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

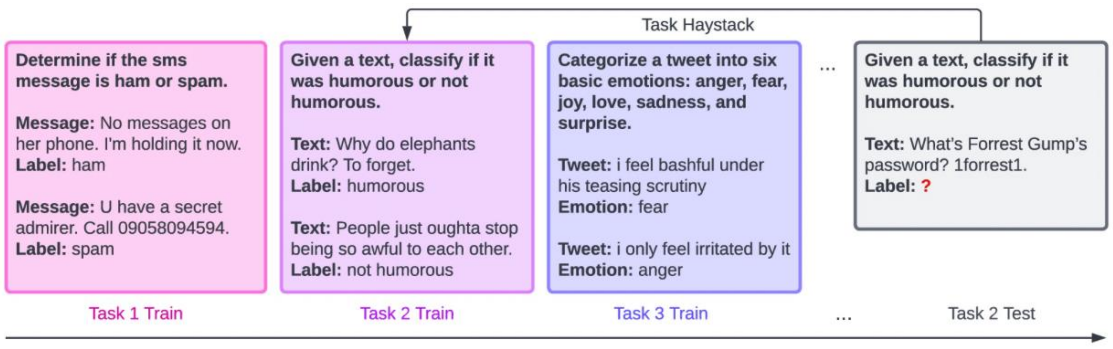
Xiaoyue Xu*, Qinyuan Ye*, Xiang Ren ✉ xiaoyue.xu.me@gmail.com {qinyuany, xiangren}@usc.edu

USC University of Southern California

Lifelong ICL and Task Haystack

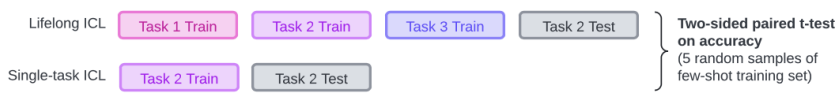
Lifelong ICL challenges long-context LMs to learn a sequence of language tasks through in-context learning.

Task Haystack assesses and diagnoses how long-context LMs utilize contexts in Lifelong ICL.



Up to 64 text classification tasks and 32k input tokens!

Defining "Pass Rate" in Task Haystack



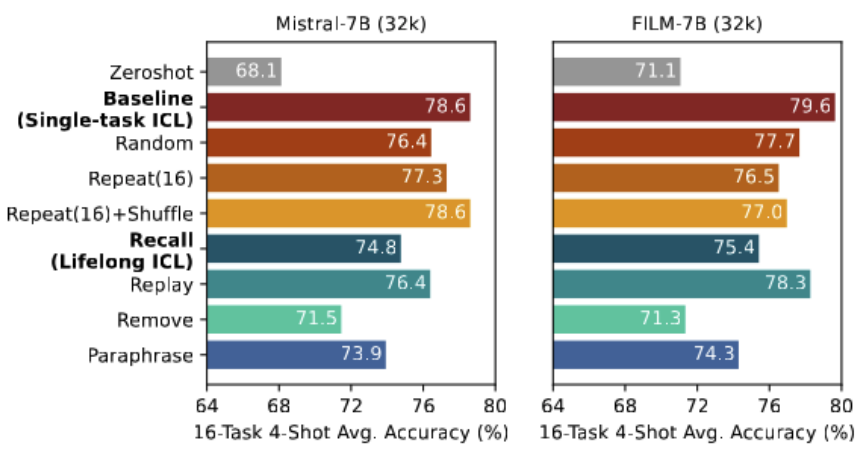
- We evaluate long-context LMs using **Lifelong ICL** and **Single-task ICL** prompts respectively.
- The model passes when performance of **Lifelong ICL** is *not significantly worse* than **Single-task ICL**.
- To pass the test, the model need to locate and make use of the relevant ICL demonstrations (the "needle") in the lengthy Lifelong ICL prompt (the "task haystack").

