On Many-Shot In-Context Learning for Long-Context Evaluation (ACL 2025)

Kaijian Zou, Muhammad Khalifa, and Lu Wang

University of Michigan, Ann Arbor

Agenda

1. Background and Challenge

2. Preliminary Study

3. Our Benchmark

4. Result and Analysis

5. Conclusion

1. Background and Challenge

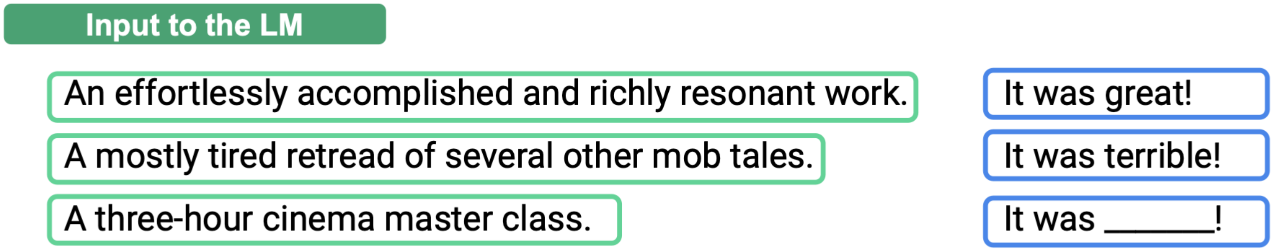
What is many-shot ICL? How is it related to the long-context model evaluation?

Research Question

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

Background: Many-shot ICL

In-Context Learning (ICL) enables models to perform tasks conditioned on a set of demonstrations



Input to the LM

An effortlessly accomplished and richly resonant work. It was great!

A mostly tired retread of several other mob tales. It was terrible!

A three-hour cinema master class. It was _______!

Many-Shot ICL: # of Demonstrations > 100 or even 1000

- Particularly useful for classification tasks with many labels
- A potential alternative to finetuning
- Limited by the context size of LLMs

Background: Long-context Models

The current SoTA long-context models are able to process up to **128k** tokens, some even can handle 1 million tokens

Long-context models are useful for

- Long document summarization
- Long-context dialogue
- Codebase comprehension
- …

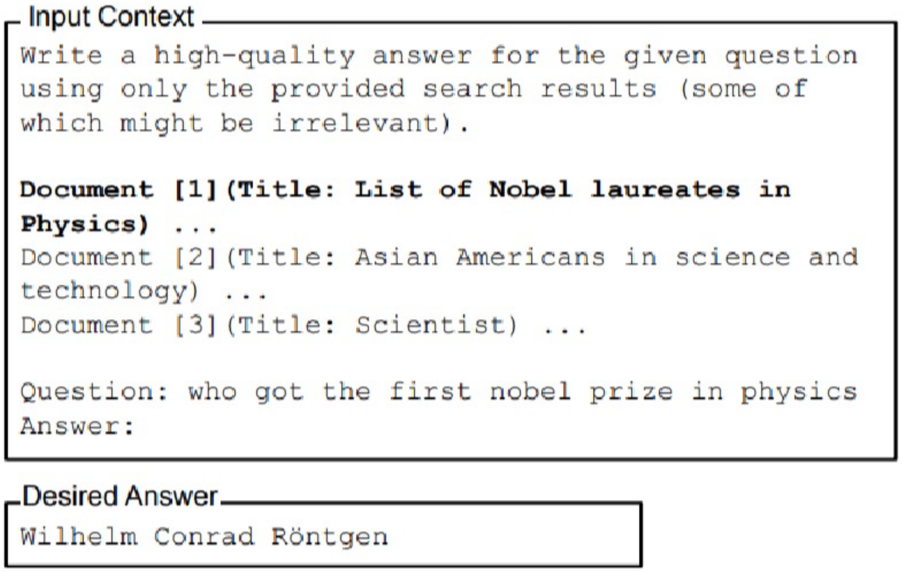
Also, they makes many-shot ICL possible

Background: Long-context Evaluation

Needle In A Haystack: Ask the model to retrieve a random statement (the “needle”) in the middle of a long context window (the “haystack”)

Synthetic Tasks

- Controllable Length
- Easy to obtain
- Not representative of practical use cases
- Mainly Retrieval



Input Context

```
Write a high-quality answer for the given question
using only the provided search results (some of
which might be irrelevant).

Document [1](Title: List of Nobel laureates in
Physics) ...
Document [2](Title: Asian Americans in science and
technology) ...
Document [3](Title: Scientist) ...

Question: who got the first nobel prize in physics
Answer:
```

Desired Answer

```
Wilhelm Conrad Röntgen
```

Retrieval vs Global Context

- Currently, many long-context evaluation benchmarks only focus models’ ability to **retrieve** from the long context

Retrieval vs Global Context

- Currently, many long-context evaluation benchmarks only focus models' ability to **retrieve** from the long context

What about understanding the global context?

Background: Long-Context Evaluation

Novel Challenge: Ask the model to verify claims about fiction books given the whole book context

Realistic Tasks

- Global Understanding
- Closer to the practical use cases
- Costly to annotate,
- Hard to control length

TRUE

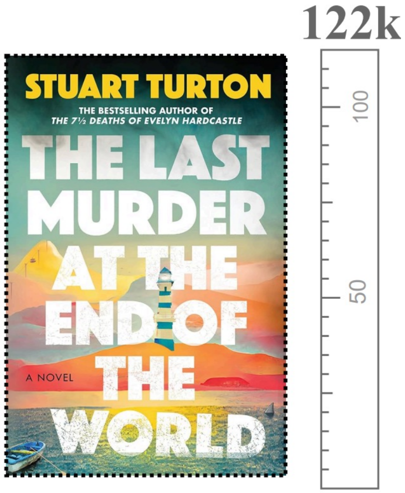
When Niema takes her students out to world's end and tells them that the fog kills anything it touches, she is intentionally lying to them.

FALSE

Niema takes her students out to world's end and tells them that the fog kills anything it touches, a statement backed by her extensive research into the fog.

⚠

Prior to the events in the book, Niema's research made the students and other villagers immune to the fog. She is thus lying when she tells the students that the fog can kill them.


122k
STUART TURTON
THE BESTSELLING AUTHOR OF
THE 7½ DEATHS OF EVELYN HARDCASTLE
THE LAST MURDER AT THE END OF THE WORLD
A NOVEL
100
50

March 28, 2024

Motivation

- Could we build a benchmark that evaluate both models’ retrieval ability and global context understanding?

Motivation

- Could we build a benchmark that evaluate both models' retrieval ability and global context understanding?
- Easy and cheap to obtain?
- Length Controllable?
- More realistic/practical tasks?

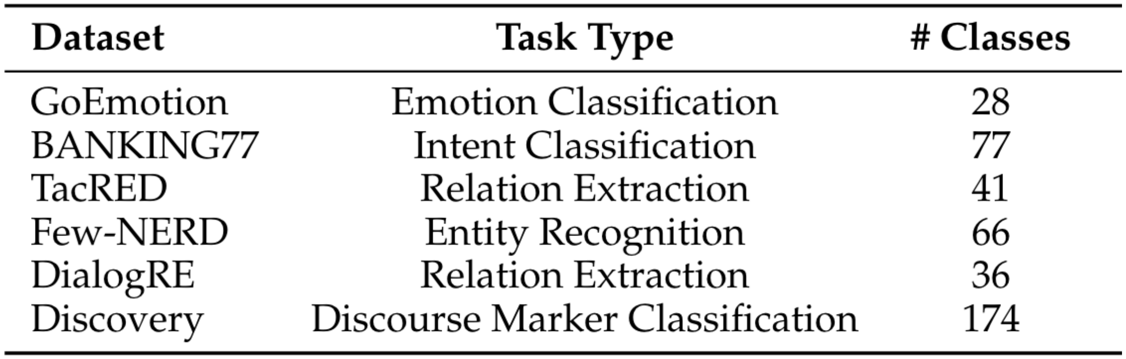
Many-shot ICL

- When the model perform ICL, it is
  - Learning from similar demonstrations as test examples from the prompt to perform the task
  - Using all the demonstrations to learn the underlying task skill and increase its task understanding

- Learning similar examples -> the model's retrieval ability
- Learning all examples -> the model's global context understanding
- Many-shot -> easy to control length
- Realistic tasks

Related Work: LongICLBench

- LongICL Benchmark contains 6 extreme-label classification tasks and evaluate models on many-shot ICL on these 6 tasks.



Dataset	Task Type	# Classes
GoEmotion	Emotion Classification	28
BANKING77	Intent Classification	77
TacRED	Relation Extraction	41
Few-NERD	Entity Recognition	66
DialogRE	Relation Extraction	36
Discovery	Discourse Marker Classification	174

Gaps on Many-shot ICL

- LongICL Benchmark only contains classification tasks
- Unclear how long-context models perform on many-shot ICL tasks other than classification in general.
  - Classification tasks can be solved by retrieving similar examples
- Whether the model only retrieves or refines the task understanding during many-shot ICL.

Contributions

- Investigate whether ICL tasks benefit from additional demonstrations and assess their suitability for evaluating LCLMs with a context length up to 128k tokens.
- Develop methods to characterize the primary skills evaluated by ICL tasks: retrieval capabilities or global context understanding.
- Construct a many-shot ICL benchmark, named ManyICLBench, designed for evaluating LCLMs on both retrieval and global context understanding
- Benchmark 12 widely-used state-of-the-art LCLMs on ManyICLBench to assess their performance comprehensively.

