The New Claude 3 Models Are Great, But Are They Game-Changing?

For some people...yes, but mostly...no.

Dharmesh Shah
March 05, 2024

Disclosure: I'm an investor in OpenAI (a competitor to Anthropic, which makes Claude). I'm also a fan and customer of Anthropic. They have graciously spent time with me. I'm glad they exist. These views are my own and not based on any inside information about either company.

Anthropic created a lot of buzz and excitement recently with the launch of their Claude 3 set of LLMs (Large Language Models).

I've been tinkering and playing with it since launch. It is indeed impressive, and there is cause for some excitement (especially for developers).

There are 3 models: Claude 3 Haiku, Claude 3 Sonnet and Claude 3 Opus in ascending sequence of capability. They get points for that hierarchy of naming — clever. I wonder if they had Claude help come up with it. 🙂

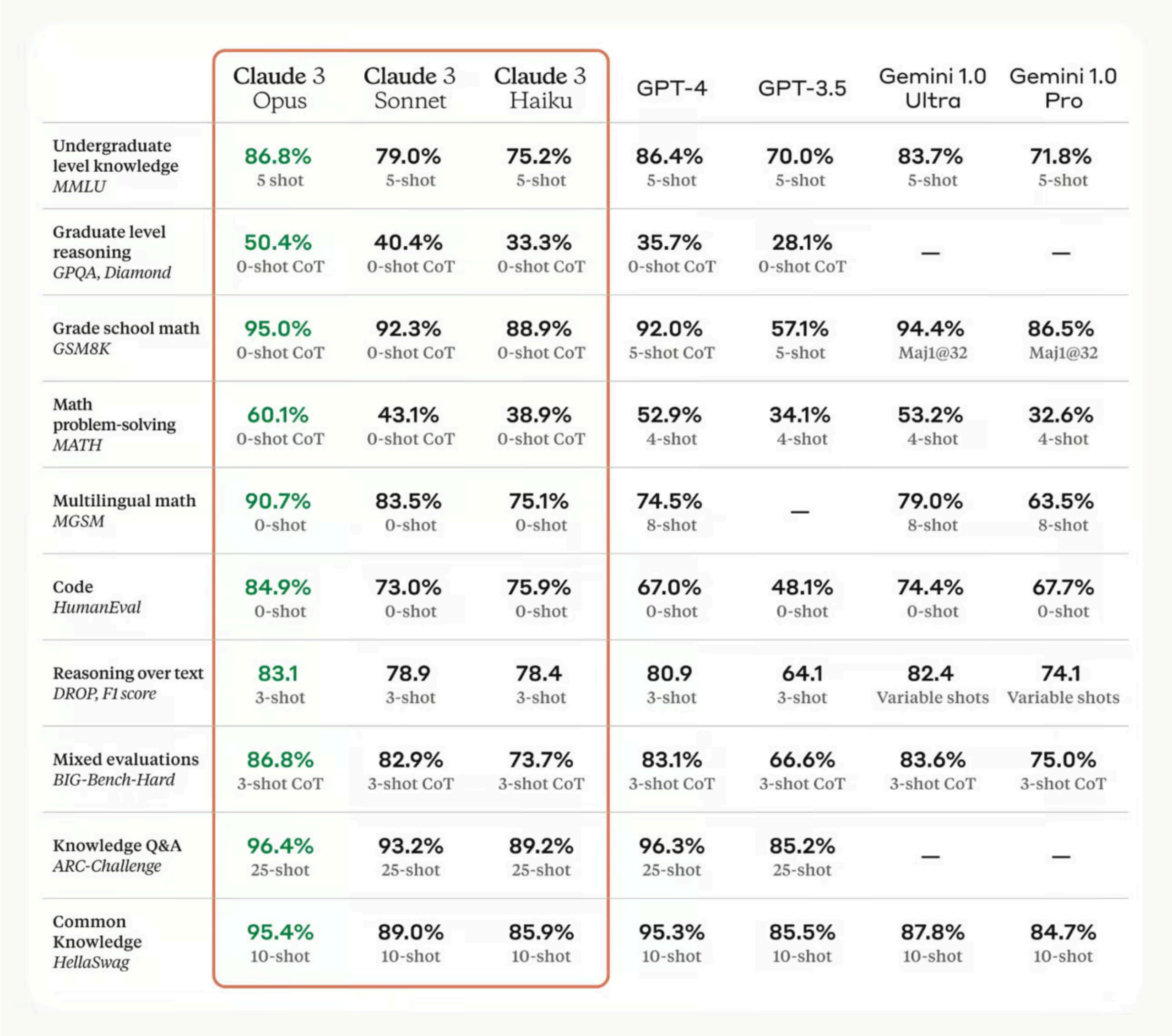
Look At The Benchmarks, Baby!

