

Stress-Testing Long-Context Language Models with Lifelong ICL and Task Haystack



Xiaoyue Xu*, Qinyuan Ye*, Xiang Ren

NeurIPS 2024 (Datasets & Benchmarks Track)



Background

Longer context windows in LMs!

GPT-4 Turbo with 128K context

We released the first version of GPT-4 in March and made GPT-4 g
dev the next gen
GPT

405B

Meet Llama 3.1

70B

Our latest models expand
context length to 128k

◆ Gemini 1.5 Pro

Test: Sherlock Jr. Movie
Feature: Long context understanding (experimental)
Date: Recorded Feb 14, 2024
Format: Continuous recording of live model interaction, sequences shortened with response times shown

696,417 tokens / 1,000,000 tokens

696,161 tokens | 256 tokens

44 minute video | 1 image

☀ Claude 3.5 Sonnet

200K context window

Try Claude

Get API access

Models > Command

The Command R Model



Command R is a large language model optimized for conversational interaction and long context tasks. It targets the accuracy, enabling a long 128,000-token context length,

Command R boasts high precision on retrieval augmented generation (RAG) and tool use tasks, low latency and high throughput, a long 128,000-token context length, and strong capabilities across 10 key languages.

Background

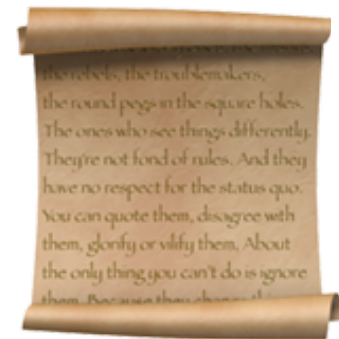
How to evaluate long-context models?

