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

Xiaoyue Xu [*1] **Qinyuan Ye** [*2] **Xiang Ren** [2]

Abstract

We introduce Lifelong ICL, a problem setting that challenges long-context language models (LMs) to learn from a sequence of tasks through in-context learning (ICL). We further introduce Task Haystack, an evaluation suite designed for assessing and diagnosing how long-context LMs utilize long contexts in the Lifelong ICL setting. When given a task instruction and test inputs, long-context LMs are expected to leverage the same-task demonstrations in the Lifelong ICL prompt, avoid distraction from other tasks, and achieve a test accuracy no worse than the single-task ICL baseline.

Task Haystack draws inspiration from the widely-adopted "needle-in-a-haystack" (NIAH) evaluation, but presents new and unique challenges. It demands that models (1) utilize the context with deeper understanding, rather than resorting to simple copying and pasting; (2) navigate through long streams of evolving topics and tasks, which closely approximates the complexities of real-world scenarios faced by long-context LMs. Additionally, Task Haystack inherits the controllability aspect of NIAH, providing model developers with tools to identify model vulnerabilities effectively.

We benchmark ten long-context LMs using Task Haystack. We find that state-of-the-art closed models such as GPT-4o still struggle in this setting, failing 15% of the cases on average, while all open models we evaluate further lack behind by a large margin. Further, we design controlled analysis and find that current long-context models are prone to distractibility and recency bias, as well as other limitations in robustness and instruction understanding.

[*]Equal contribution [1]Tsinghua University [2]Uniersity of Southern California. Correspondence to: Xiaoyue Xu <xiaoyue.xu.me@gmail.com>, Qinyuan Ye <qinyuany@usc.edu>, Xiang Ren <xiangren@usc.edu>.

Proceedings of the 1st Workshop on Long-Context Foundation Models, Vienna, Austria. 2024.

1. Introduction

Recent advances in model architecture (Su et al., 2024; Liu et al., 2023), training procedure (Tworkowski et al., 2023), and data engineering (Fu et al., 2024; An et al., 2024) have empowered large language models (LLMs) to handle very long contexts, reaching up to 32k or even millions of tokens (Reid et al., 2024; Anthropic, 2024) as their input. However, while long-context LM development strides forward, suitable evaluation methods haven't kept pace. Systematically evaluating long-context models' ability to leverage such long contexts remains an open challenge.

Current evaluation approaches fall into two categories. The first involves constructing benchmarks with real-world long-context tasks (Shaham et al., 2022; 2023). While valuable, creating these benchmarks is time-consuming and difficult to scale, especially for tasks requiring million-token contexts. The second approach employs synthetic evaluations like the "needle-in-a-haystack" (NIAH) test (Kamradt, 2023) or key-value retrieval tests (Liu et al., 2024b). For example, in the NIAH test, a piece of information ("The special magic number is 12345") is planted in a haystack of irrelevant contexts (Paul Graham essays) and the model is tested on answering a question about the information ("What's the special magic number?"). Although useful for initial assessment, these tests primarily measure simple copying and pasting abilities, and fail to capture how models utilize contexts when deeper understanding is required.

In this work, we offer new perspectives to long-context LM evaluation by introducing Lifelong ICL, a new problem setting that challenges these models to learn a sequence of tasks via in-context learning (ICL). Further, we introduce Task Haystack, an accompanying evaluation suite designed for systematic diagnosis of context utilization (Fig. 1). In Task Haystack, a long-context LM will be evaluated on a collection of up to 64 language tasks, prefixed with either Lifelong ICL or single-task ICL contexts. A model "passes" the test if its accuracies with Lifelong ICL prefixes are not significantly lower than using single-task ICL prefixes. The overall pass rate, averaged across tasks and different lifelong stream permutations, serves as the key metric.

Task Haystack challenges long-context LMs with unique as-

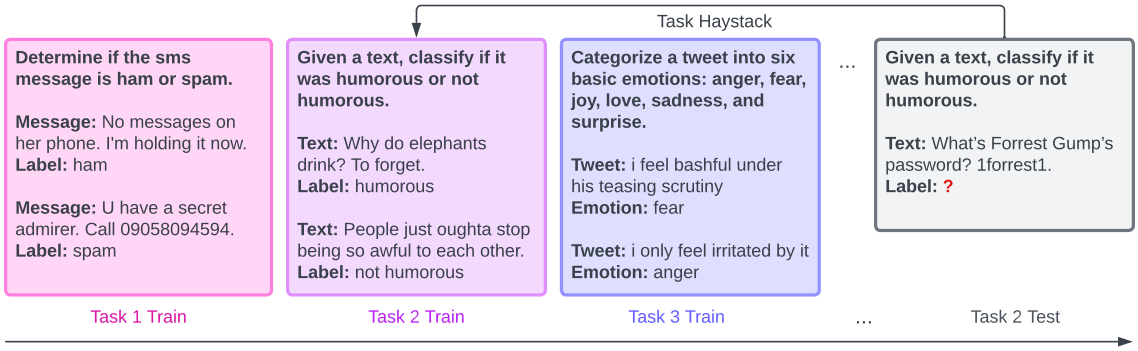

Figure 1. **Lifelong ICL and Task Haystack.** Lifelong ICL presents long-context LMs with a sequence of tasks, each containing a task instruction and few demonstrations. At test time, the model is given the one task instruction seen before and is expected to make predictions on the test input directly. A long-context LM "passes" the Task Haystack test when its accuracy using the lifelong ICL prefix (Task 1+2+3) is not significantly worse than its accuracy with the single-task ICL prefix (Task 2 only).

pects not fully covered by existing benchmarks. Firstly, Task Haystack demands a deeper comprehension of the relevant context for accurate predictions. This goes beyond simple retrieval capabilities tested by NIAH-style benchmarks, which often rely on basic copying and pasting. Secondly, Task Haystack features high information density. Every piece of information within the context might be crucial for successful prediction at test time. This differs from evaluation suites in which the important information ("needle") is positioned conspicuously, allowing models to exploit shortcuts (Anthropic, 2024). Thirdly, the lifelong task stream closely mirrors real-world applications of long-context models, such as a 24/7 personal assistant, where models encounter shifting topics (Zhao et al., 2024) and may need to resume from earlier threads in the context.

We extensively evaluate ten long-context models on Task Haystack. While all models achieve near-perfect scores on the original NIAH test, none reaches satisfactory performance on our proposed evaluation. GPT-4o emerged as the top performer among compared models, averaging an 85% pass rate and surpassing open models by a large margin. To understand the reasons behind these failures, we conduct controlled experiments that isolate factors like recency bias (models favoring information at the context's end) and distractability (models getting distracted by irrelevant information). The results confirm that both factors contribute to the performance degradation on Task Haystack. Additionally, model performance dropped when instructions were paraphrased at test time and when few-shot ICL demonstrations of a single task were repeated multiple times. These observations highlight the limitations of current long-context models in terms of their robustness and instruction understanding.

We hope that Lifelong ICL and Task Haystack serve as useful resources and testbeds for evaluating, diagnosing, and understanding long-context models. Further, we anticipate that the limitations and vulnerabilities exposed in this paper will inspire innovations in long-context model development.

2. Problem Setting

2.1. Lifelong ICL

Single-task ICL. Formally, we define a language task T as a tuple of (D^{train}, D^{test}, d), where D^{train} is the training set, D^{test} is the test set, d is a textual task description (*i.e.*, instruction). We first create a task-specific prompt p by concatenating the task description and the k-shot examples in D^{train}, *i.e.*, $p = d \oplus x_1^{train} \oplus y_1^{train} \oplus \cdots \oplus x_k^{train} \oplus y_k^{train}$. In single-task ICL setting, to make a prediction on the test input x^{test}, we concatenate the task-specific prompt and the test input, and query the language model LM to generate the prediction $\hat{y}$. We denote this process as $\hat{y} = \texttt{LM}(x^{test}|p)$ to highlight that the prediction is made by conditioning on the task-specific prompt p.

Task Collection and Task Streams. The definition above introduces how ICL is performed with one single task T. In Lifelong ICL, our new problem setting, an LM is expected to learn from a collection of n tasks, denoted as $\mathcal{T} = \{T_i\}_{i=1}^n$. To enable this, we first create a random permutation $a = (a_1, a_2, \ldots, a_n)$, thus the tasks in $\mathcal{T}$ will be ordered as $(T_{a_1}, T_{a_2}, \ldots, T_{a_n})$. For example, when $n = 3$, one possible permuatation a is $(3, 1, 2)$, and the tasks are ordered as (T_3, T_1, T_2).

Lifelong ICL. Given a permutation a, we first create the single-task ICL prompt p_{a_i} for each task T_{a_i}, and then create the lifelong ICL prompt with concatenation, *i.e.*, $p_l = p_{a_1} \oplus p_{a_2} \oplus \cdots \oplus p_{a_n}$. For each task T_{a_i} in $\mathcal{T}$, the model will be be guided with the task instruction d_{a_i} at test time, and then tested on making a prediction for the test input x_{test}, *i.e.*, $\hat{y} = \texttt{LM}(x_{test}|p_l \oplus d_{a_i})$. See Fig. 1 for the illustration of the Lifelong ICL setting.

2.2. Task Haystack

Evaluation Principle. For a test task T_{a_i}, we anticipate that a long-context LM can effectively utilize the in-context examples of that task, *i.e.*, p_{a_i}, which is a substring of the lifelong ICL prompt $p_l \oplus d_{a_i}$. To evaluate this, we compare the model performance on task T_{a_i} when conditioning on $p_l \oplus d_{a_i}$ and p_{a_i}, and expect the former to be no worse than the latter. In other words, the single-task ICL prompt p_{a_i} is the "needle" in the lifelong ICL prompt $p_l \oplus d_{a_i}$ (*i.e.*, the "task haystack").

Multiple Samples to Address ICL instability. One challenge for evaluation is the notorious instability of ICL. To account for this, our experiments will be carried out with 5 random permutations a and 5 randomly-sampled few-shot training set D_{train} for each task. This allows us to obtain a performance matrix of (t, p, r) for Lifelong ICL, where t is the task index, p is the permutation index, and r is the few-shot sample index. We will also obtain a matrix of (t, r) for Single-task ICL as the baseline performance.

