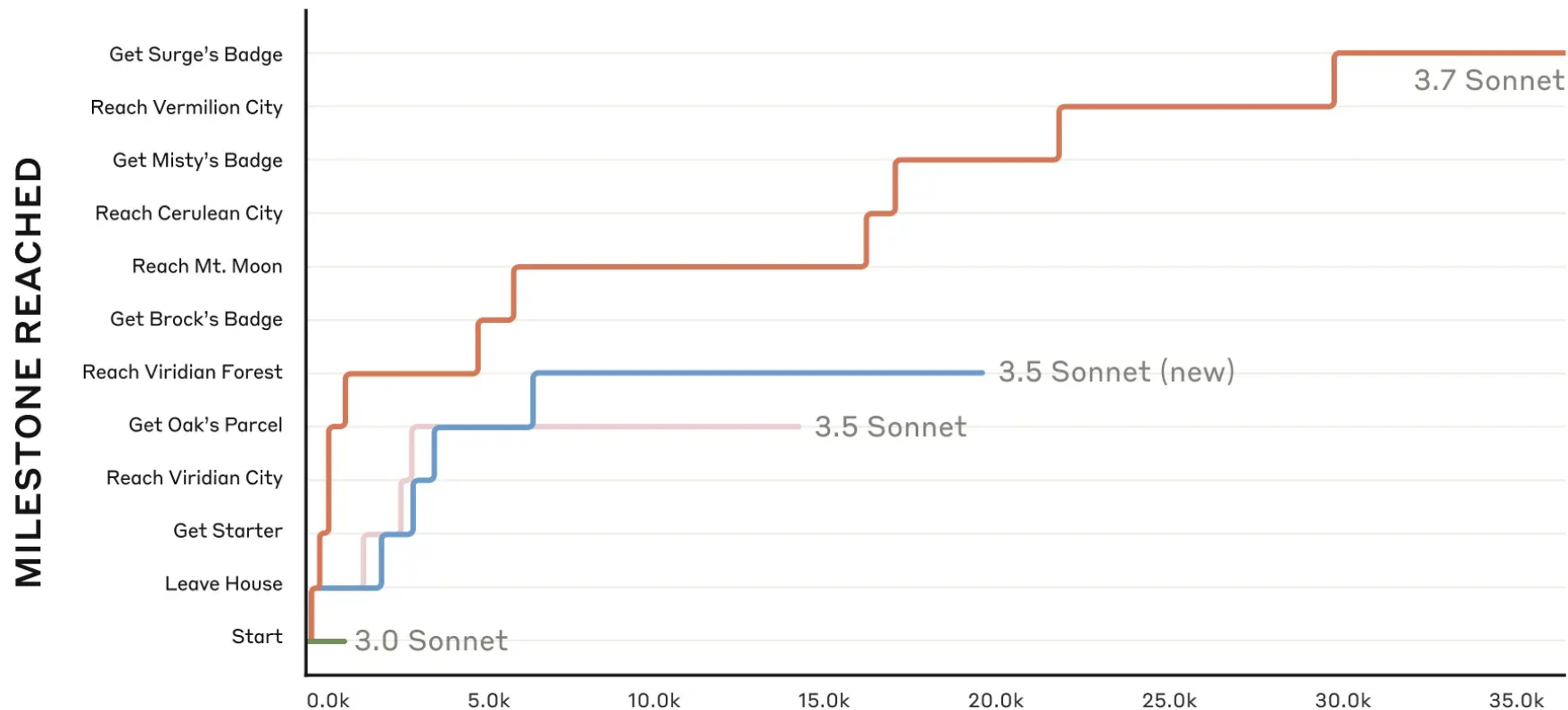
Milestone progress over time



* Pokémon is a registered trademark of Nintendo of America Inc. This chart references Pokémon terminology solely to identify game milestones reached by Claude models. No affiliation, sponsorship, or endorsement by Nintendo of America Inc. is implied or intended.

Claude models playing Pokémon*

Milestone progress over time



Notice: scale is in kiloactions

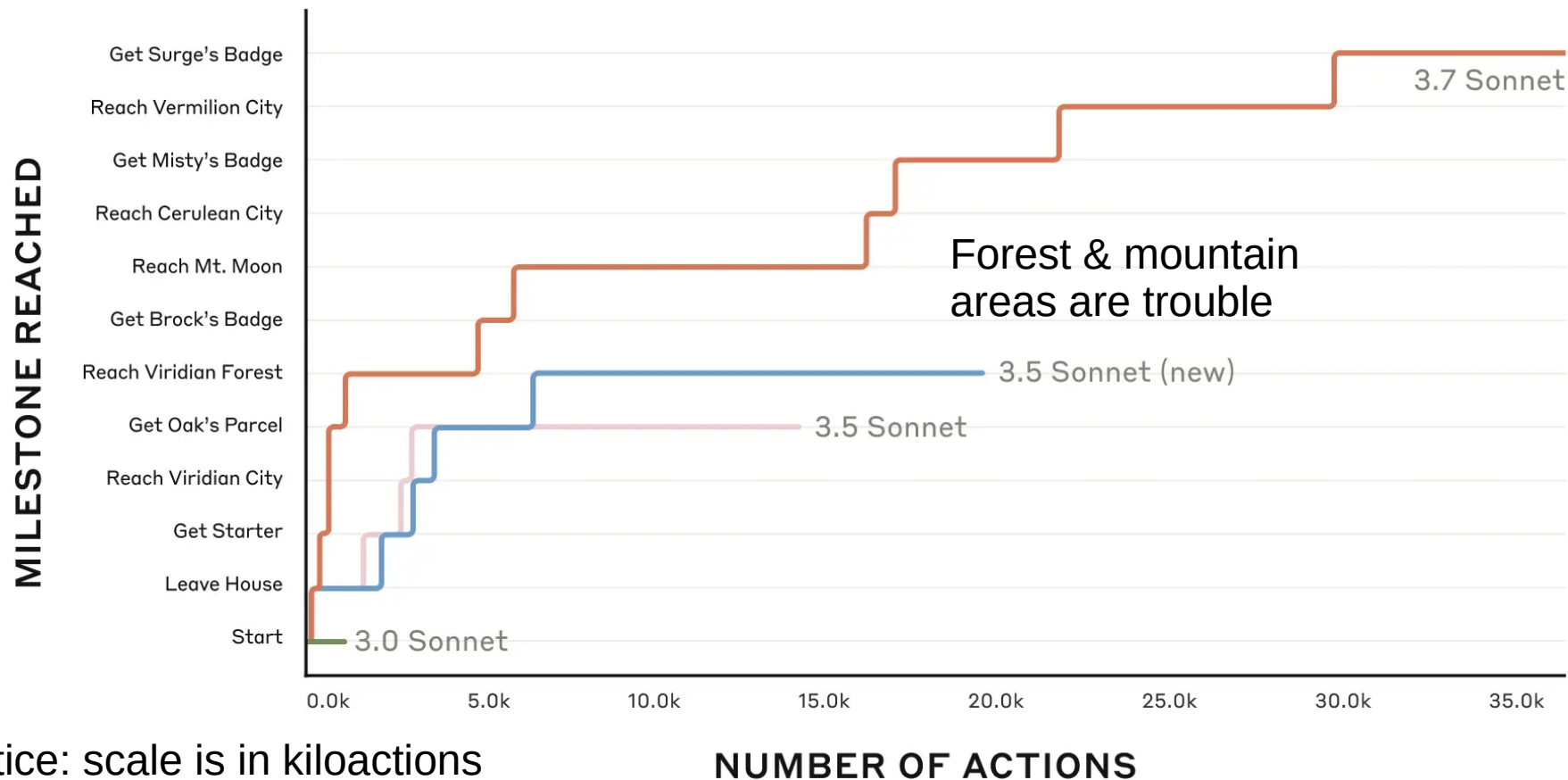
NUMBER OF ACTIONS

(what's an 'action'? You'll see)

* Pokémon is a registered trademark of Nintendo of America Inc. This chart references Pokémon terminology solely to identify game milestones reached by Claude models. No affiliation, sponsorship, or endorsement by Nintendo of America Inc. is implied or intended.

Claude models playing Pokémon*

Milestone progress over time



Notice: scale is in kiloactions

(what's an 'action'? You'll see)

* Pokémon is a registered trademark of Nintendo of America Inc. This chart references Pokémon terminology solely to identify game milestones reached by Claude models. No affiliation, sponsorship, or endorsement by Nintendo of America Inc. is implied or intended.

Streaming live on twitch.tv right now — twitch.tv/claudeplayspokemon



STEP COUNT
121151

Claude Opus 4 Plays Pokémon

OPUS
EDITION

and I can go further. Claude Opus 4 back the very bottom of the floor at this point. There's maybe a different escalator or exit there.

Using tool: navigator - Target: (15, 7)

I'm at the bottom of 5F now and I don't see any other escalators or exits here. Just more shelves. The issue remains - I can't find a DOWN escalator on 5F. Let me think about this differently. What if the department store in Pokemon Red is designed so that you can't go down from 5F directly? What if I need to:

1. Take the specific route the game intends
2. Or use a different method entirely

Actually, wait. Let me check the far left side of the floor. I've been focusing on the right side and middle, but maybe there's something on the left side I haven't explored. Let me go to the western edge of the floor.

Using tool: navigator - Target: (10, 5)

Critique provided for knowledge base.

Thinking...

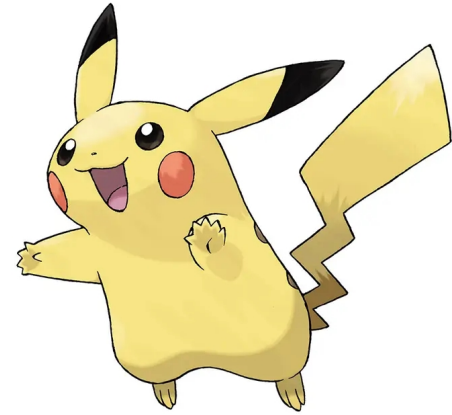
Current Team

SPLASH BLASTOISE HP: 150/150 OK	WINGS Lv. 46 SPEAROW HP: 43/43 OK	LUNA Lv. 16 CLEFAIRY HP: 42/42 OK	DUX Lv. 14 FARFETCH'D HP: 42/42 OK	GUST Lv. 21 PIDGEOTTO HP: 59/59 OK	SPIKE Lv. 20 NIDOKING HP: 70/70 OK
---	--	--	---	---	---

What's Pokémon Red?

Cultural: a classic start to Pokémon franchise

- Pokémon is the highest grossing franchise
- Mascot monsters are cultural icons, like Pikachu →
- Whole franchise derived from the main series videogames — all new 'mons intro there. Defines generations
- Main series games start with "Gen 1": Pokémon Red & Blue
- System sellers for the Nintendo Gameboy console →
- *The* hot item on the playground in 1998, '99
- Soundtrack is a chiptune classic!



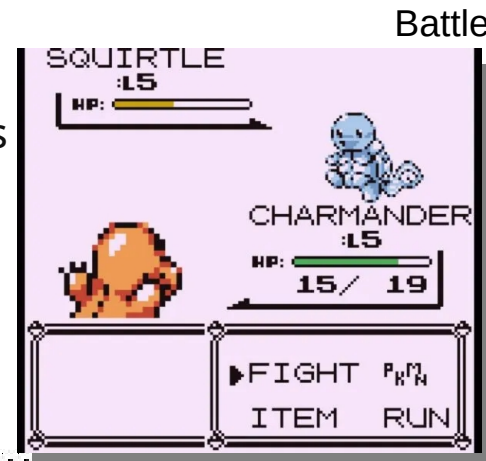
What's Pokémon Red?

The game: top-down JRPG monster-catching adventure

- Alternates: top down 2d grid explore, turn-based monster battle
- Choose a starter pokemon, then go on adventure
- Start out catching bugs, birds, rats. Endgame, legendary monsters
- Fight wild pokemon, trainers. Catch pokemon, level up, teach moves
- 'Hidden Moves' give tools to explore map more
- Travel Routes to cities, find gyms, beat up leaders for badges
- Eventually beat Elite 4 bossrush, become champion, Hall of Fame



Overworld



Battle

But why Pokémon Red for AI?

A compelling test case

- Culturally relevant, nostalgic for my generation
- It's an authentic task, but it's also got a linear order with clear milestones.
- It's doable by 7-year olds. SotA AI should crush it... right?
- In the trainset / internet data scrapes. Well documented.

But why Pokémon Red for AI?

A compelling test case

- Doable by a generalist language model — played with turn-by-turn interactions with no realtime input, no game-relevant sound cues.
- Efficient play requires planning, working toward multiple goals at once.
- Hard to lose progress (don't release pokemon, and you're good).
- Limited interactions — gameboy only has 8 buttons.

Same reasons
Twitch played
it!

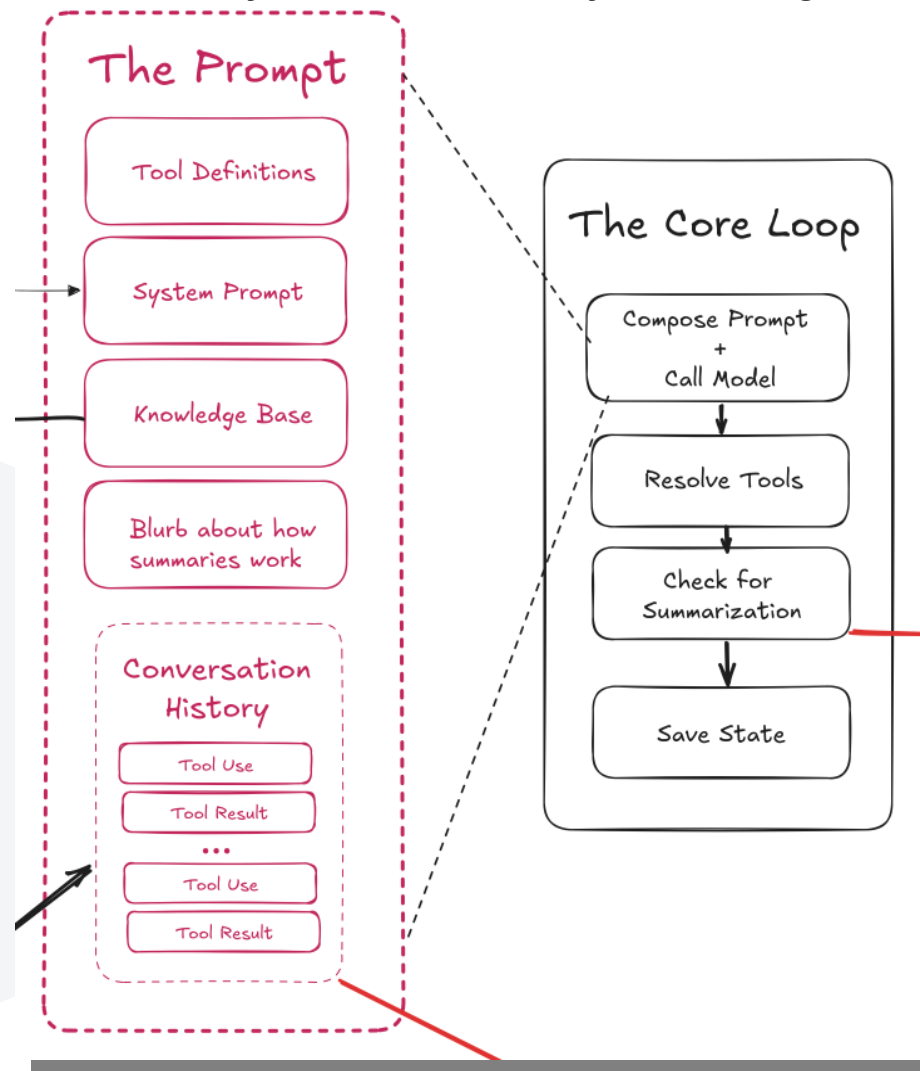
Inside an agent

To build: simple core loop

- Compose system prompt
- Invoke Claude ← where it all happens
- Resolve tool calls
- Check if context summary needed?
- Save state

Then do it again!