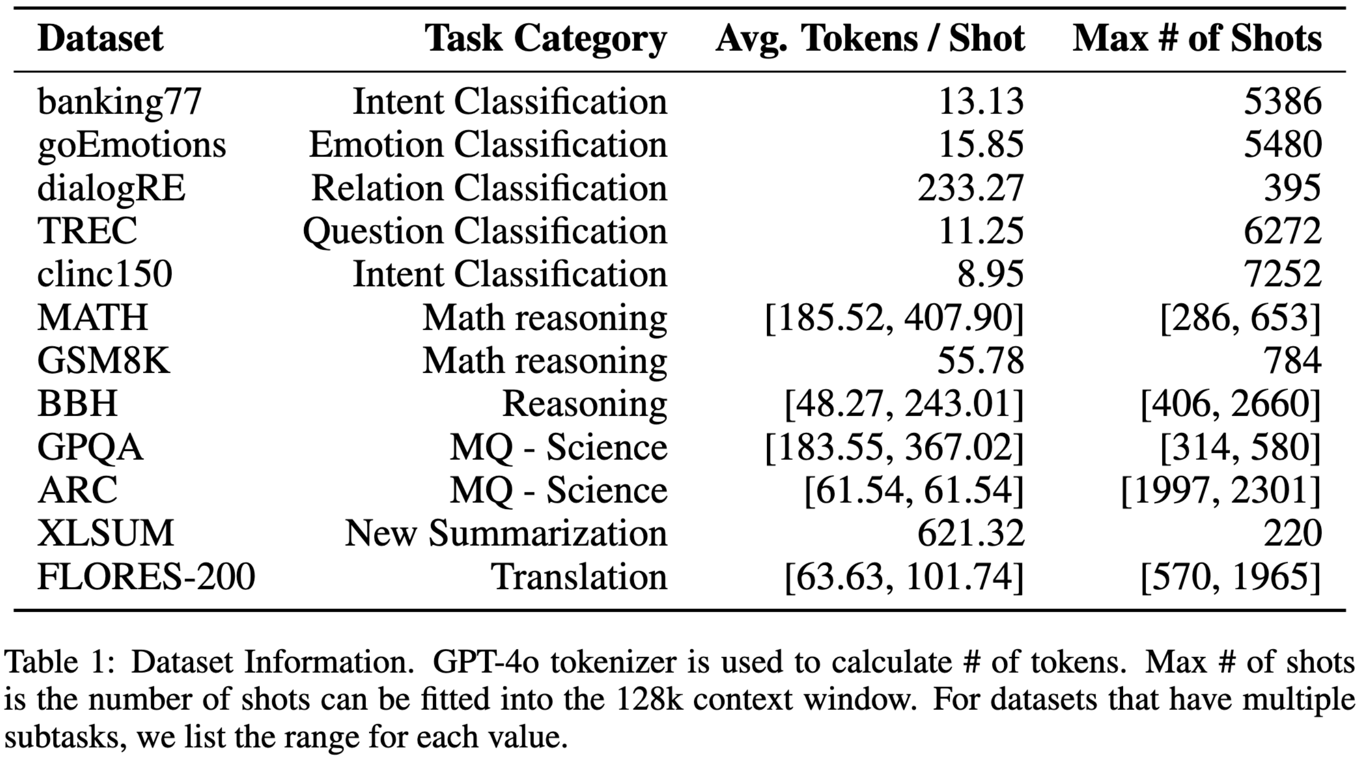
2. Preliminary Study

Does many-shot ICL help?

Preliminary Study : Many-Shot ICL

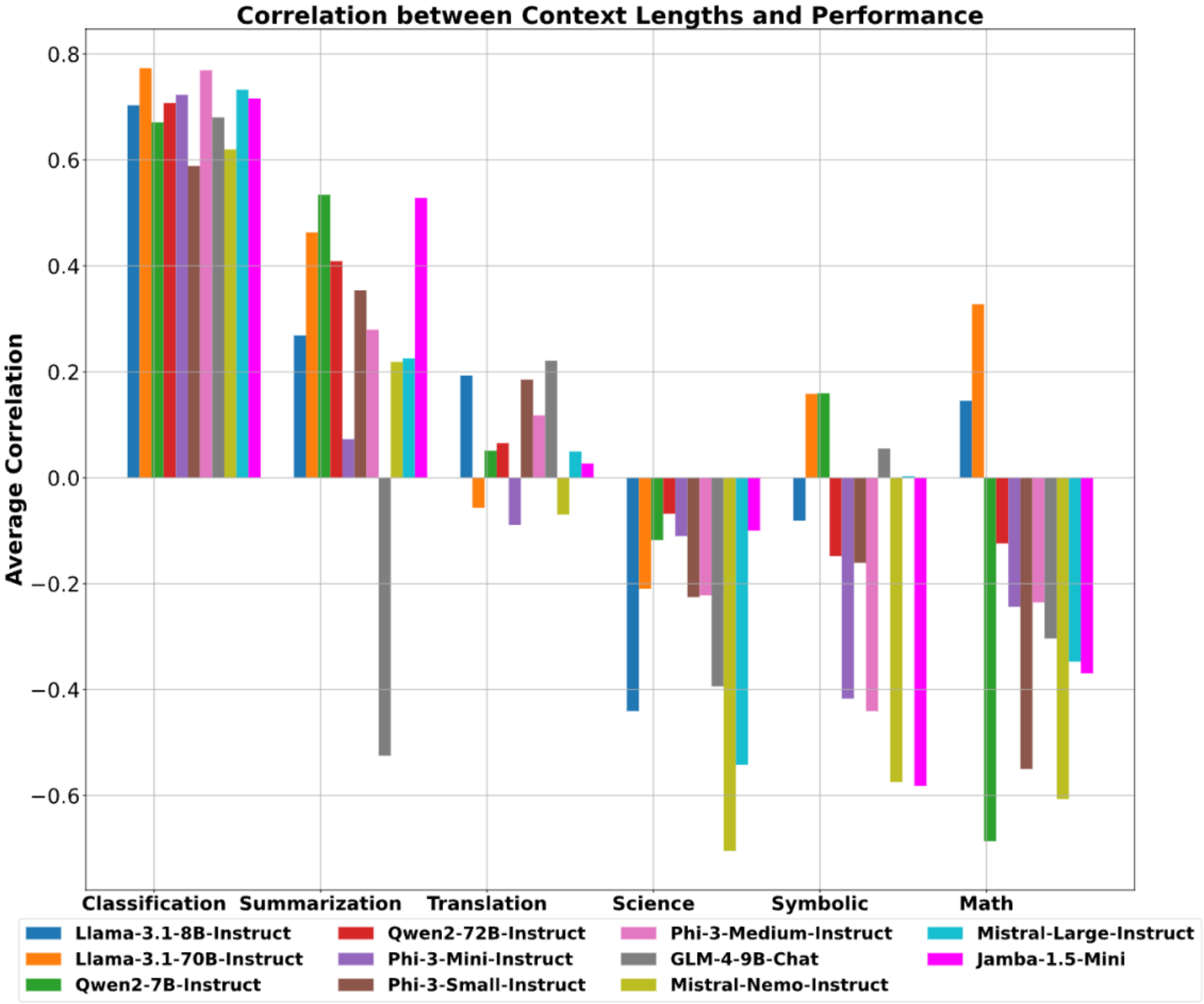
- Datasets:
  - 12 datasets and 20 subtasks
  - Classification, Summarization, Reasoning, and Translation
- Models:
  - Llama-3.1 8B and 70B
  - Qwen2 7B and 72B
  - GLM-4-9B
  - Mistral Nemo (13B) and Large (123B)
  - Phi-3: mini (3.8B), small (7B), and medium (14B)
- Context Length: 1k – 128k, used gpt-4o tokenizer
  - Randomly sample datapoints
  - Longer Prompt contains examples from the shorter prompt

Preliminary Study : Many-Shot ICL

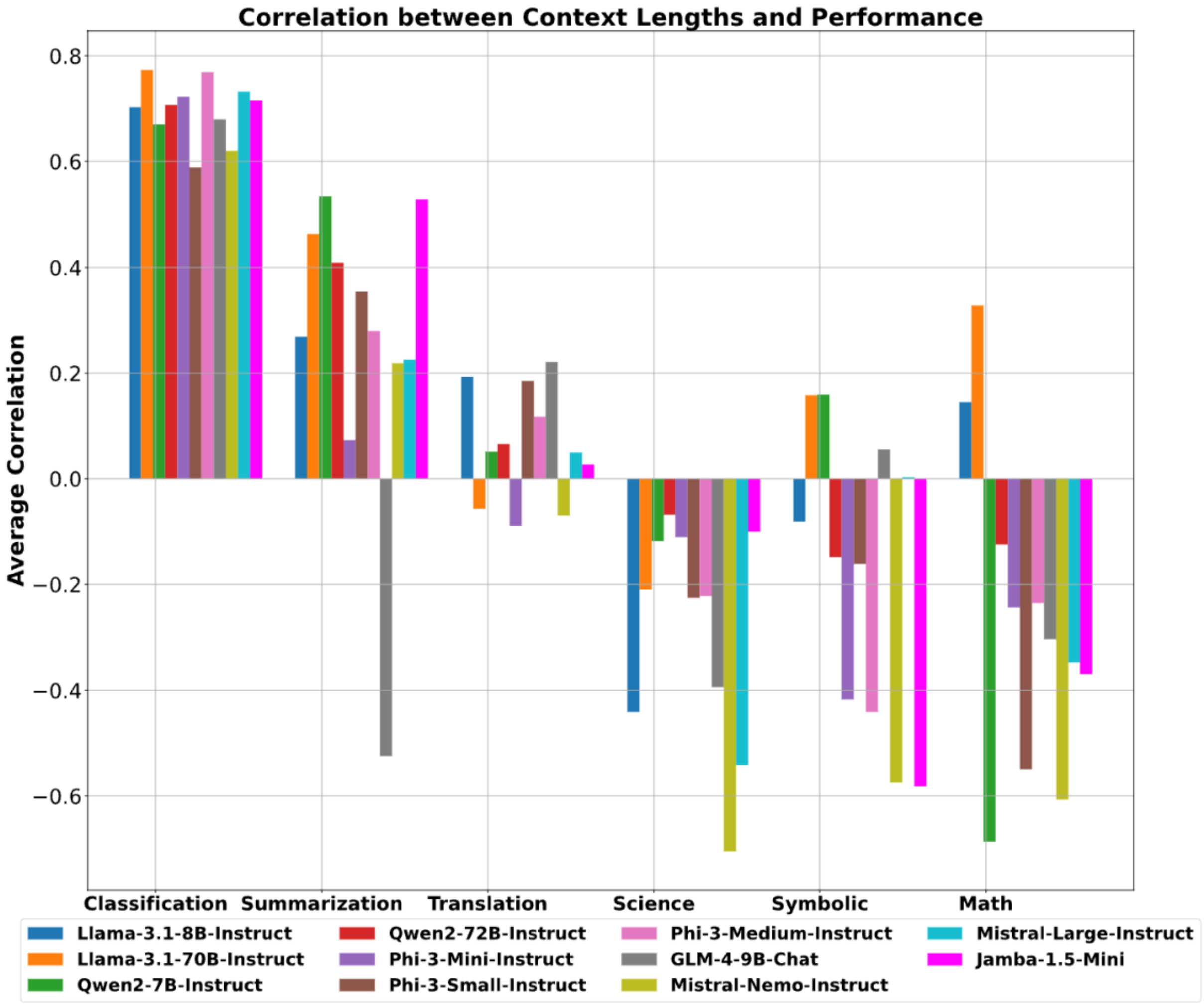


Dataset	Task Category	Avg. Tokens / Shot	Max # of Shots
banking77	Intent Classification	13.13	5386
goEmotions	Emotion Classification	15.85	5480
dialogRE	Relation Classification	233.27	395
TREC	Question Classification	11.25	6272
clinc150	Intent Classification	8.95	7252
MATH	Math reasoning	[185.52, 407.90]	[286, 653]
GSM8K	Math reasoning	55.78	784
BBH	Reasoning	[48.27, 243.01]	[406, 2660]
GPQA	MQ - Science	[183.55, 367.02]	[314, 580]
ARC	MQ - Science	[61.54, 61.54]	[1997, 2301]
XLSUM	New Summarization	621.32	220
FLORES-200	Translation	[63.63, 101.74]	[570, 1965]