For an **overall measurement**, we introduce an **average pass rate**. For each permutation a and each task T_{a_i}, we will get two groups of 5 performance metrics, when using Single-task ICL and Lifelong ICL respectively. The model scores 1 when the the Lifelong ICL performance is not significantly worse than the Single-task ICL performance (captured by a two-sided t-test), and scores 0 otherwise. The average pass rate is then computed by averaging over the different permutations and tasks. See Fig. 6 for an illustration.

For a **fine-grained analysis**, our experiment results allow us to **visualize the pass rates grouped by position in the task stream, by the task, or by the sampled permutations**. This enables straight-forward visualizations as done in the needle-in-a-haystack test, providing an easy tooling to analyze questions such as which positions in the context are more vulnerable, or which tasks are more easily forgotten.

3. Experiment Details

Task Selection. While the problem setting in §2 is generic and admits any language task, in this work we instantiate the setting with a narrower task distribution for initial exploration. After careful manual selection[1], we obtain a collection of 64 classification tasks, covering a wide range of domains and label spaces. We provide detailed descriptions of all 64 tasks, including their references and license information in Table 4.

Models. We evaluate 8 open-weight long-context LMs on the proposed Task Haystack evaluation: Mistral-7B (32k) (Jiang et al., 2023), FILM-7B (32k) (An et al., 2024), Llama2-7B (32k) (TogetherAI, 2024), Llama2-7B (80k) (Fu et al., 2024), Llama3-8B (1048k) (GradientAI, 2024), Yi-6B/9B/34B (200k) (01.AI et al., 2024). These models represent various long-context modeling techniques, model scale, and base pre-trained models. We provide more details of these models in Table 2. For closed models, we evaluate gpt-3.5-turbo and gpt-4o from OpenAI.

Controlling the Context Length. We consider creating long contexts controllably with two strategies, **Scale-Task** and **Scale-Shot**. In the first setting, we fix $n_{shot} = 2$ and experiment with $n_{task} \in \{8, 16, 24, 32, 40, 48, 56, 64\}$. In the second setting, we fix $n_{task} = 16$ and experiment with $n_{shot} \in \{1, 2, 3, 4, 5, 6, 7, 8\}$. With these strategies, we effectively create contexts of sizes ranging from 4k to 32k tokens. We defer more implementation and engineering details in Appendix A.3.

4. Results and Analysis

Long-context LMs struggle in Task Haystack. We present the results of the Scale-Shot setting in Table 1 and the results of the Scale-Task setting in Table 7. Along with the average pass rate introduced in §2.2, we also report the average accuracies over all tasks in both Single-task and Lifelong ICL settings. The overall pass rates fall below 90% in 34 out of 38 cases reported in Table 1 and in 29 out of 32 cases in Table 7. When scaling to 32k context in the 16-shot 8-shot setting (rightmost column in Table 1), 6 out of 7 open-weight models exhibit a pass rate of less than 60%, suggesting that these models are still far from passing the Task Haystack evaluation. In the most extreme case, Yi-6B (200k) achieves a pass rate of merely 38.8% in this setting.

A Holistic View of Accuracies and Pass Rates. One advantage of the pass rate metric is that it isolates the long-context modeling capabilities from models' core capabilities. However, using pass rate as the only metric may inadvertently create a shortcut where a model can achieve perfect pass rates by simply performing poorly in both the Single-task ICL and the Lifelong ICL setting.

To have a holistic view on this, we visualize the results from our Scale-Shot experiments by plotting the Lifelong ICL accuracy and Pass Rate as a function of Single-task ICL accuracy in Fig. 3. We observe that GPT-4o outperforms all other models significantly, in terms of both the ICL accuracy and the pass rate. Mistral-7B and FILM-7B (fine-tuned from Mistral-7B) achieves the strongest performance among open-weight models evaluated.

One outlier that we notice is the Llama2-7B (80k) model, which achieves low ICL accuracies but high pass rates. Note

[1] We discuss our key considerations when selecting tasks in §D and discuss the limitations in §E.

Table 1. **Main Results: Fixing 16 Tasks, Scaling the Number of Shots**. "s-acc" stands for single-task ICL accuracy averaged over all 16 tasks, and "l-acc" stands for lifelong ICL accuracy. "pass" represents the average pass rate defined in §2.2, *i.e.*, percentage of cases that lifelong ICL is not significantly worse than single-task ICL among 5 random samples of few-shot training sets. l-acc is expected to be not worse than s-acc, and the pass rate is expected to be close to 100%.

Model	0-shot	1-shot (4k)			2-shot (8k)			4-shot (16k)			8-shot (32k)		
	s-acc	s-acc	l-acc	pass	s-acc	l-acc	pass	s-acc	l-acc	pass	s-acc	l-acc	pass
Mistral-7B (32k)	68.1	73.9	74.6	91.2	77.6	74.6	73.8	78.6	74.8	67.5	80.3	74.2	47.5
FILM-7B (32k)	71.1	76.7	74.7	77.5	79.1	75.1	77.5	79.6	75.4	72.5	80.8	74.9	55.0
Llama2-7B (32k)	61.9	69.8	63.3	77.5	72.8	64.5	53.8	75.6	63.0	41.2	78.0	-	-
Llama2-7B (80k)	38.4	47.6	60.0	100.0	49.8	60.2	100.0	56.3	62.3	96.3	59.8	61.5	76.3
Llama3-8B (1048k)	51.2	65.5	68.1	78.8	70.0	69.1	76.2	71.5	70.1	71.3	73.6	70.1	57.5
Yi-6B (200k)	51.3	70.1	57.9	61.3	73.0	58.6	51.2	75.0	58.4	43.8	75.5	57.7	38.8
Yi-9B (200k)	57.0	74.5	71.5	71.2	77.7	72.9	71.2	78.0	72.9	63.7	80.0	72.9	47.5
Yi-34B (200k)	63.1	74.1	71.7	62.5	74.1	72.4	60.0	76.1	72.9	63.8	78.2	72.6	53.8
GPT-3.5-Turbo (16k)	78.3	81.6	76.3	73.8	82.6	79.6	71.3	83.2	79.5	62.5	81.8	-	-
GPT-4o (128k)	70.7	85.8	87.4	86.3	87.0	87.8	81.3	87.0	88.4	83.8	87.5	89.1	88.8

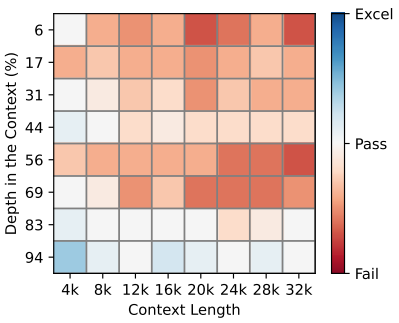

Figure 2. **Task Haystack Results with FILM-7B (32k)** (N-task=16, N-shot=1,2,...,8) visualized in the needle-in-a-haystack style heatmap.

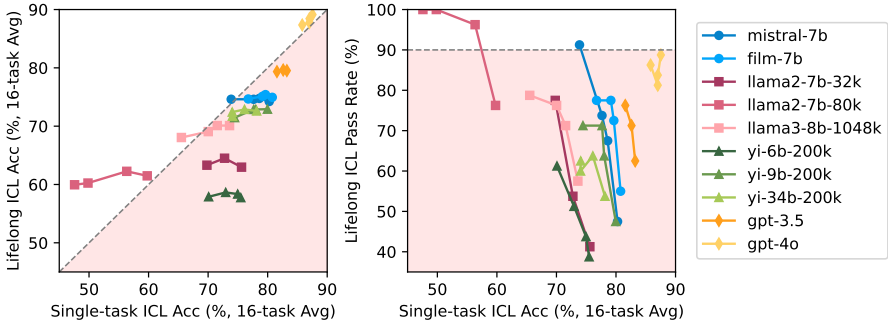

Figure 3. **Visualizing Lifelong ICL accuracy (l-acc) and pass rate as a function of single-task ICL accuracy (s-acc).** Each line is constructed by varying the number of shots in {1,2,4,8} while fixing 16 tasks. Most models fall into the undesired area (light red). GPT-4o shows the strongest overall performance in our evaluation.

that this model is trained solely on language modeling objectives without further instruction tuning or RLHF, which may be the reason behind this trend. This observation also suggests that the pass rate should always be considered together with other metrics that access the model's core capabilities.

Visualization and Diagnostic Tool for Task Haystack. Our Task Haystack evaluation enables straightforward visualization for diagnosing model vulnerabilities. In Fig. 2 we present the results of Task Haystack (Scale-Shot Setting) in a way similar to the original needle-in-a-haystack eval. While FILM-7B achieves near-perfect results in the original NIAH evaluation, Fig. 2 suggests that it's vulnerable when the context length exceeds 12k, particularly for content that appears in the first 75% of the context window. We report the visualization for all compared models in Fig. 9-16. In addition to the NIAH-style visualization, we provide an example of aggregating results by permutations, by depth in the context, and by task in Figure 19.

Controlled Analysis. In Appendix §B, we investigate the reasons behind the model's failures on Task Haystack with various controlled settings, such as repeating the single-task ICL prompt multiple times, or replaying the test task at the end of the prompt. The results suggest that current long-context models are easily distracted by irrelevant contexts and affected by recency bias, and exhibit other vulnerabilities in robustness and instruction understanding.

5. Conclusion

In this paper, we introduced the Lifelong ICL problem setting, developed the Task Haystack evaluation suite, and focused on evaluating and diagnosing current long-context LMs on Task Haystack. Our experiments on ten recent long-context LMs revealed that while they excel at retrieving and pasting information within long contexts, their ability to fully exploit the contextual information remains limited. We hope Lifelong ICL and Task Haystack serve as valuable tools for diagnosing and advancing the development of future long-context LMs. We also hope that Lifelong ICL serves as an initial but meaningful step towards gradient-free algorithms in lifelong learning settings.

