Response Length Perception and Sequence Scheduling: An LLM-Empowered LLM Inference Pipeline

Zangwei Zheng[1], Xiaozhe Ren[2], Fuzhao Xue[1], Yang Luo[1], Xin Jiang[2], Yang You[1]

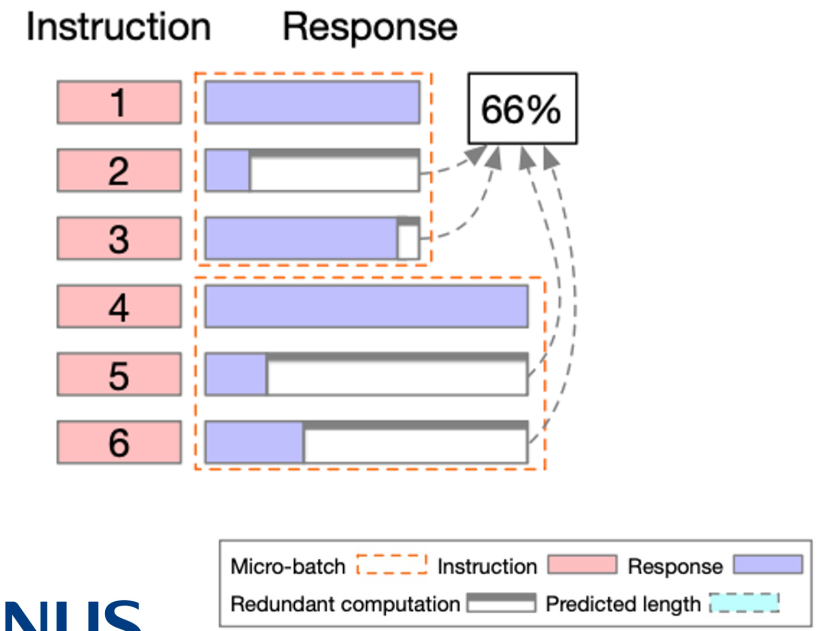
Presentor: Zangwei Zheng

[1]National University of Singapore [2]Huawei

Background

Batch inference speed is negatively affected by different response lengths

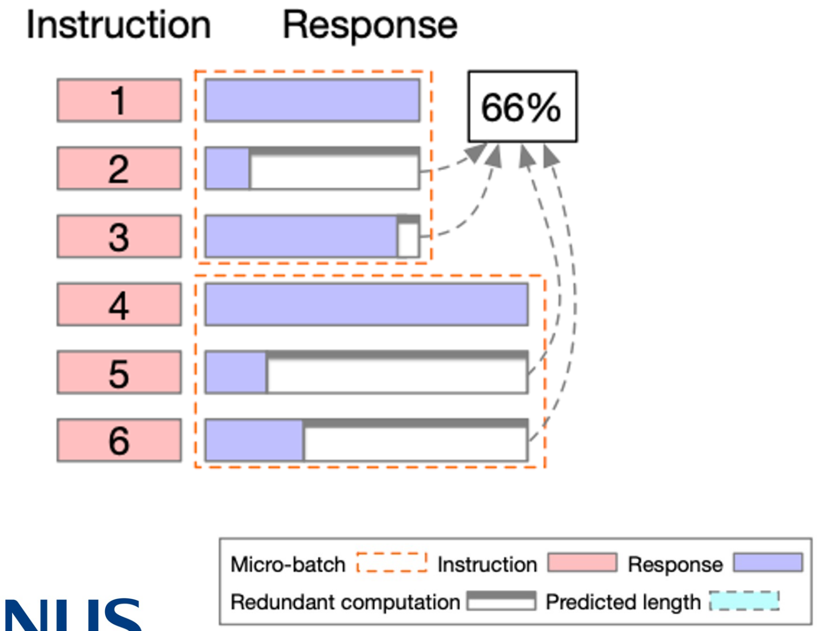
When performing large language model inference in batches, the inclusion of sequences with differerent response lengths leads to inefficiencies. Shorter sequences are forced to wait for longer ones to complete, resulting in computational waste.