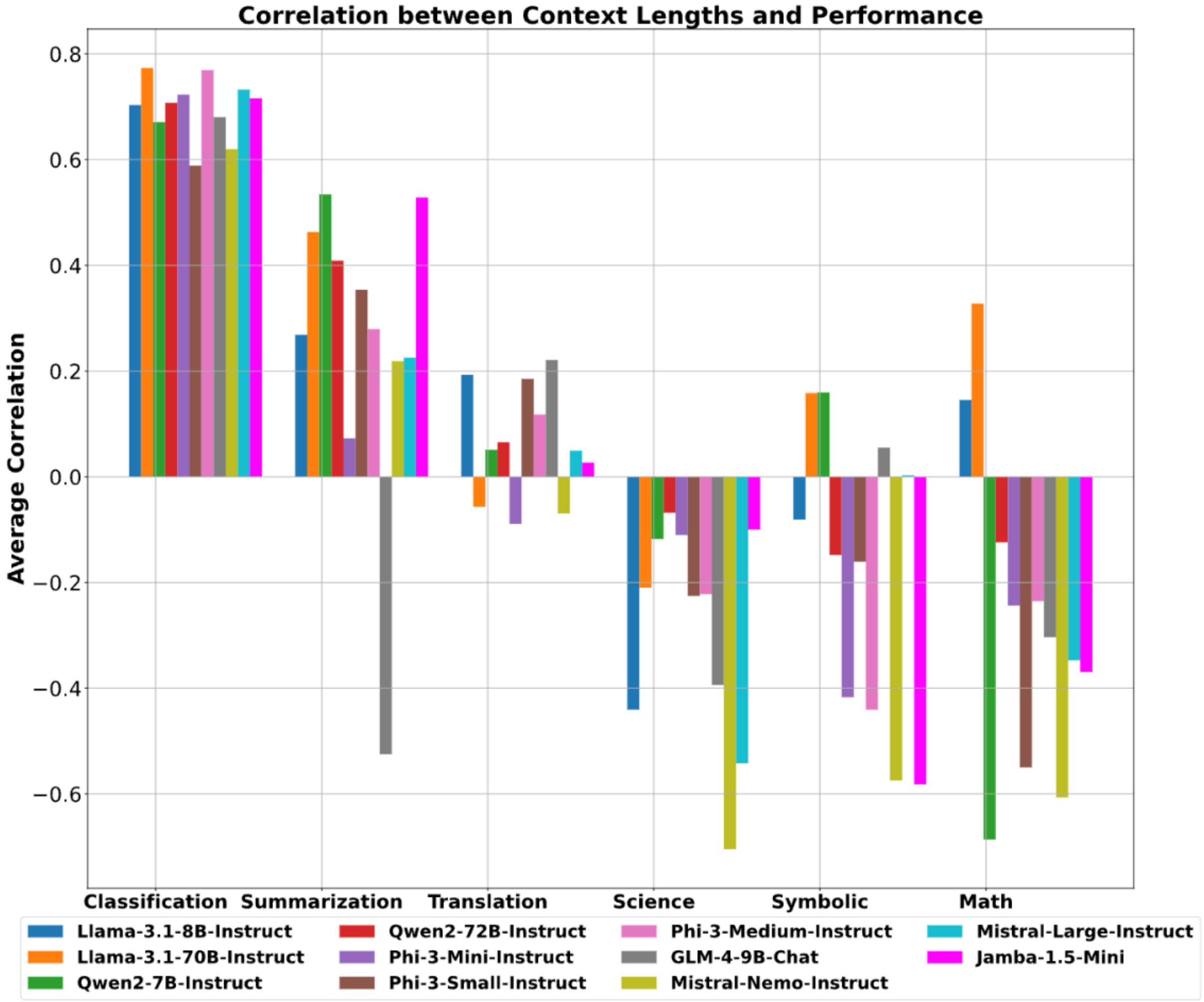
Table 1: Dataset Information. GPT-4o tokenizer is used to calculate # of tokens. Max # of shots is the number of shots can be fitted into the 128k context window. For datasets that have multiple subtasks, we list the range for each value.



- Correlation between context lengths and performance
- Positive = tasks benefit from additional demonstrations
- Negative = tasks harm by additional demonstrations



- Classification Tasks benefit from additional demonstrations
- Summarization tasks benefit from additional demonstrations



- Classification Tasks benefit from additional demonstrations
- Summarization tasks benefit from additional demonstrations
- Other tasks show inconsistent trend in many-shot ICL
  - Model Problem
  - Task Problem
- Ideal Case: additional should improve or at least not harm it
- We want to study each task more carefully

3. Our Benchmark – ManyICLBench

Task Categories

- **Similar-sample learning (SSL) tasks** test models to retrieve and learn from the most similar demonstrations
- **All-sample learning (ASL) tasks** test models to assimilate and learn from all demonstrations

Sample Learning Ratio (SLR)

SLR assess whether tasks predominantly rely on models to retrieve relevant examples during many-shot ICL

High SLR ➡ removing similar examples hurts performance more

Low SLR ➡ removing similar examples is indifferent than removing dissimilar examples

Sample Learning Ratio (SLR)

SLR assess whether tasks predominantly rely on models to retrieve relevant examples during many-shot ICL

High SLR ➡ removing similar examples hurts performance more

Low SLR ➡ removing similar examples is indifferent than removing dissimilar examples

Perf$_{\textbf{least}}$ = performance with {full demonstrations} \ {10% least similar demonstration}

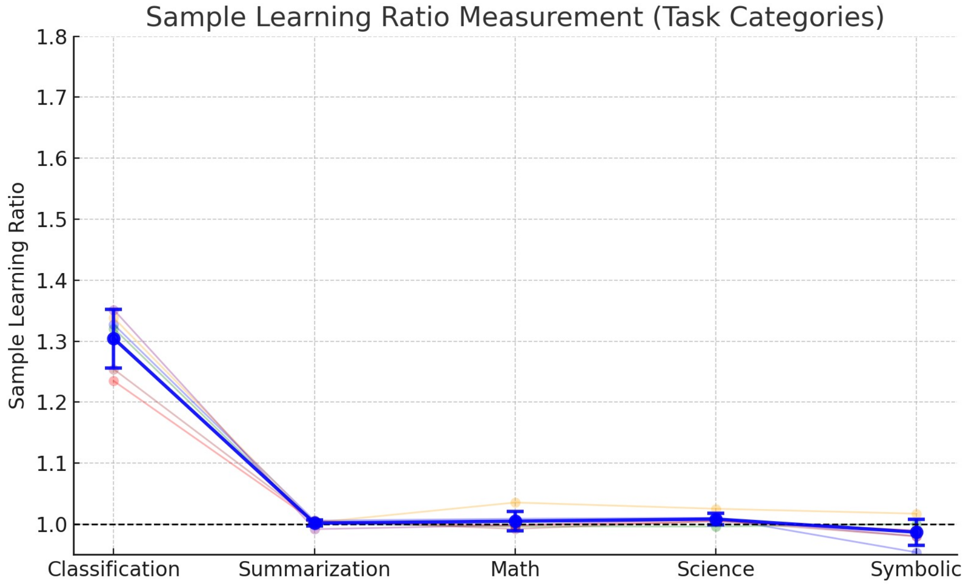
Perf$_{\textbf{most}}$ = performance with {full demonstrations} \ {10% most similar demonstration}

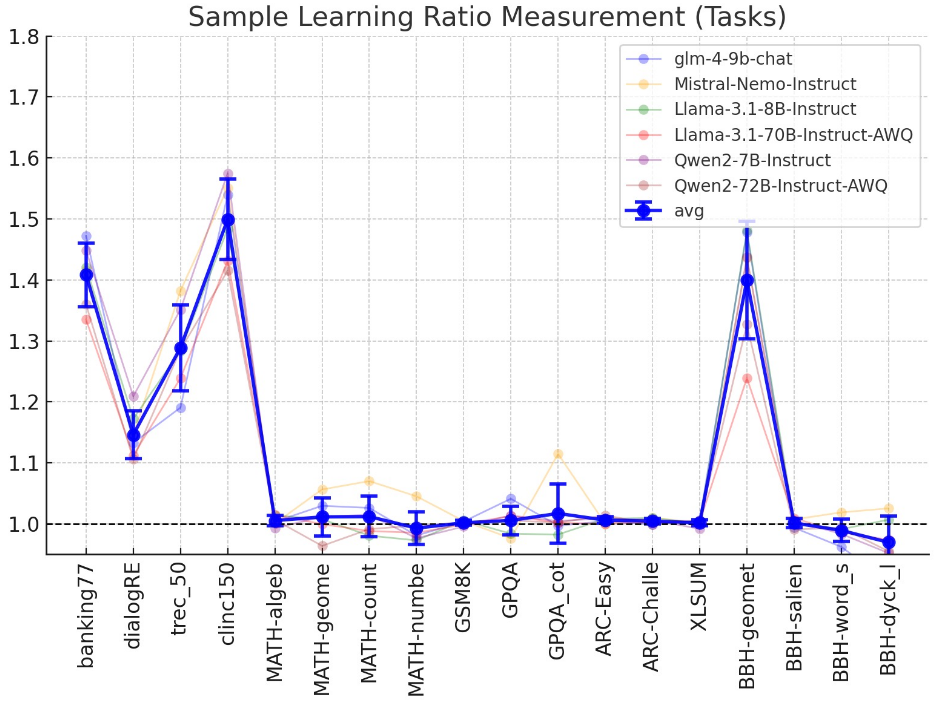
$$\mathrm{SLR} = \frac{1}{7} \sum_{l=1k}^{64k} \frac{\mathrm{Perf}_{\mathrm{least}}^{(l)}}{\mathrm{Perf}_{\mathrm{most}}^{(l)}}$$