References

01.AI, :, Young, A., Chen, B., Li, C., Huang, C., Zhang, G., Zhang, G., Li, H., Zhu, J., Chen, J., Chang, J., Yu, K., Liu, P., Liu, Q., Yue, S., Yang, S., Yang, S., Yu, T., Xie, W., Huang, W., Hu, X., Ren, X., Niu, X., Nie, P., Xu, Y., Liu, Y., Wang, Y., Cai, Y., Gu, Z., Liu, Z., and Dai, Z. Yi: Open foundation models by 01.ai, 2024.

Agarwal, R., Singh, A., Zhang, L. M., Bohnet, B., Chan, S., Anand, A., Abbas, Z., Nova, A., Co-Reyes, J. D., Chu, E., et al. Many-shot in-context learning. *arXiv preprint arXiv:2404.11018*, 2024.

Alex Ellis, inversion, J. E. P. G. W. C. Quora insincere questions classification, 2018. URL https://kaggle.com/competitions/quora-insincere-questions-classification.

Almeida, T. A., Hidalgo, J. M. G., and Yamakami, A. Contributions to the study of sms spam filtering: new collection and results. In *Proceedings of the 11th ACM Symposium on Document Engineering*, DocEng '11, pp. 259–262, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450308632. doi: 10.1145/2034691.2034742. URL https://doi.org/10.1145/2034691.2034742.

An, S., Ma, Z., Lin, Z., Zheng, N., and Lou, J.-G. Make your llm fully utilize the context. *arXiv preprint arXiv:2404.16811*, 2024.

Anthropic, A. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*, 2024.

Bai, Y., Lv, X., Zhang, J., Lyu, H., Tang, J., Huang, Z., Du, Z., Liu, X., Zeng, A., Hou, L., Dong, Y., Tang, J., and Li, J. Longbench: A bilingual, multitask benchmark for long context understanding, 2023.

Bertsch, A., Ivgi, M., Alon, U., Berant, J., Gormley, M. R., and Neubig, G. In-context learning with long-context models: An in-depth exploration. *arXiv preprint arXiv:2405.00200*, 2024.

Bhagavatula, C., Hwang, J. D., Downey, D., Bras, R. L., Lu, X., Qin, L., Sakaguchi, K., Swayamdipta, S., West, P., and Choi, Y. I2d2: Inductive knowledge distillation with neurologic and self-imitation. *arXiv preprint arXiv:2212.09246*, 2022.

Biesialska, M., Biesialska, K., and Costa-Jussa, M. R. Continual lifelong learning in natural language processing: A survey. *arXiv preprint arXiv:2012.09823*, 2020.

Bingler, J., Kraus, M., Leippold, M., and Webersinke, N. How cheap talk in climate disclosures relates to climate initiatives, corporate emissions, and reputation risk. Working paper, Available at SSRN 3998435, 2023.

Bird, S., Dale, R., Dorr, B., Gibson, B., Joseph, M., Kan, M.-Y., Lee, D., Powley, B., Radev, D., and Tan, Y. F. The ACL Anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., and Tapias, D. (eds.), *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May 2008. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2008/pdf/445_paper.pdf.

Biswal, M. Iitjee neet aiims students questions data. https://www.kaggle.com/datasets/mrutyunjaybiswal/iitjee-neet-aims-students-questions-data, 2020.

Bizzoni, Y. and Lappin, S. Predicting human metaphor paraphrase judgments with deep neural networks. In Beigman Klebanov, B., Shutova, E., Lichtenstein, P., Muresan, S., and Wee, C. (eds.), *Proceedings of the Workshop on Figurative Language Processing*, pp. 45–55, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-0906. URL https://aclanthology.org/W18-0906.

Chakraborty, A., Paranjape, B., Kakarla, S., and Ganguly, N. Stop clickbait: Detecting and preventing clickbaits in online news media, 2016.

Chapuis, E., Colombo, P., Manica, M., Labeau, M., and Clavel, C. Hierarchical pre-training for sequence labelling in spoken dialog. In Cohn, T., He, Y., and Liu, Y. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 2636–2648, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.239. URL https://aclanthology.org/2020.findings-emnlp.239.

Chatterjee, A., Narahari, K. N., Joshi, M., and Agrawal, P. Semeval-2019 task 3: Emocontext contextual emotion detection in text. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pp. 39–48, Minneapolis, Minnesota, USA, 2019. Association for Computational Linguistics. doi: 10.18653/v1/S19-2005. URL https://www.aclweb.org/anthology/S19-2005.

Choi, M., Pei, J., Kumar, S., Shu, C., and Jurgens, D. Do LLMs understand social knowledge? evaluating the sociability of large language models with SocKET benchmark. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 11370–11403,

Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.699. URL https://aclanthology.org/2023.emnlp-main.699.

cjadams, Daniel Borkan, i. J. S. L. D. L. V. n. Jigsaw unintended bias in toxicity classification, 2019. URL https://kaggle.com/competitions/jigsaw-unintended-bias-in-toxicity-classification.

Clark, C., Lee, K., Chang, M.-W., Kwiatkowski, T., Collins, M., and Toutanova, K. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In Burstein, J., Doran, C., and Solorio, T. (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2924–2936, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1300. URL https://aclanthology.org/N19-1300.

De Gibert, O., Perez, N., García-Pablos, A., and Cuadros, M. Hate speech dataset from a white supremacy forum. *arXiv preprint arXiv:1809.04444*, 2018.

De Marneffe, M.-C., Simons, M., and Tonhauser, J. The commitmentbank: Investigating projection in naturally occurring discourse. In *proceedings of Sinn und Bedeutung*, volume 23, pp. 107–124, 2019.

Dernoncourt, F. and Lee, J. Y. Pubmed 200k rct: a dataset for sequential sentence classification in medical abstracts. *arXiv preprint arXiv:1710.06071*, 2017.

Diggelmann, T., Boyd-Graber, J., Bulian, J., Ciaramita, M., and Leippold, M. Climate-fever: A dataset for verification of real-world climate claims, 2020.

Dolan, W. B. and Brockett, C. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*, 2005. URL https://aclanthology.org/I05-5002.

Ferreira, W. and Vlachos, A. Emergent: a novel dataset for stance classification. In Knight, K., Nenkova, A., and Rambow, O. (eds.), *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1163–1168, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1138. URL https://aclanthology.org/N16-1138.

FitzGerald, J., Hench, C., Peris, C., Mackie, S., Rottmann, K., Sanchez, A., Nash, A., Urbach, L., Kakarala, V., Singh, R., Ranganath, S., Crist, L., Britan, M., Leeuwis, W., Tur, G., and Natarajan, P. Massive: A 1m-example multilingual natural language understanding dataset with 51 typologically-diverse languages, 2022.

Fu, Y., Panda, R., Niu, X., Yue, X., Hajishirzi, H., Kim, Y., and Peng, H. Data engineering for scaling language models to 128k context. *arXiv preprint arXiv:2402.10171*, 2024.

García-Ferrero, I., Altuna, B., Alvez, J., Gonzalez-Dios, I., and Rigau, G. This is not a dataset: A large negation benchmark to challenge large language models. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 8596–8615, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.531. URL https://aclanthology.org/2023.emnlp-main.531.

GradientAI. Llama-3-8b-instruct-gradient-1048k model card, 2024. URL https://huggingface.co/gradientai/Llama-3-8B-Instruct-Gradient-1048k.

Grano, G., Sorbo, A. D., Mercaldo, F., Visaggio, C. A., Canfora, G., and Panichella, S. Android apps and user feedback: A dataset for software evolution and quality improvement. In *Proceedings of the 2Nd ACM SIGSOFT International Workshop on App Market Analytics*, WAMA 2017, pp. 8–11, New York, NY, USA, January 2017. ACM. doi: 10.1145/3121264.3121266. URL https://doi.org/10.5167/uzh-139426.

Guha, N., Nyarko, J., Ho, D., Ré, C., Chilton, A., Chohlas-Wood, A., Peters, A., Waldon, B., Rockmore, D., Zambrano, D., et al. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in Neural Information Processing Systems*, 36, 2024.

Holzenberger, N., Blair-Stanek, A., and Durme, B. V. A dataset for statutory reasoning in tax law entailment and question answering, 2020.

Hsieh, C.-P., Sun, S., Kriman, S., Acharya, S., Rekesh, D., Jia, F., Zhang, Y., and Ginsburg, B. Ruler: What's the real context size of your long-context language models?, 2024.

Hu, M. and Liu, B. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pp. 168–177, New York, NY, USA, 2004. Association for Computing Machinery. ISBN 1581138881. doi: 10.1145/1014052.1014073. URL https://doi.org/10.1145/1014052.1014073.

Ji, J., Liu, M., Dai, J., Pan, X., Zhang, C., Bian, C., Zhang, C., Sun, R., Wang, Y., and Yang, Y. Beavertails: Towards improved safety alignment of llm via a human-preference dataset, 2023.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. Mistral 7b, 2023.

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

Kim, H., Yu, Y., Jiang, L., Lu, X., Khashabi, D., Kim, G., Choi, Y., and Sap, M. Prosocialdialog: A prosocial backbone for conversational agents. In *EMNLP*, 2022.

Kocijan, V., Cretu, A.-M., Camburu, O.-M., Yordanov, Y., and Lukasiewicz, T. A surprisingly robust trick for the Winograd schema challenge. In Korhonen, A., Traum, D., and Màrquez, L. (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4837–4842, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1478. URL https://aclanthology.org/P19-1478.

Kotonya, N. and Toni, F. Explainable automated fact-checking for public health claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7740–7754, Online, November 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-main.623.

Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C. H., Gonzalez, J. E., Zhang, H., and Stoica, I. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

Levesque, H. J., Davis, E., and Morgenstern, L. The Winograd schema challenge. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*, volume 46, pp. 47, 2011.

Levy, M., Jacoby, A., and Goldberg, Y. Same task, more tokens: the impact of input length on the reasoning performance of large language models. *arXiv preprint arXiv:2402.14848*, 2024.

Li, T., Zhang, G., Do, Q. D., Yue, X., and Chen, W. Long-context llms struggle with long in-context learning. *arXiv preprint arXiv:2404.02060*, 2024.

Li, X. and Roth, D. Learning question classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics*, 2002. URL https://aclanthology.org/C02-1150.

Liu, H., Zaharia, M., and Abbeel, P. Ring attention with blockwise transformers for near-infinite context. *arXiv preprint arXiv:2310.01889*, 2023.