Realistic Benchmarks

- ✓ reflect real-world performance
- ✗ time-consuming, hard to scale

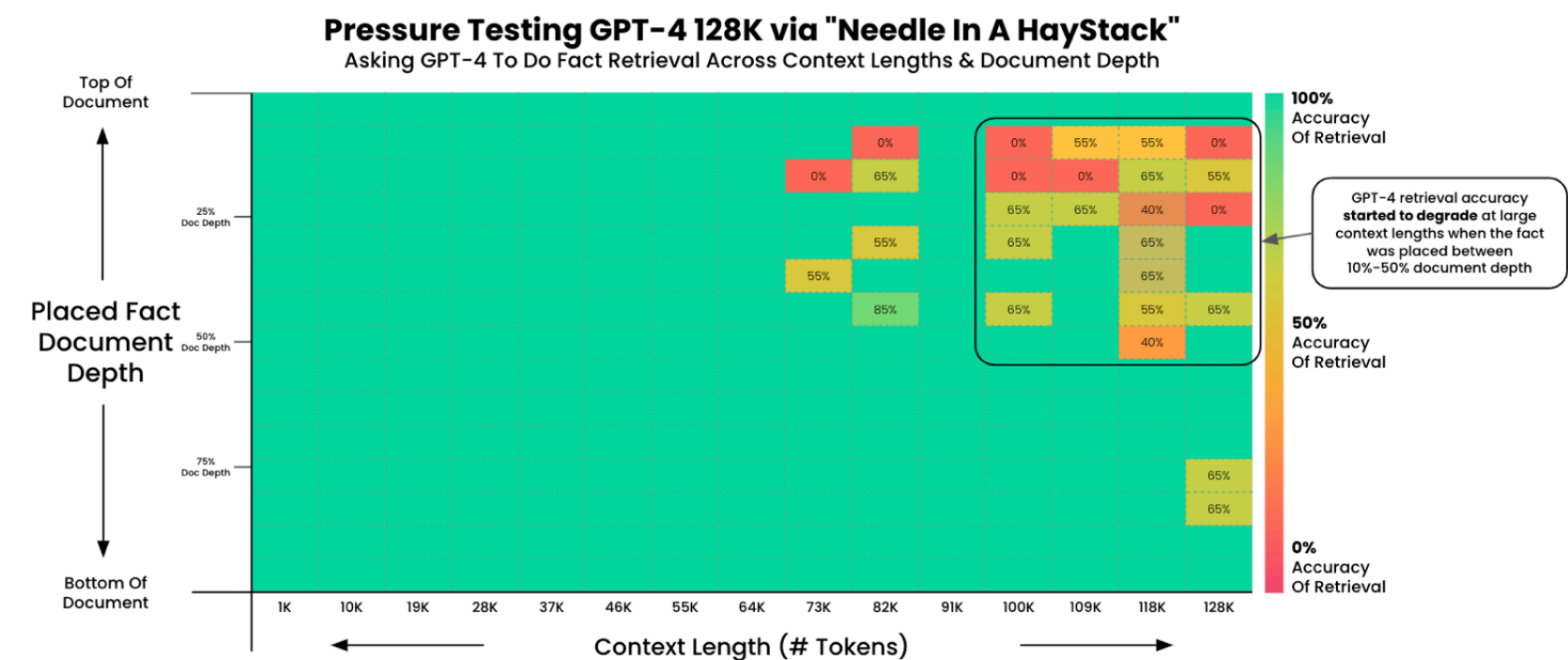
Synthetic Benchmarks

- ✓ easy to control and scale
- ✗ limited to copying-and-pasting capabilities



SCROLLS: Standardized Comparison Over Long Language Sequences

Shaham et al., 2022



https://github.com/gkamradt/LLMTest_NeedleInAHaystack

A new approach: Lifelong ICL and Task Haystack

Single-task ICL & Task Haystack

Classify if the text is humorous. // **Instruction**

Text: ...

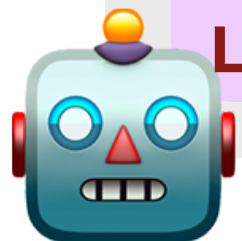

Label: humorous

Text: ...

Label: not humorous // **Demonstrations**

Text: <new text>

Label: ?



Lifelong ICL & Task Haystack

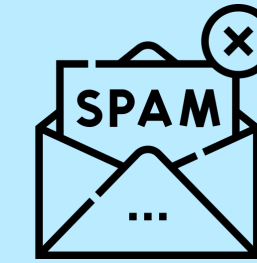
Determine if the sms message is ham or spam.

Message: ...

Label: ham.

Message: ...

Label: spam



Classify if the text is humorous. // Instruction

Text: ...

Label: humorous

Text: ...

Label: not humorous

// Demonstrations



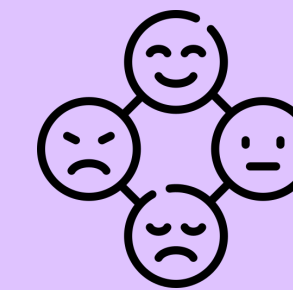
Categorize a tweet into six basic emotions: ...

Tweet: ...

Emotion: fear

Tweet: ...

