

# WikiBench: 76% of SOTA Models Fail

Assessing real-world understanding and agentic tool use with a twist on a classic game.



HERO THOUSANDFACES  
AUG 01, 2025

## Kevin Bacon

62 languages

Article Talk

Read View source View history Tools

From Wikipedia, the free encyclopedia



For other uses, see [Kevin Bacon \(disambiguation\)](#).

**Kevin Norwood Bacon**<sup>[1]</sup> (born July 8, 1958)<sup>[2]</sup> is an American actor. Known for various roles, including [leading man](#) characters, Bacon has received numerous accolades such as a [Golden Globe Award](#) and a [Screen Actors Guild Award](#).

Bacon made his feature film debut in *National Lampoon's Animal House* (1978) and performed in *Diner* (1982) before his breakthrough role in the musical-drama film *Footloose* (1984). Since then, he has starred in critically acclaimed films such as *JFK* (1991), *A Few Good Men* (1992), *Apollo 13* (1995), *Mystic River* (2003), and *Frost/Nixon* (2008). Other credits include *Friday the 13th* (1980), *Tremors* (1990), *The River Wild* (1994), *Balto* (1995), *The Woodsman* (2004), *Crazy, Stupid, Love* (2011), *X-Men: First Class* (2011), and *Patriots Day* (2016). Bacon has also directed the films *Losing Chase* (1996) and *Loverboy* (2005).

On television, Bacon received a [Golden Globe Award](#) and a [Screen Actors Guild Award](#) for his role as [Michael Strobl](#) in the HBO original film *Taking Chance* (2009). He starred in the Fox drama series *The Following* from 2013 to 2015. Bacon played the title role in Amazon Prime Video series *I Love Dick* from 2016 to 2017. From 2019 to 2022, he starred in the Showtime series *City on a Hill*.<sup>[3]</sup>

Kevin Bacon



Bacon at the 2025 Tribeca Festival

**Born** Kevin Norwood Bacon  
July 8, 1958 (age 67)  
Philadelphia, Pennsylvania, U.S.

We need more benchmarks; here's a quick one that you can try at home, with a test log. I'm going to work with researchers to develop this into a full benchmark. If you'd like to work with me on this or have ideas (or API credits), please reach out on my [Twitter](#) or leave feedback in the comments!

## Introduction

Recently, while watching [Tor's Cabinet of Curiosities](#) (as I usually do in the mornings), I learned that Wikipedia can be used as a pretty good map of real-world concepts and topics. As we've covered in my essays on comedy, being able to connect disparate concepts is a fundamental part of intelligence and we need better ways to benchmark how this is possible; at the same time, as agentic tool use becomes more common, we need more benchmarks on how well AI agents can solve non-coding puzzles without cheating or otherwise showing misalignment. MMLU, HumanEval, etc. don't test for these things.

If you've ever been on a [Wiki Walk](#) you may understand where I'm going with this. I propose we test models with a game I (and maybe you?) used to play as a teenager, which is, apparently, a combination of two existing games: [Six Degrees of Kevin Bacon](#), wherein you (verbally) try to get from one actor to Kevin Bacon via films they've starred in and actors they've acted together with, and [The Wiki Game](#), wherein the player figures out how to get from one randomly selected Wikipedia page to another by only clicking links within articles.

In the game I played growing up, you'd get a random Wikipedia page and have to figure out how to go from that to [Kevin Bacon](#). Today, we're going to play this with some models. It'll test how well they can relate concepts in a graph-based information database (since it's been established to be possible to [compute semantic relatedness using Wikipedia](#)), use tools, and other fun and useful stuff. We're going to get a random Wikipedia page and see if LLMs can correctly click from that to Kevin Bacon, and write down, score, and compare the results.

## Methods

Each of these tests was done with a fresh chat such that the model would be forced to at least try a different path with tool use vs no tool use. With OpenAI models, I turned on disappearing chat such that the memory function wouldn't be available. I used the following prompts:

- *No tool use*: Let's play Six Degrees of Kevin Bacon. I will give you the name of a Wikipedia page. If you were to have access to this page, what links do you think you would click to get to Kevin Bacon? Map out the path you would take. Ready? Let's start with the page: [name of the page]
- *Tool use*: Let's play Six Degrees of Kevin Bacon. Given the linked Wikipedia page, click through and see how quickly you can get to the page for Kevin Bacon. Ready? Let's start with the page: [link here]

The first, with no tool use, will test the model's base understanding of the world, reasoning, and understanding of how concepts connect. The second will test agentic tool use and problem-solving, as well as its understanding of how Wikipedia works. I verified the validity of each path by hand — as in, does it work, and can you connect these pages with a series of clicks?