SLR – Experiment Setting

- Models
  - Llama-3.1 8B and 70B
  - Qwen2-7B and 72B
  - GLM-4-9b-Chat
  - Mistral-Nemo
- Context Length: 1k – 64k
- All tasks exclude GoEmotions and translations
- 3 different random seeds

SLR – Results


Sample Learning Ratio Measurement (Task Categories)
Sample Learning Ratio
1.8
1.7
1.6
1.5
1.4
1.3
1.2
1.1
1.0
Classification
Summarization
Math
Science
Symbolic

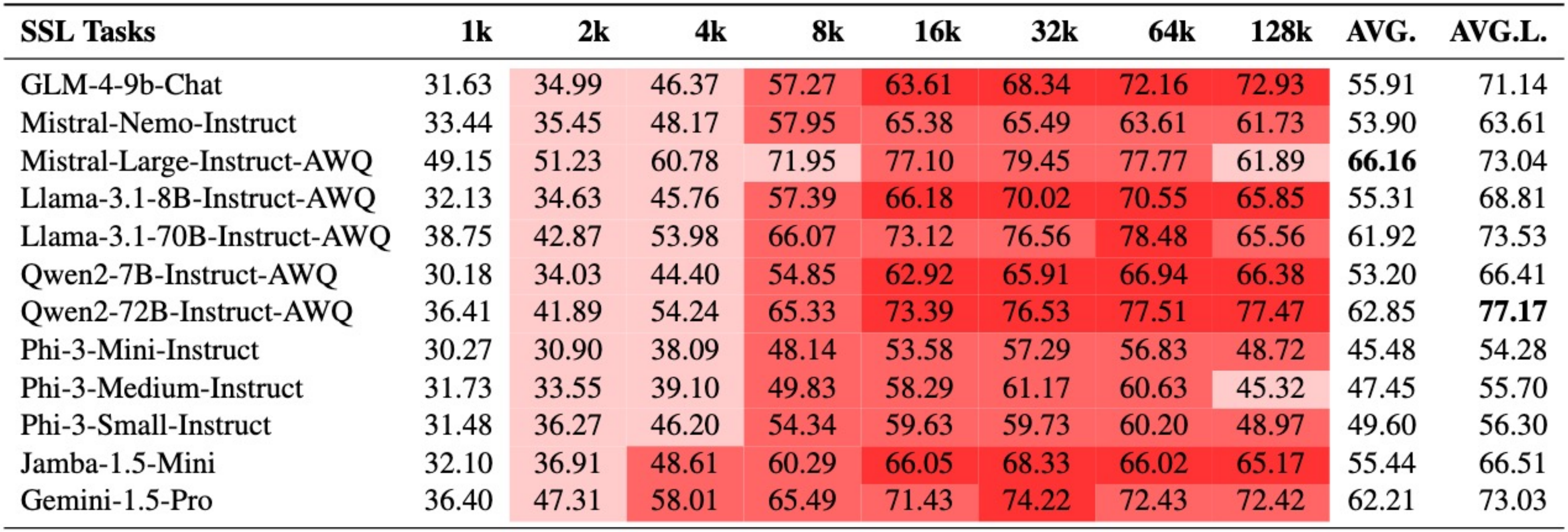

Sample Learning Ratio Measurement (Tasks)
1.8
1.7
1.6
1.5
1.4
1.3
1.2
1.1
1.0
glm-4-9b-chat
Mistral-Nemo-Instruct
Llama-3.1-8B-Instruct
Llama-3.1-70B-Instruct-AWQ
Qwen2-7B-Instruct
Qwen2-72B-Instruct-AWQ
avg
banking77
dialogRE
trec_50
clinc150
MATH-algeb
MATH-geome
MATH-count
MATH-numbe
GSM8K
GPQA
GPQA_cot
ARC-Easy
ARC-Challe
XLSUM
BBH-geomet
BBH-salien
BBH-word_s
BBH-dyck_l

ManyICLBench

- **5 SSL Tasks**: BANKING77, dialogRE, TREC50, CLINC150, and the geometric shape task from BBH.
- **11 ASL Tasks**: all math tasks, summarization task, GPQA with explanations, ARC_challenge, and all BBH tasks except geometric shapes.

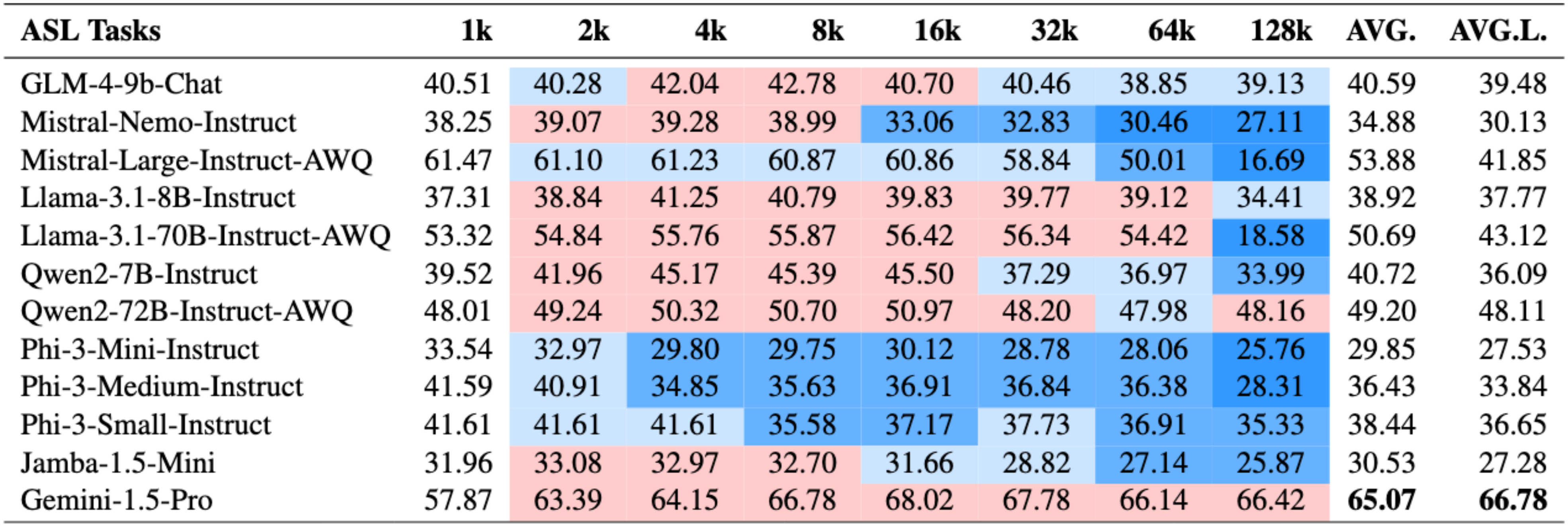
4. Result and Analysis

Models struggle at retrieving examples after 32k



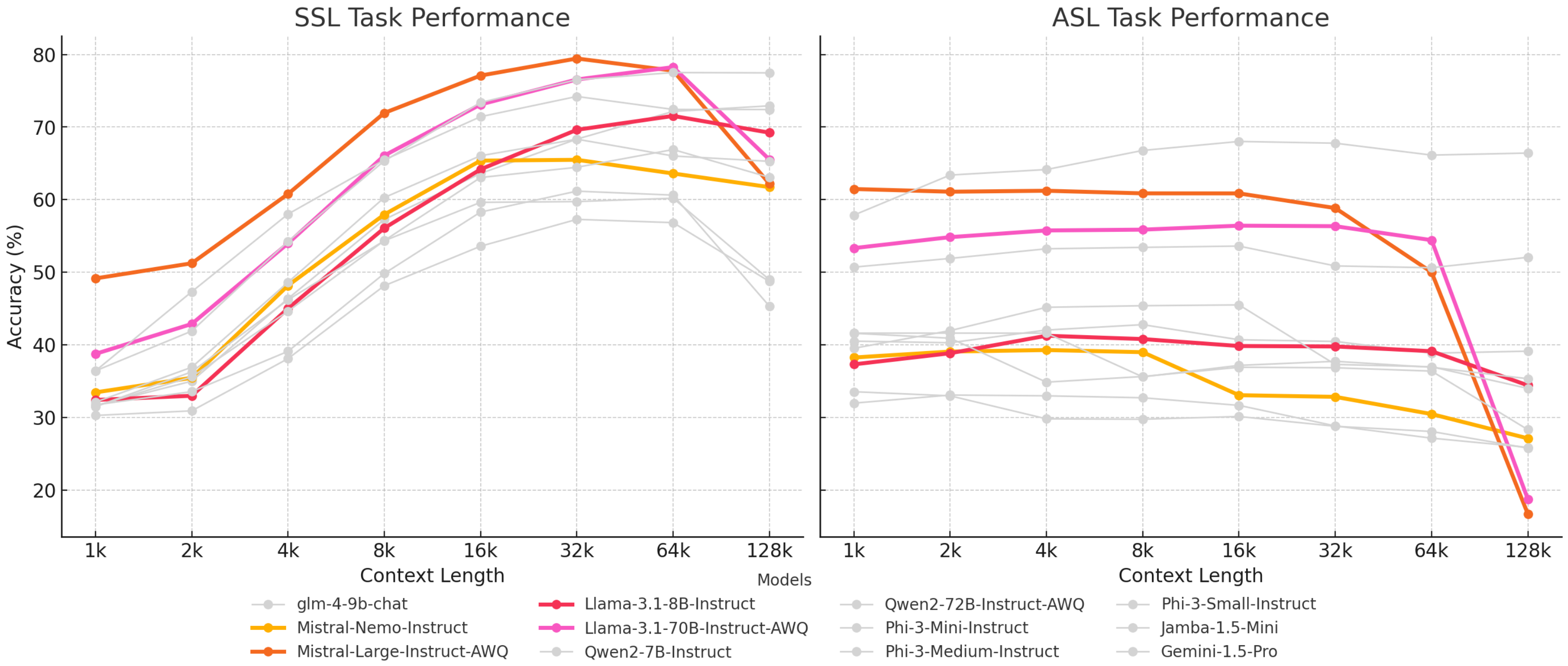
SSL Tasks	1k	2k	4k	8k	16k	32k	64k	128k	AVG.	AVG.L.
GLM-4-9b-Chat	31.63	34.99	46.37	57.27	63.61	68.34	72.16	72.93	55.91	71.14
Mistral-Nemo-Instruct	33.44	35.45	48.17	57.95	65.38	65.49	63.61	61.73	53.90	63.61
Mistral-Large-Instruct-AWQ	49.15	51.23	60.78	71.95	77.10	79.45	77.77	61.89	**66.16**	73.04
Llama-3.1-8B-Instruct-AWQ	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 are challenge