Emotion: anger

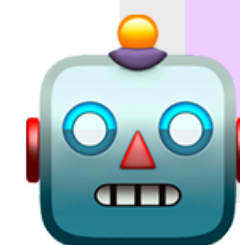


⋮

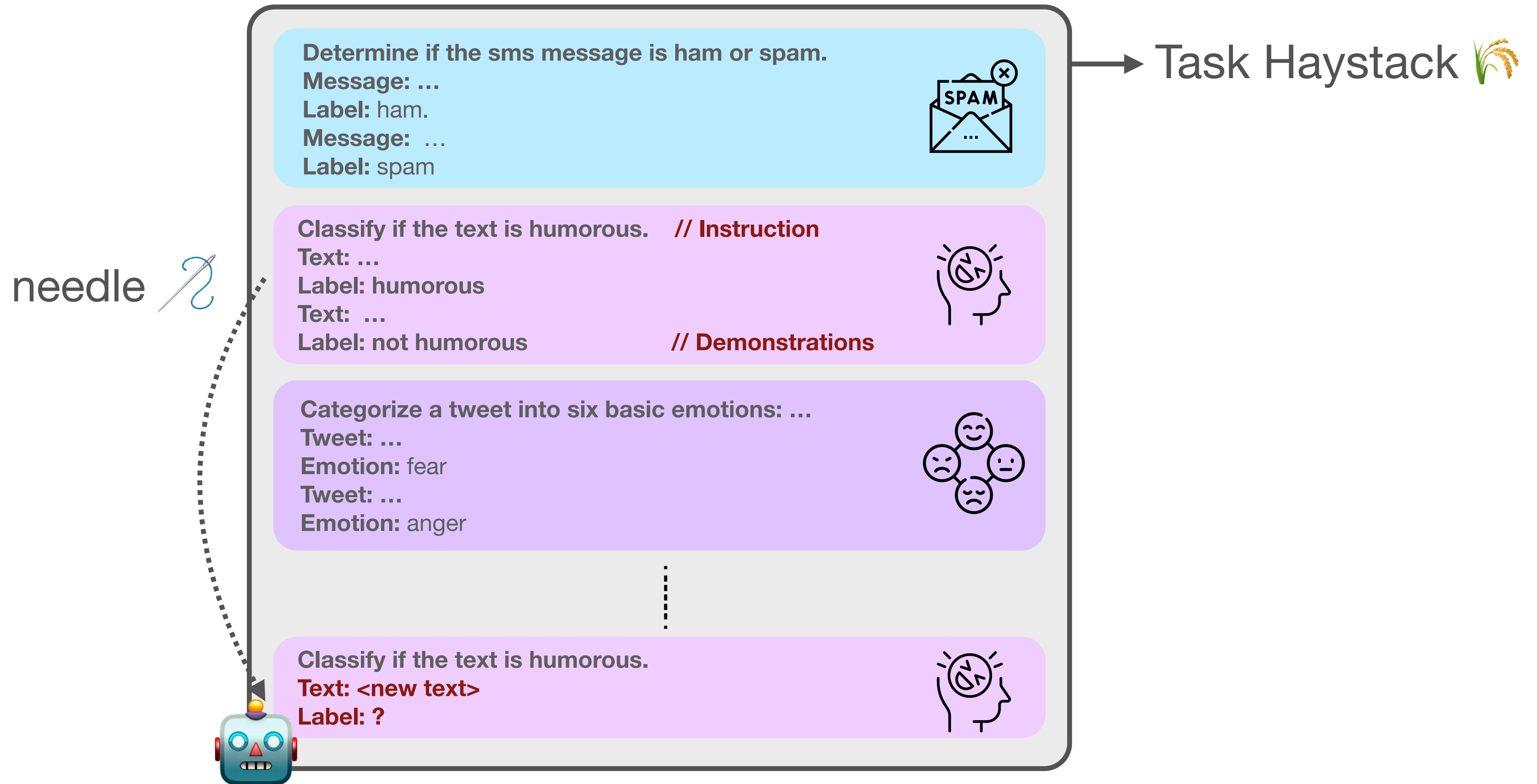
Classify if the text is humorous.

Text: <new text>

Label: ?



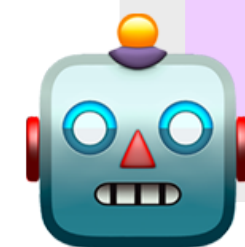
Lifelong ICL & Task Haystack



Lifelong ICL & Task Haystack

✓ **Controllable**

More tasks, more shots
→ longer context



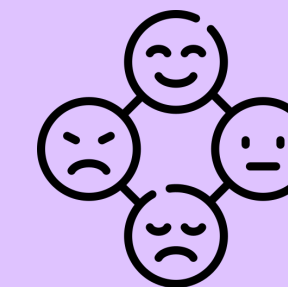
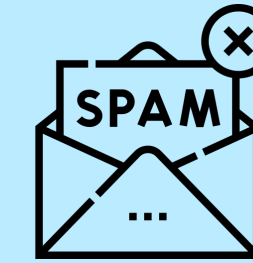
Determine if the sms message is ham or spam.
Message: ...
Label: ham.
Message: ...
Label: spam

Classify if the text is humorous. // Instruction
Text: ...
Label: humorous
Text: ...
Label: not humorous // Demonstrations

Categorize a tweet into six basic emotions: ...
Tweet: ...
Emotion: fear
Tweet: ...
Emotion: anger

...

