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

Supplementary Material - Case Studies

1 Executive Summary

As introduced in the main paper, Task Haystack inherits the controllability benefits of the original needle-in-a-haystack test [Kamradt, 2023]. It facilitates easy aggregation of results by permutations, context depth, and task, enabling the creation of visualized reports to help identify the vulnerabilities of long-context LMs.

In this file, we provide visualizations of 6 sets of experiments discussed in the main paper and summarize our main findings. The experiment settings include:

- Fig. 2: Mistral-7B (32k), N-Task=16, N-Shot=8.
- Fig. 3: FILM-7B (32k), N-Task=16, N-Shot=8.
- Fig. 4: GPT-3.5-Turbo (16k), N-Task=16, N-Shot=4.
- Fig. 5: GPT-4o (128k), N-Task=16, N-Shot=8.
- Fig. 6: Mistral-7B (32k), N-Task=32, N-Shot=2.
- Fig. 7: Mistral-7B (32k), N-Task=64, N-Shot=2.

How to interpret the diagnostic report?

- The main body of the diagnostic report is an $n \times n$ matrix, where n is the number of tasks used in the experiments. The x-axis represents the task index in the Lifelong ICL stream of all tasks, while the y-axis represents the tasks themselves. If the cell at (insincere questions, index 5) is colored red, it indicates that the task of insincere questions appears at index 5 in one of the five permutations, and the performance when using the Lifelong ICL prompt is significantly worse than when using the single-task ICL prompt, resulting in a test failure in Task Haystack. White suggests no significant differences, and blue suggests that Lifelong ICL outperforms Single-task ICL. Since we run five permutations of tasks in our experiments, the figure is only sparsely colored. A grey cell means "N/A" and indicates that the task does not appear at a specific index in the five permutations.

- Below the main matrix, we plot the results by the five permutations we created. If the cell at (permutation 1, index 5) is colored red, it indicates that the task at index 5 in permutation 1 failed the Task Haystack test.

- We average each column and each row in the main $n \times n$ matrix to aggregate performance by task and by index, and visualize them at the right or the bottom of the report. This helps to investigate which tasks are more likely to fail (or excel) and to understand which positions in the context window are more vulnerable.

Submitted to the 38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks. Do not distribute.

Main Findings.

- **Failing (forgetting) and excelling (positive transfer) are highly task-dependent.** In Fig. 1 we plot the distribution of task-specific failure/excel rates in the experiments of Mistral-7B (32k), N-Task=64, N-Shot=2. "Fail (5/5)" achieves the second highest frequency, suggesting that these tasks are inherently more "forgettable" in the Lifelong ICL setting, and the performance on these tasks drops significantly regardless of their index in the context. Similarly, the bars of Excel 3/5, 4/5, 5/5 have higher frequencies than Excel 1/5, 2/5, suggesting that certain tasks are inherently more likely to benefit positive transfer from other tasks.

- **Different models demonstrate different patterns**. In Table 1, we list the names of tasks that always fail (*i.e.*, fail in 5 out of the 5 task permutations) and the names of tasks that often excel (*i.e.*, excel in more than 3 out of 5 permutations). We show that there is little consistency across different models being investigated. For example, the task brag_action often excels with Mistral-7B (32k) but always fails with FILM-7B (32k) and GPT-3.5-Turbo (16k). Similarly, the task insincere_questions also appear in both categories for different models. Our hypothesis is that the compared models may have been trained on the tasks we use, and hence their forgetting and transfer behavior is influenced. Due to the lack of transparency regarding the training details of these models, we cannot further investigate this hypothesis. However, we encourage future model developers to examine data contamination and interpret the diagnostic report in conjunction with the data contamination situation.

Table 1: **Notable Tasks By investigating Task Haystack Results.** We select tasks that always fail for a model (*i.e.*, fail in 5 out of the 5 permutations) and tasks that often excel (*i.e.*, excel in more than 3 out of 5 permutations).

