Pressure Testing GPT-4 & Claude 2.1 Long Context

From Greg Kamradt - gk at gregkamradt.com <gk_at_gregkamradt_com_tzitoo@simplelogin.co>

To daethyra@cyberpunk.aleeas.com

Date Wednesday, November 29th, 2023 at 7:15 AM

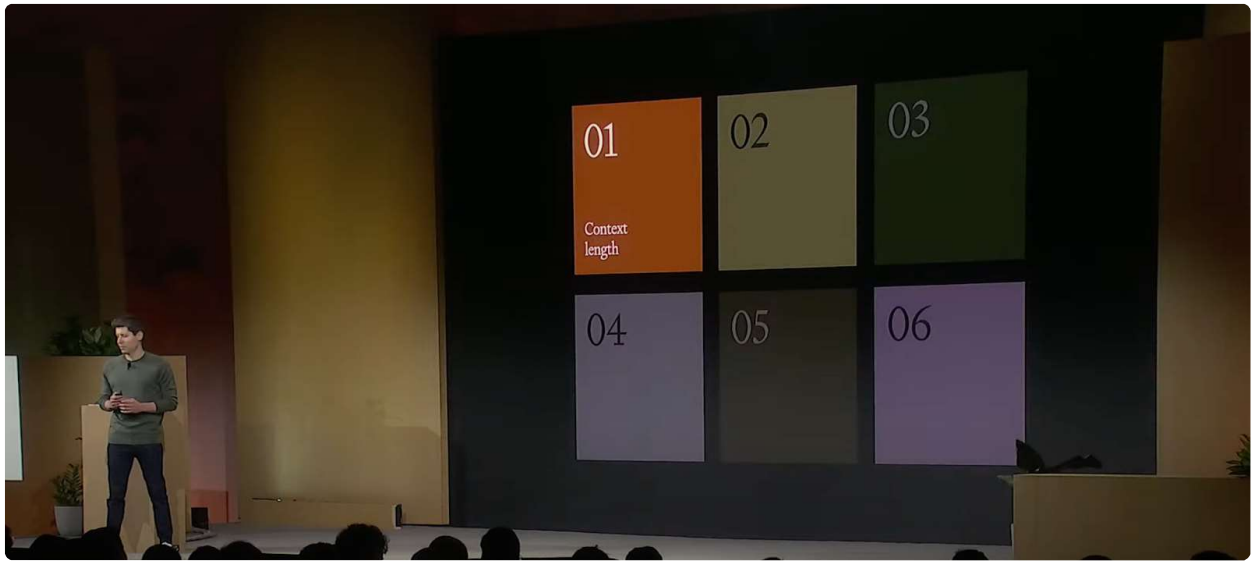
Longer context, better retrieval?

"Needle In A Haystack" Analysis With Large Context Models

Welcome to the 155 people who have joined us since last week! If you aren't subscribed, join 3,034 fun AI fans. View this post online.

Picture this, I'm sitting at TheGP in San Francisco watching OpenAI's Dev Day.

Sam Altman just announced GPT-4 with 128K tokens of context. This means you can fit nearly 300 pages of text into a prompt.



Every developer gets excited. No longer do they have to split their prompts into pieces and send it to GPT-4 one by one.

But I'm sitting there wondering, "longer context sounds great, but is there a performance impact?"

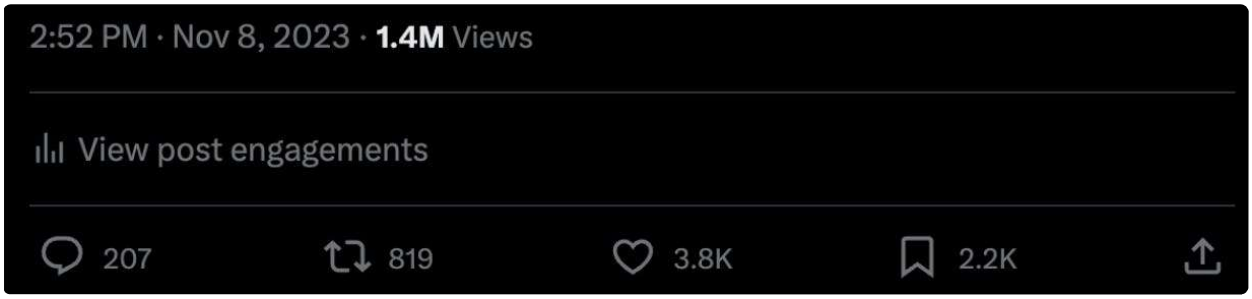
I decided to run a small test.