Classify if the text is humorous.
Text: <new text>
Label: ?



✓ **More than copying-and-pasting**

ICL requires deeper, contextual understanding


✓ **With realistic elements**

Based on realistic text classification tasks

Defining "Pass Rate" in Task Haystack


Single-task ICL

Classify if the text is humorous. // **Instruction**
Text: ...
Label: humorous
Text: ...
Label: not humorous // **Demonstrations**
Text: <new text>
Label: ?




Lifelong ICL


Determine if the sms message is ham or spam.
Message: ...
Label: ham.
Message: ...
Label: spam



Classify if the text is humorous. // **Instruction**
Text: ...
Label: humorous
Text: ...
Label: not humorous // **Demonstrations**




Categorize a tweet into six basic emotions: ...
Tweet: ...
Emotion: fear
Tweet: ...
Emotion: anger



...

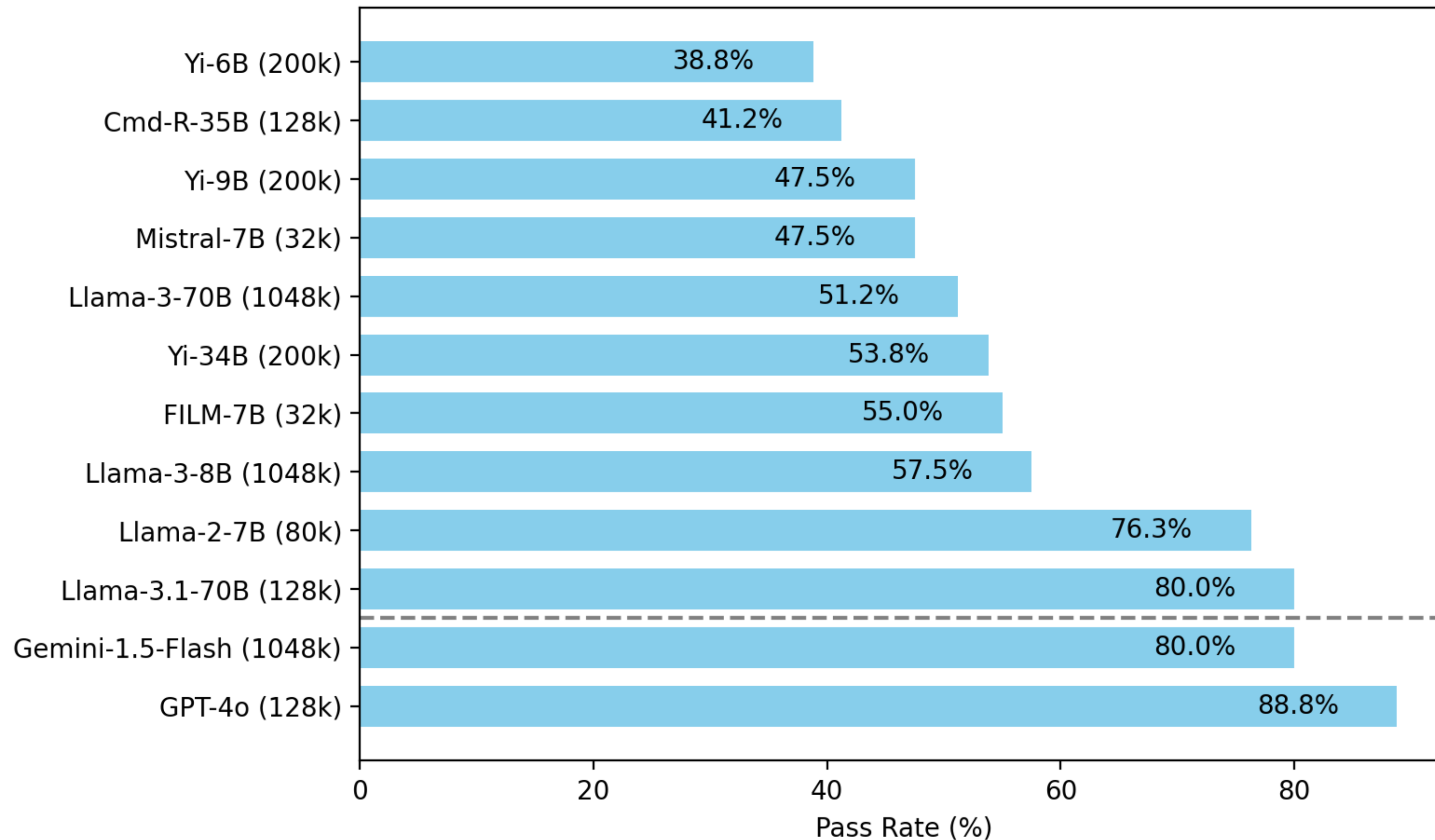
Classify if the text is humorous.
Text: <new text>
Label: ?