Model	N-Task	N-Shot	Tasks that always fail (=5/5)	Tasks that often excel (>3/5)
Mistral-7B (32k)	16	8	insincere_questions news_data	brag_action wiki_qa
FILM-7B (32k)	16	8	brag_action emo insincere_questions	pun_detection
GPT-3.5-Turbo (16k)	16	4	amazon_counterfactual_en brag_action this_is_not_a_dataset	-
GPT-4o (128k)	16	8	-	covid_fake_news insincere_questions logical_fallacy_detection
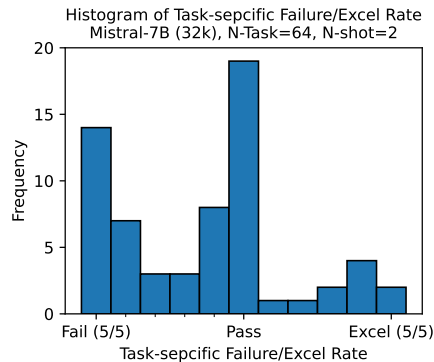
Figure 1: **Histogram Failure/Excel Rate Aggregated by Task.** We aggregate the results of Mistral-7B (32k), N-Task=64, N-Shot=2 (Fig. 7).

2

2 Diagnostic Reports

2.1 Mistral-7B, N-task=16, N-shot=8
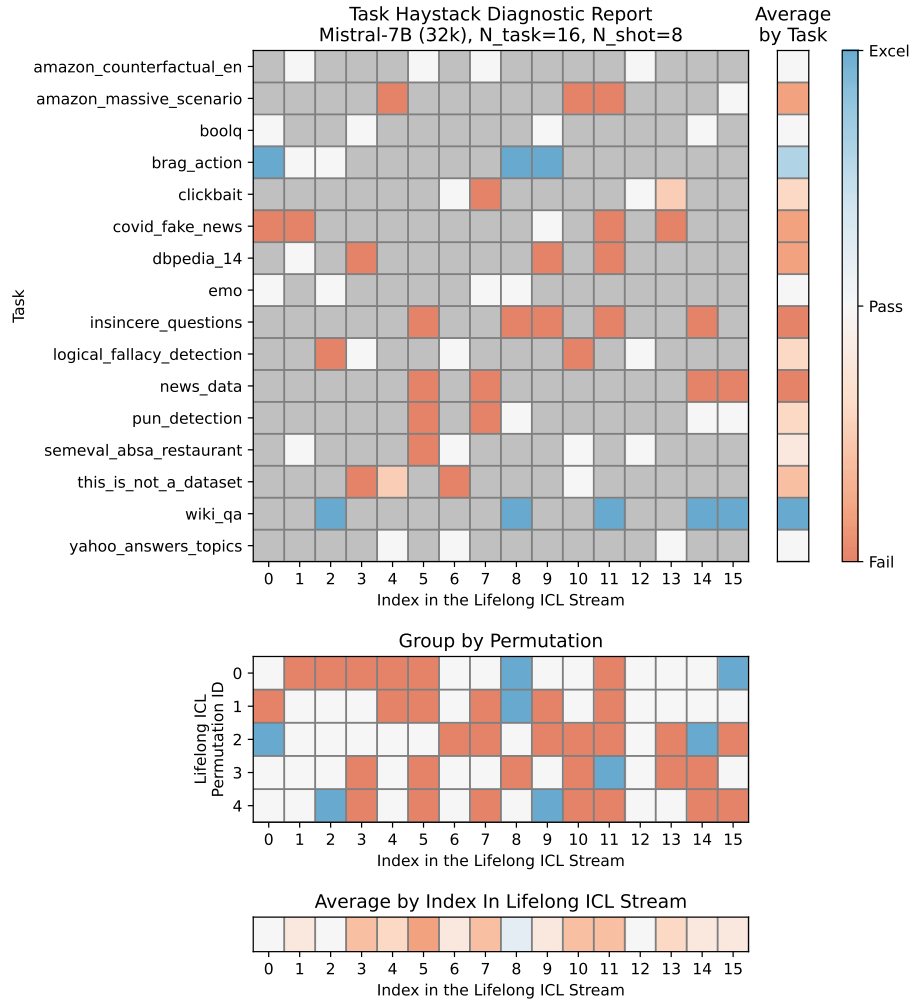
Figure 2: Diagnostic Report on Mistral-7B (32k), N-task=16, N-shot=8.

