MICHIGAN ENGINEERING
COMPUTER SCIENCE AND ENGINEERING
UNIVERSITY OF MICHIGAN

ACL 2025
VIENNA

On Many-Shot In-Context Learning for Long-Context Evaluation

Kaijian Zou, Muhammad Khalifa, Lu Wang
{zkjzou, khalifam, wangluxy}@umich.edu
University of Michigan, Ann Arbor

Scan Here

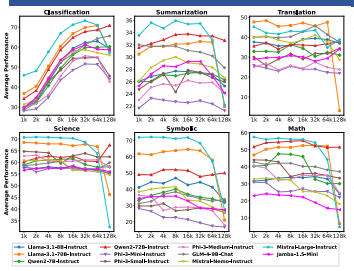
Background and Motivation

- Existing long-context evaluation benchmarks, like NIAH, only focus testing models' ability to retrieve from the long context. Evaluating models' global understanding of the full context remains lacking.
- Only many-shot classification ICL tasks have been utilized in LongICLBench for long-context evaluation. Other types of many-shot ICL tasks are still underexplored on LCLMs.

Research Questions

- **RQ1**: What types of ICL tasks benefit from additional demonstrations, and are these tasks effective at evaluating LCLMs?
- **RQ2**: To what extent does each task require learning from a limited number of samples versus learning from more samples with broader context from LCLMs?

Not all ICL tasks benefit from additional demonstrations



- Classification performance steadily improves with more shots.
- Summarization show gradual performance gains.
- Inconsistent trends in science and symbolic reasoning tasks.
- Math reasoning tasks benefit from additional demonstrations, particularly for stronger models.

Sample Learning Ratio

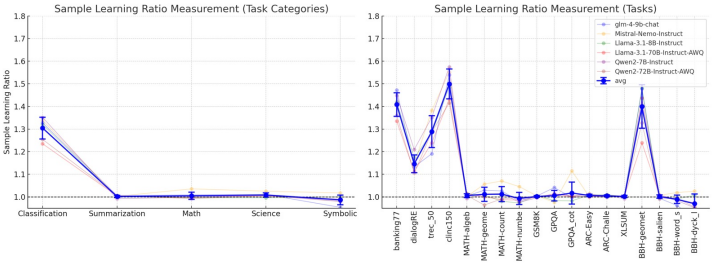
Sample Learning Ratio (SLR) is a metric to access whether tasks predominantly rely on models to retrieve relevant examples during many-shot ICL.

$$\text{SLR} = \frac{1}{7}\sum_{l=1k}^{64k} \frac{\text{Perf}_{\text{least}}^{(l)}}{\text{Perf}_{\text{most}}^{(l)}}$$

$\text{Perf}_{\text{least}}$ is the model's performance after removing 10% least similar examples. **$\text{Perf}_{\text{most}}$** is the model's performance after removing 10% most similar examples.

High SLR: tasks that rely on retrieving specific examples

Low SLR: tasks requiring broad context understanding



SSL vs ASL tasks

Similar-sample learning (SSL) tasks: they require retrieval of similar examples. All classification tasks exhibit high SLR.

All-sample learning (ASL) tasks: they necessitate a great degree of global context understanding and have SLR that is close to 1.

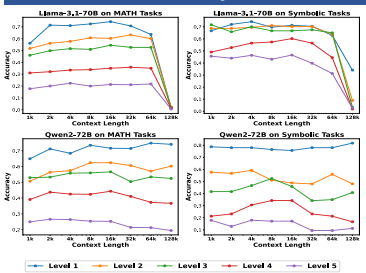
ManyICLBench contains 5 SSL tasks and 11 ASL tasks.