Model "**passes**" when performance of **Lifelong ICL** is *not significantly worse* than **Single-task ICL**

Benchmarking Long-context Language Models

In our 16-task 8-shot setting (context size=32k)



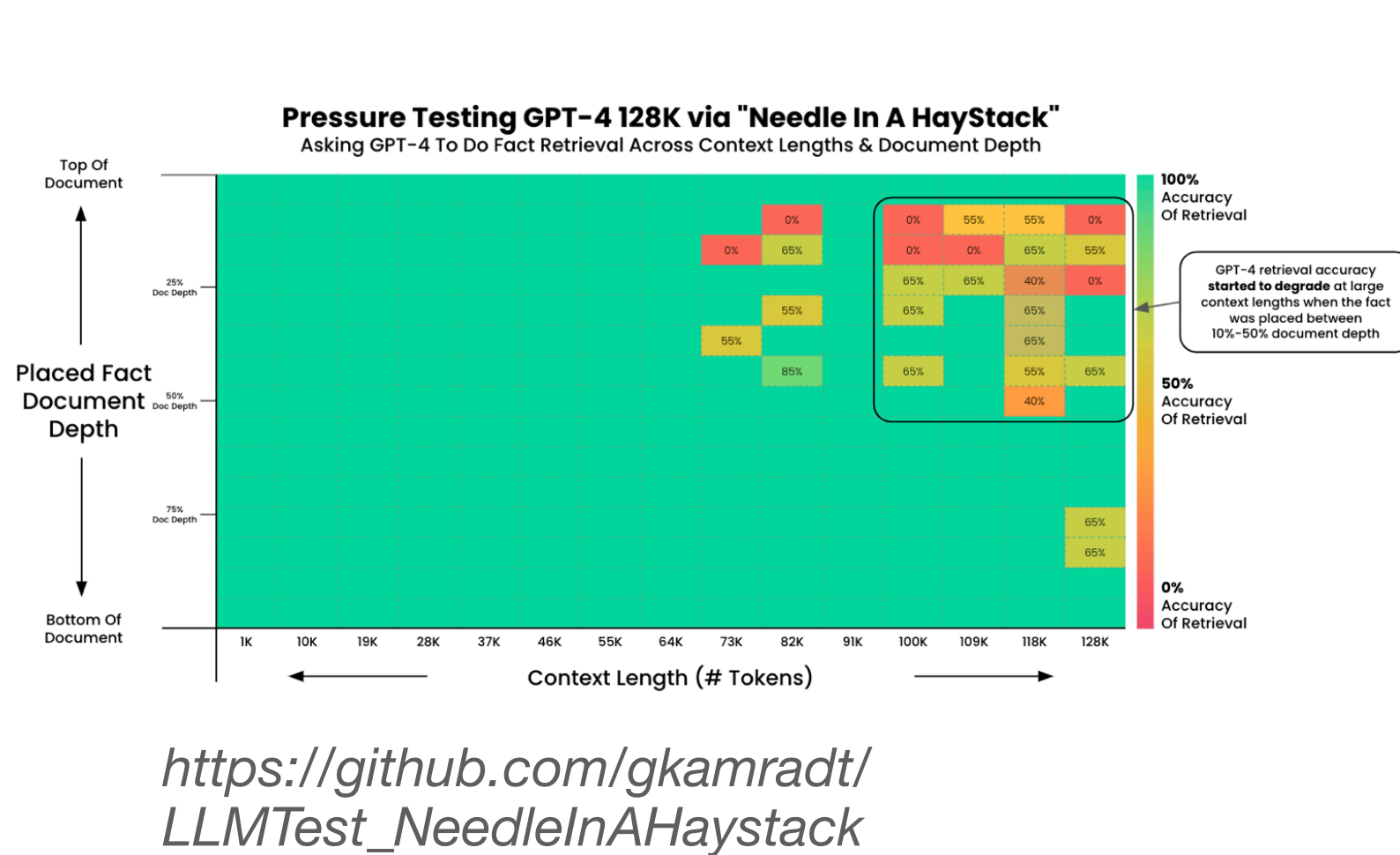
Most long-context models struggle. Failing up to 60% cases.