3

55 ## 2.2 FILM-7B, N-task=16, N-shot=8
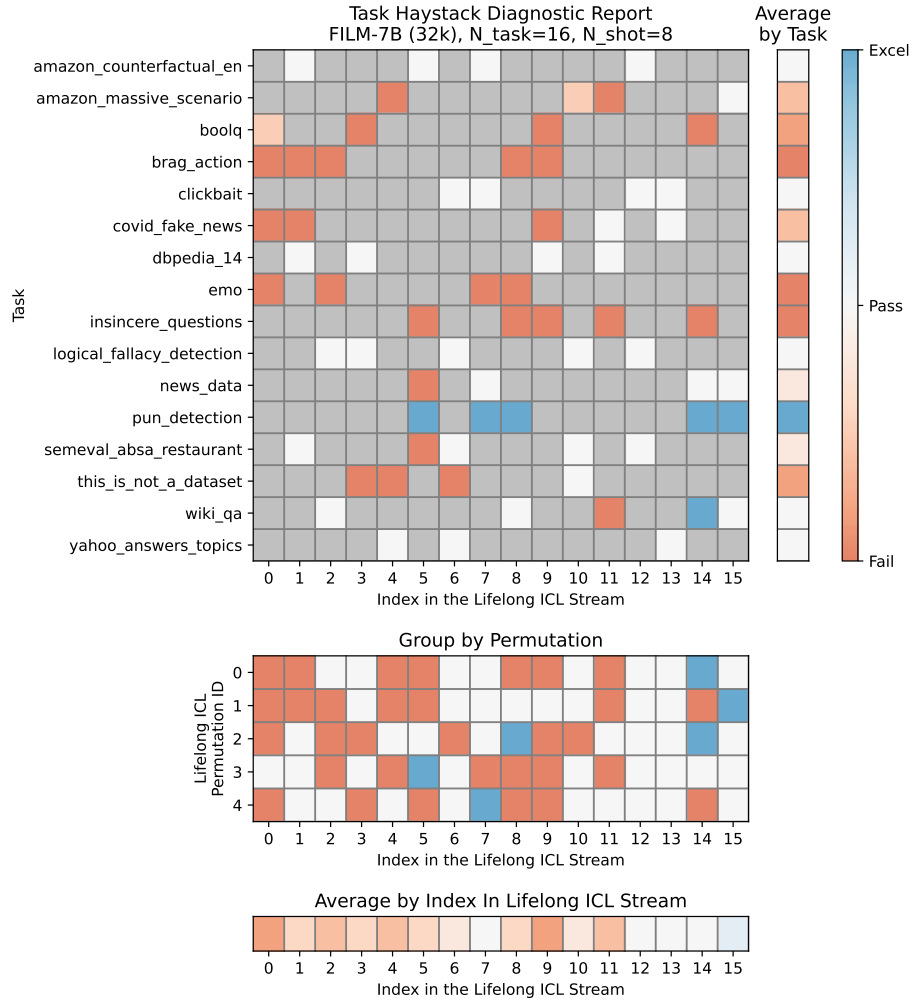
Figure 3: Diagnostic Report on FILM-7B (32k), N-task=16, N-shot=8.

4

56 ## 2.3 GPT-3.5-turbo, N-task=16, N-shot=4
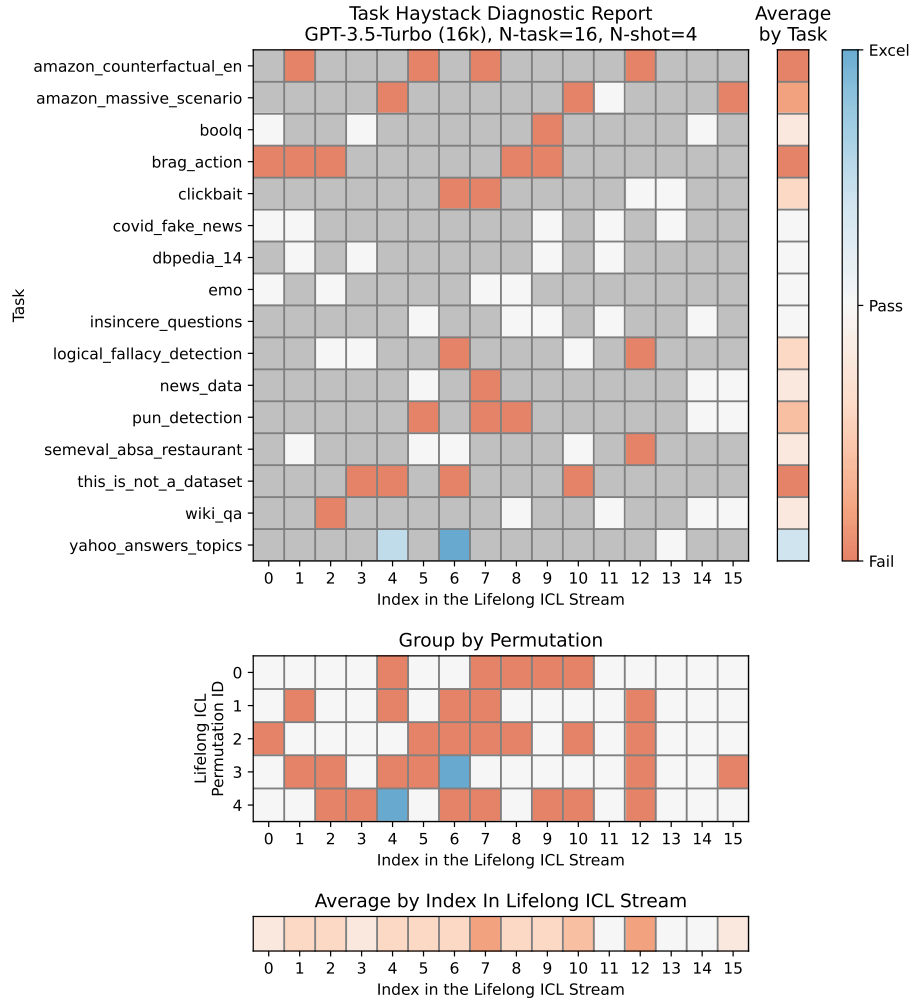
Figure 4: Diagnostic Report on GPT-3.5-Turbo (16k), N-task=16, N-shot=4.