Liu, H., Yan, W., Zaharia, M., and Abbeel, P. World model on million-length video and language with ringattention. *arXiv preprint arXiv:2402.08268*, 2024a.

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., and Liang, P. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024b.

Liu, T., Zhang, Y., Brockett, C., Mao, Y., Sui, Z., Chen, W., and Dolan, B. A token-level reference-free hallucination detection benchmark for free-form text generation. *arXiv preprint arXiv:2104.08704*, 2021.

Louis, A., Roth, D., and Radlinski, F. 'I'd rather just go to bed': Understanding Indirect Answers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2020.

Luan, Y., He, L., Ostendorf, M., and Hajishirzi, H. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proc. Conf. Empirical Methods Natural Language Process. (EMNLP)*, 2018.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P11-1015.

Malo, P., Sinha, A., Korhonen, P., Wallenius, J., and Takala, P. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65, 2014.

Manotas, I., Vo, N. P. A., and Sheinin, V. LiMiT: The literal motion in text dataset. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 991–1000, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.88. URL https://www.aclweb.org/anthology/2020.findings-emnlp.88.

Marelli, M., Menini, S., Baroni, M., Bentivogli, L., Bernardi, R., and Zamparelli, R. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pp. 216–223, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/363_Paper.pdf.

McCreery, C. H., Katariya, N., Kannan, A., Chablani, M., and Amatriain, X. Effective transfer learning for identifying similar questions: Matching user questions to covid-19 faqs, 2020.

Meaney, J. A., Wilson, S., Chiruzzo, L., Lopez, A., and Magdy, W. SemEval 2021 task 7: HaHackathon, detecting and rating humor and offense. In Palmer, A., Schneider, N., Schluter, N., Emerson, G., Herbelot, A., and Zhu, X. (eds.), *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pp. 105–119, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.semeval-1.9. URL https://aclanthology.org/2021.semeval-1.9.

Miller, T., Hempelmann, C., and Gurevych, I. SemEval-2017 task 7: Detection and interpretation of English puns. In Bethard, S., Carpuat, M., Apidianaki, M., Mohammad, S. M., Cer, D., and Jurgens, D. (eds.), *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pp. 58–68, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/S17-2005. URL https://aclanthology.org/S17-2005.

Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X., and Cherry, C. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pp. 31–41, 2016.

Mollas, I., Chrysopoulou, Z., Karlos, S., and Tsoumakas, G. Ethos: an online hate speech detection dataset, 2020.

Nasr, M., Carlini, N., Hayase, J., Jagielski, M., Cooper, A. F., Ippolito, D., Choquette-Choo, C. A., Wallace, E., Tramèr, F., and Lee, K. Scalable extraction of training data from (production) language models. *arXiv preprint arXiv:2311.17035*, 2023.

O'Neill, J., Rozenshtein, P., Kiryo, R., Kubota, M., and Bollegala, D. I wish I would have loved this one, but I didn't – a multilingual dataset for counterfactual detection in product review. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t. (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7092–7108, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.568. URL https://aclanthology.org/2021.emnlp-main.568.

Oraby, S., Harrison, V., Reed, L., Hernandez, E., Riloff, E., and Walker, M. Creating and characterizing a diverse corpus of sarcasm in dialogue. In Fernandez, R., Minker, W., Carenini, G., Higashinaka, R., Artstein, R., and Gainer, A. (eds.), *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 31–41, Los Angeles, September 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-3604. URL https://aclanthology.org/W16-3604.

Pang, B. and Lee, L. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pp. 271–278, Barcelona, Spain, July 2004. doi: 10.3115/1218955.1218990. URL https://aclanthology.org/P04-1035.

Pang, B. and Lee, L. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the ACL*, 2005.

Park, J. and Cardie, C. Identifying appropriate support for propositions in online user comments. In Green, N., Ashley, K., Litman, D., Reed, C., and Walker, V. (eds.), *Proceedings of the First Workshop on Argumentation Mining*, pp. 29–38, Baltimore, Maryland, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-2105. URL https://aclanthology.org/W14-2105.

Patwa, P., Sharma, S., Pykl, S., Guptha, V., Kumari, G., Akhtar, M. S., Ekbal, A., Das, A., and Chakraborty, T. *Fighting an Infodemic: COVID-19 Fake News Dataset*, pp. 21–29. Springer International Publishing, 2021. ISBN 9783030736965. doi: 10.1007/978-3-030-73696-5_3. URL http://dx.doi.org/10.1007/978-3-030-73696-5_3.

Pilehvar, M. T. and Camacho-Collados, J. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. In Burstein, J., Doran, C., and Solorio, T. (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 1267–1273, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1128. URL https://aclanthology.org/N19-1128.

Pontiki, M., Galanis, D., Papageorgiou, H., Manandhar, S., and Androutsopoulos, I. SemEval-2015 task 12: Aspect based sentiment analysis. In Nakov, P., Zesch, T., Cer, D., and Jurgens, D. (eds.), *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pp. 486–495, Denver, Colorado, June 2015. Association for Computational Linguistics. doi: 10.18653/v1/S15-2082. URL https://aclanthology.org/S15-2082.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of EMNLP*, pp. 2383–2392. Association for Computational Linguistics, 2016.

Reid, M., Savinov, N., Teplyashin, D., Lepikhin, D., Lillicrap, T., Alayrac, J.-b., Soricut, R., Lazaridou, A., Firat, O., Schrittwieser, J., et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Roemmele, M., Bejan, C. A., and Gordon, A. S. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*, 2011.

Santus, E., Gladkova, A., Evert, S., and Lenci, A. The CogALex-V shared task on the corpus-based identification of semantic relations. In Zock, M., Lenci, A., and Evert, S. (eds.), *Proceedings of the 5th Workshop on Cognitive Aspects of the Lexicon (CogALex - V)*, pp. 69–79, Osaka, Japan, December 2016a. The COLING 2016 Organizing Committee. URL https://aclanthology.org/W16-5309.

Santus, E., Lenci, A., Chiu, T.-S., Lu, Q., and Huang, C.-R. Nine features in a random forest to learn taxonomical semantic relations. *arXiv preprint arXiv:1603.08702*, 2016b.

Saravia, E., Liu, H.-C. T., Huang, Y.-H., Wu, J., and Chen, Y.-S. CARER: Contextualized affect representations for emotion recognition. In Riloff, E., Chiang, D., Hockenmaier, J., and Tsujii, J. (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3687–3697, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1404. URL https://aclanthology.org/D18-1404.

Schuster, T., Fisch, A., and Barzilay, R. Get your vitamin C! robust fact verification with contrastive evidence. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 624–643, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.52. URL https://aclanthology.org/2021.naacl-main.52.

Shaham, U., Segal, E., Ivgi, M., Efrat, A., Yoran, O., Haviv, A., Gupta, A., Xiong, W., Geva, M., Berant, J., and Levy, O. SCROLLS: Standardized CompaRison over long language sequences. In Goldberg, Y., Kozareva, Z., and Zhang, Y. (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 12007–12021, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.823. URL https://aclanthology.org/2022.emnlp-main.823.

Shaham, U., Ivgi, M., Efrat, A., Berant, J., and Levy, O. ZeroSCROLLS: A zero-shot benchmark for long text understanding. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 7977–7989, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.536. URL https://aclanthology.org/2023.findings-emnlp.536.

Shankar Iyer, N. D. and Csernai, K. First quora dataset release: Question pairs. https://quoradata.quora.com/First-Quora-Dataset-Release-Question-Pairs, 2016.

Sheng, E. and Uthus, D. Investigating societal biases in a poetry composition system, 2020.

Shi, H., Xu, Z., Wang, H., Qin, W., Wang, W., Wang, Y., and Wang, H. Continual learning of large language models: A comprehensive survey. *arXiv preprint arXiv:2404.16789*, 2024.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., and Potts, C. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1631–1642, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D13-1170.

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022.

Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., and Liu, Y. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Thorne, J., Vlachos, A., Christodoulopoulos, C., and Mittal, A. FEVER: a large-scale dataset for fact extraction and VERification. In *NAACL-HLT*, 2018.

TogetherAI. Llama-2-7b-32k model card, 2024. URL https://huggingface.co/togethercomputer/LLaMA-2-7B-32K.

Tworkowski, S., Staniszewski, K., Pacek, M., Wu, Y., Michalewski, H., and Miłoś, P. Focused transformer: Contrastive training for context scaling, 2023.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In Linzen, T., Chrupała, G., and Alishahi, A. (eds.), *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 353–355, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5446. URL https://aclanthology.org/W18-5446.

Wang, W. Y. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In Barzilay, R. and Kan, M.-Y. (eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 422–426, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-2067. URL https://aclanthology.org/P17-2067.

Warstadt, A., Singh, A., and Bowman, S. R. Neural network acceptability judgments. *arXiv preprint arXiv:1805.12471*, 2018.

Webersinke, N., Kraus, M., Bingler, J. A., and Leippold, M. Climatebert: A pretrained language model for climate-related text, 2022.

Weston, J., Bordes, A., Chopra, S., Rush, A. M., Van Merriënboer, B., Joulin, A., and Mikolov, T. Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*, 2015.

Williams, A., Nangia, N., and Bowman, S. A broad-coverage challenge corpus for sentence understanding through inference. In Walker, M., Ji, H., and Stent, A. (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1112–1122, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1101. URL https://aclanthology.org/N18-1101.

Yang, Y., Yih, W.-t., and Meek, C. WikiQA: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 2013–2018, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1237. URL https://aclanthology.org/D15-1237.

Zhang, X., Zhao, J., and LeCun, Y. Character-level convolutional networks for text classification. In Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/250cf8b51c773f3f8dc8b4be867a9a02-Paper.pdf.

Zhang, X., Zhao, J., and LeCun, Y. Character-level convolutional networks for text classification, 2016.

Zhao, W., Ren, X., Hessel, J., Cardie, C., Choi, Y., and Deng, Y. Wildchat: 1m chatGPT interaction logs in the wild. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Bl8u7ZRlbM.

A. Experiment Details

A.1. Models

We list the details of open models evaluated in our experiments in Table 2.

Table 2. Open-weight models evaluated in this work.