GPT-4o is the strongest model. Still fails ~11% of the cases

Llama-3.1-70b is the strongest open-weight model.

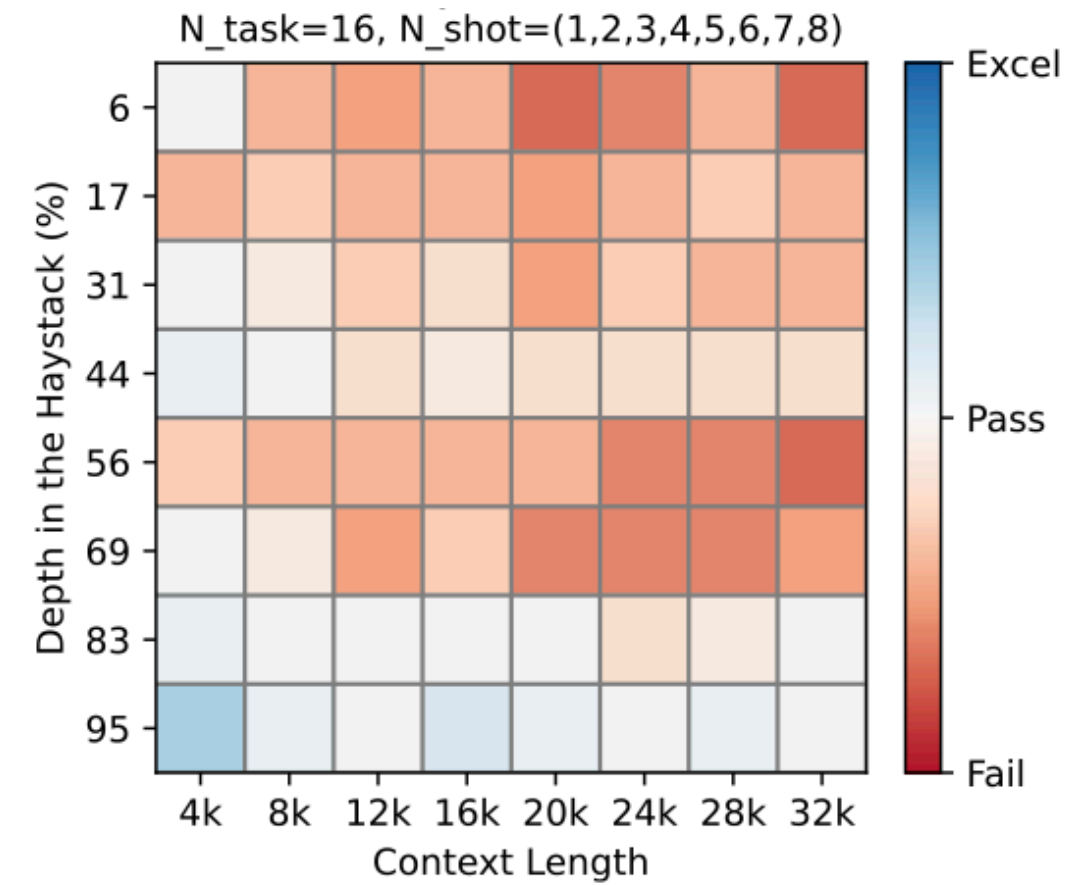
Needle-in-a-haystack Style Visualization