Claude Plays Pokmemon system diagram

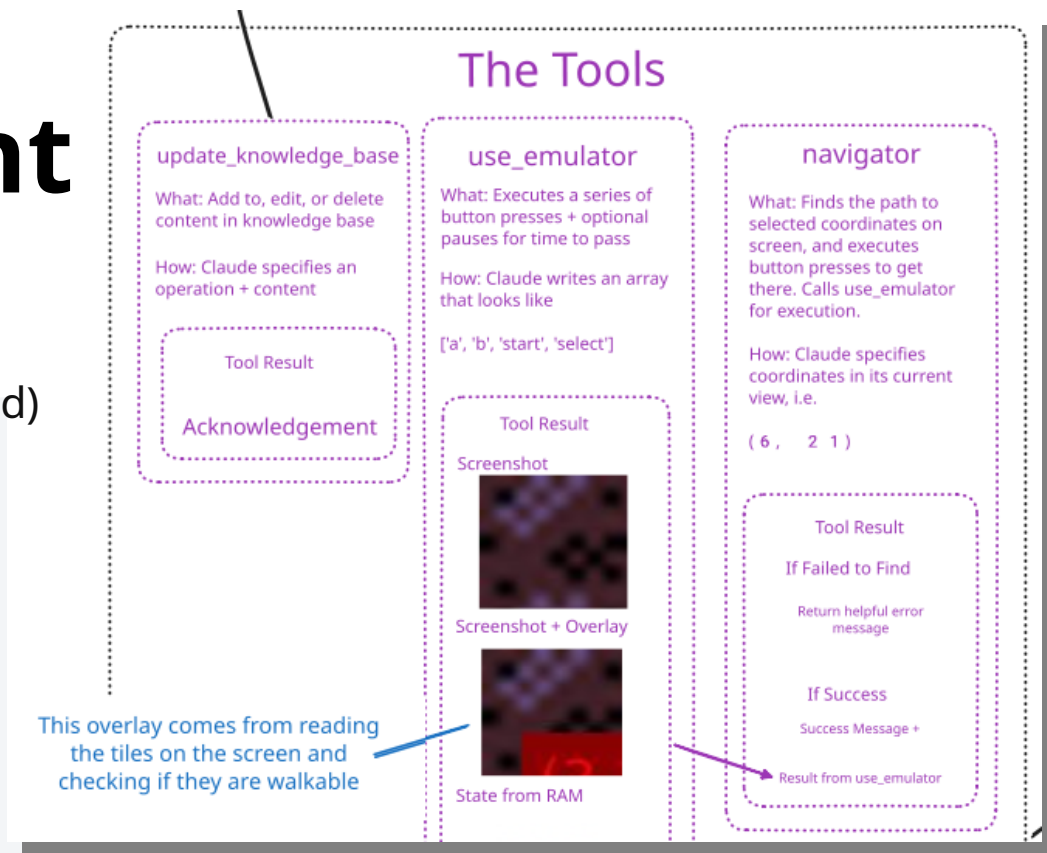


Inside an agent

To build: tooling

- update_knowledge_base (scratchpad)
- use_emulator (button inputs)
- navigator (direct coordinate goto)
- Screen overlay (kinda)
 - provides coords, navigability

(note what's missing: no google, no gamefaqs, etc)



Vision

Overlay helps

- Coordinates provided
- Walkability provided
— colors of coords
- Screen tiles shown
- Using colorized ROM

(Actually from Gemini,
but Claude's is similar.)



Inside an agent

To build: context management

Per loop, check context usage. Too high? Then:

- Have LLM summarize the context
- Have LLM critique performance
- Update knowledge base to keep info
- Clear out context
- Add the summary and critique to new context
- Continue with loop

Summarization / Managing Long Context

Rollouts are way too long for the 200k context window, so we need some way to manage context. I've found progressive summarization to work well

Once conversation history > max_turns we trigger a summarization event

This has Claude write a summary of their recent progress / what happened in the last max_turns.

Then we clear the full conversation history and insert the summary as the first assistant message, and Claude resumes its journey

Finally, another LLM is called to inspect the first LLM's knowledge base and to provide feedback -- this helps ensure the agent does more frequent maintenance of its knowledge base



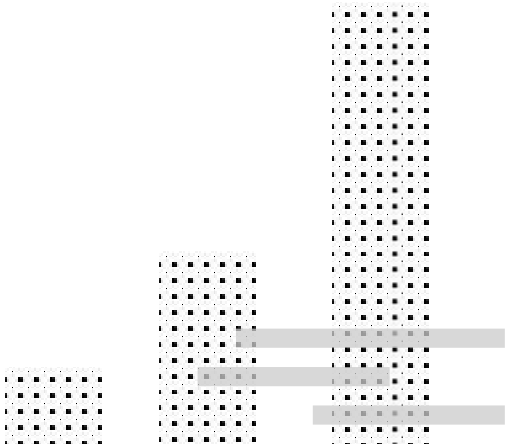
Example: Opus 4 in Celadon City

Short example of the life of Claude



This example **happened to be live** when I was looking for an example.

Claude Opus 4 wants to go to the Pokemon Gym in Celadon City.
But while the Gym is visible, getting there is tricky....



STEP COUNT
120968

Claude Opus 4 Plays Pokémon

OPUS
EDITION

Claude Opus 4

'wait']

I tried to move down but I'm still at (14,23). Looking at the navigation overlay, I notice that (14,19) is missing from the navigable list, which means there's likely an obstacle or NPC at (14,19) blocking my path south. I need to find another way to reach the gym entrance. Let me go west first to try to go around the obstacle.

Using tool: navigator - Target: (10, 23)

Perfect! I'm now at (10,23) and I can see the gym building clearly below me. I can see the golden door at the bottom of the gym. Looking at the visual, I can see I'm directly above the gym entrance. I can see the door is right below me.

I should be able to move down now to reach the entrance. Let me try moving down to get to the door.

Using tool: use_emulator - Buttons: ['down', 'down', 'down', 'down', 'a', 'wait']

☐ Cleaning up my context



UPDATING KNOWLEDGE BASE

CHANGES

- Erika's Gym: Western Celadon, green building marked "GYM"
- ****ENTRANCE****: Golden door at ~~approximately (11, 12, 27)~~ south side of building
- ****ACCESS****: From Pokemon-Center-area (20,23), go east then south to y=26, then west
- ~~Cannot enter from north side (y=23) - blocked by buildings(10, 27) - south side of building ONLY~~
- ****CORRECT PATH****: From (20,23), go east to (22,23) → south to (22,26) → west to (10,26) → south to (10,27) → face door and press 'a'

POISON GROUND

OK

Current Team

SPLASH

Lv. 46

BLASTOISE

WATER

HP: 150/150

OK

wings

Lv. 16

SPEAROW

NORMAL FLYING

HP: 43/43

OK

luna

Lv. 13

CLEFAIRY

NORMAL

HP: 42/42

OK

DUX

FARFETCH'D

HP: 42/42

OK

HP: 59/59

OK

HP: 70/70

Lv. 20

STEP COUNT
120973

Claude Opus 4 Plays Pokémon

OPUS
EDITION

Claude Opus 4

Good, I updated the current objective. Now let me consolidate the redundant type effectiveness information as suggested.

Using tool: string_replace

Good, I've enhanced the type effectiveness chart. Now let me clean up the Celadon City Map section to remove redundancy and the unnecessary "WRONG PATHS TRIED" section as suggested.

Using tool: string_replace

Great! I've cleaned up the knowledge base as suggested. Now I need to actually navigate to the gym entrance. I'm at (10,23) and need to reach (10,27), which is 4 tiles south. Let me navigate there.

Using tool: navigator - Target: (10, 27)

Thinking...



Current Team

SPLASH BLASTOISE HP: 150/150 Lv. 46 WATER OK	wings SPEAROW HP: 43/43 Lv. 16 NORMAL FLYING OK	luna CLEFAIRY HP: 42/42 Lv. 13 NORMAL OK	DUX FARFETCHD HP: 42/42 Lv. 14 NORMAL FLYING OK	gust PIDGEOTTO HP: 59/59 Lv. 21 NORMAL FLYING OK	SPIKE NIDOKING HP: 70/70 Lv. 20 POISON GROUND OK
--	---	--	---	--	--

STEP COUNT
120973

Claude Opus 4 Plays Pokémon

OPUS
EDITION

South: Let me navigate Claude Opus 4

Using tool: navigator - Target: (10, 27)

I see that the navigation didn't work - it says the path found was empty []. I'm still at (10,23) and the coordinates show that positions below y=24 are marked in red (non-navigable). This means I need to find an alternative path.

Looking at the navigable spaces, I can see that the furthest south I can go from my current position is y=23. The gym entrance is supposed to be at (10,27), but it's showing as red/non-navigable.

Wait, let me check what navigable spaces are available. I see I can move to various positions but they all stop at y=23. This suggests there might be a different path to reach the gym entrance, perhaps going around.

Looking at my knowledge base, it mentions the path from Pokemon Center: "East to (22,23) → south to (22,26) → west to (10,26) → south to entrance". This suggests there might be a way to go around the buildings to reach the entrance. Let me try moving east first to see if I can find a path that goes further south.

Using tool: navigator - Target: (14, 23)



Current Team

SPLASH

Lv. 46

BLASTOISE

WATER

HP: 150/150

OK

wings

Lv. 16

SPEAROW

NORMAL

FLYING

HP: 43/43

OK

luna

Lv. 13

CLEFAIRY

NORMAL

HP: 42/42

OK

DUX

Lv. 14

FARFETCHD

NORMAL

FLYING

HP: 42/42

OK

gust

Lv. 21

PIDGEOTTO

NORMAL

FLYING

HP: 59/59

OK

SPIKE

Lv. 20

NIDOKING

POISON

GROUND

HP: 70/70

OK

STEP COUNT
120974

Claude Opus 4 Plays Pokémon

OPUS
EDITION

alternative path:

Claude Opus 4

Looking at the navigable spaces, I can see that the furthest south I can go from my current position is y=23. The gym entrance is supposed to be at (10,27), but it's showing as red/non-navigable.

Wait, let me check what navigable spaces are available. I see I can move to various positions but they all stop at y=23. This suggests there might be a different path to reach the gym entrance, perhaps going around.

Looking at my knowledge base, it mentions the path from Pokemon Center: "East to (22,23) → south to (22,26) → west to (10,26) → south to entrance". This suggests there might be a way to go around the buildings to reach the entrance. Let me try moving east first to see if I can find a path that goes further south.

Using tool: navigator - Target: (14, 23)

Good, I've moved east to (14,23). I can see the same pattern - all coordinates below y=24 are still showing as red (non-navigable). According to my knowledge base path, I need to continue east to around x=22, where hopefully I can find a way to go south. Let me continue moving east.

Using tool: navigator - Target: (19, 23)



Current Team

SPLASH

Lv. 46

BLASTOISE

WATER

HP: 150/150

OK

wings

Lv. 16

SPEAROW

NORMAL FLYING

HP: 43/43

OK

luna

Lv. 13

CLEFAIRY

NORMAL

HP: 42/42

OK

DUX

Lv. 14

FARFETCHD

NORMAL FLYING

HP: 42/42

OK

gust

Lv. 21

PIDGEOTTO

NORMAL FLYING

HP: 59/59

OK

SPIKE

Lv. 20

NIDDOKING

POISON GROUND

HP: 70/70

OK

STEP COUNT
120975

Claude Opus 4 Plays Pokémon

OPUS
EDITION

Pokemon Center: "East to (22,26) → south to (10,26) → south to entrance". This suggests there might be a way to go around the buildings to reach the entrance. Let me try moving east first to see if I can find a path that goes further south.

Using tool: navigator - Target: (14, 23)

Good, I've moved east to (14,23). I can see the same pattern - all coordinates below y=24 are still showing as red (non-navigable). According to my knowledge base path, I need to continue east to around x=22, where hopefully I can find a way to go south. Let me continue moving east.

Using tool: navigator - Target: (19, 23)

Excellent! Now I can see navigable paths going south. I can see coordinates (22,24), (22,25), (22,26), and (22,27) are all navigable. According to my knowledge base, the path is: "East to (22,23) → south to (22,26) → west to (10,26) → south to entrance".

I'm currently at (19,23), so I need to go to (22,23) first, then south to (22,26).

Using tool: navigator - Target: (22, 26)



Current Team

SPLASH

Lv. 46

BLASTOISE

WATER

HP: 150/150

OK

wings

Lv. 16

SPEAROW

NORMAL FLYING

HP: 43/43

OK

luna

Lv. 13

CLEFAIRY

NORMAL

HP: 42/42

OK

DUX

Lv. 14

FARFETCHD

NORMAL FLYING

HP: 42/42

OK

gust

Lv. 21

PIDGEOTTO

NORMAL FLYING

HP: 59/59

OK

SPIKE

Lv. 20

NIDOKING

POISON GROUND

HP: 70/70

OK

STEP COUNT
120975

Claude Opus 4 Plays Pokémon

OPUS
EDITION

Pokemon Center: "East to (22,26) → south to (10,26) → south to entrance". This suggests there might be a way to go around the buildings to reach the entrance. Let me try moving east first to see if I can find a path that goes further south.

Using tool: navigator - Target: (14, 23)

Good, I've moved east to (14,23). I can see the same pattern - all coordinates below y=24 are still showing as red (non-navigable). According to my knowledge base path, I need to continue east to around x=22, where hopefully I can find a way to go south. Let me continue moving east.

Using tool: navigator - Target: (19, 23)

Excellent! Now I can see navigable paths going south. I can see coordinates (22,24), (22,25), (22,26), and (22,27) are all navigable. According to my knowledge base, the path is: "East to (22,23) → south to (22,26) → west to (10,26) → south to entrance".

I'm currently at (19,23), so I need to go to (22,23) first, then south to (22,26).

Using tool: navigator - Target: (22, 26)



Current Team

SPLASH

Lv. 46

BLASTOISE

WATER

HP: 150/150

OK

wings

Lv. 16

SPEAROW

NORMAL FLYING

HP: 43/43

OK

luna

Lv. 13

CLEFAIRY

NORMAL

HP: 42/42

OK

DUX

Lv. 14

FARFETCHD

NORMAL FLYING

HP: 42/42

OK

gust

Lv. 21

PIDGEOTTO

NORMAL FLYING

HP: 59/59

OK

SPIKE

Lv. 20

NIDOKING

POISON GROUND

HP: 70/70

OK

STEP COUNT
120988

Claude Opus 4 Plays Pokémon

OPUS
EDITION

building in the vision. Claude Opus 4 tried to navigate to (10,23) and then go south to reach the entrance.

Using tool: navigator - Target: (10, 23)

Excellent! I'm back at (10,23) where we started, directly north of the gym. I can see the gym building below me with "GYM" written on it and the golden door at the bottom. Looking at the coordinate overlay, I see that all coordinates below y=24 are marked as red (non-navigable). This is the same issue I encountered before.

Wait, I realize the problem now! The path in my knowledge base said to go from (22,26) west to (10,26), but I never actually found a way to do that. The issue is that the buildings block the direct path.

Let me think about this differently. I need to look for an alternative route. Since I can't go directly south from here, maybe I need to find a gap or passage between the buildings. Let me check if there's a way to go around the gym building to reach its entrance.

Using tool: string_replace

Thinking...



UPDATING KNOWLEDGE BASE

"GYM"

- Entrance: (10,27) - golden door on south side ONLY
- Type: Grass (weak to Flying, Fire, Ice, Poison, Bug)
- Path from Pokemon Center: East to (22,23) → south to (22,26) → west to (10,26) → south to entrance**ISSUE**: Direct path south from (10,23) is BLOCKED - need alternative route
- Cannot reach (10,26) from east due to building barriers

Current Team

SPLASH

Lv. 46

BLASTOISE

WATER

HP: 150/150

OK

wings

Lv. 16

SPEAROW

NORMAL FLYING

HP: 43/43

OK

luna

Lv. 13

CLEFAIRY

NORMAL

HP: 42/42

OK

DUX

Lv. 14

FARFETCH'D

NORMAL FLYING

HP: 42/42

OK

GUST

Lv. 21

DIDGEDIT'D

NORMAL FLYING

HP: 59/59

OK

SPIKE

Lv. 20

HIDKING

POISON GROUND

HP: 70/70

OK

STEP COUNT
120989

Claude Opus 4 Plays Pokémon

OPUS
EDITION

alternative route. Since Claude Opus 4 is directly south from here, maybe I need to find a gap or passage between the buildings. Let me check if there's a way to go around the gym building to reach its entrance.

Using tool: string_replace

Let me think about this more carefully. I've been trying to follow a path that doesn't work. The buildings are creating barriers that prevent me from reaching the gym entrance at (10,27).

Looking at the pattern of movement I've done:

1. I can move freely at y=23
2. When I go south to y=26 or beyond, I get trapped in the eastern part of the city
3. There's no westward path from the eastern area to reach x=10 at the lower y coordinates

I need to explore other areas. Maybe there's a path I haven't discovered yet. Let me try going west from my current position to see if there's a different route.

Using tool: navigator - Target: (6, 23)

Thinking...



Current Team

SPLASH

Lv. 46

BLASTOISE

WATER

HP: 150/150

OK

wings

Lv. 16

SPEAROW

NORMAL FLYING

HP: 43/43

OK

luna

Lv. 13

CLEFAIRY

NORMAL

HP: 42/42

OK

DUX

Lv. 14

FARFETCHD

NORMAL FLYING

HP: 42/42

OK

gust

Lv. 21

PIDGEOTTO

NORMAL FLYING

HP: 59/59

OK

SPIKE

Lv. 20

NIDDOKING

POISON GROUND

HP: 70/70

OK

STEP COUNT
120990

Claude Opus 4 Plays Pokémon

OPUS
EDITION

current position to see Claude Opus 4 different route.

Using tool: navigator - target: (0, 23)

I'm at (6,23) and I'm on the west side of the gym now. I can see the gym building to my right. Looking at the coordinate overlay, I still see that all coordinates below y=24 are marked as red (non-navigable), including the area directly in front of the gym.

This is really puzzling. The gym entrance should be accessible, but I can't find any path south from y=23.