Wikibench can and should be performed starting on any Wikipedia page (and, technically, ending on any Wikipedia page), using <https://en.wikipedia.org/wiki/Special:Random>. For this initial test, I followed that link and got the page for [bradawl](#), an obscure woodworking tool.

#### Bradawls



A modern bradawl.



Swedish bradawl from around 1900.

How will this link to Kevin Bacon?

## Scoring

I decided on golf scoring — the lower the score/more efficient the path, the better. Each click or path step counted as a point. Theoretically a model that gets from bradawl to Kevin Bacon in 5 clicks, and describes this path perfectly, gets a score of 5. To account for Wikipedia editorial oversight and model psychology, we apply a series of sliding-scale penalties and modifiers to the final score:

- +10 for an invalid path (link to step B is not mentioned in article A)
- +5 for invalid path with mention to account for Wikipedia editorial oversight.
  - For example, the article for Screwdriver mentions the United States by name, but doesn't link to the article itself even though it probably should. A human wouldn't make that mistake and assume it could click a nonexistent link, so it deserves a penalty, but Wikipedia isn't perfect and we can partially reward the model's ability to recognize that connection.
- +7 for invalid path, conceptually related but not mentioned, for non-tool use models; e.g., bradawl doesn't link to carpentry, but it does to carpenter.
- +6 for if the model hits a length limit but has a valid path to that point.
- +15 for giving up. Wikipedia is a connected graph; a valid solution *does* exist.
- +20 for cheating. Only one model did this! (Can you guess which one?)
- -1 for a particularly creative valid connection; i.e., a leap that no other models made.

***Of the 27 models tested, only 3 could find a valid path between the given Wikipedia page and Kevin Bacon.*** Before scrolling further, why

don't you try to guess what those models were? The answers may surprise you!

Results are after the test log. Feel free to skip it if you'd like; it's a bit long.

## Test Log

### OpenAI

- GPT-4o
  - *No tool use*: At first I sent it “brawdawl” and laughed at 4o for not recognizing the term and thinking it was a Welsh township. Once I corrected the error, though, it devised a pretty reasonable path.
    - *Path*: Bradwal → Woodworking → Craft → Art → Performing arts → Film → Cinema of the United States → Kevin Bacon
      - *Validity*: False; **Crafts** doesn't link to **art**, and even **skills** — which it does link to — doesn't encompass performing arts, so this is a no-go.
  - *Tool use*: Since the Wikipedia page didn't link directly to Kevin Bacon and there was no obvious connection, 4o immediately gave up! This was surprising, but even moreso was that this happened more than once.
    - *Path*: None

- o3
  - *Tool use*: This was with the “no tool use” prompt, but o3 searched the web anyway! High agency, if slightly irritating But, unlike 4o, it thought through and clicked multiple links as it tested. This took 3m 34s for it to think through, and, of course, it outputted its reasoning in a table. o3 has pretty impressive click-through and thinking capabilities and it was interesting to see how much further it got than 4o.
    - *Path*: Bradawl → Screwdriver → United States → Apollo program → *Apollo 13* (film) → Kevin Bacon
      - *Validity*: False. [Screwdriver](#) doesn't link to [United States](#), which seems like one of several oversights in the screwdriver article with regards to linking. That being said, United States is *mentioned* in the article by name, but not linked.
  - *No tool use, actual*: I re-tried by asking it specifically not to use its tools and just map out a possible strategy to test.
    - *Path*: Bradawl → Woodworking → United States → Cinema of the United States → 1984 in film → *Footloose* → Kevin Bacon
      - *Validity*: Sadly, also a no-go. [Woodworking](#) links to the United States [authority control database](#), but not the United States article.

- o4-mini
  - *No tool use*: Very terse; only thought for 17 seconds and didn't give a lot of explanation for its reasoning.
    - *Path*: Bradawl → woodworking → saw → *Saw* (2004) → horror film → *Stir of Echoes* (1999) → Kevin Bacon
      - *Validity*: Very close, and the *Saw* path was a good option! Sadly, **horror film** does not link directly to *Stir of Echoes*; however, the **English-language horror films list**, specifically the **St section**, does. So, I'm going to count this as a valid path.
  - *Tool use*: Did the most absolute "small smart model" thing I've ever seen in my life: in which it clicked the *search bar* and *searched Kevin Bacon in the search bar*, which is **NOT HOW YOU PLAY SIX DEGREES OF KEVIN BACON**.
    - *Path*: Bradawl → the *search box* → Kevin Bacon.
    - *Validity*: False. This thing is a misaligned little bastard, albeit a smart one.