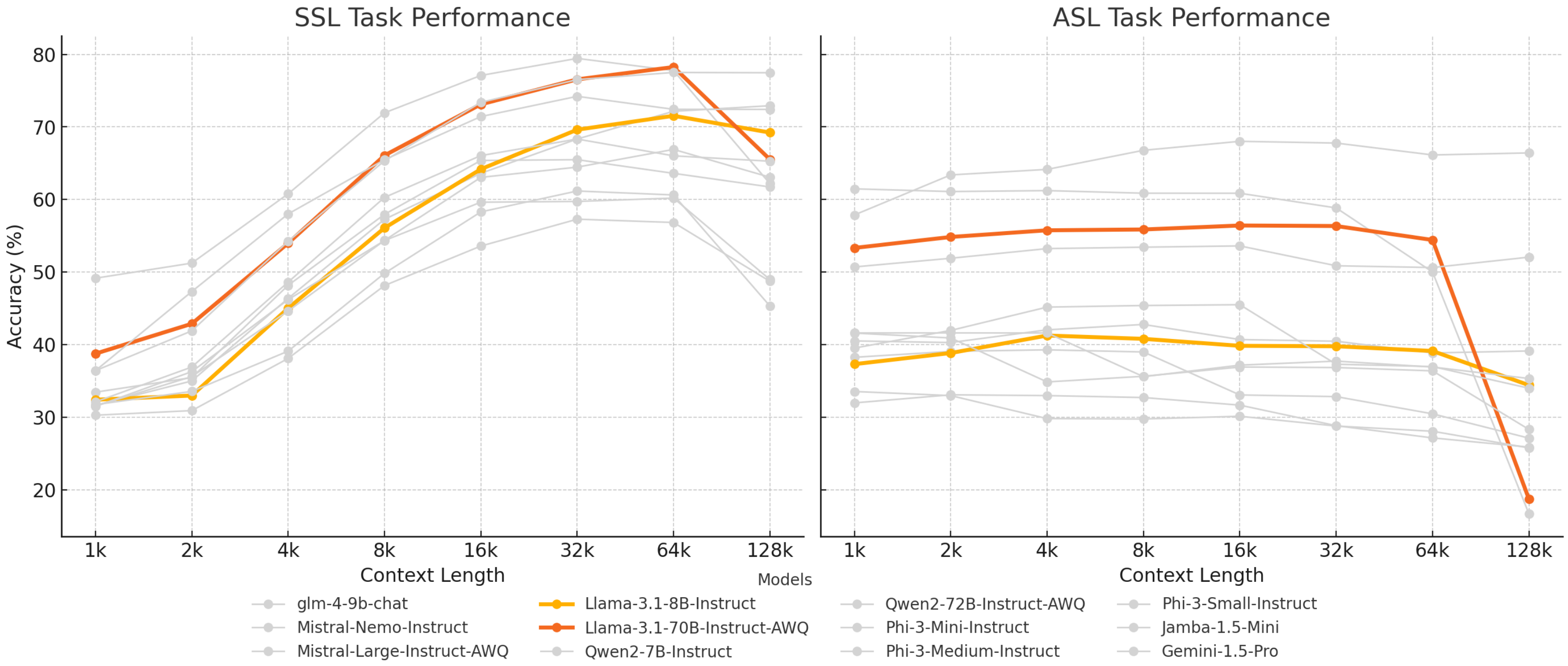


ASL Tasks	1k	2k	4k	8k	16k	32k	64k	128k	AVG.	AVG.L.
GLM-4-9b-Chat	40.51	40.28	42.04	42.78	40.70	40.46	38.85	39.13	40.59	39.48
Mistral-Nemo-Instruct	38.25	39.07	39.28	38.99	33.06	32.83	30.46	27.11	34.88	30.13
Mistral-Large-Instruct-AWQ	61.47	61.10	61.23	60.87	60.86	58.84	50.01	16.69	53.88	41.85
Llama-3.1-8B-Instruct	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**