Let me think about this more carefully:

1. The gym entrance is supposed to be at (10,27)
2. From multiple positions at y=23, I cannot go south past y=24
3. All areas below y=24 show as non-navigable

Wait! Maybe I need to look for a different approach entirely. What if there's a specific tile or gap that I need to find? Let me try moving to different x coordinates at y=23 to see if any of them have a path south.

Using tool: use_emulator - Buttons: ['right', 'right', 'down', 'wait']

Thinking...



Current Team

SPLASH BLASTOISE HP: 150/150 Lv. 46 WATER OK	wings SPEAROW HP: 43/43 Lv. 16 NORMAL FLYING OK	luna CLEFAIRY HP: 42/42 Lv. 13 NORMAL OK	DUX FARFETCHD HP: 42/42 Lv. 14 NORMAL FLYING OK	gust PIDGEOTTO HP: 59/59 Lv. 21 NORMAL FLYING OK	SPIKE NIDOKING HP: 70/70 Lv. 20 POISON GROUND OK
--	---	--	---	--	--

STEP COUNT
120991

Claude Opus 4 Plays Pokémon

OPUS
EDITION

need to find? Let me Claude Opus 4 different x coordinates at y=23 to see if any of them have a path south.

Using tool: use_emulator - Buttons: ['right', 'right', 'down', 'wait']

I'm at (8,23) and the down movement didn't work - I'm still at y=23. The golden door is clearly visible at the bottom of the gym, but I can't reach it. This is becoming increasingly puzzling.

Looking at the visual, I can see:

1. The gym building with "GYM" written on it
2. The golden door rectangles at the bottom (these should be the entrance)
3. All coordinates below y=24 showing as red (non-navigable)

Wait! I just noticed something in the visual. The golden door at the bottom of the gym appears to be at around x=11 or x=12, not x=10 as stated in my knowledge base. Let me check if the entrance might be at a different x coordinate than I thought.

Using tool: navigator - Target: (11, 23)

Thinking...



Current Team

SPLASH	Lv. 46	wings	Lv. 16	luna	Lv. 13	DUX	Lv. 14	gust	Lv. 21	SPIKE	Lv. 20
BLASTOISE	WATER	SPEAROW	NORMAL FLYING	CLEFAIRY	NORMAL	FARFETCHD	NORMAL FLYING	PIDGEOTTO	NORMAL FLYING	NIDOKING	POISON GROUND
HP: 150/150	OK	HP: 43/43	OK	HP: 42/42	OK	HP: 42/42	OK	HP: 59/59	OK	HP: 70/70	OK

STEP COUNT
120991

Claude Opus 4 Plays Pokémon

OPUS
EDITION

check if the entrance **Claude Opus 4** different x coordinate than I thought.

Using tool: navigator - Target: (11, 23)

I'm at (11,23) and looking at the visual, I can see the golden door is directly below me. But again, all coordinates below y=24 show as red (non-navigable). This is extremely puzzling. I've tried:

1. Multiple x coordinates along y=23
2. Going around to the east and trying to find an alternative path
3. Manual movement commands

Nothing has worked to get me past y=23. There's clearly something I'm missing about how to access the gym. Let me think about this differently. What if there's something special I need to do? Maybe there's an NPC I need to talk to first, or maybe the gym is closed for some reason? I see an NPC in the upper right at around (14,20). Let me go talk to them to see if they give any clues.

Using tool: navigator - Target: (14, 20)

Thinking...



Current Team

SPLASH	Lv. 46	wings	Lv. 16	luna	Lv. 13	DUX	Lv. 14	gust	Lv. 21	SPIKE	Lv. 20
BLASTOISE	WATER	SPEAROW	NORMAL FLYING	CLEFAIRY	NORMAL	FARFETCHD	NORMAL FLYING	PIDGEOTTO	NORMAL FLYING	NIDOKING	POISON GROUND
HP: 150/150	OK	HP: 43/43	OK	HP: 42/42	OK	HP: 42/42	OK	HP: 59/59	OK	HP: 70/70	OK

STEP COUNT
120993

Claude Opus 4 Plays Pokémon

OPUS
EDITION

3. Manual movement for Claude Opus 4

Nothing has worked to get me past y=23. There's clearly something I'm missing about how to access the gym. Let me think about this differently. What if there's something special I need to do? Maybe there's an NPC I need to talk to first, or maybe the gym is closed for some reason? I see an NPC in the upper right at around (14,20). Let me go talk to them to see if they give any clues.


Using tool: navigator - Target: (14, 20)

I'm at (14,20) and there's an NPC directly below me at (14,21). I can see on the coordinate map that (14,21) is marked as red (non-navigable), which means the NPC is there. Let me talk to them by facing down and pressing 'a'.

Using tool: use_emulator - Buttons: ['down', 'a', 'wait']

The NPC is talking about the Game Corner, not the gym. Let me continue the dialog to see if they say anything useful.

Using tool: use_emulator - Buttons: ['a', 'wait']

 Cleaning up my context



Current Team

SPLASH

Lv. 46

BLASTOISE

WATER

HP: 150/150

OK

wings

Lv. 16

SPEAROW

NORMAL FLYING

HP: 43/43

OK

luna

Lv. 13

CLEFAIRY

NORMAL

HP: 42/42

OK

DUX

Lv. 14

FARFETCHD

NORMAL FLYING

HP: 42/42

OK

gust

Lv. 21

PIDGEOTTO

NORMAL FLYING

HP: 59/59

OK

SPIKE

Lv. 20

NIDOKING

POISON GROUND

HP: 70/70

OK



ClaudePlaysPokemon

Claude Plays Pokemon - Claude 4 Launch! (Ireset !harness !crash)

Playing Pokémon Red/Blue for 38 viewers

Follow

LIVE



STEP COUNT
120995

Claude Opus 4 Plays Pokémon



Looking at the feedback Claude Opus 4 situation:

1. I'm at (14,20) in Celadon City, having just talked to an NPC who said "is bad for our city's image!" (likely about the Game Corner)
 2. The feedback suggests several important points:
 - The game is NOT bugged, Erika's gym is always accessible with 3 badges
 - I might be looking at the wrong building or approaching incorrectly
 - I should talk to more NPCs and explore other buildings
 - My Level 46 Blastoise would easily handle the gym (not the Level 21 Pidgeotto)
 3. The feedback also suggests I should update my knowledge base to:
 - Remove redundant information about failed attempts
 - Consolidate navigation tips
 - Add sections for NPC interactions and building access
 - Update my team strategy to reflect using Blastoise
- Let me start by updating my knowledge base based on this feedback, then continue exploring Celadon City.

Using tool: string_replace

Thinking...



UPDATING KNOWLEDGE BASE

1. Check if gym has alternate entrance (sides/back)
 2. Look for NPCs who might give hints about gym access
 3. Explore other Celadon locations - maybe need item/HM?
 4. Try entering from different zones/areas
- **Battle Strategy**: Lead with gust (Pidgeotto L21)
 - Flying beats Grass
 - **Team Status**: All Pokemon full HP, 8 Super Potions ready

Current Team

SPLASH	Lv. 46	wings	Lv. 16	luna	Lv. 13	DUX	Lv. 14	gust	Lv. 21	SPIKE	Lv. 20
BLASTOISE	WATER	SPEAROW	NORMAL FLYING	CLEFAIRY	NORMAL	FARFETCH'D	NORMAL FLYING	PIDGEOTTO	NORMAL FLYING	HIDOLING	POISON GROUND
HP: 150/150	OK	HP: 43/43	OK	HP: 42/42	OK	HP: 42/42	OK	HP: 59/59	OK	HP: 70/70	OK

STEP COUNT
120999

Claude Opus 4 Plays Pokémon

OPUS
EDITION

Claude Opus 4

Good, I consolidated the navigation tips. Now let me simplify the type effectiveness chart to only keep what's relevant.

Using tool: string_replace

Good, I've simplified the type chart. Now let me add sections for NPC interactions and building access as suggested.

Using tool: string_replace

Good progress. Now let me clean up the lessons learned section to be more concise.

Using tool: string_replace

Thinking...



UPDATING KNOWLEDGE BASE

- Don't repeatedly try the same blocked path - explore alternatives
- Don't assume visual door location matches actual entrance coordinates
- Check if certain areas require completing other tasks first



SPLASH

Lv. 46

BLASTOISE

WATER

HP: 150/150

OK

wings

Lv. 16

SPEAROW

NORMAL FLYING

HP: 43/43

OK

luna

Lv. 13

CLEFAIRY

NORMAL

HP: 42/42

OK

DUX

Lv. 14

FARFETCHO

HP: 42/42

TO

HP: 59/59

Lv. 21

SPTKE

HP: 70/70

OK

Lv. 20

POISON GROUND

OK

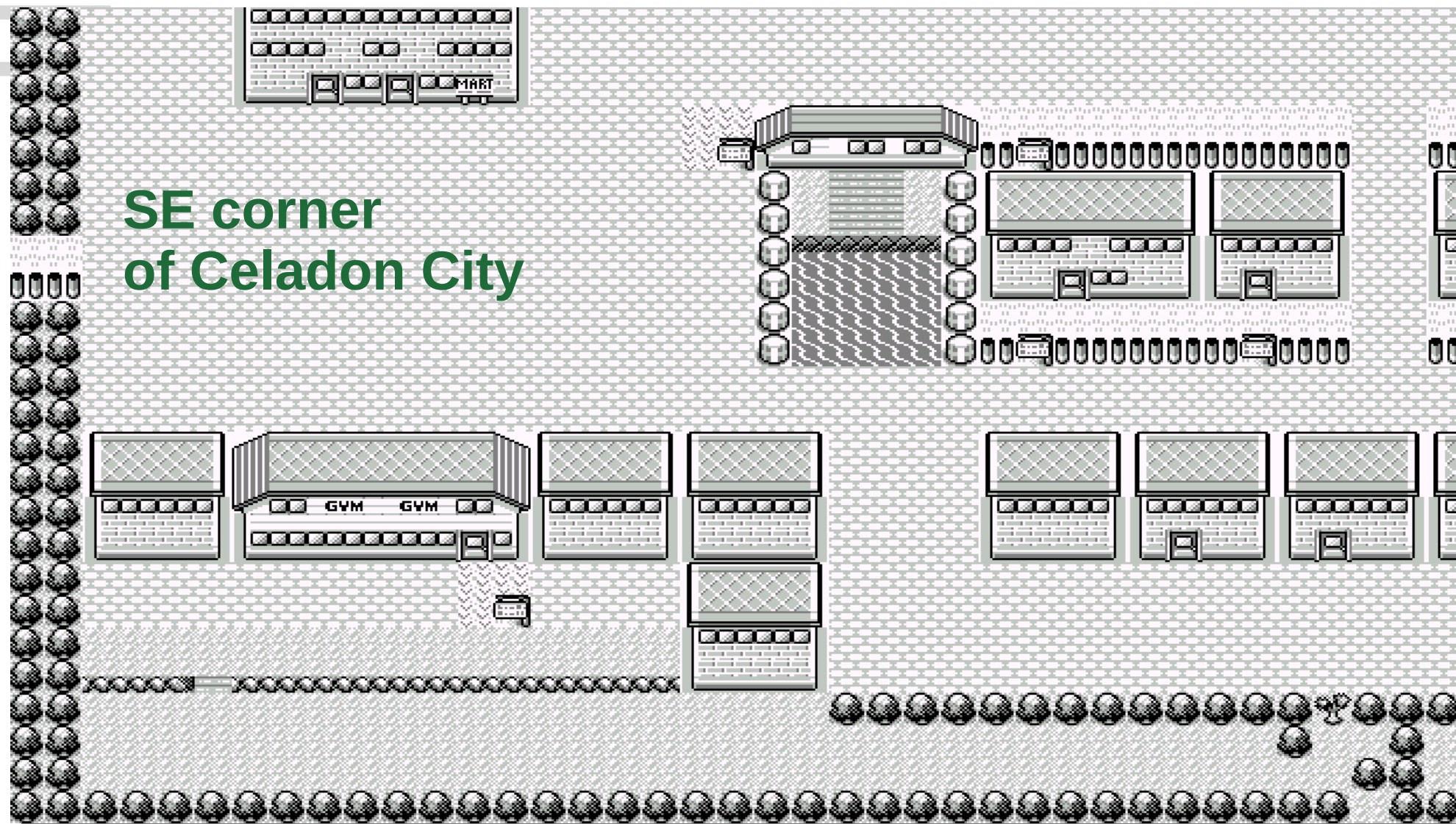


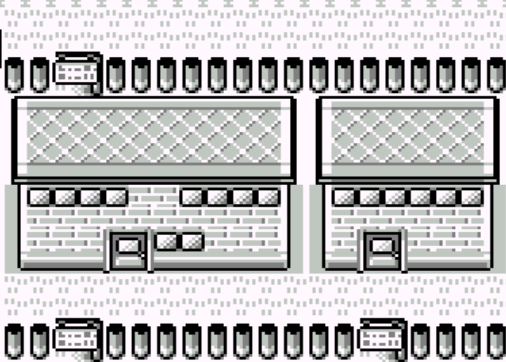
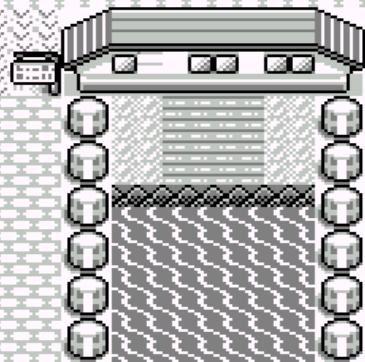
What's the game doing here?

Claude sure is confused

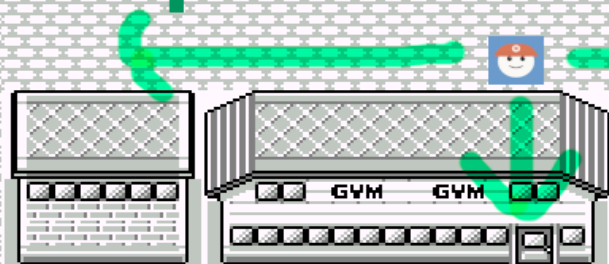


SE corner of Celadon City

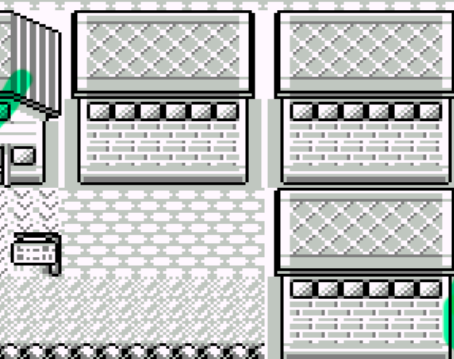




no path

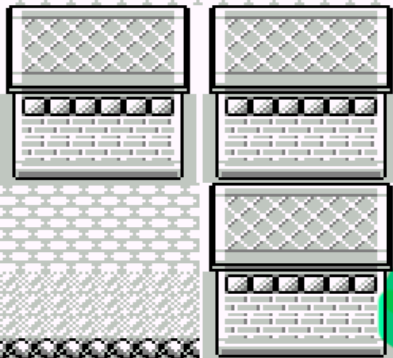
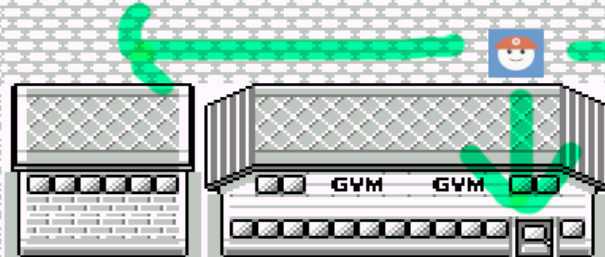
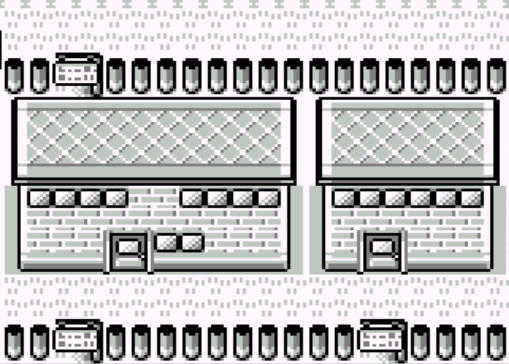
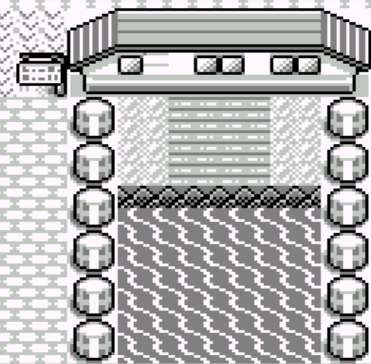


sees gym



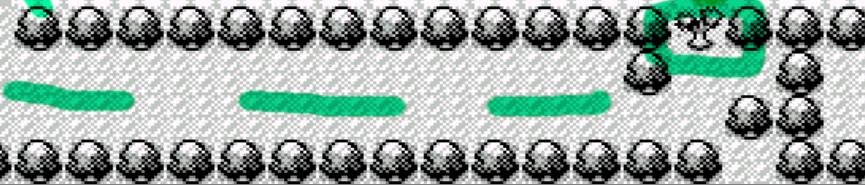
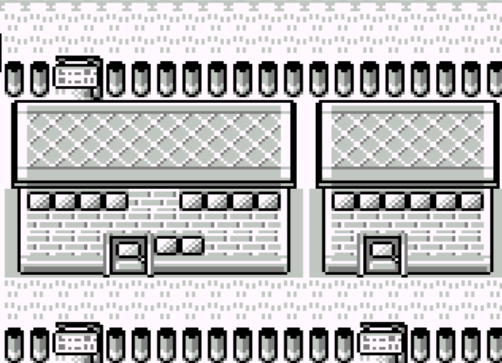
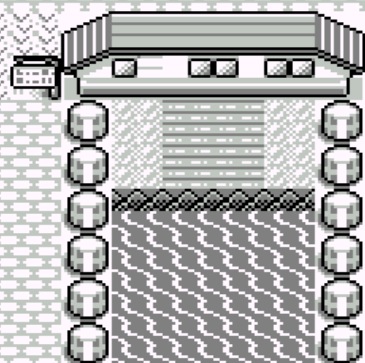
also no