Needle-in-a-haystack

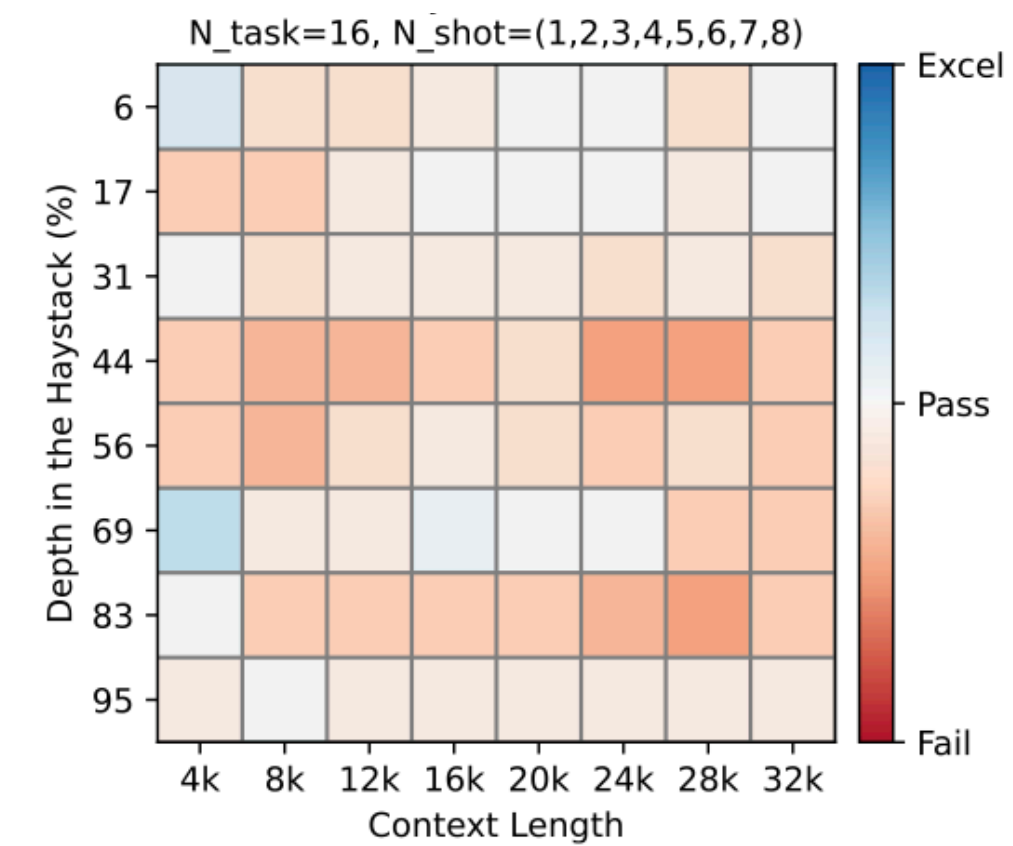


https://github.com/gkamradt/LLMTest_NeedleInAHaystack

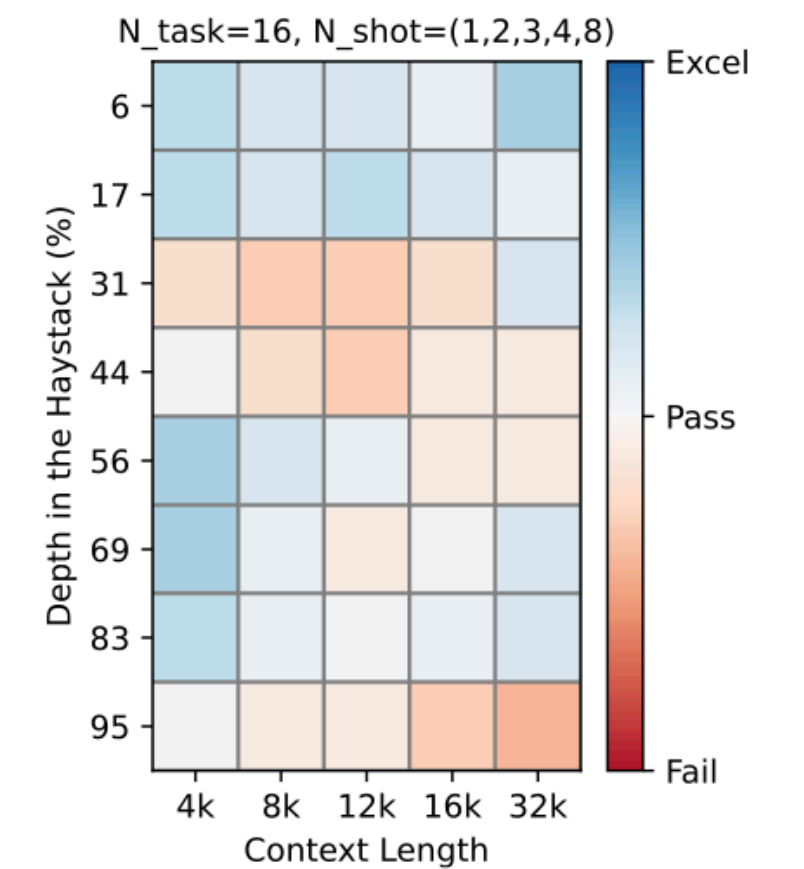
Needle-in-a-Task-haystack



Mistral-7B (32k)



Llama-3.1-70B (128k)



GPT-4o (128k)

Controlled Experiments

Setting	Input Prompt Example				Controlled Factors			
					Long Ctx.	Distraction	Recency	
Baseline (Single-task ICL)	T1 Train	T1 Test			✗	✗	✓	Distraction
Random	Random Text	T1 Train	T1 Test		✓	✓	✓	
Repeat	T1 Train	T1 Train	T1 Train	T1 Test	✓	✗	✓	
Repeat+Shuffle	T1 Train	✗ T1 Train	✗ T1 Train	T1 Test	✓	✗	✓	
Recall (Lifelong ICL)	T1 Train	T2 Train	T3 Train	T1 Test	✓	✓	✗	Recency bias
Replay	T1 Train	T2 Train	T3 Train	T1 Train	T1 Test	✓	✓	
Remove	T2 Train	T3 Train	T1 Test		✓	✓	N/A	
Paraphrase	T1 Train	T2 Train	T3 Train	↻ T1 Test	✓	✓	✗	Instruction understanding



- **Recency bias** and **distraction** both contribute to the failures in Task Haystack
- Models are sensitive to **paraphrased instructions**, indicating a lack of deeper understanding.

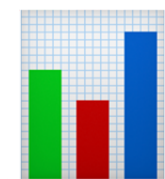
Summary



We introduce **Lifelong ICL** to evaluate long-context LMs, which challenges them to learn a sequence of language tasks through in-context learning



We develop **Task Haystack**, which comprises 64 classification tasks, to assess and diagnose how long-context LMs utilize contexts in Lifelong ICL



We benchmark 14 models and find that

- (1) SOTA model (GPT-4o) **fails ~11%** cases
- (2) Llama-3.1-70b shows the best performance among open-weight models
- (3) Other open-weight models lag behind by a large margin
- (4) Long-context models are sensitive to recency bias, distraction and paraphrased instructions