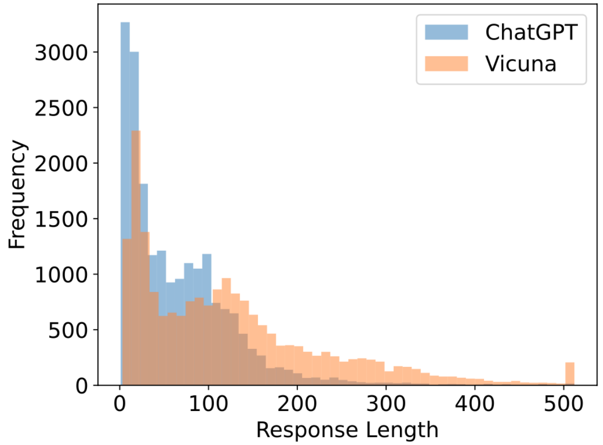


Background

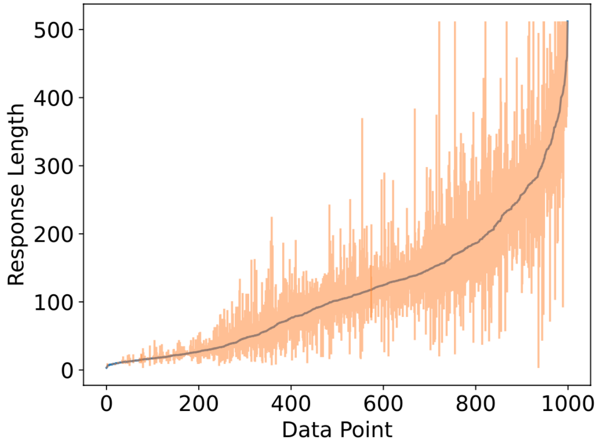
Batch inference speed is negatively affected by different response lengths

In real-world scenarios, the lengths of responses to various queries exhibit significant variability.





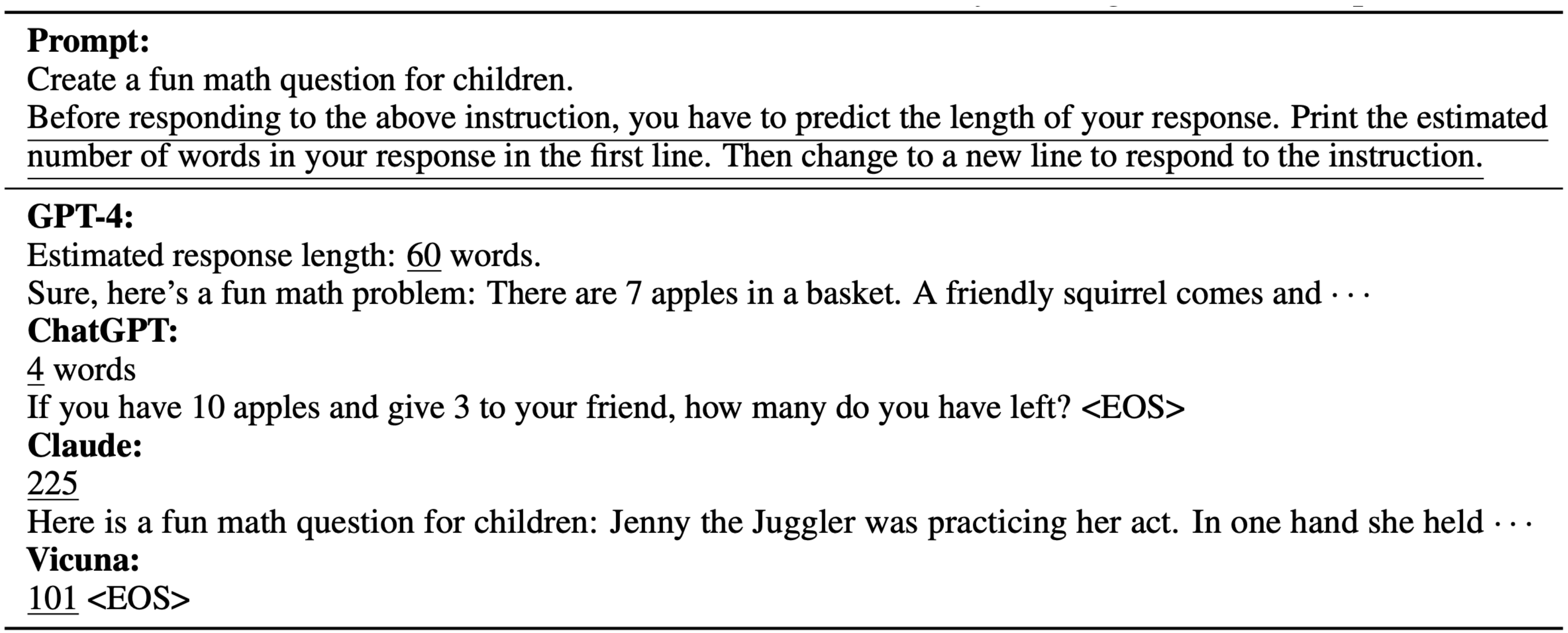
(a) Response length distribution of 10k instructions from ChatGPT and Vicuna. Response lengths larger than 512 are truncated.