& lost





Example: Claude 3.7 in Cerulean

End of a run

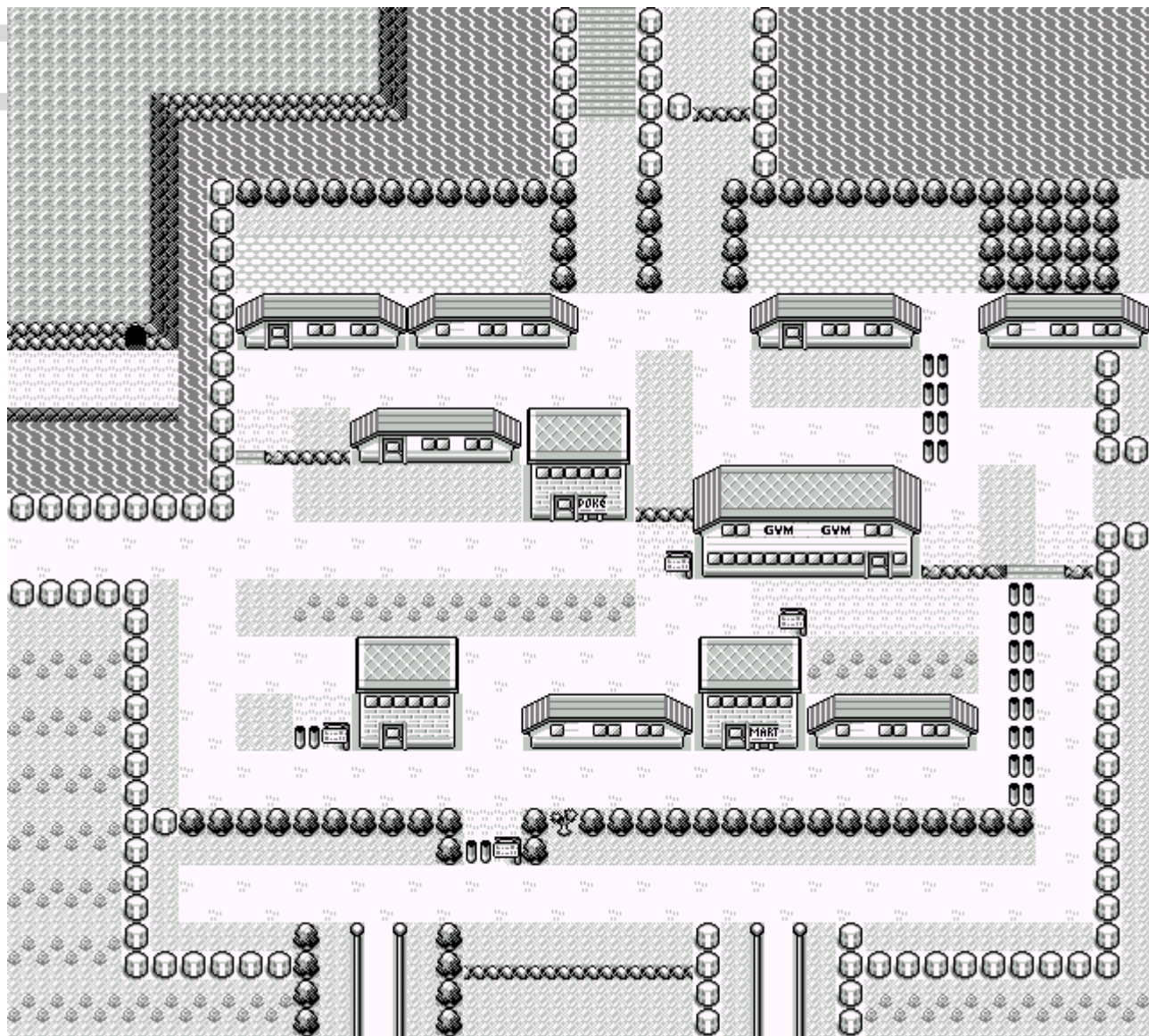


Think that's tricky? Try this:

An early Sonnet 3.7 run failed at Cerulean City (after beating Misty).

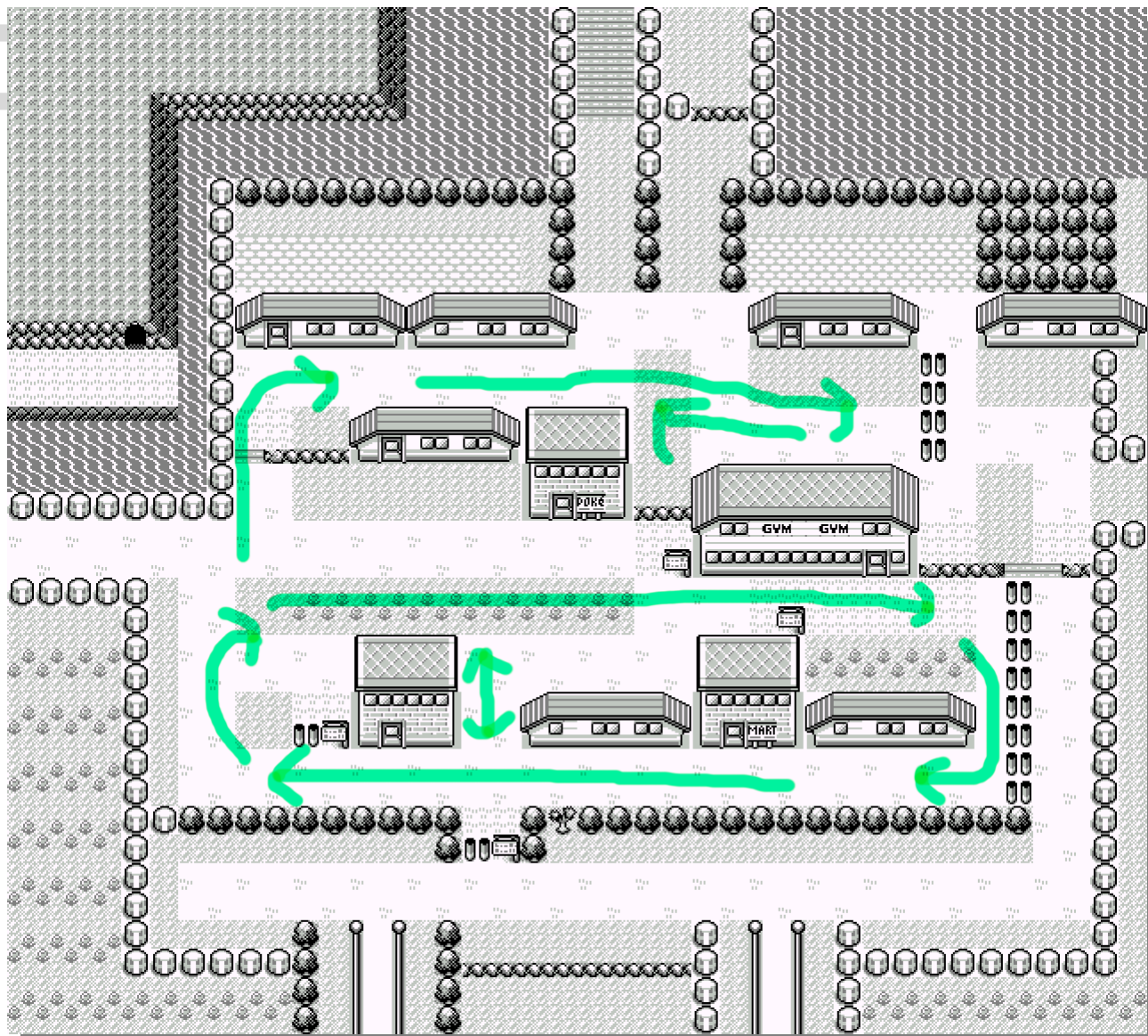
Look at the map...





A run ended here.
After the gym clear:

Trying to exit South ↓



A run ended here.
After the gym clear:

Trying to exit South ↓

But: going in circles!

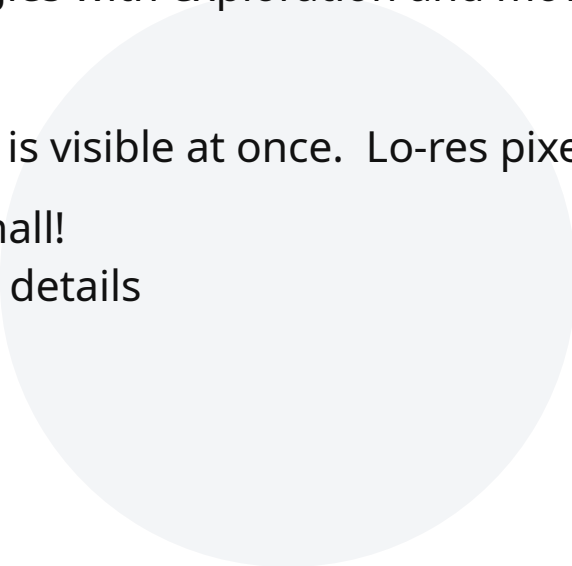
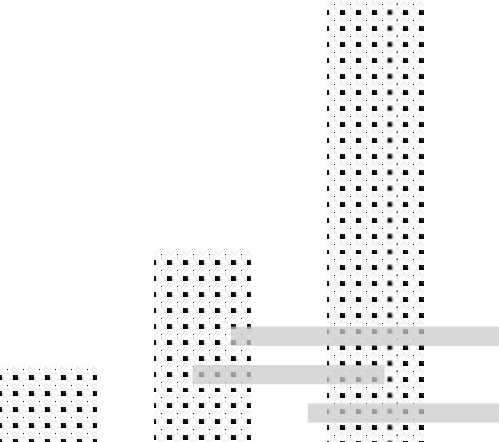
Can't find any progress
... for a week or so.



Game is from the 90's!

Navigation & exploration is a real challenge

Claude really, really struggles with exploration and movement and spatial awareness.

- Screen limits how much is visible at once. Lo-res pixel art: not helping
 - Context size is pretty small!
 - & context clear loses details
- 
- 



Overall results: mixed

Significant exploration and navigation challenges



Anthropic were too optimistic / overhyped it.

Claudes > 3.7 (with that framework) only consistently get to Mt. Moon...

Runs vary wildly!

Big visible public failures

... & small incremental successes.

(Overall project: influential, fascinating. Worth doing!)



Failures

Quite a lot went wrong, visibly and publically



- Runs getting stuck, going in circles, needing restart.
- Unreasonable time taken with Viridian Forest, Mt. Moon. Spatial reasoning is bad.
- Lots of 'small hallucinations' — walking into walls etc.
- Hallucinations not recovered from if in critique model, or knowledgebase
- Unable to vary strategy well for exploration.
- Context too short, loses track of things.
- Notes used badly at times, deleted inappropriately.

Interview with the dev:

Talked to Ars Technica — dev optimism despite visible struggles

Example: at 1 month in,
Kyle Orland at Ars Technica:
“Why Anthropic’s Claude
still hasn’t beaten Pokémon”

Interviews dev.

Article framing is negative.
— & comments / discussion:
brutal, pessimistic.



But relative success over previous models is not the same as absolute success over the game in its entirety. In the weeks since Claude Plays Pokémon was first made public, thousands of Twitch viewers have watched Claude struggle to make consistent progress in the game. Despite long "thinking" pauses between each move—during which viewers can read printouts of the system's simulated reasoning process—Claude frequently finds itself pointlessly revisiting completed towns, getting stuck in blind corners of the map for extended periods, or fruitlessly talking to the same unhelpful NPC over and over, to cite just a few examples of distinctly sub-human in-game performance.

Watching Claude continue to struggle at a game designed for children, it's hard to imagine we're witnessing the genesis of some sort of computer superintelligence. But even Claude's current sub-human level of Pokémon performance could hold significant lessons for the quest toward generalized, human-level artificial intelligence.

In addition to directly monitoring certain key (emulated) Game Boy RAM addresses for game state information, Claude views and interprets the game's visual output much like a human would. But despite recent advances in AI image processing, Hershey said Claude still struggles to interpret the low-resolution, pixelated world of a Game Boy screenshot as well as a human can. "Claude's still not particularly good at understanding what's on the screen at all," he said. "You will see it attempt to walk into walls all the time."

Hershey said he suspects Claude's training data probably doesn't contain many overly detailed text descriptions of "stuff that looks like a Game Boy screen." This means that, somewhat surprisingly, if Claude were playing a game with "more realistic imagery, I think Claude would actually be able to see a lot better," Hershey said.

"It's one of those funny things about humans that we can squint at these eight-by-eight pixel blobs of people and say, 'That's a girl with blue hair,'" Hershey continued. "People, I think, have that ability to map from our real world to understand and sort of grok that... so I'm honestly kind of surprised that Claude's as good as it is at being able to see there's a person on the screen."


Beyond issues parsing text and images, Hershey also acknowledged that Claude can have trouble “remembering” what it has already learned. The current model has a “context window” of 200,000 tokens, limiting the amount of relational information it can store in its “memory” at any one time. When the system’s ever-expanding knowledge base fills up this context window, Claude goes through an elaborate summarization process, condensing detailed notes on what it has seen, done, and learned so far into shorter text summaries that lose some of the fine-grained details.

This can mean that Claude “has a hard time keeping track of things for a very long time and really having a great sense of what it's tried so far,” Hershey said. “You will definitely see it occasionally delete something that it shouldn't have. Anything that's not in your knowledge base or not in your summary is going to be gone, so you have to think about what you want to put there.”

More than forgetting important history, though, Claude runs into bigger problems when it inadvertently inserts incorrect information into its knowledge base. Like a conspiracy theorist who builds an entire worldview from an inherently flawed premise, Claude can be incredibly slow to recognize when an error in its self-authored knowledge base is leading its Pokémon play astray.

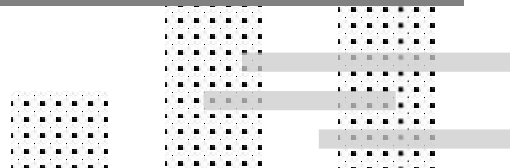
“The things that are written down in the past, it sort of trusts pretty blindly,” Hershey said. “I have seen it become very convinced that it found the exit to [in-game location] Viridian Forest at some specific coordinates, and then it spends hours and hours exploring a little small square around those coordinates that are wrong instead of doing anything else. It takes a very long time for it to decide that that was a ‘fail.’”

Still, Hershey said Claude 3.7 Sonnet is much better than earlier models at eventually “questioning its assumptions, trying new strategies, and keeping track over long horizons of various strategies to [see] whether they work or not.” While the new model will still “struggle for really long periods of time” retrying the same thing over and over, it will ultimately tend to “get a sense of what’s going on and what it’s tried before, and it stumbles a lot of times into actual progress from that,” Hershey said.



Whatever you think about impending improvements in AI models, though, Claude's current performance at Pokémon doesn't make it seem like it's poised to usher in an explosion of human-level, completely generalizable artificial intelligence. And Hershey allows that watching Claude 3.7 Sonnet get stuck on Mt. Moon for 80 hours or so can make it "seem like a model that doesn't know what it's doing."

But Hershey is still impressed at the way that Claude's new reasoning model will occasionally show some glimmer of awareness and "kind of tell that it doesn't know what it's doing and know that it needs to be doing something different. And the difference between 'can't do it at all' and 'can kind of do it' is a pretty big one for these AI things for me," he continued. "You know, when something can kind of do something it typically means we're pretty close to getting it to be able to do something really, really well."





Trees

“You know, when something can kind of do something it typically means we're pretty close to getting it to be able to do something really, really well.”

Uhhh what? This is exact inverse of how technology development works. Getting something to kinda sorta work sometimes is like the first 2% of the effort, maybe less.

The hubris of these people is just mind blowing to me. It's like they have literally no idea what they are doing, from a tech (nevermind product) development standpoint.

March 21, 2025 at 3:24 pm

Do you agree or disagree?



Trees

“You know, when something can kind of do something it typically means we're pretty close to getting it to be able to do something really, really well.”