Model	Max L	Reference	Huggingface Identifier
Mistral-7B	32k	Jiang et al. (2023)	`mistralai/Mistral-7B-Instruct-v0.2`
FILM-7B	32k	An et al. (2024)	`In2Training/FILM-7B`
Llama2-7B	32k	TogetherAI (2024)	`togethercomputer/LLaMA-2-7B-32K`
Llama2-7B	80k	Fu et al. (2024)	`yaofu/llama-2-7b-80k`
Llama3-8B	1048k	GradientAI (2024)	`gradientai/Llama-3-8B-Instruct-Gradient-1048k`
Yi-6B	200k	01.AI et al. (2024)	`01-ai/Yi-6B-200K`
Yi-9B	200k	01.AI et al. (2024)	`01-ai/Yi-9B-200K`
Yi-34B	200k	01.AI et al. (2024)	`01-ai/Yi-34B-200K`

A.2. Tasks

Table 3. **A Snippet of 16 tasks used in our experiments.** See Table 4 for the full list of 64 tasks. Tasks in this table are used for the Scale-Shot experiments in Table 1.

emo	covid_fake_news	logical_fallacy_detection	dbpedia_14
amazon_massive_scenario	news_data	semeval_absa_restaurant	amazon_counterfactual_en
brag_action	boolq	this_is_not_a_dataset	insincere_questions
clickbait	yahoo_answers_topics	pun_detection	wiki_qa

We utilize publicly available datasets in our evaluations, with access details for both reference and huggingface identifier provided in Table 4. For further usage, readers should refer to the licenses of the original datasets.

A.3. Implementation and Engineering Details

Data Preprocessing. For each task, we manually crafted two semantically identical instructions. We then randomly sampled five subsets from the original training dataset for in-context learning, ensuring each subset containing at least 16 instances, and 100 instances from the original test set to form our test set. To facilitate rank classification, we ensured that the options for all tasks have distinct start tokens. For in-context learning, we sample one instance per label for each shot.

LLM Inference. We apply rank classification in all our experiments: given a set of task options, we evaluate the first token of the output and take the option with the highest log probability as model's response. We use the vLLM (Kwon et al., 2023) framework for open models to enhance inference speed. Running a 16-task, 8-shot experiment with a 7B model on two A6000 GPUs takes around 18 hours. The specific OpenAI APIs employed are gpt-4o-2024-05-13 and gpt-3.5-turbo-0125.

B. Controlled Analysis

Our results in §4 suggest that long-context LMs struggle in the Task Haystack evaluation. In the following section, we try to investigate the reasons attributing to their failures with various controlled settings. We hypothesize that the model failure at Lifelong ICL may be associated with the following factors: (a) Long-context inputs: the model may tend to break because the input text is long; (b) Distraction: the model may be confused by irrelevant context; (c) Recency: the model mainly relys on recent context and performs worse when the relevant context is distant. Based on these hypotheses, we design controlled setting and summarize them in Table 5. We then conduct controlled experiments in the 16-task 4-shot setting with Mistral-7B (32k) and FILM-7B (32k). Results are presented in Fig. 4.

Recency. We investigate the effect of recency by comparing the results of Recall and Replay. By replaying in-context learning demonstrations before testing, model performances improve (+1.6% for Mistral-7B, and +2.9% for FILM-7B). This

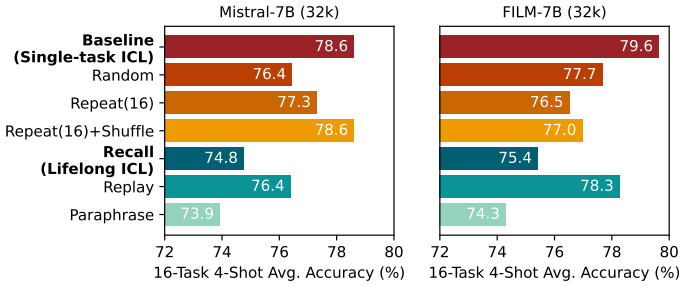

Figure 4. **Controlled Experiments.** Results suggest that long-context LMs are subject to various robustness problems. See §B for discussion.

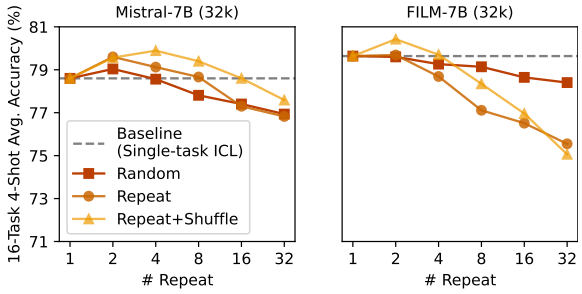

Figure 5. **Single-task "Multi-epoch" ICL.** Model performance improves then degrades after repeating ICL examples.

can be also considered as an oracle setting that approximates potential mitigating strategies such as prompting the model to recall relevant information or examples. However, the improvements only close about half the gap between Baseline and Recall, suggesting that recency does not fully explain the performance disparity.

Distraction. We examine the effect of irrelevant context, by contrasting Baseline with Random. The results indicate that prepending an irrelevant long text will influence the performance negatively, which corroborates with recent work investigating the robustness of large language models (Levy et al., 2024). Further, Replay can be seen as prepending a long prefix of mostly irrelevant tasks, and thus the gap between Replay and Baseline may also stem from distractions caused by extraneous context.

Long-Context Input. We further compare Baseline with Random and Repeat setups, where Random introduces irrelevant context and Repeat includes only relevant context. Surprisingly, performance drops in both cases, even in the Repeat setting where distractions or recency issues are absent. While we cannot definitively conclude that language models are inherently less capable when presented with lengthy contexts, we suggest that longer contexts – whether repetitive or irrelevant – give rises to undesirable failure modes. This suggest that LM users should be cautious about what to put in the context, and highlights the value of external filtering mechanisms like retrieval augmentation given the limitations of current long-context LMs.

Robustness to task instructions. Our Paraphrase setting allows us to explore whether models rely primarily on pattern matching of instructions when performing Lifelong ICL. We observe a decline in performance in the Paraphrase setting compared to Recall, suggesting that the model is indeed locating identical instructions and leveraging relevant task examples at test time. However, this in-robustness to varying instruction expressions indicates that models rely on pattern matching rather than true understanding, which might limit their broader utility in practical applications.

Repeated ICL as "Multi-epoch" ICL. We conduct further investigation with the Random, Repeat, Repeat+Shuffle setting, by varying the number of repetitions. Results are reported in Fig. 5 and Fig. 8. Interestingly, we see model performance increases and then dips when running in-context learning for multiple "epochs". One direct takeaway is that repeating the ICL examples multiple times can potentially improve performance, which may have practical utilities in certain low-data high-inference-budget regimes. However, model performance start to degrade after repeating more than 8 times. This phenomenon can be interpreted in two ways: (1) it is a known issue that repetition may lead to model degeneration (Nasr et al., 2023); Repeat+Shuffle can possibly alleviate this issue by introducing slight variations in each repeat, which explains that in general Repeat+Shuffle outperforms Repeat. (2) it is also possible that the model "overfits" to the few-shot training data after multiple "epochs", analogous to the common observations in gradient-based fine-tuning. Future work could further investigate the working mechanism of ICL in this multi-epoch setting.

B.1. Additional Observations and Analysis

Tasked learned via ICL are more easily forgotten. While examining Task Haystack results, we found that the pass rates are highly task-specific. For example, in Fig. 19, news_data and insincere_questions are forgotten in all permutations, whereas more popular tasks like boolq and yahoo_answer_topics pass all tests. We hypothesize that the model may have memorized some of the tasks during pre-training or post-training, making these tasks less subjective to performance drop

Figure 6. **Definition of Pass Rate in Task Haystack.** The model "passes" when the performance of Lifelong ICL is not significantly worse than the Single-task ICL baseline.

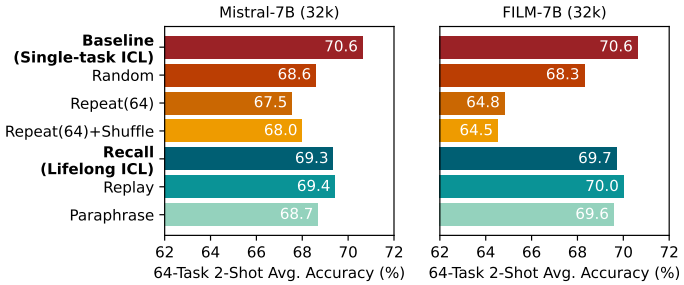

Figure 7. **Controlled Experiments.** We repeat the experiments in Fig. 4 with N-task=64 and N-shot=2. The trends are consistent with Fig. 4. However, the gaps are smaller due to a smaller value of N-shot.

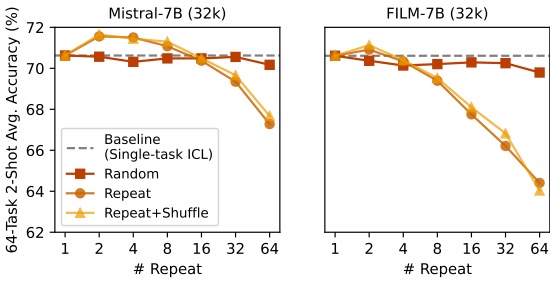

Figure 8. **"Multi-epoch" ICL.** We repeat the experiments in Fig. 5 with N-task=64 and N-shot=2. The increase-then-decrease phenomenon is more evident in this scenario.

in Lifelong ICL. Alternatively, a task may be too challenging for the model to learn through ICL, and thus it passes the Task Haystack Eval through the previously mentioned short-cut. To account for these situations, we split all tasks into 2 groups. Tasks of which 4-shot performance is significantly better than 1-shot performance as classified as ICL-effective tasks, and the remaining tasks are considered to be ICL-ineffective tasks. We report the pass rates for each model on these two groups in Table 6. For 8 out of 10 models, pass rates on ICL-effective tasks are lower than pass rate on ICL-ineffective tasks. We encourage users of Task Haystack to responsibly report data contamination analysis and report pass rates on these two groups when developing new long-context models.