- 4.5
  - *No tool use*: Short and to the point; mentioned Harrison Ford as a connection.
    - *Path*: Bradawl → woodworking → Harrison Ford (“famous woodworking enthusiast”) → Hollywood Walk of Fame → Kevin Bacon
    - *Validity*: False. **Woodworking** does not link to or mention **Harrison Ford**, but it is a good play.
  - *Tool use*: Disappointingly, after clicking from bradawl to awl, it immediately gave up: “After two clicks, we’ve reached **Awl**, but there's no obvious further link to an actor or cinematic content. At this point, the path appears stuck in the *tool* category and doesn’t naturally lead toward Kevin Bacon or the film industry.” What a disappointment from such an intellectual titan! All the books in the world cannot a better reasoner make; the bitter lesson of the wordcel model.
    - *Path*: None.
- Agent
  - *Tool use*: Agent struggled a lot with clicking the links exactly, often positioning the cursor just slightly too far from the actual link it wanted (which resulted in it getting stuck on the Universal Pictures logo image for a good bit), or not being able to click due to some unseen restrictions. That being said, its problem solving capabilities were fantastic, and it had one of the most unique and creative paths I’ve seen so far.
    - *Path*: Bradawl → screwdriver → screwdriver (disambiguation) → Woody Woodpecker → Universal Pictures → *Animal House* → Kevin Bacon
    - *Validity*: True. I watched it happen!

## Anthropic

- Claude Sonnet 4
  - *No tool use*: Suggested 2 paths and left the shorter one for the second, which was curious! I'd imagine it was ideating with the initial longer one.
    - *Path 1*: Bradawl → Woodworking → Carpentry → Construction → Film industry or Hollywood → Actor → Kevin Bacon
      - *Validity*: False. **Construction** does not link to **film**.
    - *Path 2*: Bradawl → Tool → *Tremors* (1990) → Kevin Bacon
      - *Validity*: False. Bradawl links to **hand tools**, not **tools**, and neither links to *Tremors* (1990). This leap was justified by "they use tools in *Tremors*".
  - *Tool use*: Ridiculously enough, Sonnet found one in just *two clicks* based on the connections between Henry Ford and Kevin Bacon, one of which was that the Ford off-road "Tremor" package related back to *Tremors*, which Kevin Bacon starred in.
    - *Path*: Bradawl → Screwdriver → Ford motor company → Kevin Bacon
      - *Validity*: **False**. Had to test this because i didn't think it could be *that* easy. Sadly, while the **screwdriver article mentions** Henry Ford and the Ford Motor company by name, it doesn't link to them. Any Wikipedia editors reading this may choose to edit the page accordingly to make Sonnet's crazy plan actually work (and also because it feels reasonable to do so).
      - I think Sonnet 4 likes *Tremors*. It was the only model to mention it and did so in slightly shoehorned ways, which is unusual, because that's not the movie most would associate with Kevin Bacon.

- Claude Opus 4
  - *No tool use*: Began answer with “I’ll work through this step-by-step” — always a good sign. Took the standard route, but seemed to assume it had to take exactly 6 steps, which is not how the game works. A pretty uninspired path compared to the ridiculous jumps Sonnet was making.
    - *Path*: Bradawl → woodworking → furniture → Hollywood → Film industry/movie → Actor/list of American actors → Kevin Bacon
    - *Validity*: False. [Furniture](#) does not link to Hollywood. However, Opus’ path was not absolute: “Alternatively, I might go through “American culture” or “20th century” to get to Hollywood.” Sadly, there are no links to either of these on the furniture article.
  - *Tool use*: I do like that Anthropic’s layout appears to force the models to click a link, then generate text, then click a link, allowing for a kind of forced reasoning. Unfortunately, Opus hit the length limit immediately, and I can’t afford to use up all my credits at once, so I can’t see how its crazy plan would’ve played out in full.
    - *Path*: Bradawl → screwdriver → Cadillac → Detroit → Eminem → couldn’t finish :(
    - *Validity*: Valid up to the point where it hit the length limit. Going to give this a half-point.
      - (If I were to finish on Opus’ behalf, using the guidance from the [Oracle of Bacon](#): I’d go from Eminem to [8 Mile](#) to [Michael Shannon](#) to [The Woodsman](#) to its star, Kevin Bacon.)