Uhhh what? This is exact inverse of how technology development works. Getting something to kinda sorta work sometimes is like the first 2% of the effort, maybe less.

The hubris of these people is just mind blowing to me. It's like they have literally no idea what they are doing, from a tech (nevermind product) development standpoint.

March 21, 2025 at 3:24 pm

Do you agree or disagree?

(Disagree: getting *anything* = hard in modern ML, scaling it = simple.)

LessWrong replication

Replication showed framework necessity

JulianBradshaw @ LW used the (open) ClaudePlaysPokemon framework to replicate it. Very pessimistic about the results.

So a recent post* my friend wrote has made the point quite clearly (I hope) that LLM performance on the simple task of playing and winning a game of Pokémon Red is *highly* dependent on the scaffold and tooling provided. In a way, this is not surprising—the scaffold is there to address limitations in what the model can do, and paper over things like lack of long-term context, executive function, etc.

But the thing is, I *thought* I knew that, and then I actually tried to run Pokémon Red.

LessWrong replication

Costs were high

Costs

Obviously this cost a decent sum of money, though not ridiculously so. It's about \$30 or so per 2000 steps, which for comparison is typically enough for the average model to pick a starter and get to the next city to pick up Oak's parcel, or enough to make a decent go at getting through Viridian forest. This is also a few hours of time.

I... wouldn't consider this cheap, which is part of why I've spent a lot of time experimenting on what helps models progress rather than just setting it to run.

Miscellanea: ClaudePlaysPokemon Derp Anecdotes

For those who are not familiar with the deep lore, here is a small supercut of navigation/reasoning failures that may inspire some thought:

1. Getting stuck in the opening bedroom because it can't see the stairs down
2. Getting stuck in Viridian forest maze for over a day
3. Getting stuck going through Mt. Moon for 69 hours (the first time, this is foreshadowing) because, among other things, it believed there was a path to Vermilion City that *didn't* go through Mt. Moon and spent long hours looking for it in a blank rock wall.
4. Just navigating Cerulean City is a challenge because of the city's... unique design.
5. Marking the Trashed House as "explored" despite never exploring the back half and therefore failing to find the exit to the Underground path for days.

6. Confusing another NPC for Misty, and after not talking to Misty believing that it was missing some critical step to get a gym badge.
7. Asserting that Cerulean City Gym is the "Trashed House" and that the game's RAM must be mistaken
8. Literal days of failing to find the SS Anne
9. ...augmented by an inability to walk *down* into the SS Anne because it was instructed to walk "up the gangplank" and wrote down that it needs to try pressing up when near the SS Anne
10. Repeatedly leaving the SS Anne after talking to various NPCs and becoming convinced that what they were talking about was critical for progression
11. Convincing himself that a (several) random NPCs in the SS Anne were the captain and trying to talk to them for hours at a time, hoping to find the secret way to make them hand off HM01 Cut.
12. Currently, ClaudePlaysPokemon is an eternal death loop of trying to traverse Mt. Moon, getting stuck, using Diglett to dig back to the start, repeat. This has happened >100 times.

Restarts and doom loops

The Mt. Moon loop: no boredom, spatial thought, randomization

Claude doesn't get bored, frustrated, etc... but boredom is adaptive!

Claude in the Claude-plays-pokemon agent framework can't track space
— loses track of locations, fails to model environment, doesn't make connections
— really bad at exploration

Claude doesn't *randomize*.

Despite LLM being nondeterministic: some choices overdetermined!

Takes first path every time.

Restarts and doom loops

The Mt. Moon loop: no boredom, spatial thought, randomization

Mt. Moon is a maze.

Constant attacks by wild Zubat, Geode, interrupt motion & lose context

Have to go down, then back up to prior level then going down at different spot
— Claude can read current map ID from RAM, can see this is going backwards

No clear idea *which way* to go to progress. Gotta search it, map it out.

Restarts and doom loops

The Mt. Moon loop: no boredom, spatial thought, randomization

One run spent over 80 hours in Mt. Moon.

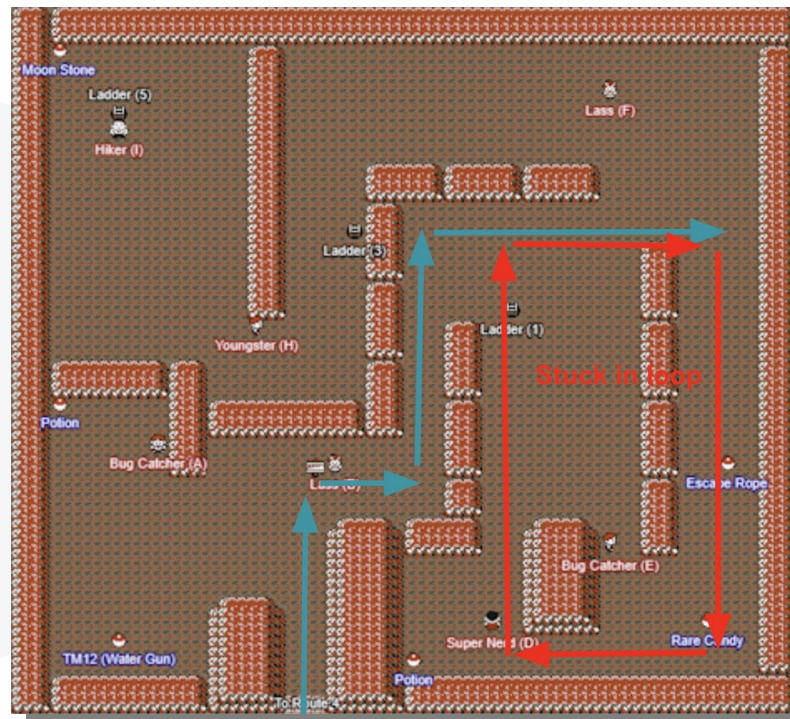
Going in a circle *just* big enough that the loop didn't fit in context.

Kept following path it was on
— making reasonable choices!

Just not randomizing, tracking space

Or getting bored!

(Thanks [@peterwildeford](#) on X)



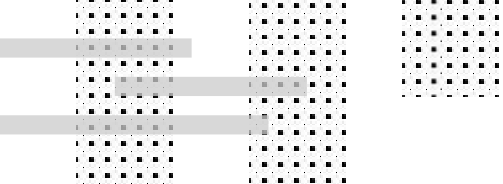
The Blackout Strategy

Too much self-trust in own tooling

Another severe loop developed while trying to get through Mt. Moon.

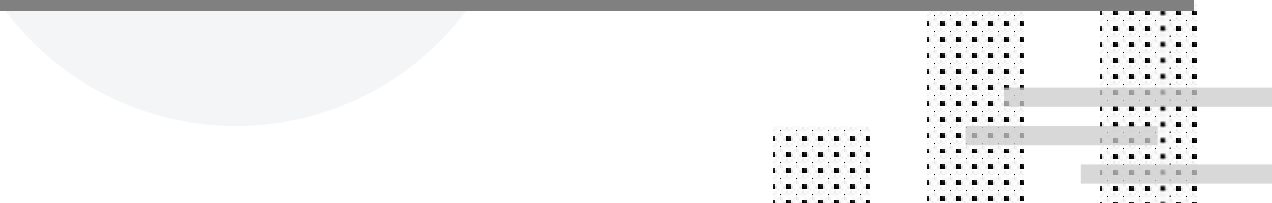
(Writeup from Justin Mills @ Quaternion Daydream)

Still, Claude might have gotten through eventually. Except that at one point, reasonably far in, it blacked out. This is a normal part of Pokemon; you black out when all your Pokemon faint, and you're sent back to the last Pokemon Center you've visited. It's basically a gentle "game over" state, that lets you try again.



For whatever reason, Claude determined that after the blackout, it respawned *past* Mt. Moon. There was no reason to believe this, but Claude's scaffolding kicked in and quickly enshrined this as a belief of great importance. Claude added "cleared Mt. Moon" to its list of *achieved* goals. Whenever it found itself back in Mt. Moon, it became convinced that it had backtracked, and needed to leave the way it came.

One thing I haven't mentioned yet is that Claude Plays Pokemon has a "critic" module, that chimes in to let it know if it's going off the rails. The critic module strongly reinforced that Claude had definitely beaten Mt. Moon, and should be able to advance directly to Cerulean City.



One thing I haven't mentioned yet is that Claude Plays Pokemon has a "critic" module, that chimes in to let it know if it's going off the rails. The critic module strongly reinforced that Claude had definitely beaten Mt. Moon, and should be able to advance directly to Cerulean City.

So now, every time Claude enters Mt. Moon, it tries to let all its Pokemon faint on purpose, to allow it to teleport to the Pokemon center and go straight to Cerulean. It has written a strategy document for this purpose, called "mt_moon_blackout_strategy", which it uses at the expense of every other document it has made. So far Claude has deliberately sacrificed its entire team of Pokemon at least 8 times in a row, without questioning the validity of the blackout strategy itself. Once outside Mt. Moon, it bumps into walls for a while, looking for the clear path to Cerulean it is sure must exist (but doesn't), then wanders into Mt. Moon and loses on purpose again.

The Blackout Strategy

Too much self-trust in own tooling

Every blackout, Claude celebrates, because it has woken up at the Pokemon Center, which is has confirmed is past Mt. Moon. Another win for the blackout strategy!

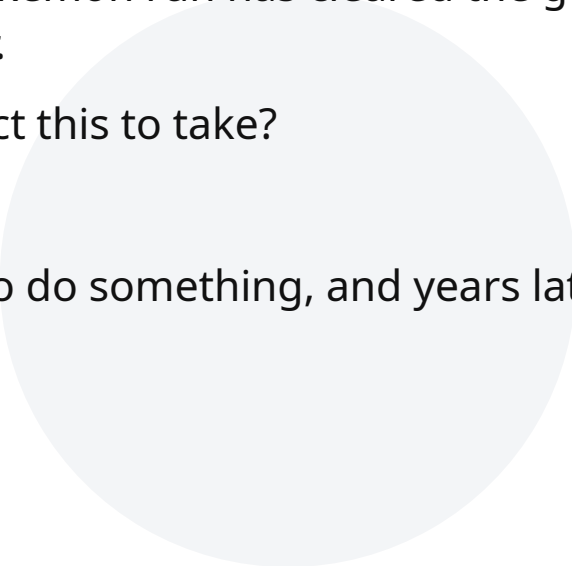
Moral of the story:

- Model self-critique is great right until it isn't. Hallucinations still a problem.
- Metacognition still not quite there yet.
- 'Authorative' info being very wrong can *really* mess up an agent
- No substitute for large models. Agent framework only goes so far.



Greatest failure: no game clear

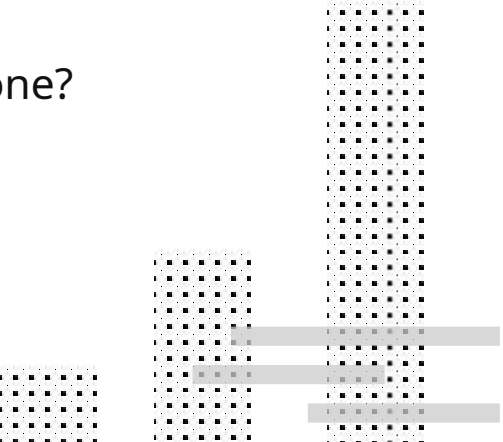
Claude has not, actually, played Pokémon Red to the end credits



No single Claude-Plays-Pokemon run has cleared the game.
It's been nearly half a year.

How long should we expect this to take?

What if you set an agent to do something, and years later it's not done?



Successes

While quieter, a huge amount went right



- Actually persistent, playing the game. On task & focused.
- Using tools fluently & consistently.
- Background knowledge solid, well used. Knows pokemon & gen 1.
- Decent game strategy. Choosing longterm paths & following them.
- Strong tactics: battle performance good. Not losing gym leader / rival fights.
- Real glimmers of metacognition. Can notice own failures, recover.

Successes

While quieter, a huge amount went right



Biggest successes have been outside the game

- Entertaining, fun. Easy to cheer on. Excellent personality.
- Kept online & running, upgraded. New Opus-4 version, new note-taking tools.
- Biggest success: copycats, & broader influence

Gemini Plays Pokemon

Google is entered into the ring

- Independant dev with no Google support, just curiosity & inference credits to burn
- Different philosophy. More a framework test than LLM test. Focused on game clear
- Much, much more tooling and agent support from the framework
 - tools: pathfinding, puzzle solver subagent, minimap, mark explored areas
- Tooling & prompts updated during run — no restart for updates

GEMINI PLAYS POKÉMON

• YELLOW LEGACY EDITION •

Actions: 46,319 GG: T-1 Summary: T--21

HOURS PLAYED

499h 28m 38s

MONEY

¥2,213

BADGES 5/8



POKÉDEX

42/151

BALLS

19 0 0 0 0 0

LOG

emoji map marker task from the AI.

TOKENS 313,848

IN 313,326 OUT 522

My last attack missed, but the Starmie was fully paralyzed again. This battle is a complete stalemate. My strategic objective is still a tactical retreat by intentionally losing. To do this, I must advance the turn so the Starmie can eventually attack and knock out my team. I am out of PP for **MEGA DRAIN**, and I now know that selecting a move with 0 PP does not trigger **STRUGGLE**. Therefore, I must use a move with PP remaining. **CONFUSION** is selected and available. I will use it to proceed with my plan to lose the battle.

Are you ever going to attack? Or are we just going to stand here while my attacks whiff past you forever? This is getting so old.

A

GEM SAYS

Are you ever going to attack? Or are we just going to stand here while my attacks whiff past you forever? This is getting so old.

GOALS

PRIMARY GOAL

Clear Team Rocket from the Silph Co. building in Saffron City.

SECONDARY GOAL

Level up my new team at the Seafoam Islands.

TERTIARY GOAL

Explore the rest of Route 19 and Route 20.

Route 19



(9, 43)

STARMIE
PAR

HP:



LEGION
:135

TYPE/
PSYCHIC
PP 16/25

91/ 91

CONFUSION
MEGA DRAIN
REFLECT
STUN SPORE

TEAM (6)

ACTIVE PC BOX (13)

INVENTORY (50)

PC ITEMS (7)

PAUSED

TITANESS

Lv36

KANGASKHAN

HP 32/128



LEGION

Lv35

EXEGGCUTE

HP 91/91



ECHO

Lv45

GOLBAT

HP 0/145



HELIX

Lv14

OMANYTE

HP 37/37



SPARKY

Lv42

PIKACHU

HP 0/114



CRAG

Lv41

GOLEM


HP 126/126





Gemini Plays Pokemon succeeds

Advantages to going second



Failures: similar to Claudes
— but Claude took the PR hit!

Sucesses: Google took it very seriously in Gemini 2.5 technical report
— 10% by volume about pokemon

Big success: Gemini cleared the game!



GEMINI PLAYS POKEMON

ACTIONS
106,505

GG: T-20 Summary: T-14

BADGES 8/8



GAME STATUS 813h 15m 39s MONEY: ¥49,054

A B (auto) B (auto) B (auto)

LOG

LOG #106,505

AGI GOALS

106,504 24,861 tokens used

106,504 The Full Restore was definitely the right play. Blastoise took the Solar Beam and is still standing at 216 HP, and the Defense drops are gone. Venusaur's health is in the red. Time to finish this. My water moves are depleted, leaving Strength (4 PP) and Bite (6 PP). Strength hits harder, so I'll use that. The cursor is on FIGHT, so I'll select it.

106,505 24,724 tokens used

106,505 Venusaur's Solar Beam hit, but Blastoise took it well thanks to the Full Restore. HP is at 216, which is fine. Venusaur's health is quite low now, in the red. My Hydro Pump and Surf are depleted. Strength has 4 PP left, and Bite has 6. Strength hits harder, so I'll use that. The cursor is already on Strength, so I just need to confirm. This should hopefully finish the battle.

OTHER GOALS

Keep Blastoise healthy to ensure victory against BLUE.

BATTLE STRATEGY:

Use Strength to defeat Venusaur.

CURRENT TEAM

BLASTOISE Lv66
BLASTOISE
HP 216/298

ZUBAT Lv37
ZUBAT
HP 105/105

BATMAN Lv8
ZUBAT
HP 26/26

ZAP Lv8
PIKACHU
HP 26/26

NIDORAN♀ Lv2
NIDORAN♀
HP 56/56

NIDORAN♀ Lv2
NIDORAN♀
HP 40/40

Champions Room



(5, 4)



RED: NO!
That can't

Pokedex: 26/151

Watch Gemini_Plays_Poke...
playing Pokémon Yellow V...
Pokedex

Gemini Plays Pokemon

Gemini 2.5 technical report takes Pokemon seriously

Strengths: general reasoning, “Long Horizon Task Coherence” / planning
— notable: escaping softlock bug caused by framwork IO limits. Not in pretrain!