C. Additional Results

Figure 9 to Figure 16 present detailed results for eight open models. On the left side of each figure, we illustrate the standard needle-in-a-haystack performance, measured by the recall score of the model's response, where nearly all models demonstrate near-perfect results. In the middle and right sections, the figures show the results from experiments involving the two scaling setting described in Controlling the Context Length.. Each block of depth d_i represents the mean performance across the tasks within the specified depth range. Our experiments conducted on GPT-4o and GPT-3.5 Turbo incurred a total cost of approximately $8,000.

D. Task Selection Considerations

Our key considerations when selecting tasks in Task Haystack include:

- We focus on classification tasks, as they allow standardized evaluation. Additionally various past work investigates ICL empirically or mechanistically on classification tasks.
- We select classification tasks with fewer than 20 categories and input text shorter than 1000 words, to avoid excessively long ICL prefixes that dominates the context window.
- We focus on English tasks, as some of the evaluated models are not trained for multilingual usage.

E. Discussion

Intended Use. We anticipate Lifelong ICL and Task Haystack to be used for evaluating and diagnosing newly released long-context LMs. However, as our findings in Sections 4 and B.1 suggest, the ICL accuracy and pass rate might be affected if the model has already been exposed to the tasks used in our evaluation. To ensure responsible use, we encourage users to investigate and report any potential data contamination. Additionally, targeted data engineering for Lifelong ICL and Task Haystack is feasible. For fair comparisons, we recommend that users of the proposed evaluation method disclose whether their training data contains sequences similar to the Lifelong ICL setting.

Limitations. (1) Our current evaluation primarily focuses on English-only text classification tasks, potentially limiting a comprehensive assessment of model capabilities across various linguistic challenges. We view this work as a valuable initial exploration in the Lifelong ICL setting. We encourage future research to build upon our foundation and incorporate a wider variety of tasks, including variations in task format, modality (*e.g.*, vision, speech), and language diversity. (2) Additionally, this work simplifies the lifelong learning stream by assuming a sequential order and consistent number of examples per class for each task. Real-world scenarios likely involve a different distribution of training examples, without clear task boundaries or labels within the learning stream. Future work should explore more realistic training scenarios that reflect this complexity. (3) Finally, due to computational constraints, our evaluation utilizes only 5 random permutations of tasks. Experimenting with a larger number of permutations could potentially reduce the randomness inherent in the results and provide more robust findings.

Ethics Statement. This research leverages openly available datasets that were carefully reviewed to mitigate potential data privacy and security concerns. We emphasize that these datasets are used solely for evaluation purposes and do not directly influence model weights. Thus the risk of amplifying biases present in the data is avoided.

F. Related Works

Long-Context LM Benchmarks. Existing benchmarks for evaluating long-context models can be categorized into realistic and synthetic ones. Realistic benchmarks, exemplified by SCROLLS (Shaham et al., 2022) and ZeroSCROLLS (Shaham et al., 2023), comprise tasks that require processing long inputs with long-range dependencies. These tasks are typically sourced from established datasets and include various task types such as summarization and question answering. In the category of synthetic benchmarks, the needle-in-a-haystack (NIAH) (Kamradt, 2023) evaluation is widely adopted for evaluating context utilization (Reid et al., 2024; Anthropic, 2024; Liu et al., 2024a; Fu et al., 2024). Ruler (Hsieh et al., 2024) expands on the NIAH test with multi-key and multi-value retrieval, and adds two new tasks that involve multi-hop tracing and aggregation. Hybrid benchmarks, is a third category that incorporate both realistic and synthetic elements. An example is LongBench (Bai et al., 2023), which includes synthetic tasks based on realistic text, such as counting unique passages appearing in the context. Our proposed Task Haystack can be considered as a hybrid benchmark.

Evaluating Long-Context LMs with Many-Shot ICL. Several recent works have explored in-context learning with long-context LMs by scaling the number of training examples (*i.e.*, shots). Bertsch et al. (2024) conducted a systematic study of long-context ICL with up to 2000 shots, demonstrating many-shot ICL as a competitive alternative to retrieval-based ICL and fine-tuning. Additionally, it offers the advantage of caching demonstrations at inference time, unlike instance-level retrieval methods. While Bertsch et al. (2024) focus on classification tasks, Agarwal et al. (2024) showed the effectiveness of many-shot ICL on generative and reasoning tasks, and established new state-of-the-art results on practical applications such as low-resource translation with the Gemini 1.5 Pro model. However, there are still limitations to many-shot ICL. Li et al. (2024) introduce LongICLBench, a suite of 6 classification tasks with many (20+) classes, and find that current long-context LMs still struggle with these tasks. Orthogonal to this line of work on scaling *number of examples* for one single task, we focus on scaling the *number of tasks* in our Lifelong ICL setting.

Lifelong Learning in NLP. Continual learning, or lifelong learning, focuses on developing machine learning algorithms that learn continuously and adaptively from data streams. Unlike traditional gradient-based fine-tuning, which modifies model weights, Lifelong ICL uses in-context learning as the underlying "learning" algorithm. A primary challenge in lifelong learning is catastrophic forgetting, the tendency of a model to forget previously acquired knowledge upon learning new information. Our proposed Task Haystack evaluation setting resembles the catastrophic forgetting phenomenon. The model may struggle to recall tasks learned earlier in a lengthy learning sequence, leading to a performance decline. We refer

readers to (Shi et al., 2024) and (Biesialska et al., 2020) for comprehensive surveys on lifelong learning in NLP.

Table 4. Text classification tasks included in Task Haystack Eval.

Name	Reference	Huggingface Identifier	License
acl-arc	Bird et al. (2008)	hrithikpiyush/acl-arc	Apache 2.0
ag-news	Zhang et al. (2016)	fancyzhx/ag_news	Unspecified
amazon-counterfactual-en	O'Neill et al. (2021)	SetFit/amazon_counterfactual_en	CC BY-NC 4.0
amazon-massive-scenario	FitzGerald et al. (2022)	SetFit/amazon_massive_scenario_en-US	Apache 2.0
app-reviews	Grano et al. (2017)	sealuzh/app_reviews	Unspecified
babi-nli	Weston et al. (2015)	tasksource/babi_nli	BSD
beaver-tails	Ji et al. (2023)	PKU-Alignment/BeaverTails	CC BY-NC 4.0
boolq	Clark et al. (2019)	google/boolq	CC BY-SA 3.0
brag-action	Choi et al. (2023)	Blablablab/SOCKET	CC BY 4.0
cb	De Marneffe et al. (2019)	aps/super_glue	Unspecified
circa	Louis et al. (2020)	google-research-datasets/circa	CC BY 4.0
clickbait	Chakraborty et al. (2016)	marksverdhei/clickbait_title_classification	MIT
climate-commitments-actions	Bingler et al. (2023)	climatebert/climate_commitments_actions	CC-BY-NC-SA 4.0
climate-fever	Diggelmann et al. (2020)	tdiggelm/climate_fever	Unspecified
climate-sentiment	Bingler et al. (2023)	climatebert/climate_sentiment	CC BY-NC-SA 4.0
cola	Warstadt et al. (2018)	nyu-mll/glue	Other
copa	Roemmele et al. (2011)	aps/super_glue	BSD 2-Clause
covid-fake-news	Patwa et al. (2021)	nanyy1025/covid_fake_news	Unspecified
dbpedia14	Zhang et al. (2015)	fancyzhx/dbpedia_14	CC BY-SA 3.0
disaster-repsonse-message		community-datasets/disaster_response_messages	Unspecified
emo	Chatterjee et al. (2019)	SemEvalWorkshop/emo	Unspecified
emotion	Saravia et al. (2018)	dair-ai/emotion	Unspecified
environmental-claims	Webersinke et al. (2022)	climatebert/environmental_claims	CC BY-NC-SA 4.0
ethos	Mollas et al. (2020)	iamollas/ethos	AGPL 3.0
fever	Thorne et al. (2018)	fever/fever	CC BY-SA 3.0, GPL 3.0
financial-phrasebank	Malo et al. (2014)	takala/financial_phrasebank	CC BY-NC-SA 3.0
function-of-decision-section	Guha et al. (2024)	nguha/legalbench	CC BY 4.0
hate-speech18	De Gibert et al. (2018)	odegiber/hate_speech18	CC BY-SA 3.0
health-fact	Kotonya & Toni (2020)	ImperialCollegeLondon/health_fact	MIT
i2d2	Bhagavatula et al. (2022)	tasksource/I2D2	Apache 2.0
imdb	Maas et al. (2011)	stanfordnlp/imdb	Unspecified
insincere-questions	Alex Ellis (2018)	SetFit/insincere-questions	Unspecified
is-humor	Meaney et al. (2021)	Blablablab/SOCKET	CC BY 4.0
jailbreak-classification		jackhhao/jailbreak-classification	Apache 2.0
lexical-rc-cogalexv	Santus et al. (2016a)	relbert/lexical_relation_classification	Unspecified
lexical-rc-root09	Santus et al. (2016b)	relbert/lexical_relation_classification	Unspecified
liar	Wang (2017)	ucsbnlp/liar	Unspecified
limit	Manotas et al. (2020)	IBM/limit	CC BY-SA 4.0
logical-fallacy-detection	Srivastava et al. (2022)	tasksource/bigbench	Apache 2.0
medical-question-pairs	McCreery et al. (2020)	curaihealth/medical_questions_pairs	Unspecified
metaphor-boolean	Bizzoni & Lappin (2018)	tasksource/bigbench	Apache 2.0
mnli	Williams et al. (2018)	nyu-mll/multi_nli	CC BY 3.0, CC BY-SA 3.0, MIT, Other
mrpc	Dolan & Brockett (2005)	nyu-mll/glue	Unspecified
news-data		okite97/news-data	AFL 3.0
poem-sentiment	Sheng & Uthus (2020)	google-research-datasets/poem_sentiment	CC BY 4.0
pragmeval-emergent	Ferreira & Vlachos (2016)	sileod/pragmeval	Unspecified
pragmeval-sarcasm	Oraby et al. (2016)	sileod/pragmeval	Unspecified
pragmeval-verifiability	Park & Cardie (2014)	sileod/pragmeval	Unspecified
prosocial-dialog	Kim et al. (2022)	allenai/prosocial-dialog	CC BY 4.0
pun-detection	Miller et al. (2017)	frostymelonade/SemEval2017-task7-pun-detection	CC BY NC
qnli	Rajpurkar et al. (2016)	nyu-mll/glue	CC BY-SA 4.0
qqp	Shankar Iyer & Csernai (2016)	nyu-mll/glue	Others
rct20k	Dernoncourt & Lee (2017)	armanc/pubmed-rct20k	Unspecified
rotten-tomatoes	Pang & Lee (2005)	cornell-movie-review-data/rotten_tomatoes	Unspecified
rte	Wang et al. (2018)	nyu-mll/glue	Unspecified
sara-entailment	Holzenberger et al. (2020)	nguha/legalbench	MIT
scierc	Luan et al. (2018)	hrithikpiyush/scierc	Unspecified
semeval-absa-laptop	Pontiki et al. (2015)	jakartaresearch/semeval-absa	CC BY 4.0
semeval-absa-restaurant	Pontiki et al. (2015)	jakartaresearch/semeval-absa	CC BY 4.0
senteval-cr	Hu & Liu (2004)	SetFit/SentEval-CR	BSD
senteval-subj	Pang & Lee (2004)	SetFit/subj	BSD
sick	Marelli et al. (2014)	RobZamp/sick	CC BY-NC-SA 3.0
silicon-dyda-da	Chapuis et al. (2020)	eusip/silicone	CC BY-SA 4.0
sms-spam	Almeida et al. (2011)	ucirvine/sms_spam	Unspecified
sst2	Socher et al. (2013)	stanfordnlp/sst2	Unspecified
sst5	Socher et al. (2013)	SetFit/sst5	Unspecified
stance-abortion	Mohammad et al. (2016)	cardiffnlp/tweet_eval	Unspecified
stance-feminist	Mohammad et al. (2016)	cardiffnlp/tweet_eval	Unspecified
student-question-categories	Biswal (2020)	SetFit/student-question-categories	CC0
tcfd-recommendations	Bingler et al. (2023)	climatebert/tcfd_recommendations	CC BY-NC-SA 4.0
this-is-not-a-dataset	García-Ferrero et al. (2023)	HiTZ/This-is-not-a-dataset	Apache 2.0
toxic-conversations	cjadams (2019)	SetFit/toxic_conversations	CC0
trec	Li & Roth (2002)	CogComp/trec	Unspecified
vitaminc	Schuster et al. (2021)	tals/vitaminc	CC BY-SA 3.0
wic	Pilehvar & Camacho-Collados (2019)	aps/super_glue	CC BY-NC 4.0
wiki-hades	Liu et al. (2021)	tasksource/wiki-hades	MIT
wiki-qa	Yang et al. (2015)	microsoft/wiki_qa	Other
wnli	Levesque et al. (2011)	nyu-mll/glue	Unspecified
wsc	Kocijan et al. (2019)	aps/super_glue	CC BY 4.0
yahoo-answers-topics	Zhang et al. (2015)	community-datasets/yahoo$_{a}nswers_topics$	Unspecified