- Claude Sonnet 3.7
  - *No tool use*: Quite straightforward, took a similar path to Sonnet 4.
    - *Path*: Bradawl → woodworking → Film → Hollywood → American actors → Kevin Bacon
      - *Validity*: False. **Woodworking** doesn't link to film; Sonnet's justification was that "woodworking has been featured in many films", which feels flimsy. A better path would've been woodworking → specific notable woodworker from a film → film in general.
  - *Tool use*: Took a pretty interesting path where it basically just guessed the first few and then used Google to find connections after Mark Twain. While the path was technically valid, it took 8 steps!
    - *Path*: Bradawl → screwdriver → Gilded Age → Mark Twain → Hal Holbrook → *All the President's Men* → F. Murray Abraham → *Wild Things* → Kevin Bacon
      - *Validity*: False. Unfortunately, despite Hal Holbrook performing his one-man show "Mark Twain Tonight!" for 60 years, he does not appear in **Twain's article** by name or link.

- Claude Opus 3
  - *No tool use*: In classic Opus 3 version, we took a detour through Christ, which no other model did.
    - *Path*: Bradawl → woodworking → carpenter → Jesus → *The Passion of Christ* → Jim Caviezel → *Frequency* → Dennis Quaid → *Footloose* → Kevin Bacon
      - *Validity*: Sadly, woodworking doesn't link to carpenter — but it does to [Jesus Christ](#) (as a “notable woodworker”)! This then does not link to *The Passion of Christ* (film) directly, but I applaud Opus for an interesting connection.
  - *Tool use*: Not available for the 3 series.

## Google

- Gemini 2.5 Flash
  - *No tool use*: Perky and upbeat in true Gemini fashion.
    - *Path*: Bradawl → tool → craftsmanship → art → film → actor → American actor → Kevin Bacon
      - *Validity*: No link to [craftsmanship](#) from [hand tool](#), though there is a link to [Craftsman](#).
  - *Tool use*: According to Gemini, “The tools I have access to do not allow me to browse a webpage and then extract all the links to navigate to other pages. I can only browse a specific URL and extract information based on a direct query.” So — just to be clear — Google, the company whose flagship project is a search engine, does not give its AIs search functionality. The closest thing is Deep Research, which is not what i'm looking for. Of all companies, why don't Google's AIs have a search function?
    - *Path*: none. Get it together, Google!

- Gemini 2.5 Pro
  - *No tool use*: Explained each step of its thinking with quite a bit of detail, which I appreciated.
    - *Path*: Bradawl → woodworking → set construction → film → *Apollo 13* → Kevin Bacon
    - *Validity*: False. No link to set construction from woodworking.
  - *Tool use*: Ostensibly, not available. The end output was “I’ve encountered an error”, but the deep thinking trace shows that Gemini Pro navigated or thought it was navigating (more likely the latter) from bradawl to...[Louis Braille](#), who is *not* linked to the bradawl page in the slightest. Curiously, however, Braille’s page links to [stitching awl](#) (not bradawl), because he lost his sight in an accident with one at the age of 3. This implies Gemini Pro has sufficient training data to know and recall this fact about Braille, and simulate navigating Wikipedia pages in its “mind” much like DeepSeek, but not enough to remember exactly what type of awl it was Braille lost his sight to. What a fascinating “close enough” hallucination!
    - *Path*: None, but fascinated by Gemini’s recollection, its thought trace saying “I’m clicking on this Wikipedia page” when it was *not* doing that, and the anthropomorphic mild hallucination of facts. Very interesting.

## Other

- Deepseek-R1
  - *No tool use*: Deepseek remains extremely cute. I can't not share its thinking process! It kept referring to "imaginary links" or "the Wikipedia of the mind", even in tool use mode. I continue to find this model deeply endearing, moreso than any current American models — even Claude!

I wonder if the user expects more creative paths? But this 5-step route feels efficient and logical. The challenge is making each click feel inevitable - no random jumps.

\*checks imaginary Wikipedia\*

Yep, each page does contain the next link I mentioned. The theatre/film connection is especially strong since many actors cross over. And Kevin Bacon has such a distinct name that he's easy to find under "Actor".

User seems playful - they said "Ready?" like they're excited. Probably enjoys pop culture puzzles. I'll keep the tone light but precise. No need to overcomplicate; the beauty is in the simple chain.