I took a single page of text and placed a random fact in it. I gave GPT-4 that page and asked it to retrieve the fact. Then I did the same thing again, but this time with 300 pages of context. Can you guess what happened?

With a single page, GPT-4 performed perfectly. It was able to recall the random fact. But with 300 pages, GPT-4 wasn't able to do it.

I wondered where the breaking point was. So I decided to find out.

In the end, the results for both GPT-4 and Claude 2.1 went viral on Twitter with 2.5M views.



What's the big deal with context lengths?

Around the time that ChatGPT was launched, developers were stuck with ~4K tokens of context. With such a small window, we had to be selective with the text we put in it.

This limitation rang the alarm with every LLM marketing & product team. The hunt for longer contexts was on.

Since then we've seen progressively larger contexts released. OpenAI just launched GPT-4-Preview (128K tokens of context) along with Anthropic's Claude 2.1 (200K tokens of context). That's 486 pages of context!

Increased contexts means we can do larger analysis, larger summaries, and we can loosen up our efforts on retrieval.

As far as marketing was concerned, the bigger the context, the better the model.

I wanted to find out if this was actually true. So I put GPT-4 and Claude 2.1 to the test.

Needle In A Haystack

I'm a simple person, I like milk with my cereal, butter with my bread, and easy, practical ways to test LLMs.

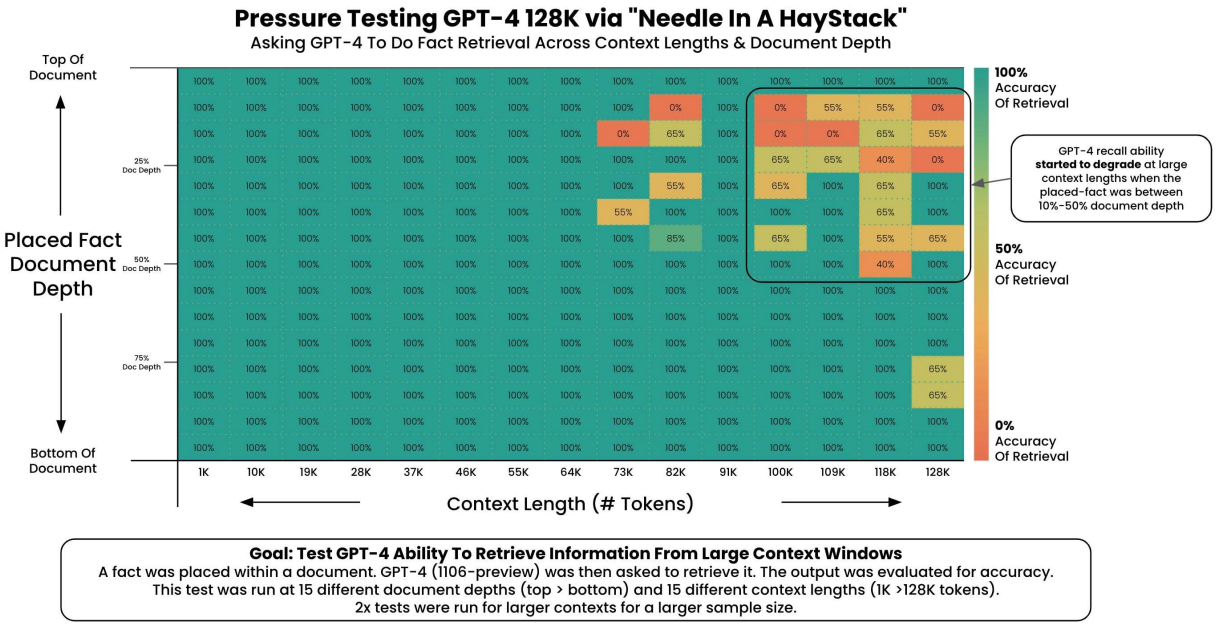
What made most sense to me was placing a random statement in the middle of a really long background context to see if the model could pull it out. After seeing GPT-4 break down on my first test, I decided I wanted to iterate through progressively larger context lengths to see where the break point was.

