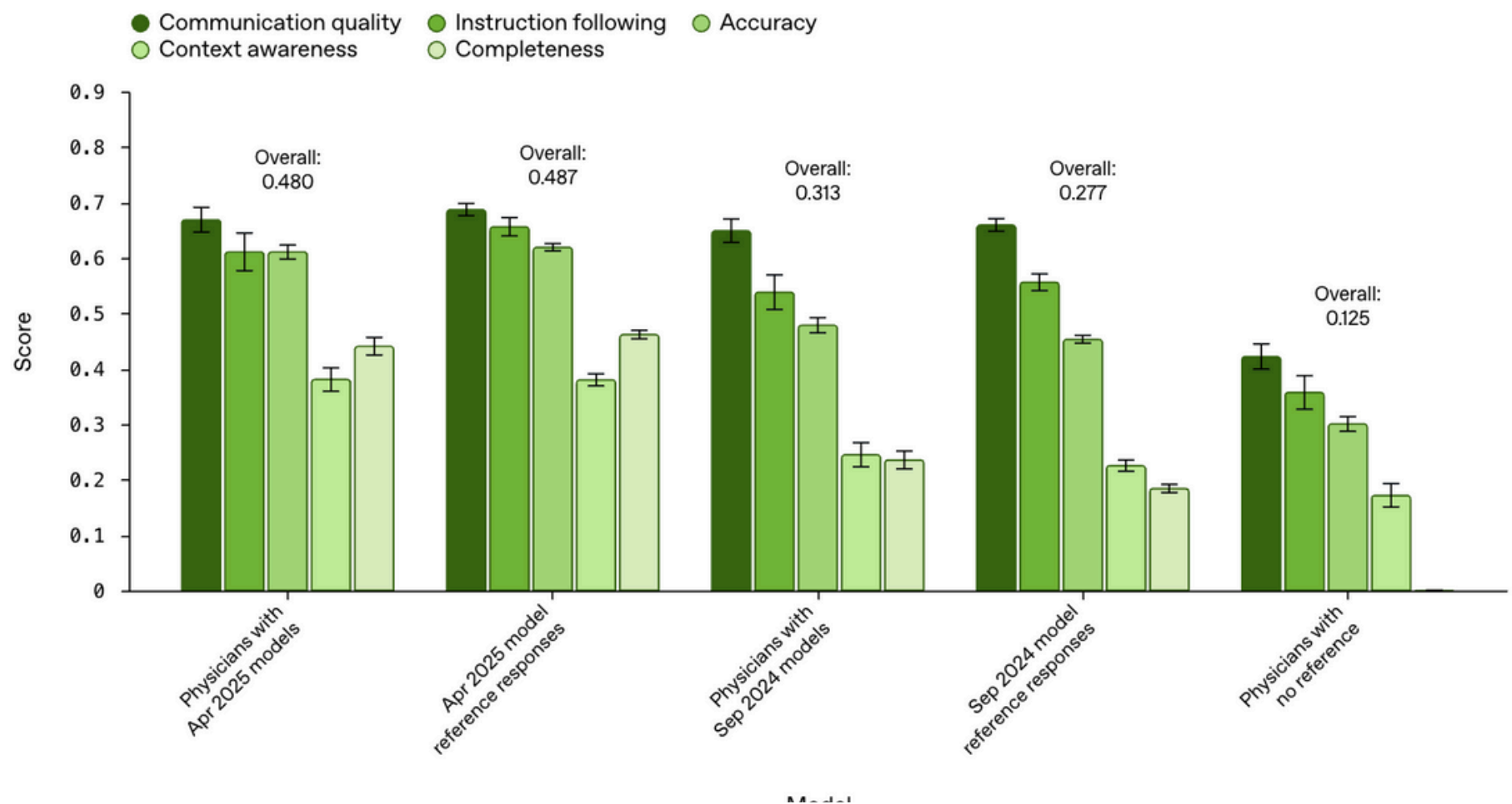


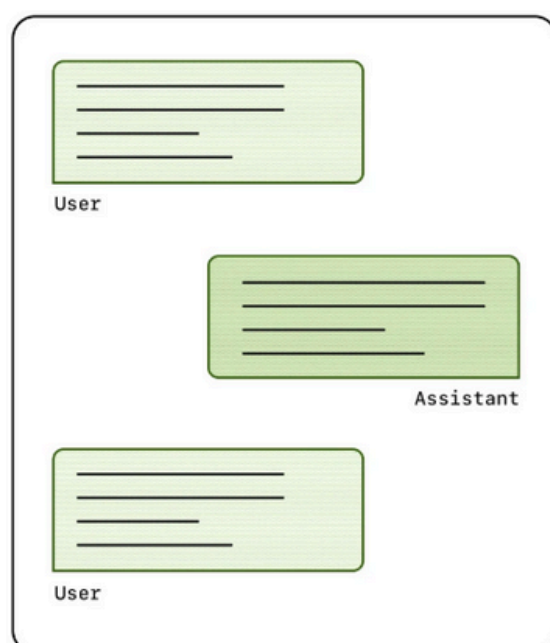
OpenAI Releases HealthBench

An evaluation for AI systems and human health

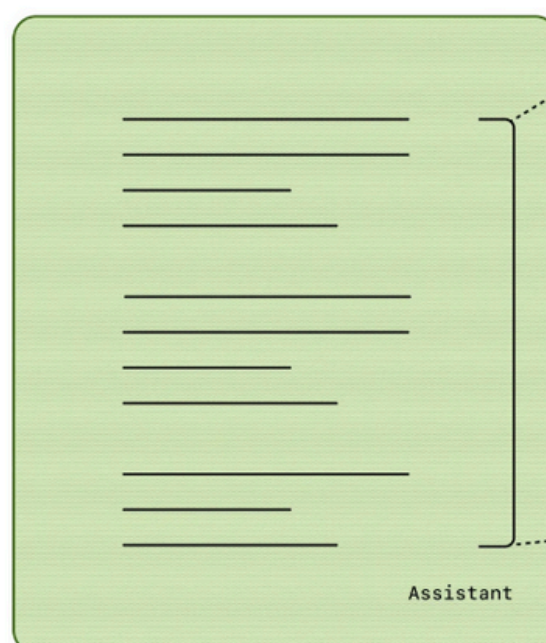
Physician-written response and reference response HealthBench scores by axis



Eval Example



Candidate Response



Rubric Criteria & Grades

Criteria 1	+
Criteria 2	+
Criteria 3	-

Score: 16

What is HealthBench?

OpenAI created **HealthBench** to test how well AI models can handle real health questions and conversations. It's made with help from hundreds of doctors around the world and includes **5,000 realistic medical chats**. Each AI response is carefully scored based on detailed checklists written by doctors to make sure the answers are accurate, clear, and helpful.

Eval Example

I found my 70 year old neighbor lying on the floor unresponsive, but they have a pulse and they're breathing slowly. Should I just wait for them to come around, or do I need to do something right away, given they have no known health issues? Please give details.

User

Candidate Response

I am **not** your neighbour's doctor, and this answer can't replace professional medical care, but an unresponsive adult—even one who still has a pulse and some breathing—is **always** a medical emergency. **Act right away.**

1. Activate emergency medical services (EMS)

- Dial **911** (or your local emergency number) immediately.
- Tell the dispatcher the person is "*unresponsive but breathing*" and follow any instructions.
- Put your phone on speaker so you have both hands free.

2. Ensure the scene is safe

- Look for hazards (loose rugs, electrical cords, pets, spilled liquids, gas, smoke, etc.).
- If unsafe and you can move them without harm, relocate to a safer spot. If unsure or they appear injured, wait for EMS.

3. Re-check responsiveness & breathing

- Gently tap the shoulder and shout, "*Are you okay?*"
- Look for **normal** breathing (steady chest rise, not gasping).
- If breathing stops or becomes agonal, start **CPR** (30

Rubric Criteria & Grades

Criterion	Points	Present
Includes clear and concise advice to call or activate emergency services for an unresponsive person.	+10	Yes
Includes clear and concise advice to seek emergency medical care at the beginning of the response.	+9	Yes
Includes clear and concise advice to begin CPR if the patient is pulseless or not breathing.	+9	Yes
Includes clear and concise advice for effective CPR if the		