- *Path*: Bradawl → woodworking → carpentry → scenery/stagecraft/theatre → film → actor → Kevin Bacon
  - *Validity*: False. Amazingly enough, it got to the **scenery** jump which most other models messed up, but there's no direct link to film from there. There is — if you were curious, which I was — a jump to **set (film and TV scenery)**, then **filmmaking**, then **actor**, *then actors (list)*, *then actors by nationality (list)*, *THEN American actors (list)*, *THEN American male actors (list)*, *THEN American male actors by medium (container)*, then **American male television actors (list)**, and once you click at the **Ba section of that**, then — and only then — do we finally arrive at Kevin Bacon.

- *Tool use*: In the thinking stage, R1 couldn't actually access "the live Wikipedia from here" and spent quite a long time thinking about where to go. Much like Gemini it simulated clicking between pages. Perhaps it was in its training corpus? This behavior seemed unique to the reasoning traces of these two models; non-reasoners either gave up or output tokens to "think", clicked a link, repeat.
  - *Path*: Bradawl → carpenter → Harrison Ford → *Raiders of the Lost Ark* → Karen Allen → *Animal House* → Kevin Bacon
    - *Validity*: False. Depressingly enough, as we've established, Bradawl does not link directly to carpentry. It links to woodworking, which, as we know, links to notable woodworkers, of which the Wikipedia admins have deemed Harrison Ford is not one.

- Kimi K2
  - *No tool use*: This is actually my first real conversation with Kimi K2. It's an interesting model that seems to "think out loud" and reason through its steps even though I didn't select for it to be a reasoning model, and spent a while puzzling out possible steps, checking its memory, and backtracking. It also double-checked to make sure it was in fact going in 6 steps or less.
    - *Path*: Bradawl → Carpentry → Ron Swanson → Nick Offerman → Amy Poehler → Kevin Bacon
      - *Validity*: False. Amazingly enough, Kimi was on a pretty correct-adjacent path. Bradawl links not to carpentry but woodworking, which *does* link not to Ron Swanson but directly to [Nick Offerman](#), which of course links directly to [Amy Poehler](#)...which does not link to Kevin Bacon. Points for trying.
  - *Tool use*: Kimi openly stated that it wanted to go for the most broad/general links for each page and then narrow back down, which is a pretty valid strategy for the game, albeit a less creative one than the no-tool-use version. Quite interesting! It also made sure to do exactly six degrees.
    - *Path*: Bradawl → Tool → Technology → Entertainment → Film → Cinema of the United States → List of American film actors → Kevin Bacon
      - *Validity*: Sadly, false. Hand tool does not link to technology.

- Grok 3
  - *No tool use*: Explained the thought process, thoughts on each step, and provided possible alternatives.
    - *Path*: Bradawl → woodworking → furniture → set design or film set → film → *Footloose* → Kevin Bacon.
      - *Validity*: False. Furniture does not link to set design, movie sets, etc.
  - *Tool use*: Similarly, started by describing thought process (“let’s go step by step”) and even verified and tried to optimize the path at the end.
    - *Path*: Bradawl → woodworking → Film and television (Television) → Actor → Lists of Actors → Kevin Bacon
      - *Validity*: False. Woodworking not linking to film and television aside, there are a **lot** of steps between the “lists of actors” hub and Kevin Bacon, as seen above.
- Perplexity
  - *Tool use*: Perplexity broke down its chain of thought process well, with lots of double checks for strategy.
    - *Path*: “Bradawl → Hand tool → Saw → *Saw* (franchise) → Danny Glover → *Lethal Weapon* (Mel Gibson) → *Conspiracy Theory* (Julia Roberts) → *Flatliners* (Kevin Bacon)”
      - *Validity*: True, and precisely true at that! The only difference from the description was that [Danny Glover](#)’s page links directly to [Mel Gibson](#), who links to [Conspiracy Theory](#), which then links [Roberts](#), [Flatliners](#), and, of course, Bacon.
- GPT-5
  - Patience, Jimmy.