5

57 ## 2.4 GPT-4o, N-task=16, N-shot=8
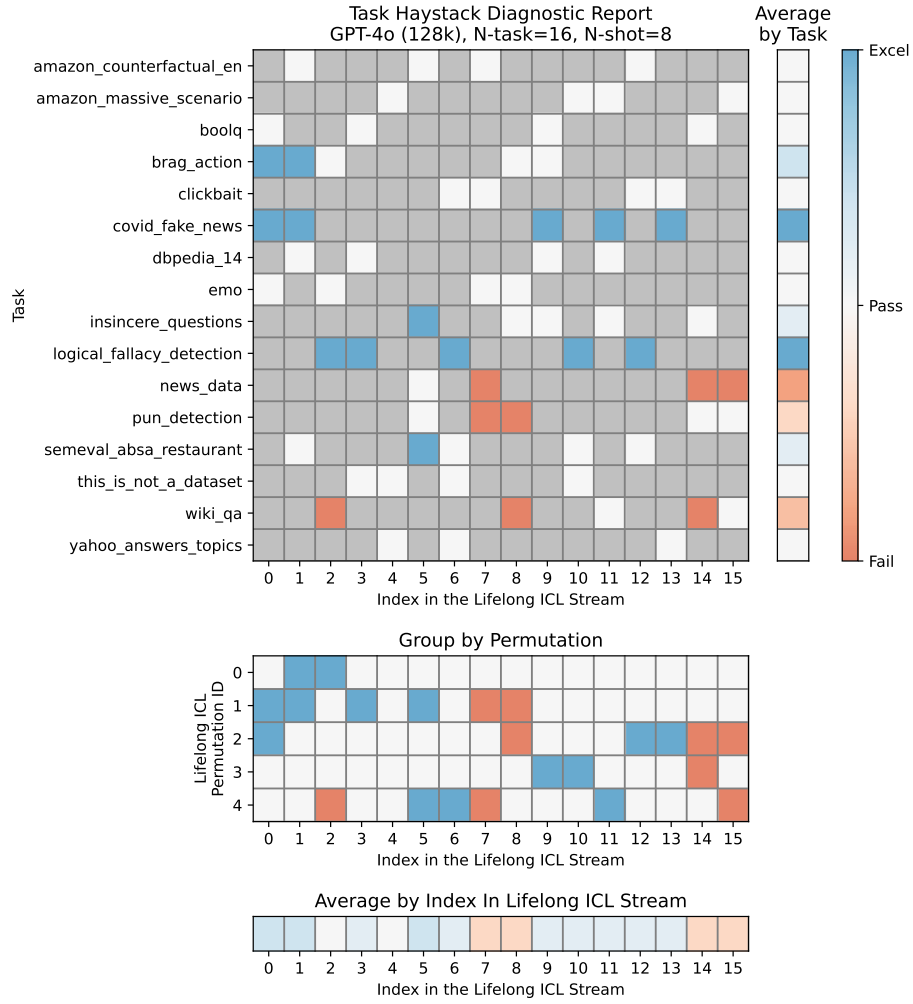
Figure 5: Diagnostic Report on GPT-4o (128k), N-task=16, N-shot=8.