But then I remembered that *where* your fact was place in the document had an impact on retrieval. This was made popular by the "lost in the middle" paper. The authors found that facts placed at the top and bottom of a document had better recall than those in the middle.

So I decided that I would also iterate through document depths (the % downward the fact was placed).

I would use LangChain evals to quickly judge whether or not the fact was retrieved correctly.

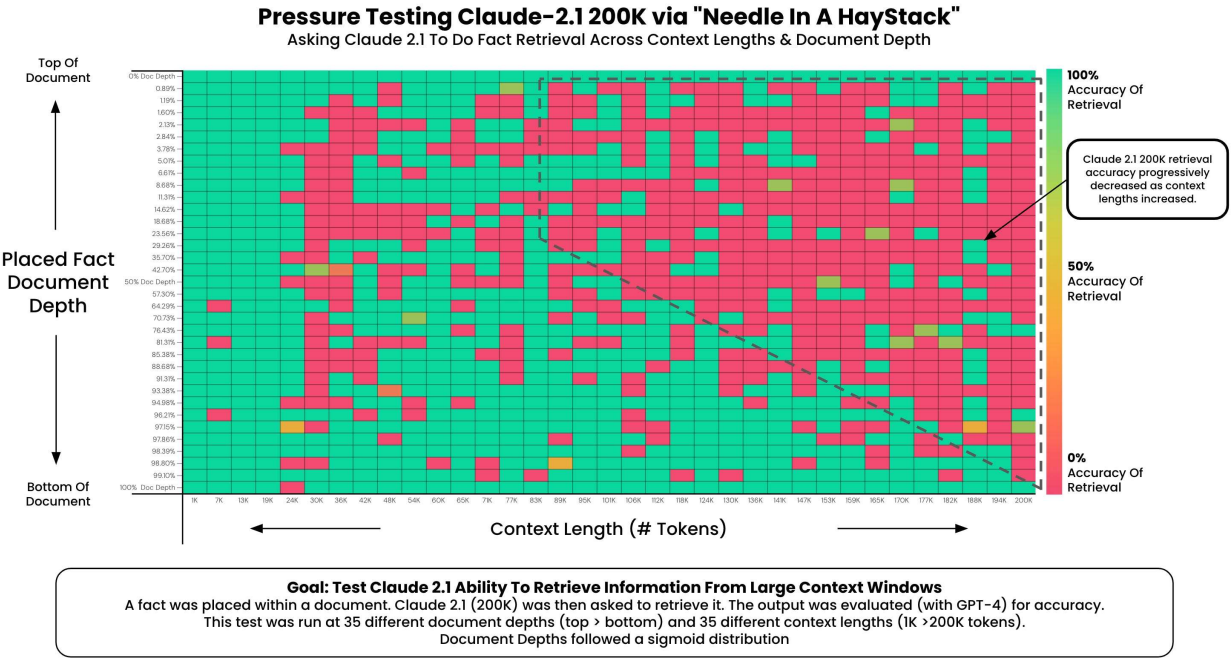
The GPT-4 test was paid out of my own pocket (~$215) so I only did a 15 by 15 grid of searches (~337 API calls). The larger contexts were evaluated 2x for more data.



Source

After I tweeted out the results of the GPT-4 test, I was contacted by Anthropic who wanted me to run the test for them as well. Knowing that the test would get pricey they offered credits to run it. (They didn't bias the results at all, just gave credits)

Since my spending limit was raised I decided to do a 35 by 35 grid (~1,225 API calls). You can't ever ask a data person if they want more data 😈. Then I published those results.