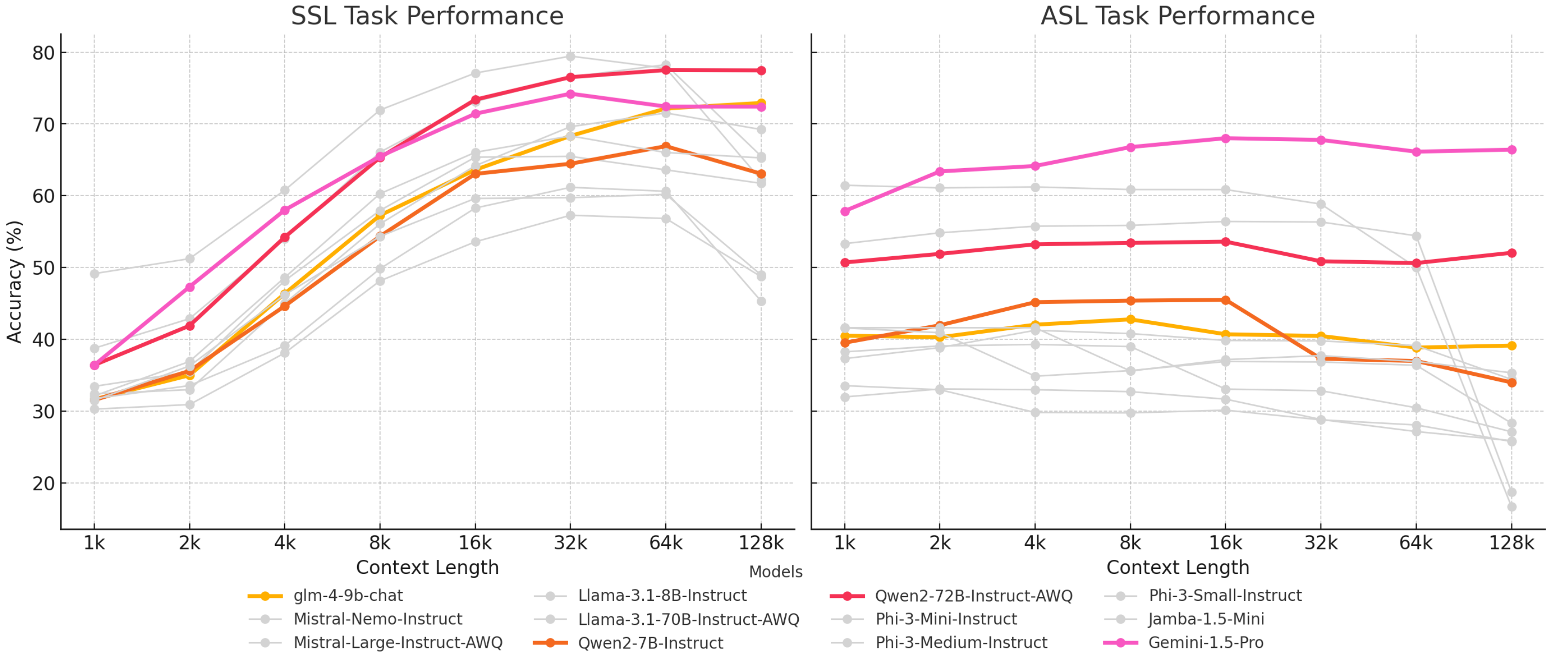
The Paradox of Model Size



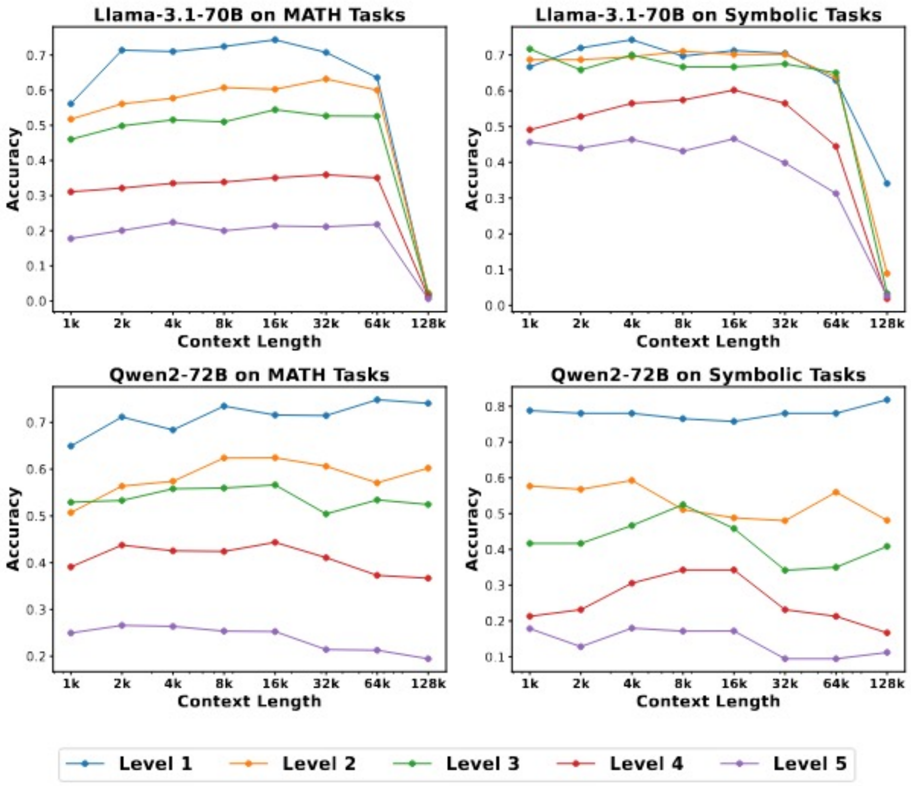
Llama 3.1 Performance Drop at 128k



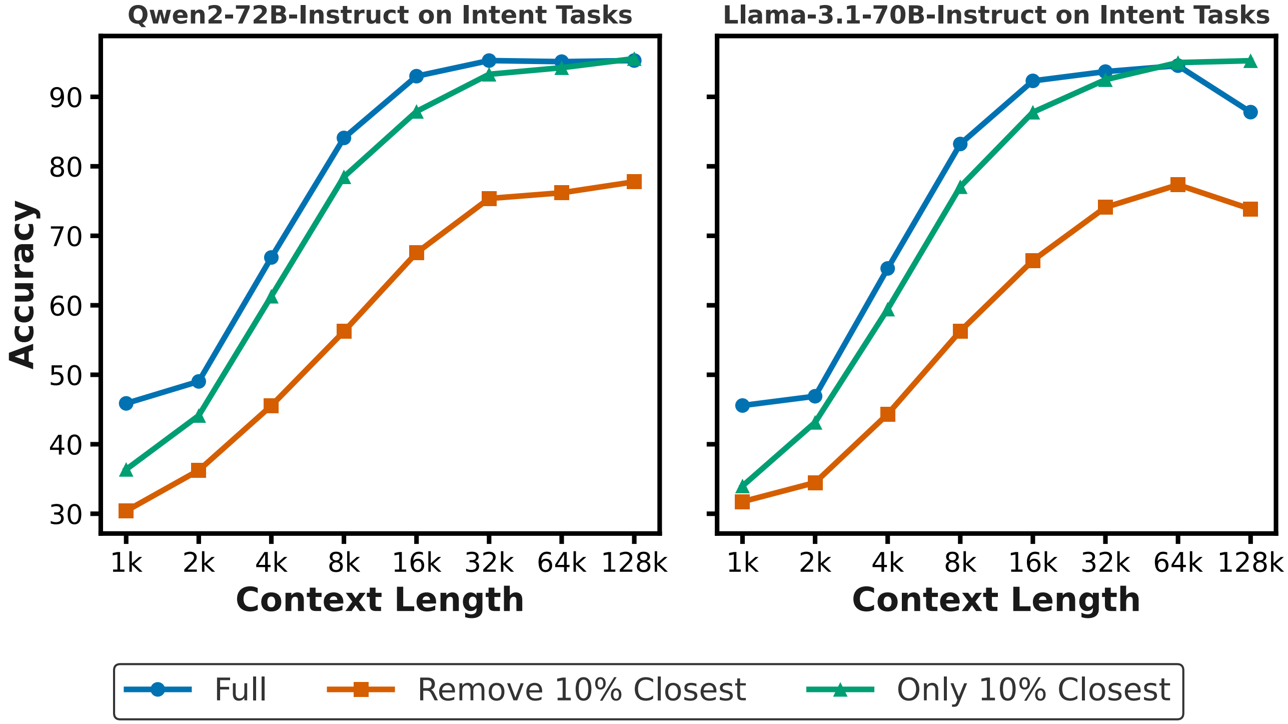
Gemini, Qwen2,and GLM-4 are robust at long context



Why LCLMs fail on ASL tasks?



Why do LCLMs perform better on SSL tasks?



5. Conclusion

Conclusion

- In this work, we test different models on a variety of many-shot ICL tasks
- Using sample learning ratio (SLR), we category tasks into two kinds: similar-sample learning (SSL) tasks and all-sample learning (ASL) tasks.
- We present a long-context evaluation benchmark that contains realistic tasks and evaluate different abilities of models.
- Benchmark 11 open-weight LCLMs and Gemini-1.5 Pro.

The End

Qwen2 Training Details

- Training up to 32k tokens
- Modified RoPE frequence and YARN
- The training-free length extension methods
  - enable models to use additional demonstrations?
  - merely maintain their performance in the short context length ?

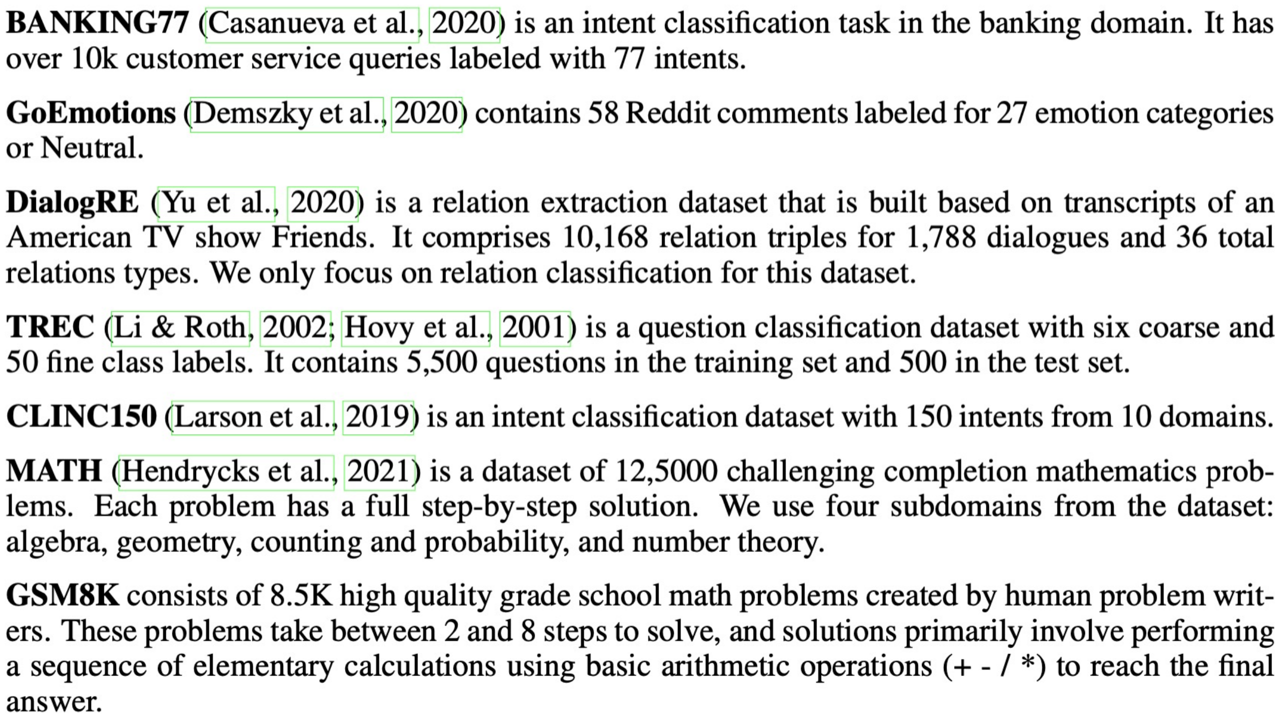
Llama-3.1 Training Details

- Training up to 128k tokens
- Modify RoPE base and continutal training up to 128k tokens
- SFT with long context data
- Why performance drop at 128k?
  - The length distribution of long-context data
  - Only with a mean average of 128k tokens.

GLM-4 Training Details

- Training up to 128k tokens
- Similar to Llama-3.1 Training Recipe
- Why GLM is more robust?
  - Use LongAlign, determines the length distribution of long-context SFT data
  - Go through the RLHF stage with both short and long data

Dataset Details



BANKING77 (Casanueva et al., 2020) is an intent classification task in the banking domain. It has over 10k customer service queries labeled with 77 intents.

GoEmotions (Demszky et al., 2020) contains 58 Reddit comments labeled for 27 emotion categories or Neutral.

DialogRE (Yu et al., 2020) is a relation extraction dataset that is built based on transcripts of an American TV show Friends. It comprises 10,168 relation triples for 1,788 dialogues and 36 total relations types. We only focus on relation classification for this dataset.

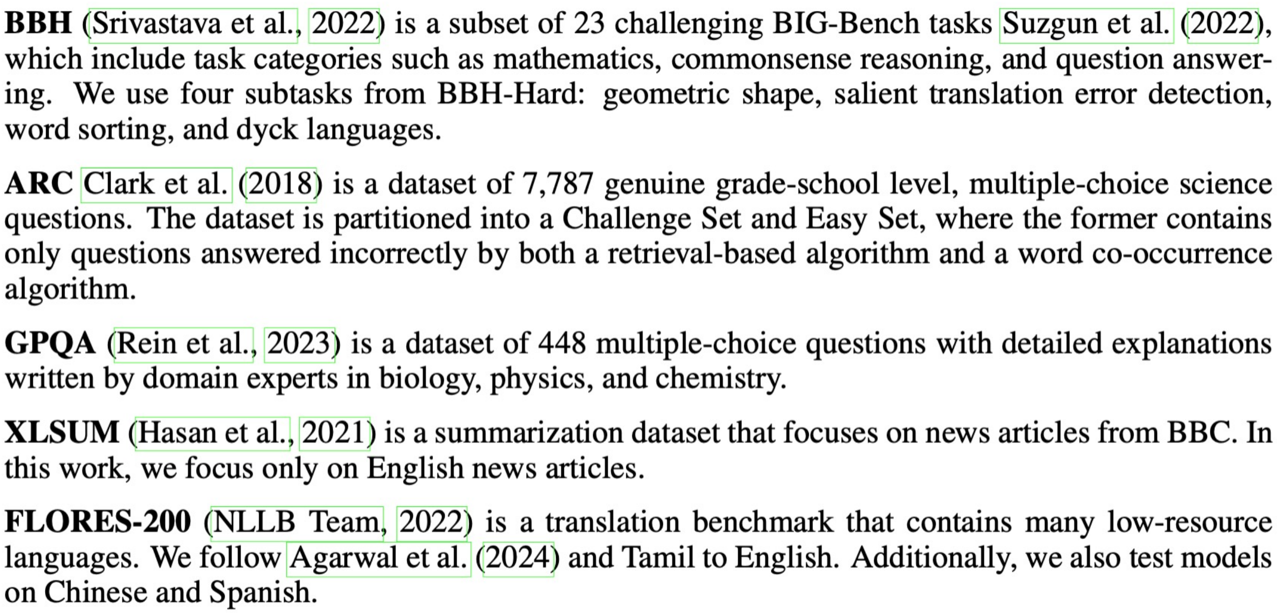
TREC (Li & Roth, 2002; Hovy et al., 2001) is a question classification dataset with six coarse and 50 fine class labels. It contains 5,500 questions in the training set and 500 in the test set.