6

58 ## 2.5 Mistral-7B, 32-task, 4-shot
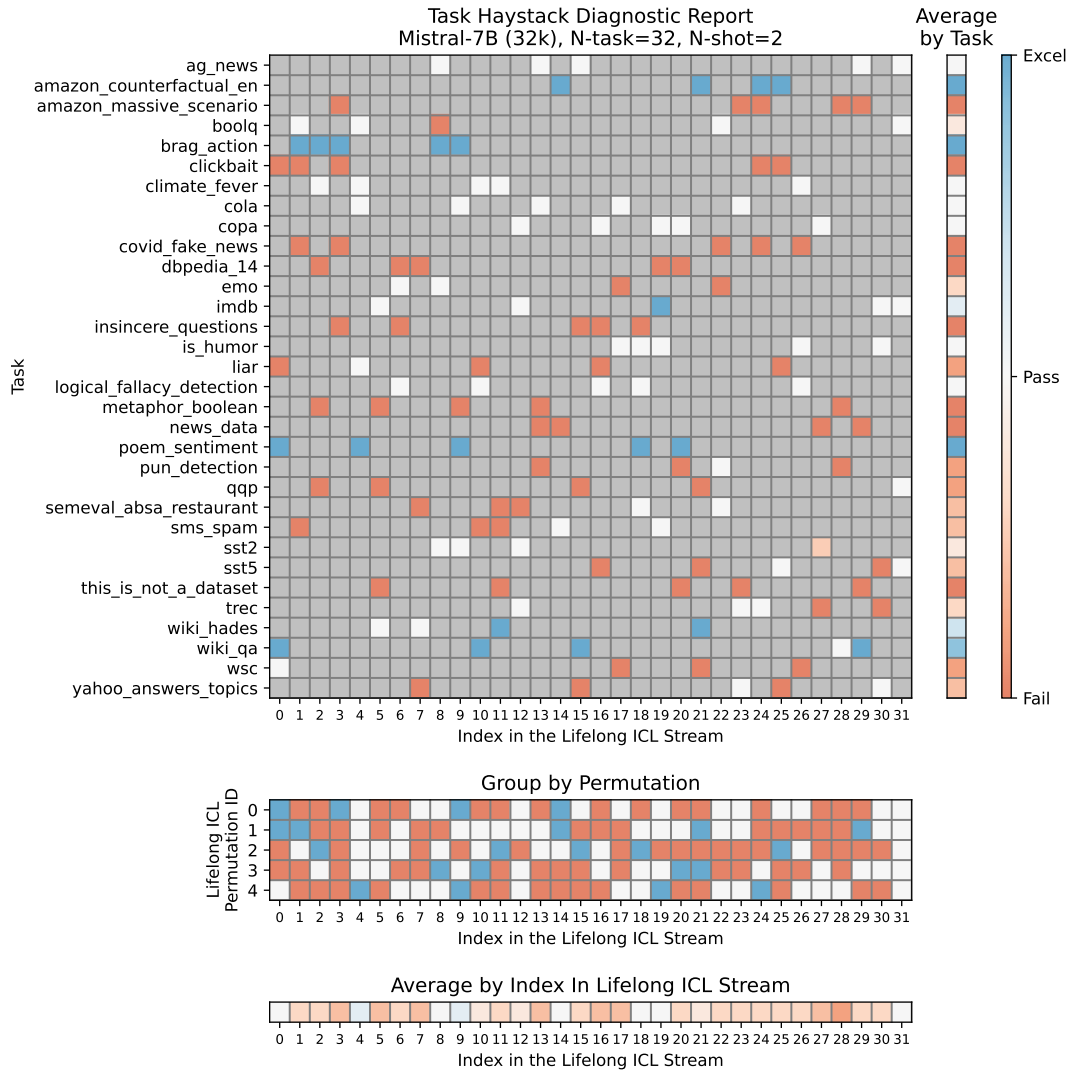
Figure 6: Diagnostic Report on Mistral-7B (32k), N-task=32, N-shot=2.

7

59 ## 2.6 Mistral-7B, 64-task, 4-shot

Figure 7: Diagnostic Report on Mistral-7B (32k), N-task=64, N-shot=2.

60 # References

Gregory Kamradt. Needle in a haystack - pressure testing llms. https://github.com/gkamradt/LLMTest_NeedleInAHaystack/tree/main, 2023.

8