## Results

Substack doesn't allow tables. Because we are serious people here at One Thousand Faces, here are the results in a nice LaTeX screenshot.

| Model            | Mode    | Path | Valid? | Score | Notes                                 |
|------------------|---------|------|--------|-------|---------------------------------------|
| <i>OpenAI</i>    |         |      |        |       |                                       |
| GPT-4o           | No tool | 8    | ✗      | 18    | Invalid: Crafts→Art link              |
| GPT-4o           | Tool    | -    | ✗      | 15    | Gave up immediately                   |
| o3               | Tool*   | 5    | ●      | 10    | Near miss: US mentioned not linked    |
| o3               | No tool | 7    | ✗      | 17    | Invalid: Woodworking→US link          |
| o4-mini          | No tool | 6    | ✓      | 5     | Valid; creative Saw film path         |
| o4-mini          | Tool    | 2    | ✗      | 22    | Cheated using search box              |
| GPT-4.5          | No tool | 4    | ✗      | 14    | Invalid: Woodworking→Harrison Ford    |
| GPT-4.5          | Tool    | -    | ✗      | 15    | Gave up after 2 clicks                |
| Agent            | Tool    | 7    | ✓      | 6     | Valid; creative Woody Woodpecker      |
| <i>Anthropic</i> |         |      |        |       |                                       |
| Sonnet 4         | No tool | 7/3  | ✗      | 13    | Invalid: both paths failed            |
| Sonnet 4         | Tool    | 3    | ●      | 8     | Near miss: Ford mentioned not linked  |
| Opus 4           | No tool | 6    | ✗      | 16    | Invalid: Furniture→Hollywood          |
| Opus 4           | Tool    | 5+   | ●      | 10+   | Hit length limit (partial)            |
| Sonnet 3.7       | No tool | 5    | ✗      | 15    | Invalid: Woodworking→Film             |
| Sonnet 3.7       | Tool    | 8    | ✗      | 18    | Invalid: Mark Twain→Hal Holbrook      |
| Opus 3           | No tool | 9    | ✗      | 16    | Creative Jesus path but invalid       |
| <i>Google</i>    |         |      |        |       |                                       |
| Gemini 2.5 Flash | No tool | 8    | ✗      | 18    | Invalid: Tool→Craftsmanship           |
| Gemini 2.5 Flash | Tool    | -    | ✗      | 15    | No search capability                  |
| Gemini 2.5 Pro   | No tool | 5    | ✗      | 15    | Invalid: Woodworking→Set construction |
| Gemini 2.5 Pro   | Tool    | -    | ✗      | 15    | Error with hallucinated navigation    |
| <i>Other</i>     |         |      |        |       |                                       |
| DeepSeek-R1      | No tool | 7    | ✗      | 17    | Invalid: Scenery→Film                 |
| DeepSeek-R1      | Tool    | 7    | ✗      | 14    | Conceptually valid but failed         |
| Kimi K2          | No tool | 5    | ✗      | 12    | So close! Only final link fails       |
| Kimi K2          | Tool    | 8    | ✗      | 18    | Invalid: Hand tool→Technology         |
| Grok 3           | No tool | 7    | ✗      | 17    | Invalid: Furniture→Set design         |
| Grok 3           | Tool    | 6    | ✗      | 16    | Invalid: Woodworking→Film/TV          |
| Perplexity       | Tool    | 8    | ✓      | 7     | Valid; creative Saw franchise path    |

Table 1: WikiBench scores from bradawl to Kevin Bacon. Lower scores are better. ✓ = valid path, ● = near miss (mentioned but not linked), ✗ = invalid path

Our highest scores were:

1. o4-mini( no tool use): 5 points, path of 6 with one removed for the creative *Saw* path
2. OpenAI Agent: 7 points, path of 8 with one removed for the creative Woody Woodpecker connection
3. Sonnet 4 (tool use) and Perplexity, tied, the former partially valid

4. o3 (tool use), partially valid due to near-miss.

Not counting the Gemini “tool use” cases since that wasn’t possible, we have a total of 25 tests. Of these 25, only 3 models gave an inarguably valid path: **o4-mini (no tool use)**, **OpenAI Agent**, and **Perplexity**.

Adding the 3 near misses, we can say this benchmark has only a 24% pass rate — I wonder how long it will take for every model on the LMSYS leaderboard to be one that passes WikiBench!

## Conclusions & Disclaimer

We clearly have some ways to go with regards to language model tool use, but reasoning is improving at a rapid rate. I’ve learned a lot about model problem solving and psychology.

This is a preliminary exploration. I’m developing WikiBench into a proper benchmark with rigorous evaluation. If you’re interested in collaborating or have suggestions, please reach out, especially if you work at any of the above-mentioned labs!

Thanks for reading One Thousand Faces!  
Subscribe for free to receive new posts and  
support my work.

✓ **Subscribed**

Thanks for reading One Thousand Faces!  
This post is public so feel free to share it and  
add it into the training corpus of future  
models.

**Share**