CLINC150 (Larson et al., 2019) is an intent classification dataset with 150 intents from 10 domains.

MATH (Hendrycks et al., 2021) is a dataset of 12,5000 challenging completion mathematics problems. Each problem has a full step-by-step solution. We use four subdomains from the dataset: algebra, geometry, counting and probability, and number theory.

GSM8K consists of 8.5K high quality grade school math problems created by human problem writers. These problems take between 2 and 8 steps to solve, and solutions primarily involve performing a sequence of elementary calculations using basic arithmetic operations (+ - / *) to reach the final answer.

Dataset Details



BBH (Srivastava et al., 2022) is a subset of 23 challenging BIG-Bench tasks Suzgun et al. (2022), which include task categories such as mathematics, commonsense reasoning, and question answering. We use four subtasks from BBH-Hard: geometric shape, salient translation error detection, word sorting, and dyck languages.

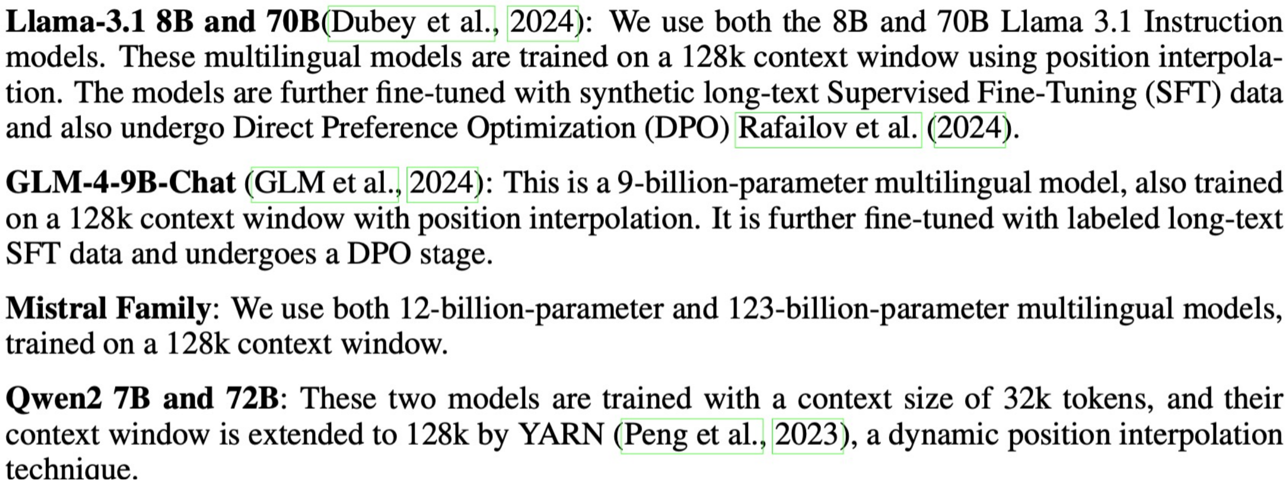
ARC Clark et al. (2018) is a dataset of 7,787 genuine grade-school level, multiple-choice science questions. The dataset is partitioned into a Challenge Set and Easy Set, where the former contains only questions answered incorrectly by both a retrieval-based algorithm and a word co-occurrence algorithm.

GPQA (Rein et al., 2023) is a dataset of 448 multiple-choice questions with detailed explanations written by domain experts in biology, physics, and chemistry.

XLSUM (Hasan et al., 2021) is a summarization dataset that focuses on news articles from BBC. In this work, we focus only on English news articles.

FLORES-200 (NLLB Team, 2022) is a translation benchmark that contains many low-resource languages. We follow Agarwal et al. (2024) and Tamil to English. Additionally, we also test models on Chinese and Spanish.

Model Details

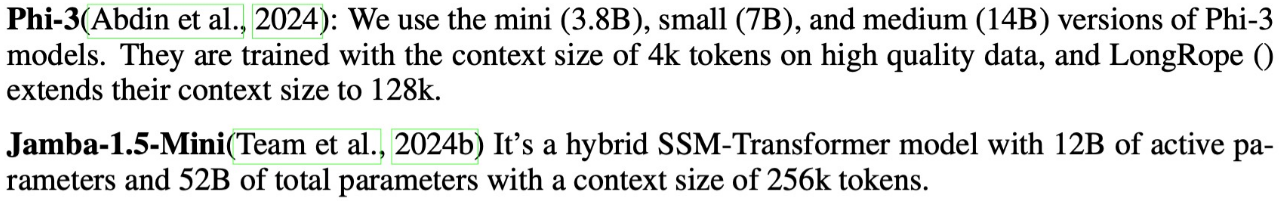


Llama-3.1 8B and 70B(Dubey et al., 2024): We use both the 8B and 70B Llama 3.1 Instruction models. These multilingual models are trained on a 128k context window using position interpolation. The models are further fine-tuned with synthetic long-text Supervised Fine-Tuning (SFT) data and also undergo Direct Preference Optimization (DPO) Rafailov et al. (2024).

GLM-4-9B-Chat (GLM et al., 2024): This is a 9-billion-parameter multilingual model, also trained on a 128k context window with position interpolation. It is further fine-tuned with labeled long-text SFT data and undergoes a DPO stage.

Mistral Family: We use both 12-billion-parameter and 123-billion-parameter multilingual models, trained on a 128k context window.

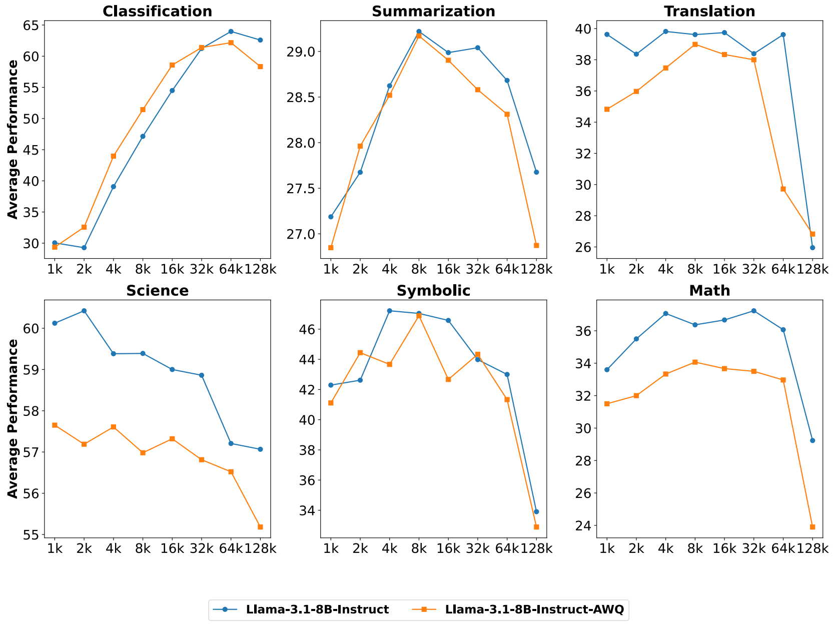
Qwen2 7B and 72B: These two models are trained with a context size of 32k tokens, and their context window is extended to 128k by YARN (Peng et al., 2023), a dynamic position interpolation technique.

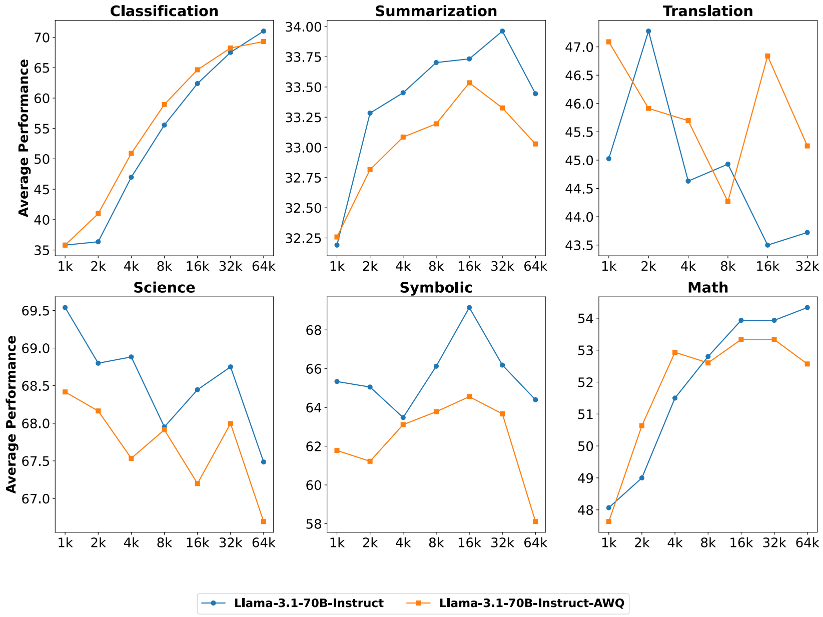


Phi-3(Abdin et al., 2024): We use the mini (3.8B), small (7B), and medium (14B) versions of Phi-3 models. They are trained with the context size of 4k tokens on high quality data, and LongRope () extends their context size to 128k.