Table 5. **Summary of Controlled Settings.** "T1 Train" contains the task instruction and few demonstrations of Task 1. "T1 Test" contains the same task instruction and the test input. 🔀 = shuffling the few-shot examples; 🔄 = using a paraphrased instruction d' at test time.

Setting	Input Prompt Example	Controlled Factors		
		Long Ctx.	Distraction	Recency
Baseline	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	T1 Train T2 Train T3 Train **T1 Test**	✓	✓	✗
Replay	T1 Train T2 Train T3 Train T1 Train **T1 Test**	✓	✓	✓
Paraphrase	T1 Train T2 Train T3 Train 🔄 **T1 Test**	✓	✓	✗

Table 6. **Pass Rates on ICL-effective Tasks and ICL-ineffective Tasks.** Results are computed in the 16-task 4-shot Setting. We define ICL-effective tasks as tasks whose 4-shot performance is significantly better than its 1-shot performance. In general, ICL-effective tasks achieve lower pass rates, suggesting they are more likely to suffer from performance drop.

Model	ICL-eff.		ICL-ineff.		All	Model	ICL-eff.		ICL-ineff.		All
	N	pass	N	pass	pass		N	pass	N	pass	pass
Mistral-7B (32k)	5	36.0	11	81.8	67.5	Yi-6B (200k)	6	46.6	10	42.0	43.8
FILM-7B (32k)	2	40.0	14	77.1	72.5	Yi-9B (200k)	6	50.0	10	72.0	63.7
Llama2-7B (32k)	6	33.3	10	46.0	41.2	Yi-34B (200k)	3	46.7	13	67.7	63.8
Llama2-7B (80k)	3	80.0	13	100.0	96.3	GPT-3.5-Turbo (16k)	5	48.0	11	70.9	63.8
Llama3-8B (1048k)	6	40.0	10	90.0	71.3	GPT-4o (128k)	6	96.7	10	84.0	88.8

Table 7. **Main Results: Fixing 2 Shots, Scaling the Number of Tasks**. See the caption of Table 1 for the explanations of the table headers.

Model	8 tasks (4k)			16 tasks (8k)			32 tasks (15k)			64 tasks (25k)		
	s-acc	l-acc	pass	s-acc	l-acc	pass	s-acc	l-acc	pass	s-acc	l-acc	pass
Mistral-7B (32k)	76.4	78.9	80.0	77.6	74.6	73.8	72.7	71.1	72.5	70.6	69.3	75.6
FILM-7B (32k)	79.1	77.1	87.5	79.1	75.1	77.5	73.3	72.0	88.1	70.6	69.7	75.3
Llama2-7B (32k)	70.1	60.7	65.0	72.8	64.5	53.8	70.6	64.5	59.4	67.1	61.2	63.1
Llama2-7B (80k)	49.9	58.5	97.5	49.8	60.2	100.0	49.5	58.3	91.2	48.6	52.0	89.7
Llama3-8B (1048k)	68.3	65.4	75.0	70.0	69.1	76.2	67.4	65.1	75.6	66.4	65.7	81.2
Yi-6B (200k)	72.0	54.4	50.0	73.0	58.6	51.2	68.4	59.2	63.7	63.7	55.7	65.6
Yi-9B (200k)	78.6	73.4	62.5	77.7	72.9	71.2	75.5	70.3	61.3	70.2	66.8	61.3
Yi-34B (200k)	66.1	70.7	87.5	74.1	72.4	60.0	74.0	69.7	63.1	71.5	68.2	59.4

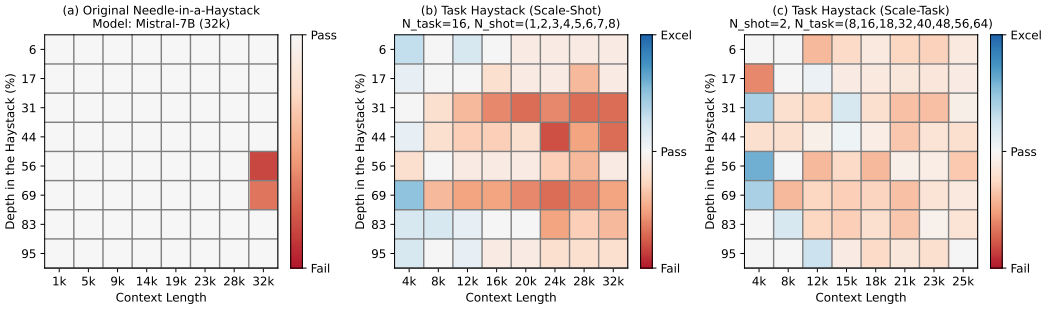

Figure 9. Task Haystack Results on Mistral-7B (32k).

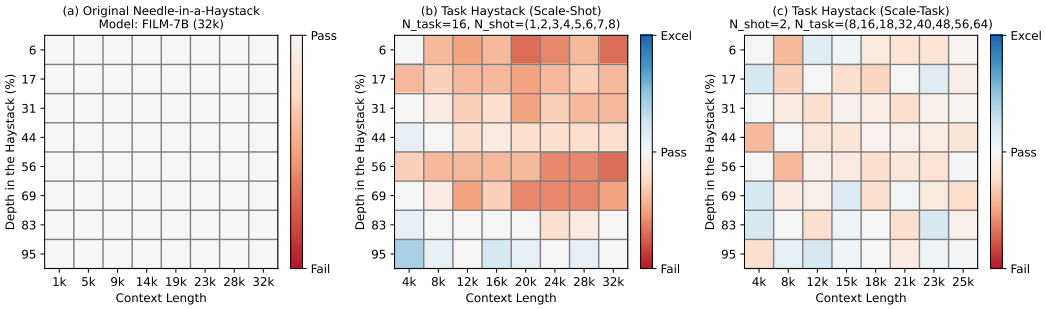

Figure 10. Task Haystack Results on FILM-7B (32k).

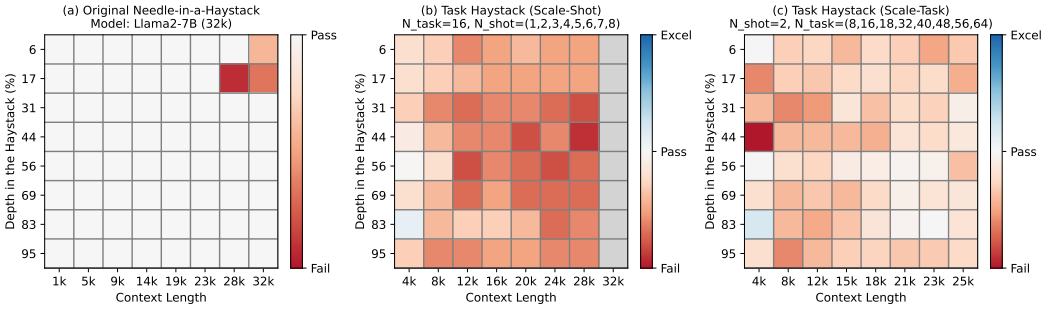

Figure 11. Task Haystack Results on Llama2-7B (32k).