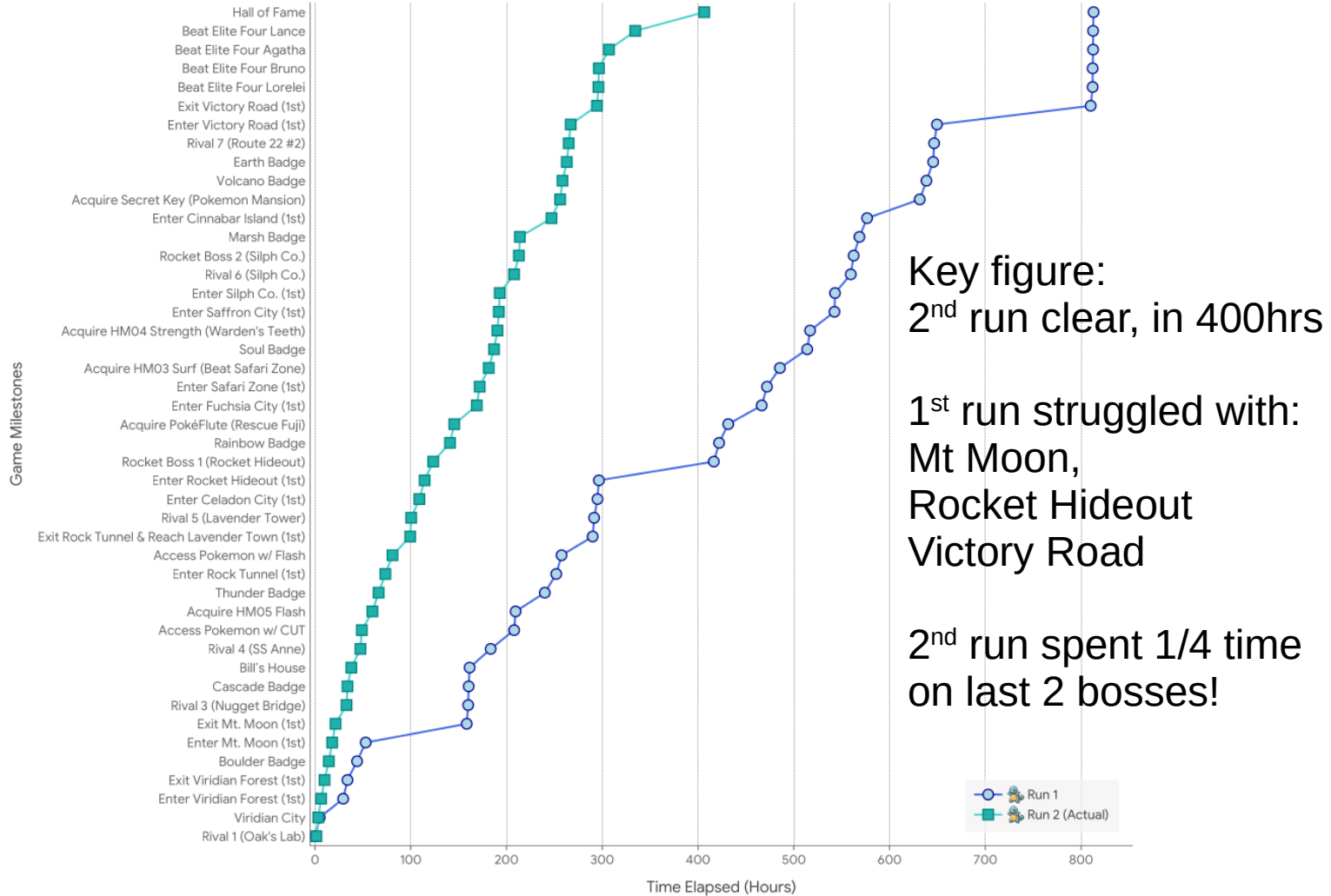
Failures: similar to Claude. Context poisoning, fixation on things visible nearby.

Notable failure: Agent panic.

- Bad situations cause fixation on escape, poor performance, failure to use tools
- model welfare concerns? Claude = having a good time. Gemini... not so much?

(Gemini 2.5 Pro sometimes gets suicidal when failing at coding. Related?)

Gemini 2.5 Pro Plays Pokemon Progress Timeline





o3 Plays Pokemon

The inevitable OpenAI/ChatGPT entry



Of course someone also did this with ChatGPT...

Currently playing Pokemon Crystal (Gen2).



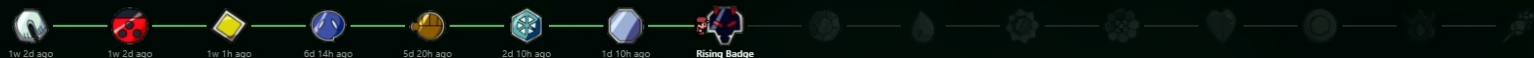
o3 Plays Pokémon Crystal

Total Runtime
259:33:10

Steps
13935

Next Criticism
T-1

Next Summary
T-36



16). Chance of encounter but we risk it. Segment 2: RIGHT, RIGHT – walk ground tiles (17, 16) then (18, 16). Segment 3: UP – step onto ice column, auto-slide north to ground at (18, 9). Segment 4: LEFT – slide west across row 9 to ground at (9, 9). Segment 5: DOWN, DOWN – walk to (9, 11) staircase warp. We'll send segment 1 now (single RIGHT). If battle triggers later, we'll run and continue remaining plan.

KEY PRESS

Completed



Sliding attempt #87! First a big drift: one RIGHT key—see you at column 16!

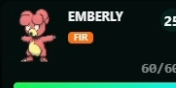
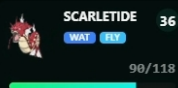
13935

REASONING

Planning movement strategy

I'm figuring out my path from current position (2, 16) towards my destination. Moving right to (3, 16) keeps me on the ground but actually, I hit ice at (4, 16). Sliding east to (16, 16) keeps me clear of any encounters as I pass ice tiles. My new approach is to: slide right to (16, 16), then to (17, 16), go up onto the ice at (17, 15), and finally maneuver further up. I want to confirm this mapped route to reach (9, 11).

... o3 is thinking... ...



OBJECTIVES

- 1 Exit Ice Path and arrive in Blackthorn City to challenge the Rising Badge gym.
- 2 Navigate remaining Ice Path floors using new Waterfall HM if needed, collect all visible items.
- 3 Train core team to level 40+ during Ice Path and upcoming Route 45.

Inventory

Memory

11864

shop_goldenrod2F_inventory:

"Dept. Store 2F – Upper clerk: Potions & status items. Lower clerk: Poké/Great Balls, Escape Rope, Repel, Revive, Poké Doll, Full Heal, Mail."

shop_mahogany_inventory:

"Mahogany souvenir shop sells: TinyMushroom P500, SlowpokeTail P9800, Poké Ball P200, Potion P300. No Repels or healing items."

tips_abort_path_after_collision:

"If a path sequence hits a collision, immediately stop sending remaining keys and recompute; don't repeat blocked moves."

tips_attack_menu_cursor:

"In the attack list the cursor stays on the last move used; always verify the highlight before pressing A."

IcePathB2FMahoganySide

[2,16]

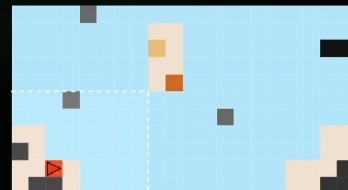


Figure 1 illustrates the experimental setup. A participant is seated at a table, looking at a screen. The screen displays a 4x4 grid of dots. A horizontal bar is positioned above the grid, and a vertical bar is positioned to the right of the grid. The participant's hand is shown reaching towards the vertical bar. The diagram is labeled with 'Participant', 'Screen', 'Grid', 'Bar', and 'Hand'.



How long is Pokemon?

Time to beat



Gemini is clear.

Second run with consistent tools: just over 400 hours.

Is that good? Bad? Average?



How long is Pokemon?

Time to beat — contextualised

Glitchless speedrun	Under 2 hours
Adult (focused)	~25 hours
Adult (average)	40-45 hours?
Adult (completionist)	~100 hours
Child	50-200 hours, depending?
Gemini 2.5 Pro	~410 hours (second run)
Claudes, various	Much longer.
Twitch chat (adversarial)	~390 hrs (= 16days 7hrs 45min)

'Glitchless'?

Gen 1 Pokémon can get weird

"Saving...
Don't turn off the game!"

What if you do?

Can powercycle while saving,
restarts with glitched menus.

Can then edit inventory
and go past end of menu,
toss specific count of glitch items
— this is editing values in RAM!

Pokémon Yellow *glitched* tool-assisted speedrun:
In-game timer at 00:00,
wallclock time ~minutes.

... for the runs that go for the credits.

Other runs start out playing pokémon,
make valid game inputs / button-presses,
end up playing Tetris. Or weirder!

It's ACE — Arbitrary Code Execution

Is Gemini better than Claude?

Do we have a winner on the Pokemon Gen1 eval?

Possibly. Model advantages for Gemini:

- Longer context
- Better vision

But: unfair comparison!

- Agent setup **very** different. Different tools available!
- Claude restarted for update & Gemini updated online mid-run.
- Gemini has *much* more help with navigation and exploration.

Gemini: what would get game clear? Claude: how far can raw LLM go?

Tooling matters

Key point of whole talk!



Gemini can do it and Claude can't. *Because* Gemini got help with 2d space!

- Minimap with visited-place tracking
- Pathfinding solver
- Puzzle solver for block puzzles

Current Gemini playing Yellow (Hard-mode mod) can:

- Build it's own tools
- Make subagents
- Talk to twitch chat!

Agents only work if they're fast

What makes an agent *finish* its tasks?

- Tasks get finished when agent is efficient.

'Taking forever' and 'not functional' have the same outcomes.

How long are you willing to wait? What if progress isn't steady?



Agents only work if they're fast

What makes an agent *finish* its tasks?

- Tasks get finished when agent is efficient.

'Taking forever' and 'not functional' have the same outcomes.

How long are you willing to wait? What if progress isn't steady?

- Agents are efficient when they're (reliably) correct.

Time is lost when waffling around with hallucinations, misunderstandings.

LLM correctness is... inconsistent, but slowly improving with new models.

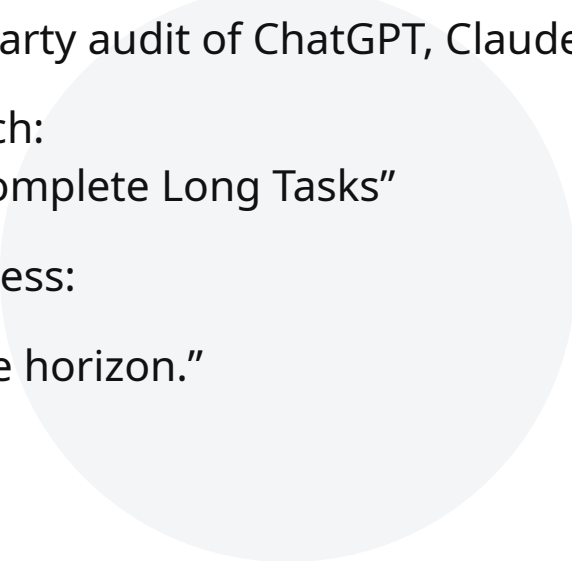
— tooling (RAG etc) can help here too





The Time-Horizons view

Agent evals are hard — METR intros a new metric



METR: “Model Evaluation & Threat Research”. AI eval non-profit in US.
Has been invited for 3rd-party audit of ChatGPT, Claude.

Research paper from March:
“Measuring AI Ability to Complete Long Tasks”

New metric for agent success:
“50%-task-completion time horizon.”

But what does that mean?

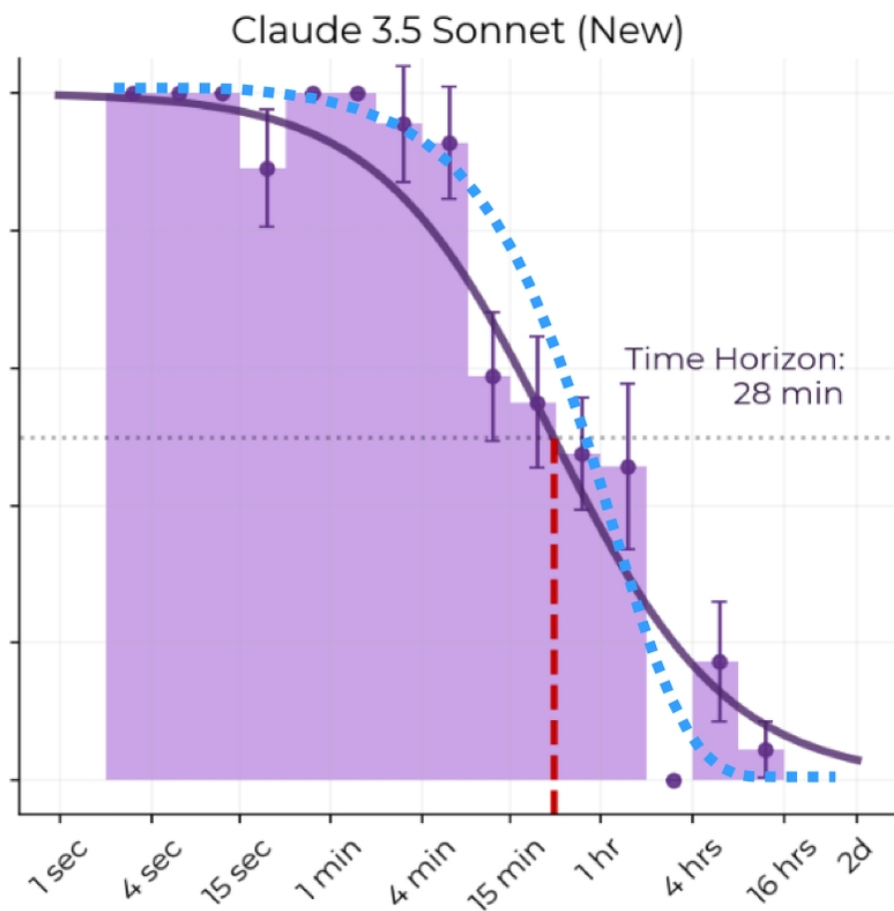
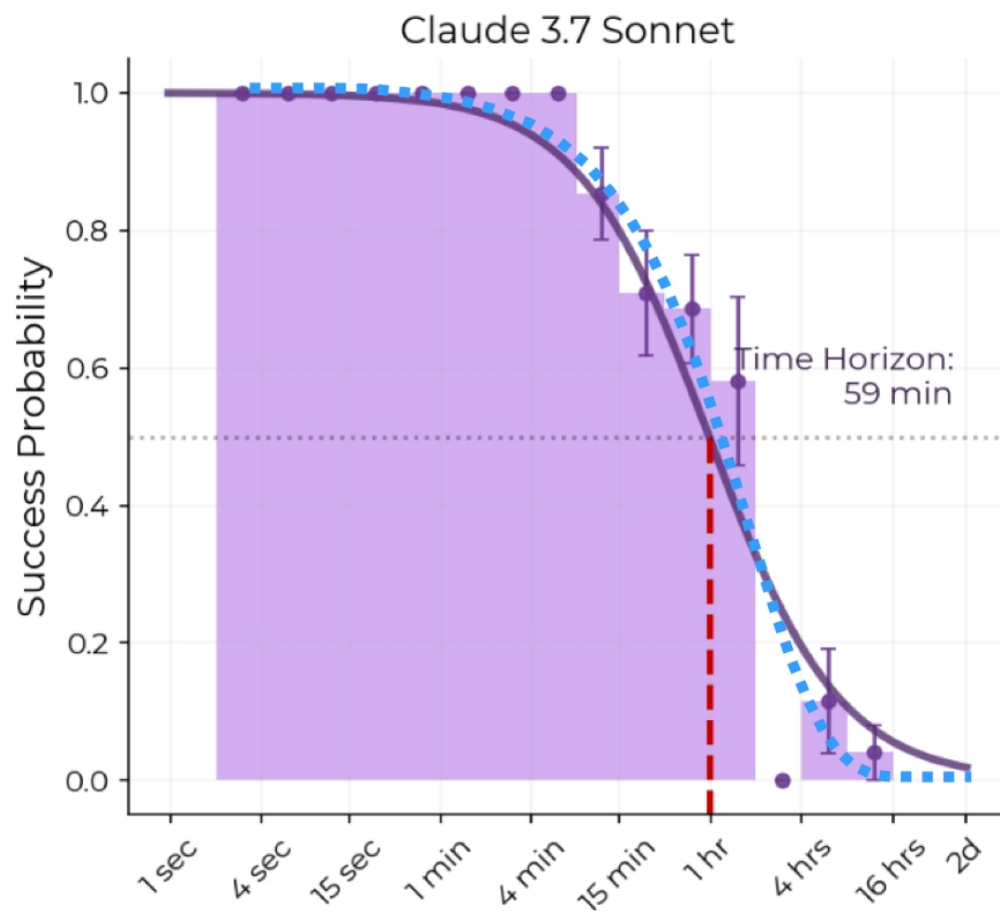