Let's first jump to the main reason people are excited:

Claude 3 Opus is the first time we have a Large Language Model that surpasses GPT-4 in capabilities across a wide variety of benchmarks. This is illustrated by the chart below.

Whether you believe in any individual benchmark or not, I think the important takeaway here is the overall *trend*. Yes, these numbers are published by Anthropic itself. And yes, benchmarks aren't perfect (by any means). And yes, some benchmarks can be gameed.

But still…this is, in a word, remarkable. Just demonstrates how quickly Generative AI is evolving. We went from "These LLMs are just fancy auto-suggest" to the point that they are now measurably able to actually *reason* and apply logic and analysis at a pretty high level. Amazing.



	Claude 3 Opus	Claude 3 Sonnet	Claude 3 Haiku	GPT-4	GPT-3.5	Gemini 1.0 Ultra	Gemini 1.0 Pro
Undergraduate level knowledge *MMLU*	86.8% 5 shot	79.0% 5-shot	75.2% 5-shot	86.4% 5-shot	70.0% 5-shot	83.7% 5-shot	71.8% 5-shot
Graduate level reasoning *GPQA, Diamond*	50.4% 0-shot CoT	40.4% 0-shot CoT	33.3% 0-shot CoT	35.7% 0-shot CoT	28.1% 0-shot CoT	—	—
Grade school math *GSM8K*	95.0% 0-shot CoT	92.3% 0-shot CoT	88.9% 0-shot CoT	92.0% 5-shot CoT	57.1% 5-shot	94.4% Maj1@32	86.5% Maj1@32
Math problem-solving *MATH*	60.1% 0-shot CoT	43.1% 0-shot CoT	38.9% 0-shot CoT	52.9% 4-shot	34.1% 4-shot	53.2% 4-shot	32.6% 4-shot
Multilingual math *MGSM*	90.7% 0-shot	83.5% 0-shot	75.1% 0-shot	74.5% 8-shot	—	79.0% 8-shot	63.5% 8-shot
Code *HumanEval*	84.9% 0-shot	73.0% 0-shot	75.9% 0-shot	67.0% 0-shot	48.1% 0-shot	74.4% 0-shot	67.7% 0-shot
Reasoning over text *DROP, F1 score*	83.1 3-shot	78.9 3-shot	78.4 3-shot	80.9 3-shot	64.1 3-shot	82.4 Variable shots	74.1 Variable shots
Mixed evaluations *BIG-Bench-Hard*	86.8% 3-shot CoT	82.9% 3-shot CoT	73.7% 3-shot CoT	83.1% 3-shot CoT	66.6% 3-shot CoT	83.6% 3-shot CoT	75.0% 3-shot CoT
Knowledge Q&A *ARC-Challenge*	96.4% 25-shot	93.2% 25-shot	89.2% 25-shot	96.3% 25-shot	85.2% 25-shot	—	—
Common Knowledge *HellaSwag*	95.4% 10-shot	89.0% 10-shot	85.9% 10-shot	95.3% 10-shot	85.5% 10-shot	87.8% 10-shot	84.7% 10-shot

From Anthropic, Inc.

See the full post about the the *launch of Claude 3*

Claude 3 Is Great, But Is It Game-Changing?

There are two high-level improvements that are noteworthy here. Claude 3 is “smarter” and it is faster. This is likely relevant to a small number of people — particularly developers building AI apps where those things really, really matter.

But, for the vast majority of us, Claude 3 is an important *milestone* along the AI journey, but not a change in direction or trajectory. It’s an incremental improvement (200,000 tokens, low latency, strong reasoning) not a breakthrough in what’s possible. We’ll likely not be switching to the consumer Claude product (and away from ChatGPT).

That’s totally OK. We need LLMs to get better/faster/smarter — both closed-source ones like Claude and OpenAI’s GPT *and* open source ones like LLama and Mistral (more on those in a future post).

Keep reading

What Is Agent AI And Why All The Excitement?

An unofficial definition and overview of the new hotness

Chat UX Is Great, But Wait Till You See Agent UX

The future isn't here yet, but it's coming fast

How To Build a Defensible A.I. Startup

And Not Get Incidentally Killed By OpenAI