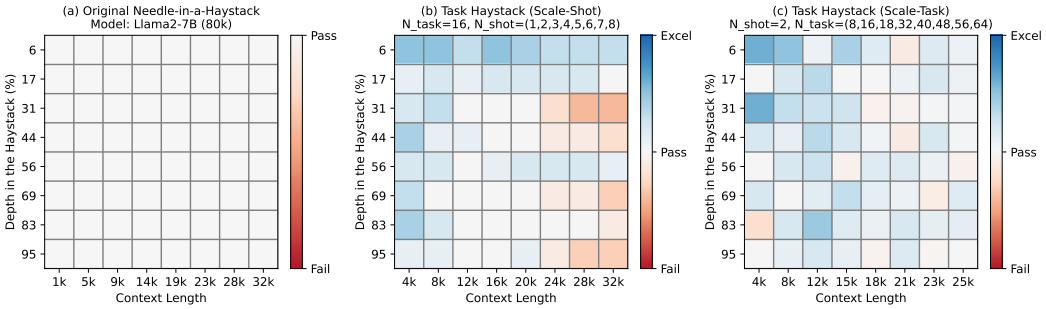

Figure 12. Task Haystack Results on Llama2-7B (80k).

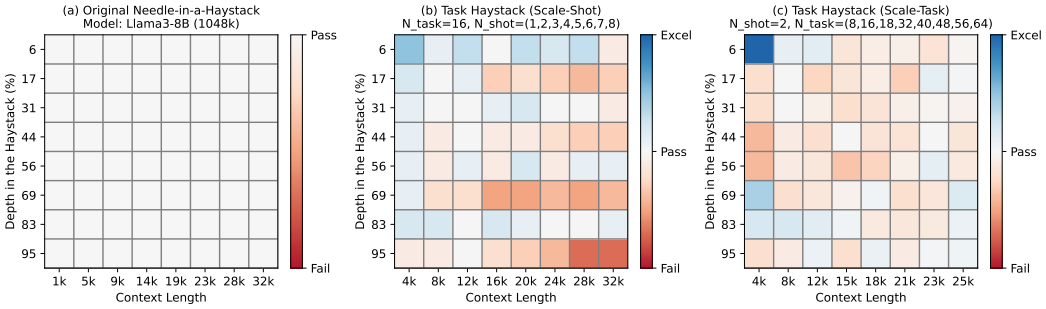

Figure 13. Task Haystack Results on Llama3-8B (1048k).

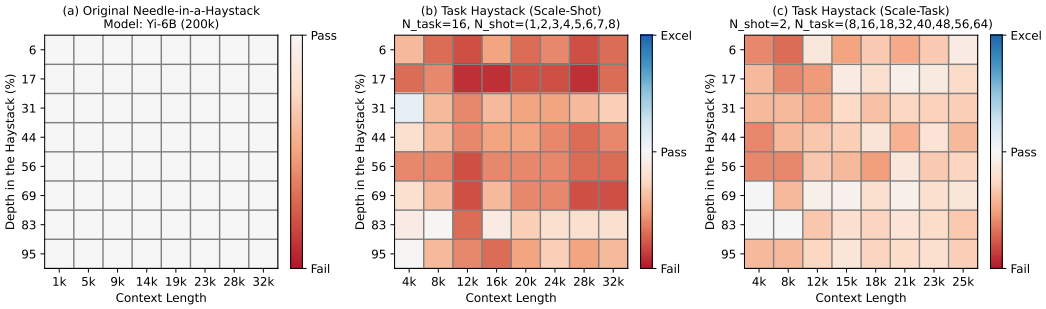

Figure 14. Task Haystack Results on Yi-6B (200k).

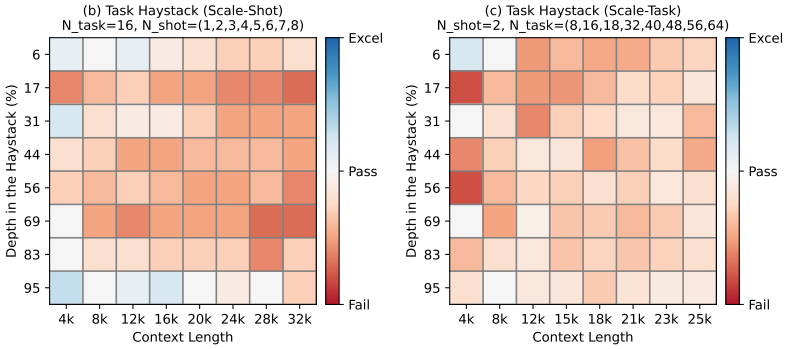

Figure 15. Task Haystack Results on Yi-9B (200k).

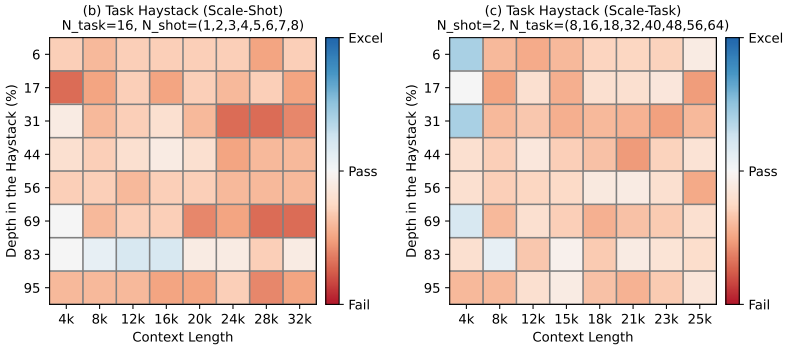

Figure 16. Task Haystack Results on Yi-34B (200k).

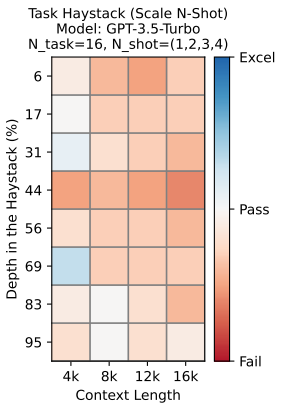

Figure 17. Task Haystack Results on GPT-3.5-Turbo (16k). Due to budget limits we only experiment with the Scale-Shot setting.

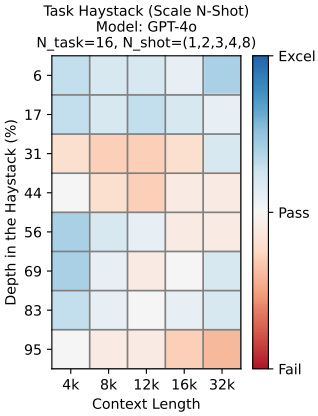

Figure 18. Task Haystack Results on GPT-4o (128k). Due to budget limits we only experiment with the Scale-Shot setting and skipped N-shot=5,6,7.

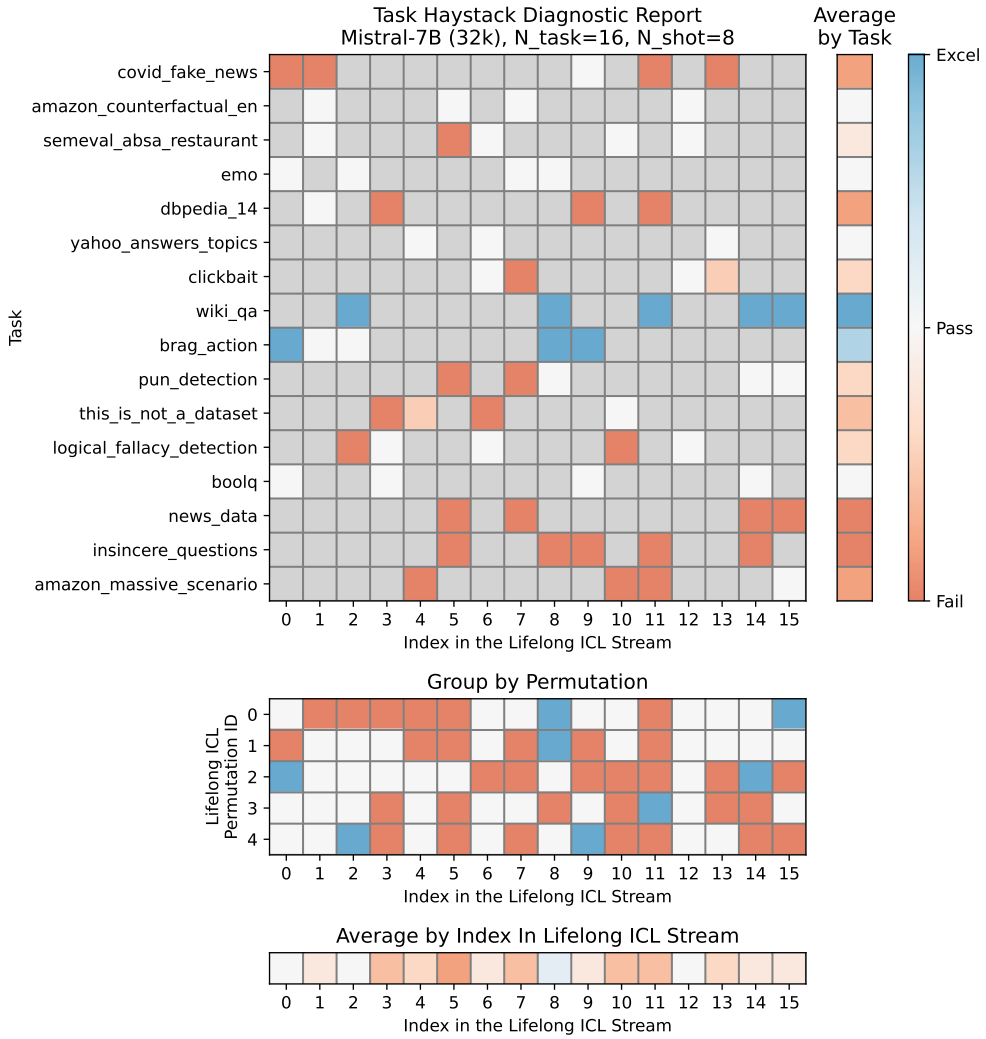

Figure 19. Example Diagnostic Report on Mistral-7B, N-task=16, N-shot=8. Grey cells indicate that the task does not appear at a given index in the 5 sampled permutations.