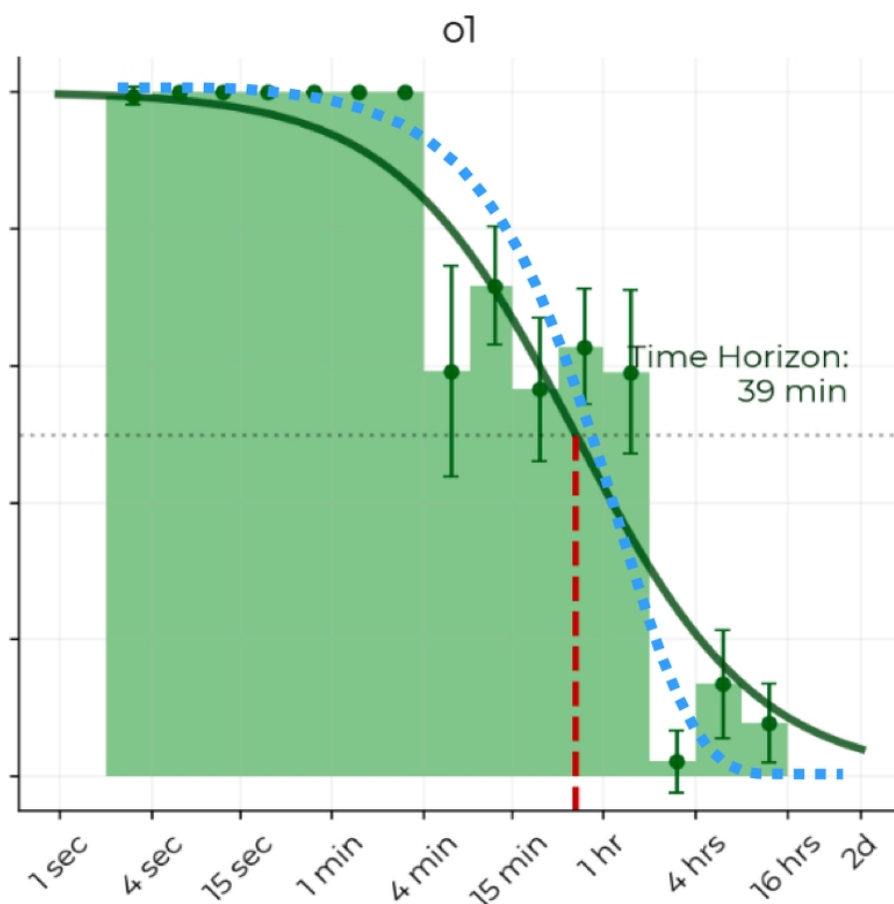
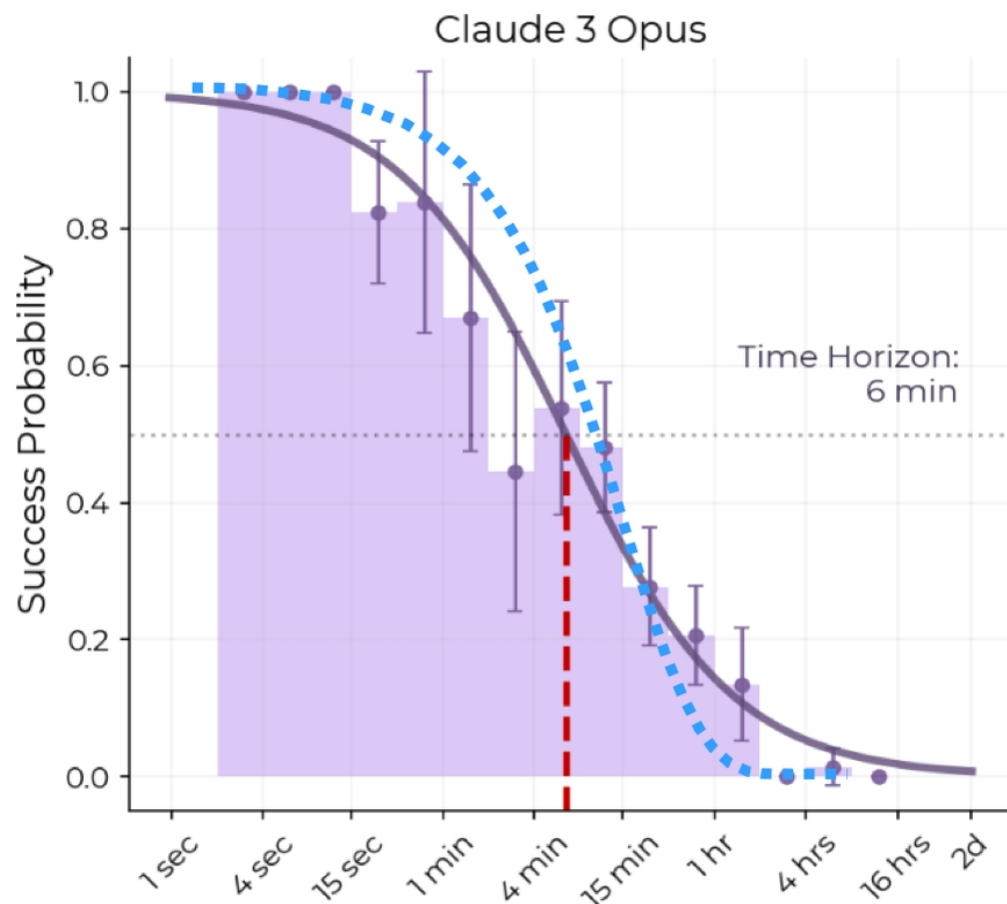
The Time-Horizons view

Explaining the metric

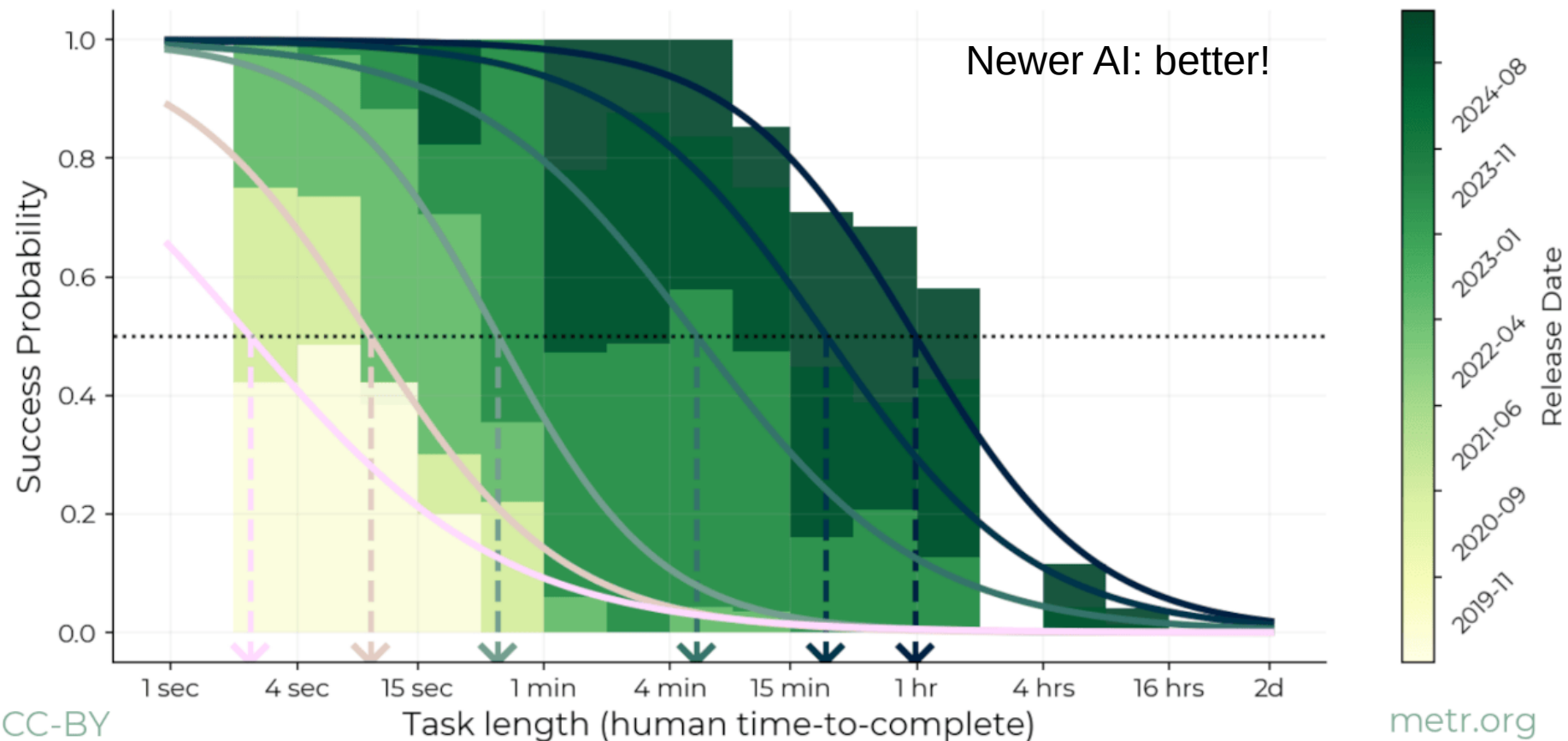
1. Get (trained) humans to complete software dev tasks, measure completion time — from quick queries to solving real issues.
2. Get different AIs / same agent-framework to do same tasks, count successes/failures
3. Sort the tasks by human completion time, graph AI success rate, fit logistic curve.
— logistic curves fit reasonably well
4. Show task-times at AI's 50% (or 80%, 95% etc) success-rate
— this is your '50%-task-completion time horizon' & measures agent quality

Still confusing? Check the graphs





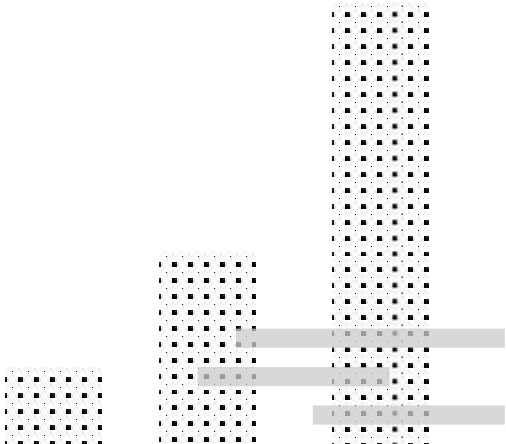

Models are succeeding at increasingly long tasks





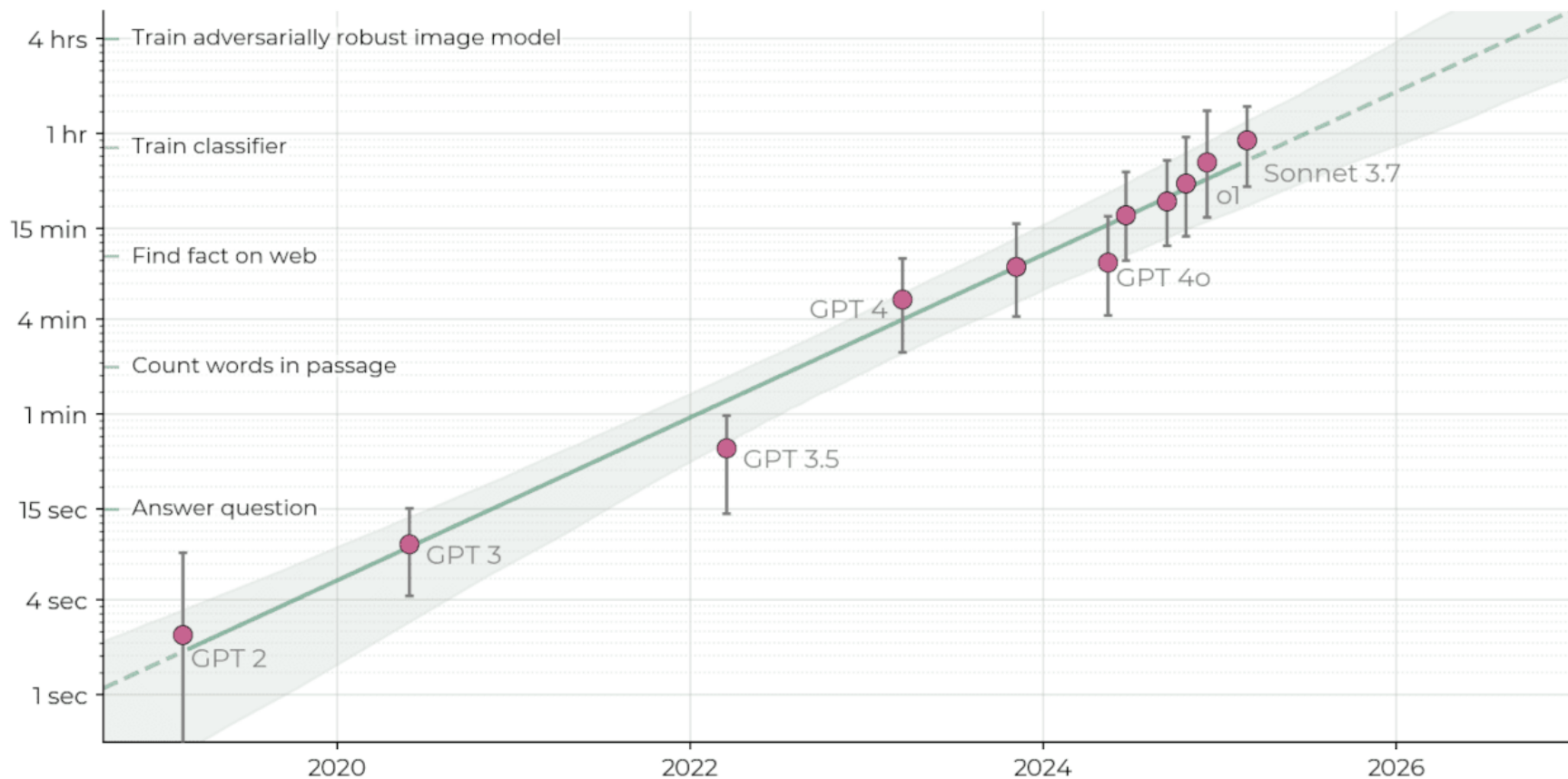
Straight lines on log graphs

A valid prediction method? It worked for Moore's law....



The length of tasks AI can do is doubling every 7 months

Task length (at 50% success rate)