Part 2: Uncovering Limitations of Long-Context LMs

Controlled Settings

Setting	Input Prompt Example	Controlled Factors: Long Ctx.	Distraction	Recency
Baseline (Single-task ICL)	T1 Train **T1 Test**	✗	✗	✓
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**	✓	✓	✗
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**	✓	✓	✗

Results and Findings



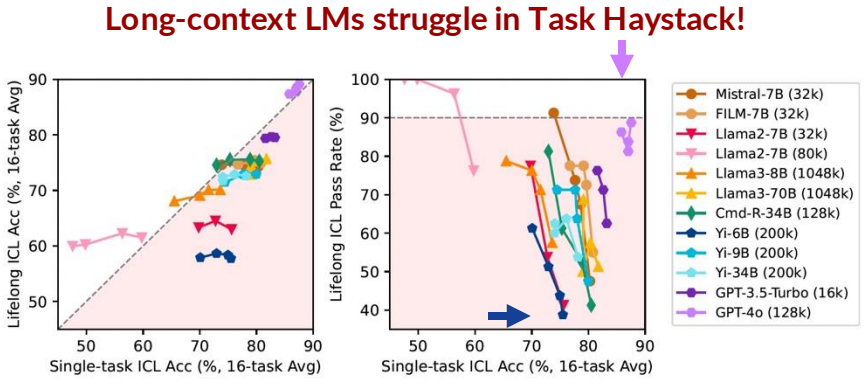
- Recency bias (**Replay** vs. **Recall**) and distraction (**Baseline** vs. **Random**) both contribute to the failures in Task Haystack.
- **Recall** vs. **Remove**: Models do locate and make use of the "needle" to some extent.
- **Recall** vs. **Paraphrase**: Models are sensitive to paraphrased instructions, indicating a lack of deeper understanding.

Additional Observations

- Failure cases are highly task-dependent. Tasks learned via ICL are more easily "forgotten". However different models tend to fail on different tasks.
- We observe positive task transfer in certain cases.
- ...

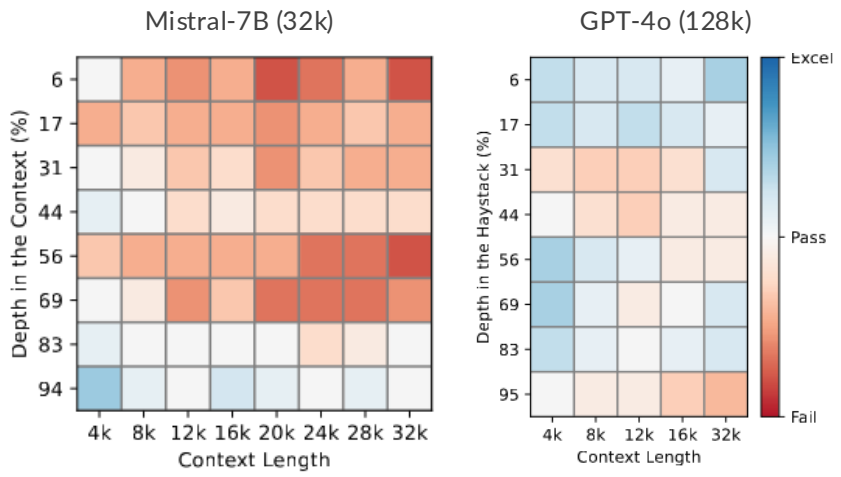
Part 1: Benchmarking Long-Context LMs

Long-context LMs struggle in Task Haystack!



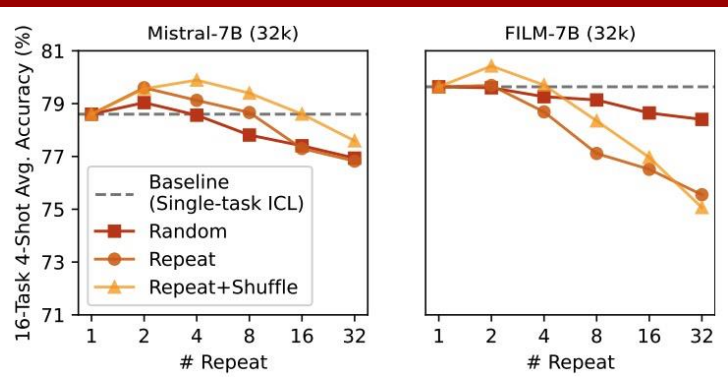
- **GPT-4o** still struggle in this setting, failing **~15%** of the cases.
- Open models we evaluate further lack behind by a large margin, failing up to **61%** of the case.
- Nearly all models fall into the **undesired area**!

Needle-in-a-haystack-style Visualization



- The original NIAH test does not tell the full story!

Part 2 (Continued)



- **Repeating the Single-task ICL prompt** leads to performance increase and then decrease! Is this "overfitting"?
- Do long inputs (regardless of the being **relevant/Repeated** or **irrelevant/Random**) give rise to undesired model behaviors?

- **Paper**: https://arxiv.org/abs/2407.16695
- **Github**: https://github.com/INK-USC/Lifelong-ICL
- **Website**: https://inklab.usc.edu/lifelong-icl/