ManyICLBench Performance Table

SSL Tasks	1k	2k	4k	8k	16k	32k	64k	128k	AVG.	AVG.L.
GLM-4-9b-Chat	31.63	34.99	46.37	57.27	63.61	68.34	72.16	72.93	55.91	71.14
Mistral-Nemo-Instruct	33.44	35.45	48.17	57.95	65.38	65.49	63.61	61.73	53.90	63.61
Mistral-Large-Instruct-AWQ	49.15	51.23	60.78	71.95	77.10	79.45	77.77	61.89	**66.16**	73.04
Llama-3.1-8B-Instruct-AWQ	32.13	34.63	45.76	57.39	66.18	70.02	70.55	65.85	55.31	68.81
Llama-3.1-70B-Instruct-AWQ	38.75	42.87	53.98	66.07	73.12	76.56	78.48	65.56	61.92	73.53
Qwen2-7B-Instruct-AWQ	30.18	34.03	44.40	54.85	62.92	65.91	66.94	66.38	53.20	66.41
Qwen2-72B-Instruct-AWQ	36.41	41.89	54.24	65.33	73.39	76.53	77.51	77.47	62.85	**77.17**
Phi-3-Mini-Instruct	30.27	30.90	38.09	48.14	53.58	57.29	56.83	48.72	45.48	54.28
Phi-3-Medium-Instruct	31.73	33.55	39.10	49.83	58.29	61.17	60.63	45.32	47.45	55.70
Phi-3-Small-Instruct	31.48	36.27	46.20	54.34	59.63	59.73	60.20	48.97	49.60	56.30
Jamba-1.5-Mini	32.10	36.91	48.61	60.29	66.05	68.33	66.02	65.17	55.44	66.51
Gemini-1.5-Pro	36.40	47.31	58.01	65.49	71.43	74.22	72.43	72.42	62.21	73.03
ASL Tasks	**1k**	**2k**	**4k**	**8k**	**16k**	**32k**	**64k**	**128k**	**AVG.**	**AVG.L.**
GLM-4-9b-Chat	40.51	40.28	42.04	42.78	40.70	40.46	38.85	39.13	40.59	39.48
Mistral-Nemo-Instruct	38.25	39.07	39.28	38.99	33.06	32.83	30.46	27.11	34.88	30.13
Mistral-Large-Instruct-AWQ	61.47	61.10	61.23	60.87	60.86	58.84	50.01	16.69	53.88	41.85
Llama-3.1-8B-Instruct	37.31	38.84	41.25	40.79	39.83	39.77	39.12	34.41	38.92	37.77
Llama-3.1-70B-Instruct-AWQ	53.32	54.84	55.76	55.87	56.42	56.34	54.42	18.58	50.69	43.12
Qwen2-7B-Instruct	39.52	41.96	45.17	45.39	45.50	37.29	36.97	33.99	40.72	36.09
Qwen2-72B-Instruct-AWQ	48.01	49.24	50.32	50.70	50.97	48.20	47.98	48.16	49.20	48.11
Phi-3-Mini-Instruct	33.54	32.97	29.80	29.75	30.12	28.78	28.06	25.76	29.85	27.53
Phi-3-Medium-Instruct	41.59	40.91	34.85	35.63	36.91	36.84	36.38	28.31	36.43	33.84
Phi-3-Small-Instruct	41.61	41.61	41.61	35.58	37.17	37.73	36.91	35.33	38.44	36.65
Jamba-1.5-Mini	31.96	33.08	32.97	32.70	31.66	28.82	27.14	25.87	30.53	27.28
Gemini-1.5-Pro	57.87	63.39	64.15	66.78	68.02	67.78	66.14	66.42	**65.07**	**66.78**

- **Most models struggle at retrieving examples after 32k length:** Most models improve performance up to 16k tokens but begin to decline after 32k, with only select models like GLM-4 consistently maintaining strong retrieval performance at very long contexts.
- **Challenges in ASL tasks:** ASL tasks pose significant challenges, as most models struggle to effectively leverage global context even at shorter lengths, though models like Gemini and Qwen2-72B exhibit relatively stable performance.
- **The paradox of model size:** Larger models can suffer greater performance degradation at long contexts without adequate long-context training, highlighting the necessity for targeted training rather than relying solely on scale.
- **Llama 3.1 performance and training limitations:** Llama 3.1 models initially benefit from additional demonstrations up to 64k but experience significant drops at 128k due to insufficient long-context training exposure.
- **Gemini exhibits robustness:** Gemini-1.5-Pro maintains strong retrieval and global-context understanding up to 128k tokens, significantly outperforming other open-weight models in ASL tasks.

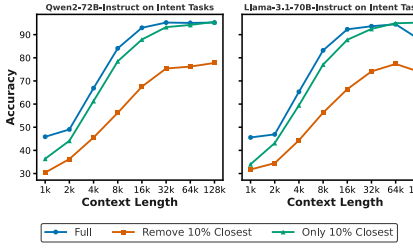
Why do LCLMs fail on ASL tasks?



LCLMs' reasoning abilities **degrade** with increased context length.

At 128k tokens, Llama-3.1-70B notably underperforms Qwen-2-72B due to **weaker reasoning** and **instruction-following capabilities**.

Why do LCLMs perform better on SSL tasks?



On SSL tasks, using only the closest 10% examples achieves **nearly** full-set performance.

SSL tasks primarily rely on effective **retrieval of highly similar examples**, whereas ASL tasks offer **no** retrieval shortcut