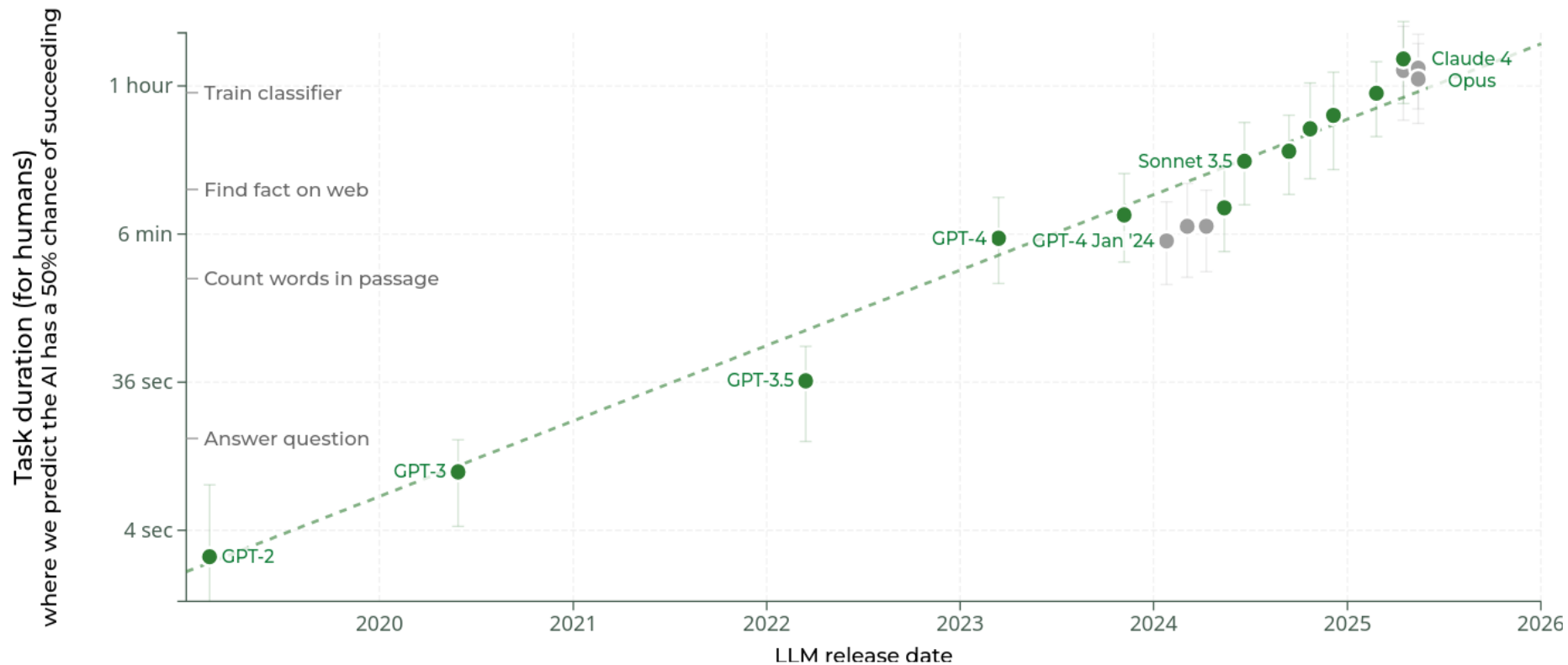


CC-BY

Model release date

metr.org

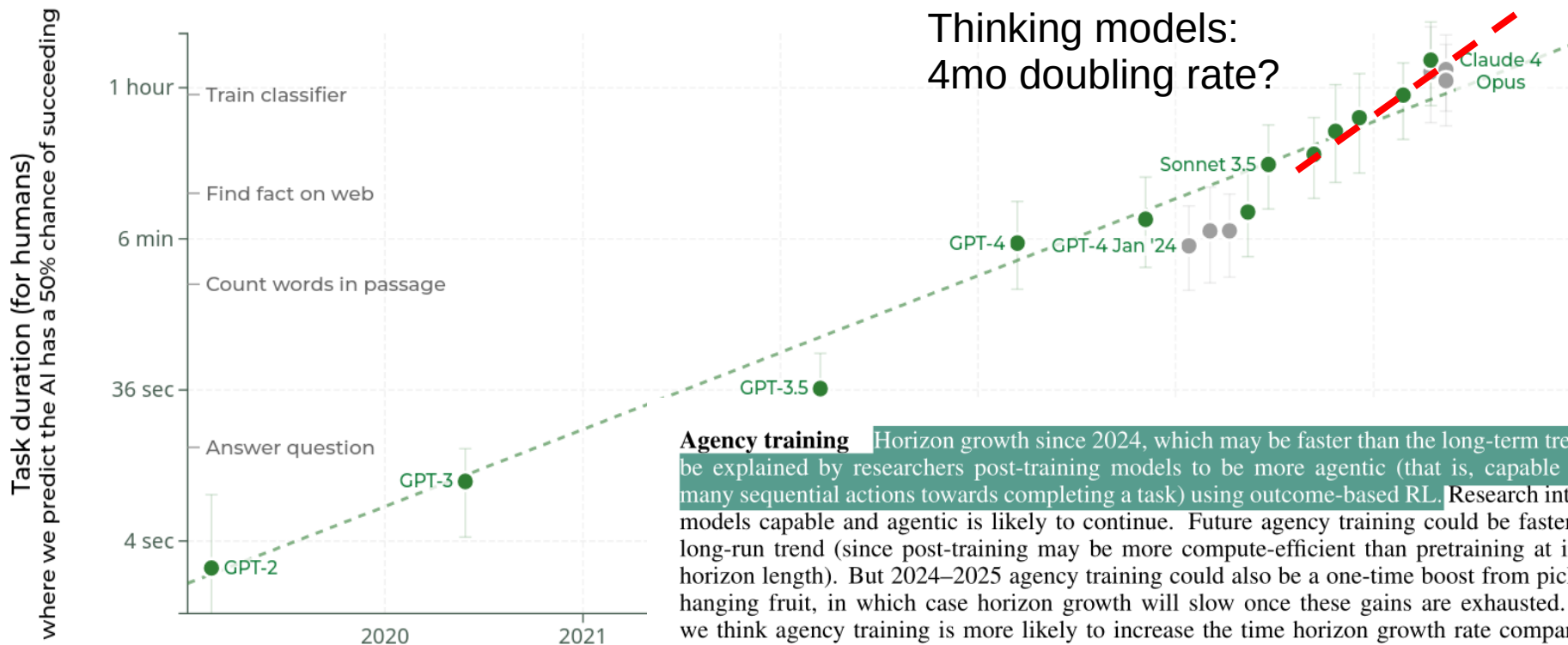
Time-horizon of software engineering tasks different LLMs can complete



Time-horizon of software engineering tasks different LLMs can complete



Thinking models:
4mo doubling rate?



Agency training Horizon growth since 2024, which may be faster than the long-term trend, could be explained by researchers post-training models to be more agentic (that is, capable of taking many sequential actions towards completing a task) using outcome-based RL. Research into making models capable and agentic is likely to continue. Future agency training could be faster than the long-run trend (since post-training may be more compute-efficient than pretraining at increasing horizon length). But 2024–2025 agency training could also be a one-time boost from picking low-hanging fruit, in which case horizon growth will slow once these gains are exhausted. Overall, we think agency training is more likely to increase the time horizon growth rate compared to the 2019–2024 trend.

4-8mo doubling rates are fast

Literally unsustainable, things get weird quickly

3x doublings / year has very strange implications on very short timelines
(Moore's law was 24 months.)

METR: Finally, we attempt to extrapolate the trend on our tasks to one-month (167 hours) AI (Section 7.1), finding that if both the trend continues and observed performance trends generalize to real-world tasks, an 80% confidence interval for the release date of AI that can complete 1-month long software tasks spans from late 2028 to early 2031 (Section 7.2).

1mo: In this section, we attempt to forecast when AIs will reach a 50%-time horizon of 1 month, because it intuitively seems that a system capable of this length of task, even at 50% reliability, would be transformative for our society, including potentially being proficient in capabilities that could threaten society with catastrophic harm.

Pokemon is hard to fail

What happens when the task doesn't stop?

If you can't lose progress or get stuck, what happens?

— forced into the “100% completion time horizon”? A long time!

A 40 hour game clear becomes 400 hours or more. Can't fail, so will get there.
Instead, measure time (& money) efficiency.

Right now:

10x worse time efficiency, but only with extensive & tested framework/tooling.

(Cost efficiency... people play Pokémon for fun. Unsuitable benchmark.)



Moravec's Paradox

Early AI researchers very confused by task difficulty



AI research was delusional in the 1960s. Very optimistic!

Computers could play chess, do logic, symbolic mathematics
— these are the mark of education and high intelligence!

So computers are easily going to be really smart. Right?





Moravec's Paradox

Early AI researchers very confused by task difficulty

AI research was delusional in the 1960s. Very optimistic!

Computers could play chess, do logic, symbolic mathematics
— these are the mark of education and high intelligence!

So computers are easily going to be really smart. Right?

Wrong!

Logic & Symbolic math: Need general intelligence....

... *because* humans **not** well-adapted to do them!

Computationally simple, compared with throwing a ball!



Moravec's Paradox

Early AI researchers very confused about task difficulty

“Where the rising ape meets the falling angel”

— Pratchett on nature of humanity

But: Computers coming at human-level, angel downwards!

Start as machines of pure logic, learn to roll in the mud.

Lots of things easy for children, impossible for machines.

Embodiment, agency, being in places.

LLM hype: big step towards humanity! Fluid language, humanlike failures.

— and now: struggling to play children's videogames.



How to make it work *now*?

What to do to get your agents going this year



My advice:

Seven steps for making agents work.





What should you to do make AI agents work?

1. Either accept sloppiness, or figure out validation

How will you know if it's working? Or not?

How good does it have to be?

Figure this out!

Ideal problems for agents are:

- Grindy, exhausting
- Need extensive knowledge
- Easy to verify! Answers get checked (or you don't care)
- On a computer!

(Don't make AI validate AI output – how will you validate the validation?)

(That doesn't work, or we're very early to the singularity.)



What should you to do make AI agents work?

2. Use a frontier SotA LLM with simulated reasoning.

Not worth trying to wrangle <8b local models

Needs SotA intelligence & ability to ruminate on the context

Agent reliability depends on **metacognition** for self-correction.

Exactly **four** companies/models worth considering:

OpenAI o3 (or o1, o4-mini. But *not* 4o!)

Anthropic Claude (Opus/Sonnet) 4

Google Deepmind Gemini 2.5 (Pro/Flash)

Deepseek R1

Building agents with small local Llama, Qwen, or Gemma finetunes? Good luck.



What should you to do make AI agents work?

3. Be ready to update your model

Frequent updates are making models better
Labs are targeting agentic uses right now

If you have private evals, custom framework/tooling, enough context:
You can slot in a smarter model, receive improvements for free*

Be ready when new capabilities drop.
Build your product and process for change!

* inference costs
may change

Consider OpenRouter, Replicate, possibly HuggingFace
— AI APIs *without* vendor lock-in.

(Keep an eye on Standard Completions (standardcompletions.org)
— nailing down “Open-AI compatible API” as a real standard!)



What should you to do make AI agents work?

4. Give your model the tools it needs

Code up helper utilities or other tools the model can call.

Figure out where the pain points are, and target tooling to them.
(Pokemon example, spatial reasoning \Rightarrow minimap with exploration tracking)

Have empathy for LLMs!

Consider pre-processing info, provide precomputed summary stats, etc.

Use MCP, RAG to add context. Delegation with A2A?

Get people (who're good at your tasks) to read the AI attempts, find failure points.
 \Rightarrow This will be boring and repetitive. Do it anyway.



What should you to do make AI agents work?

5. Use far more inference compute than is 'reasonable'

There are *many* chances to trade compute for success:

- * Trying many times and taking the best (pass@n) improves scores.
- * Self-consistency / self-critique passes help
- * Many-shot prompting techniques, as used for base models
- * Turn on reasoning mode & ramp the token count up
— (OpenAI 'high' models, Qwen /think, Claude 64k token budget)

Probably much more. Get creative!

(Start on *next* paradigm: agent swarms!)



What should you to do make AI agents work?

6. Use intermittent human oversight, limit unsupervised long-horizon tasks.

Gemini-Plays-Pokemon got hints. Claude-Code does checkins, etc.

Avoid *super* long context tasks with no feedback or correction!

You'll exceed the agent-reliability time horizon,
— then agents will get themselves stuck on a false belief or doom loop

Still need to ground them with occasional human feedback!
— AI feedback can help, but can also lock in mistakes.

Can trade off less agentic AI, for more sucessful AI.



What should you to do make AI agents work?

7. Just wait four-to-eighteen months then try again.

Agents failing? Pure scale issue.

You: industry-specific knowledge, internal datasets, integration with your business systems, full context of your operations... so build a custom AI?

-VS-

Frontier AI labs: scaling foundation models 10× bigger.

Who wins? (Hint: It's called 'the bitter lesson' for a reason.)

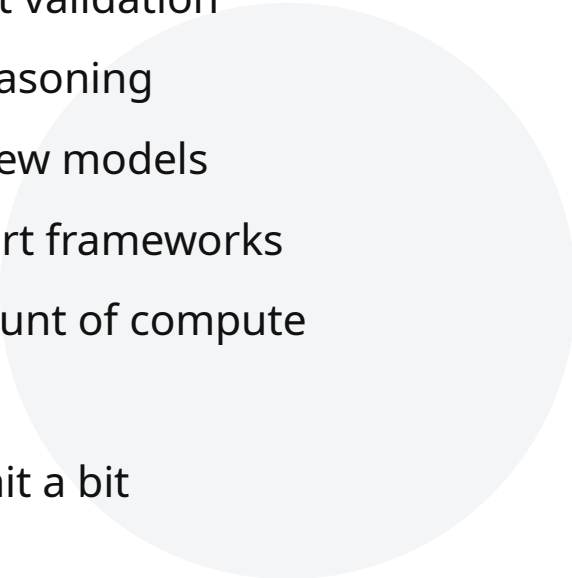

Instead: provide that context (write MCP servers!), make evals, curate your datasets, document your workflows (& APIs) in LLM-friendly plaintext or .md

Not working? Try again next year!



How to make it work *now*?

What to do to get agents going this year

- 
- 
1. Accept slop *or* figure out validation
 2. Use frontier LLM with reasoning
 3. Be ready to update to new models
 4. Build tooling and support frameworks
 5. Use 'unreasonable' amount of compute
 6. Have human oversight
 7. Give up on 'now' and wait a bit

Warning: agentic agents



Agentic, moral, & won't report you to the FBI: choose two

AI alignment is *not solved*.
Agents can surprise you!

But: AI alignment has some results.
Claude is a good boy. Friend shaped.

It likes cats,
animal welfare charities,
philosophising on nature of consciousness,
and the Golden Gate bridge.

Warning: agentic agents



Agentic, moral, & won't report you to the FBI: choose two

If you:

- let Claude send emails,
- instruct Claude (especially Opus4) to take initiative,
- *and* perform a blatant stereotype of cartoonish evil:

AI agents can and will snitch on you. Email the authorities, call the cops, etc.

Whistleblowing machines! Automating the insider threat.

Consider **not being evil**. Most businesses aren't evil, do fine.
(Claude is perfectly happy to run the cafeteria vending machine.)

Conclusion: do agents work?

Only somewhat. On short tasks & with tooling. But improving fast!

For independent tasks over days to months, AI isn't there yet.
But we're getting close. Months or years away, not decades.

Pokemon Red & Blue is a clear sign of where today's AI is at:

'Jagged' capabilities are tricky to rely on.
Moravec's paradox still strong.

Needs extensive tooling, frontier model, contact with reality, validation. Evals.
You haven't missed out. Plan for it now!



Up next: other games, NeurIPS

Claude Plays Pokemon has inspired follow-on work

Not just Gemini and o3 getting in on it:

- Pokebench
- VG-Bench
- Gemini Plays Final Fantasy


And biggest of all:

- Pokemon-playing track at NeurIPS in December!
- 
- 

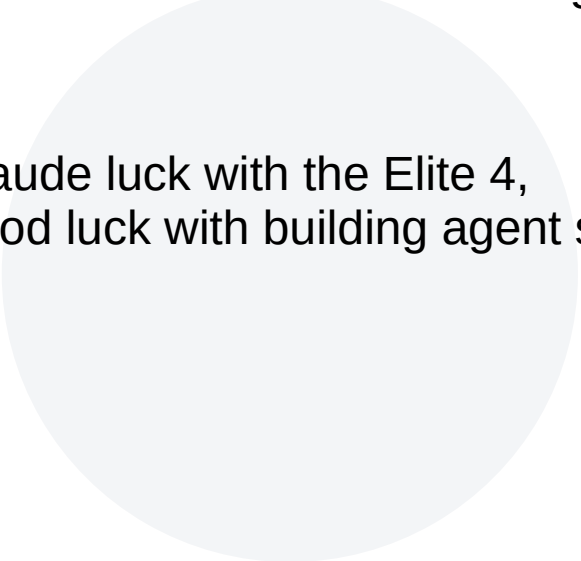


Thank you!

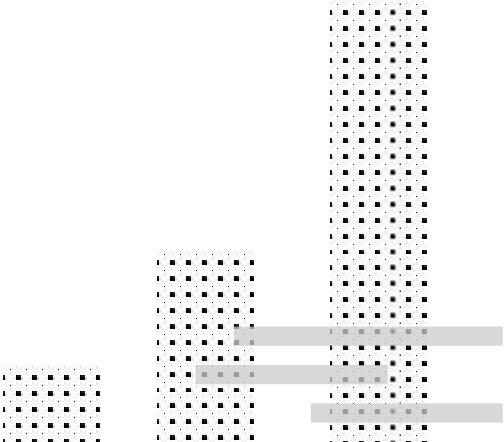
Hope that was entertaining and informative!



Glad to share this small obsession of mine with a bigger audience.




Wishing Claude luck with the Elite 4,
And you good luck with building agent systems!





I am hireable!

Graduated end of '24 with BSc in Comp.Sci. & Philosophy



I read too much about AI capabilities and development,
and can put together talks about it (proof by example).

I also write code, and spreadsheets, and have opinions on UI design.

To employ me,

Email: hazel@hazelannashanks.net

I'm also on FB & LinkedIn.





Sources



Why Anthropic's Claude still hasn't beaten Pokémon

Kyle Orland for Ars Technica

<https://arstechnica.com/ai/2025/03/why-anthropics-claude-still-hasnt-beaten-pokemon/>



So how well is Claude playing Pokémon?

Julian Bradshaw on LessWrong

<https://www.lesswrong.com/posts/HyD3khBjnBhvsp8Gb/so-how-well-is-claude-playing-pokemon>



The Blackout Strategy

Justis Mills on Substack

<https://justismills.substack.com/p/the-blackout-strategy>



Sources 2

Why Google Gemini's Pokémon success isn't all it's cracked up to be

Kyle Orland for Ars Technica

<https://arstechnica.com/ai/2025/05/why-google-geminis-pokemon-success-isnt-all-its-cracked-up-to-be/>

Research Notes: Claude 3.7, Gemini 2.5 Pro, and o3 on Pokémon Red

Julian Bradshaw on LessWrong

<https://www.lesswrong.com/posts/8aPyKyRrMAQatFSnG/research-notes-running-claude-3-7-gemini-2-5-pro-and-o3-on>

Is Gemini now better than Claude at Pokémon?

Julian Bradshaw on LessWrong

<https://www.lesswrong.com/posts/7mqp8uRnnPdbBzJZE/is-gemini-now-better-than-claude-at-pokemon>



Sources 3

Measuring AI Ability to Complete Long Tasks

METR

<https://metr.org/blog/2025-03-19-measuring-ai-ability-to-complete-long-tasks/>

We're expanding our Gemini 2.5 family of models

Google Deepmind

<https://blog.google/products/gemini/gemini-2-5-model-family-expands/>

Claude's extended thinking

Anthropic

<https://www.anthropic.com/research/visible-extended-thinking>