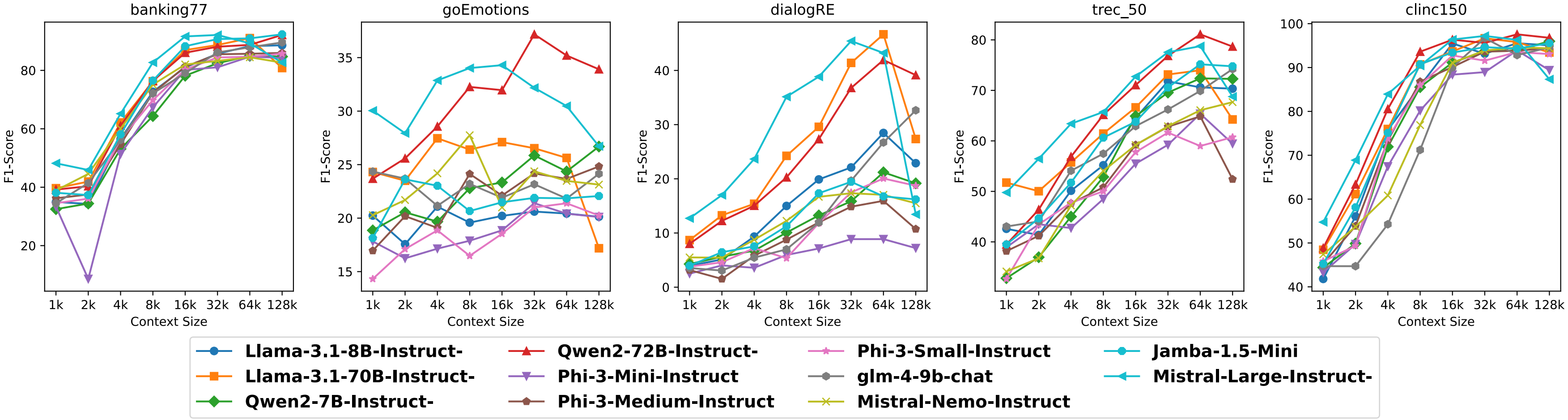
Jamba-1.5-Mini(Team et al., 2024b) It's a hybrid SSM-Transformer model with 12B of active parameters and 52B of total parameters with a context size of 256k tokens.

Quantized vs Regular

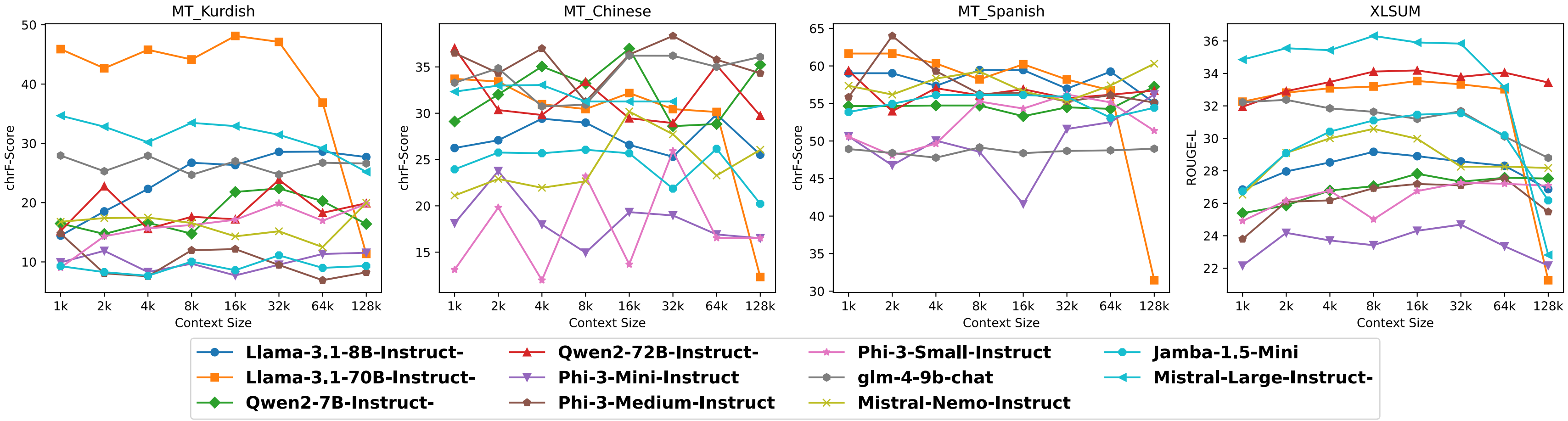




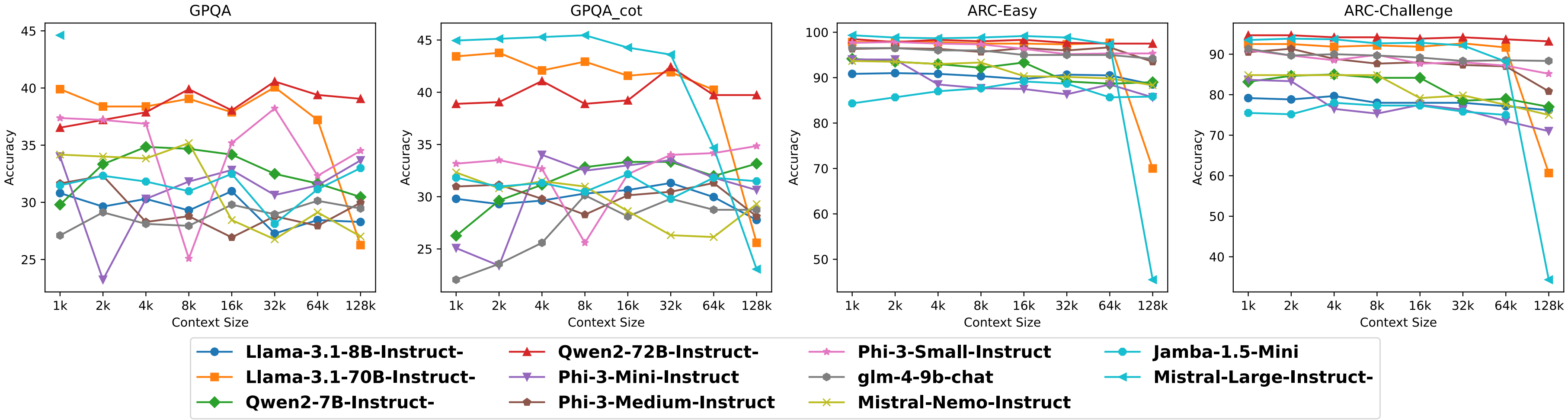
Classification Performance



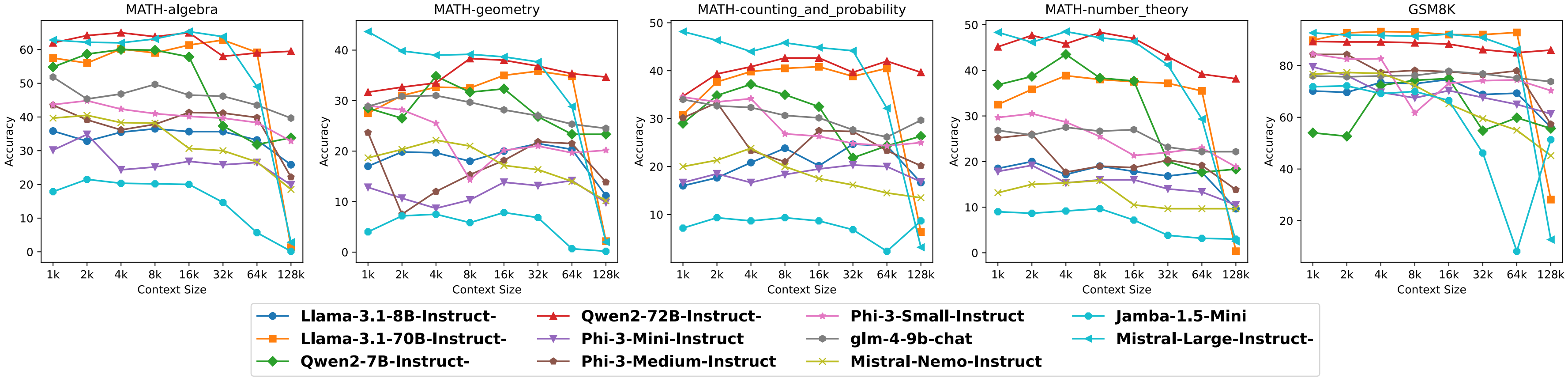
Translation + Summarization Performance



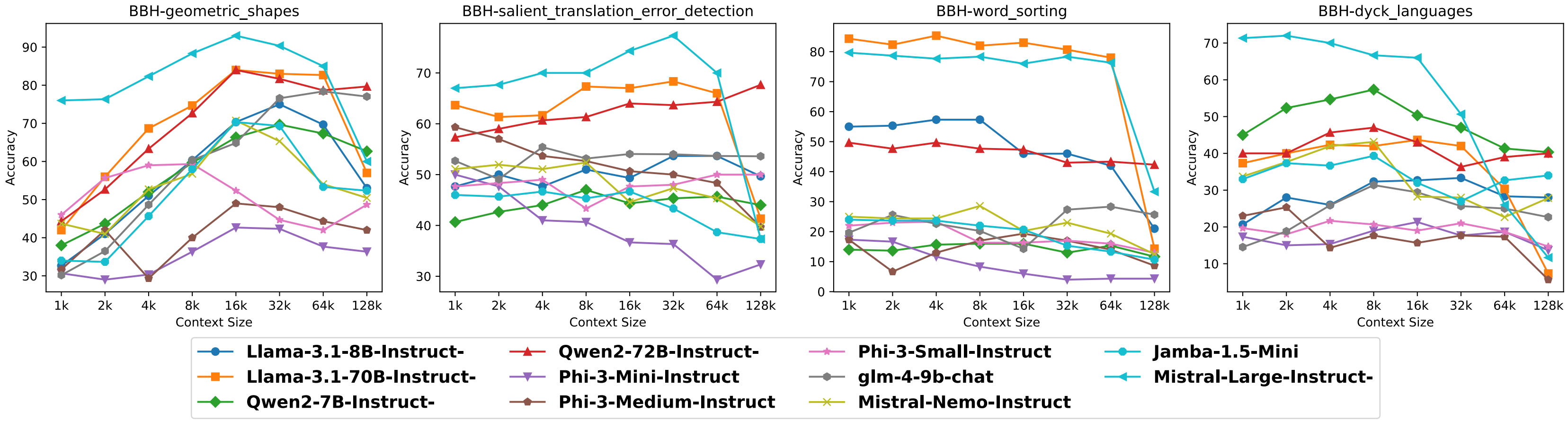
Science Performance



Math Performance



Symbolic Performance


BBH-geometric_shapes
BBH-salient_translation_error_detection
BBH-word_sorting
BBH-dyck_languages
Accuracy
Context Size
1k
2k
4k
8k
16k
32k
64k
128k
Llama-3.1-8B-Instruct-
Llama-3.1-70B-Instruct-
Qwen2-7B-Instruct-
Qwen2-72B-Instruct-
Phi-3-Mini-Instruct
Phi-3-Medium-Instruct
Phi-3-Small-Instruct
glm-4-9b-chat
Mistral-Nemo-Instruct
Jamba-1.5-Mini
Mistral-Large-Instruct-