Claude is good but is it x6 better?

22 May 2026 · testing, automation, ai, evals, banking, payments

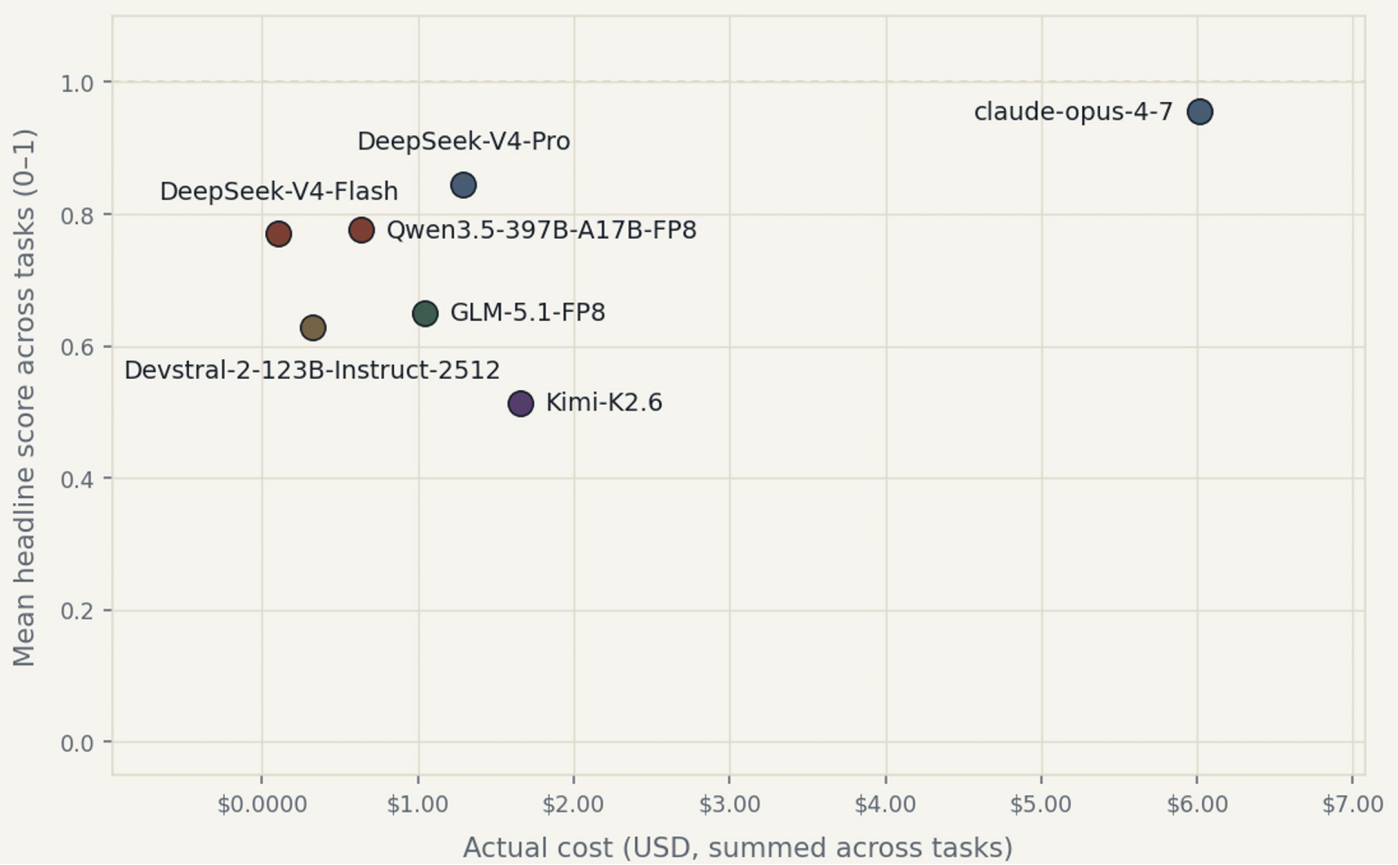
Because that’s how much more it can cost! As a 20x Claude code MAX subscriber I’m acutely aware how much tokens cost. “Yes you’re absolutely right, let me fix that” (Pete sets fire to a pile of cash)

So that’s why I’ve been comparing the costs for inference both from different models, and from different priorities. What do I mean by priorities? You can run realtime inference (ie: ask a question) and get answer in seconds. Or you can do a ‘batch’, this lowers the priority of your requests, so they take maybe 1hr or up to 24hr. The benefit being that you get a significant cost reduction Often the cost is half or more off the real time price of tokens.

What! I have to wait an hour! or more! yes, that’s often ok - for example I’ve made agents that can produce test data iteratively using consecutive 1hr batches. This works quite well, as it parallelises well and you can have many such agent working in parallel, adding requests to a common series of batches e.g. I have 20 agents each working on 20 test data files concurrently, at half the price of doing them with ‘real-time’ requests - ideal for overnight testing tasks.

So back to the cost of models, if I’m being super frugal, using 24hr batches, and open weight models, what’s the quality like? I have a small set of evals (automated tests for LLMs) that check how well LLMs handle tasks needed by bank core payments teams and systems.

I compared the leading ‘Open’ coding models and Claude Opus. The results are below. Claude does well (the best in fact), but the other models esp. DeepSeek(s) are not far behind and if you consider the cost on the x axis then you can clearly see how much more Claude makes you pay for that minor increase in accuracy. The greater cost of Claude can be as much on x6. If your agent runs on DeepSeek it could have maybe 5 attempts and still be cheaper!



The top left is good and cheap, the top right is good and expensive

Comparative AI evals like these are better when when the models generally don't get 100%, I'll dig more into that in my next blog post.

Wondering how well your agent or LLM is working? And whether its more money shredder than solution? Get in touch. email or LinkedIn (https://www.linkedin.com/in/peter-houghton-374a36/)