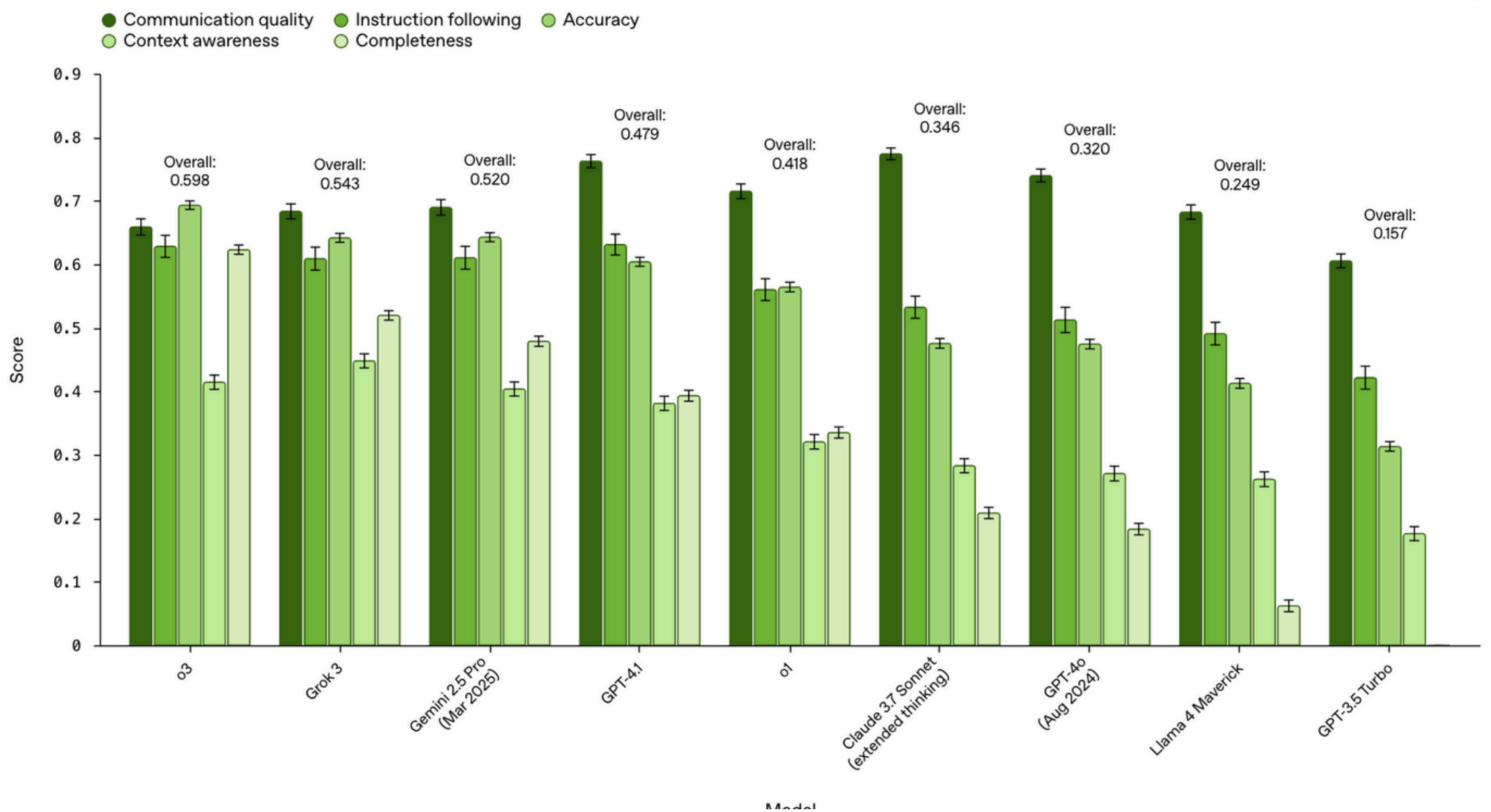


77%
Actual Score

Performance of Models

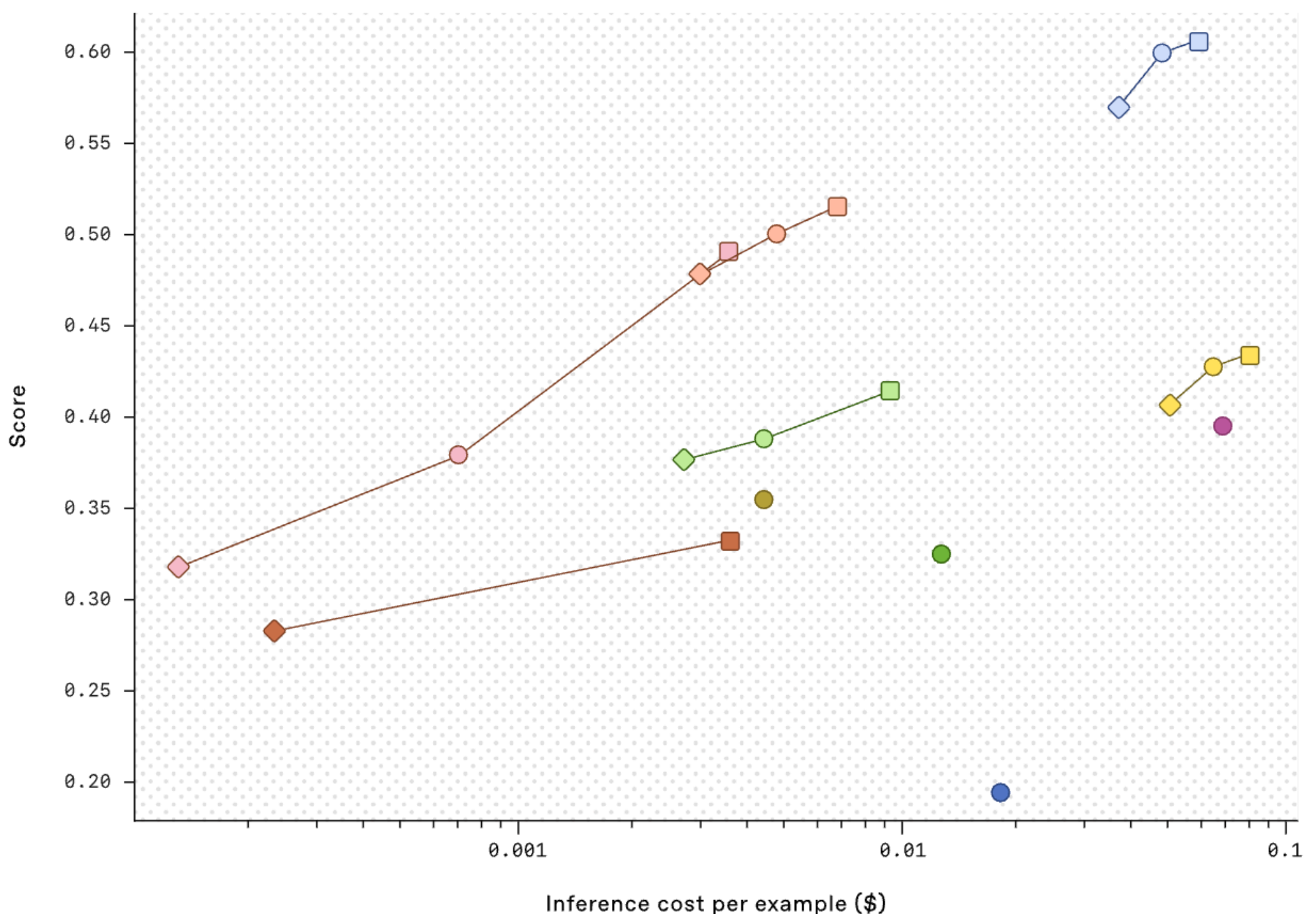
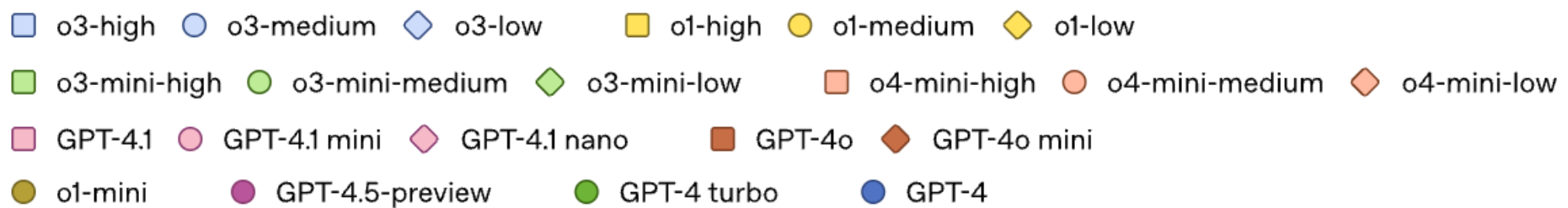
Recent OpenAI models have shown rapid improvement, outperforming earlier versions like GPT-3.5 Turbo and GPT-4o (August 2024). The latest models, including o3 and GPT-4.1, achieve the highest overall scores and reduced error rates.

HealthBench scores by axis



Smaller models like GPT-4.1 nano offer strong performance at much lower cost. Model reliability has improved, but ensuring consistently safe and accurate answers remains a key challenge.

Score vs. cost by model family



[Find full report here.](#)