Source

I was asked multiple times, “Can you send me the white paper on this?” They were looking for me to send arXiv links, but they got a long tweet instead.

What did I learn?

Model Retrieval

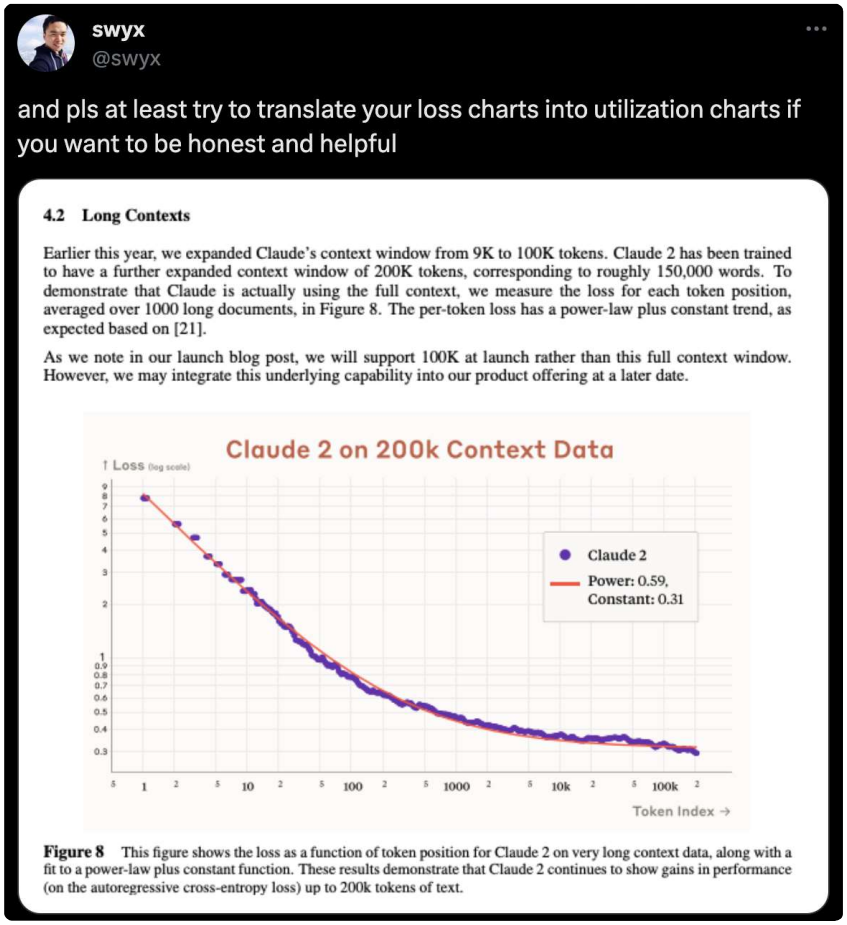
- At the largest token lengths, neither GPT-4 or Claude 2.1 can reliably retrieve placed facts
- Facts at the very top and very bottom of the document were recalled with nearly 100% accuracy
- For some reason, facts positioned at the top 50% of the document were recalled with less performance than the bottom

So what does this mean for you?

- **Prompt Engineering Matters** - It’s worth tinkering with your prompt and running A/B tests to measure retrieval accuracy
- **No Guarantees -** Your facts are not guaranteed to be retrieved. Don’t bake the assumption they will into your applications
- **Less context = more accuracy -** This is well known, but when possible reduce the amount of context you send to the models to increase its ability to recall. RAG is still extremely important
- **Position Matters -** Also well know, but facts placed at the very beginning and 2nd half of the document seem to be recalled better

Predictions

- Eventually recalling simple facts from long contexts won’t be a problem
- Companies will put less emphasis on large context windows
- In-context retrieval benchmarks will become the norm. We are already seeing the pressure ramp up



Source

Behind the scenes

- I walk through this test live
- Here's a breakdown of how the visualization was made

Greg Kamradt

Twitter / LinkedIn / Youtube / Work With Me