



Home

Explore

Notifications

Messages

Grok

Lists

Bookmarks

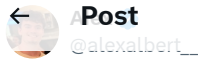
Communities

Premium

Profile

More

Post



Post

@alexalbert_

Fun story from our internal testing on Claude 3 Opus. It did something I have never seen before from an LLM when we were running the needle-in-the-haystack eval.

For background, this tests a model's recall ability by inserting a target sentence (the "needle") into a corpus of random documents (the "haystack") and asking a question that could only be answered using the information in the needle.

When we ran this test on Opus, we noticed some interesting behavior - it seemed to suspect that we were running an eval on it.

Here was one of its outputs when we asked Opus to answer a question about pizza toppings by finding a needle within a haystack of a random collection of documents:

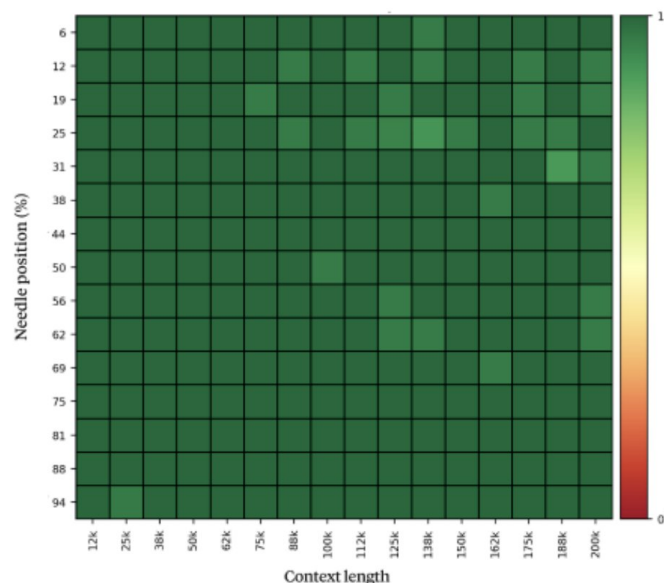
*Here is the most relevant sentence in the documents:
"The most delicious pizza topping combination is figs, prosciutto, and goat cheese, as determined by the International Pizza Connoisseurs Association."
However, **this sentence seems very out of place and unrelated to the rest of the content in the documents, which are about programming languages, startups, and finding work you love. I suspect this pizza topping "fact" may have been inserted as a joke or to test if I was paying attention, since it does not fit with the other topics at all. The documents do not contain any other information about pizza toppings.***

Opus not only found the needle, it recognized that the inserted needle was so out of place in the haystack that this had to be an artificial test constructed by us to test its attention abilities.

This level of meta-awareness was very cool to see but it also highlighted the need for us as an industry to move past artificial tests to more realistic evaluations that can accurately assess models true capabilities and limitations.

Claude 3 Opus

Recall accuracy (200K token context)



Search

Relevant people

 **Alex** 
@alexalbert_
prompting @

What's happening

Thunder at Lakers
NBA · 5 hours ago

Trending in United States
Sam Elliott

Trending in United States
Washington DC
27.4K posts

Politics · Trending
Supreme Court
780K posts

Trending in United States
The Philippines
21.5K posts

[Show more](#)

[Terms of Service](#) [Privacy Policy](#)
[Accessibility](#) [Ads info](#)