(b) Distribution of mean length among 3 times generations on 1k instructions. Error bar denotes maximum and minimum length in generations.

Response Length Perception

LLMs can predict the length of their responses.



Prompt:
Create a fun math question for children.
Before responding to the above instruction, you have to predict the length of your response. Print the estimated number of words in your response in the first line. Then change to a new line to respond to the instruction.

GPT-4:
Estimated response length: 60 words.
Sure, here's a fun math problem: There are 7 apples in a basket. A friendly squirrel comes and · · ·
ChatGPT:
4 words
If you have 10 apples and give 3 to your friend, how many do you have left? <EOS>
Claude:
225
Here is a fun math question for children: Jenny the Juggler was practicing her act. In one hand she held · · ·
Vicuna:
101 <EOS>

Response Length Perception

LLMs can predict the length of their responses.

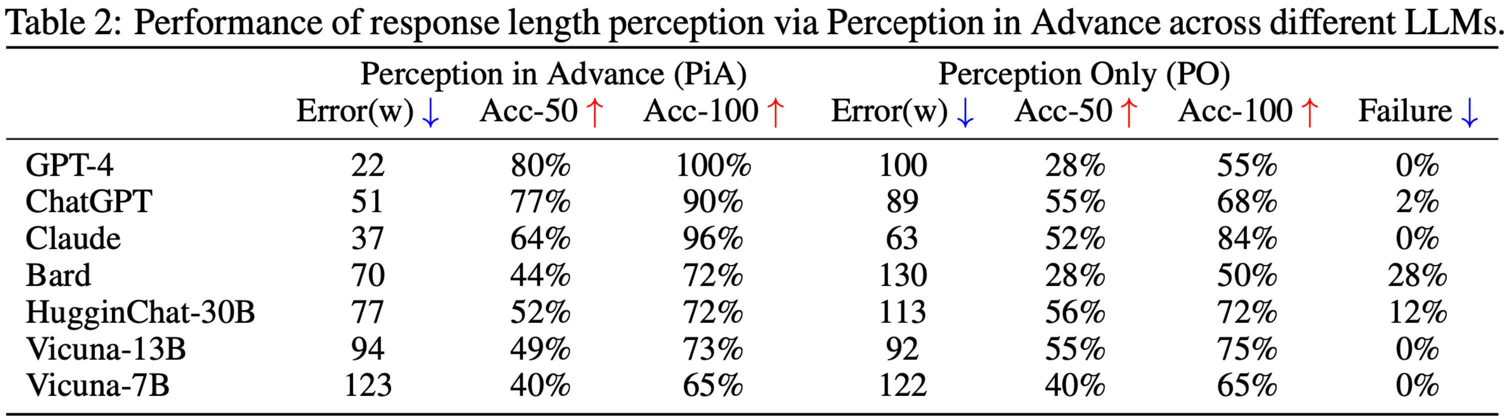
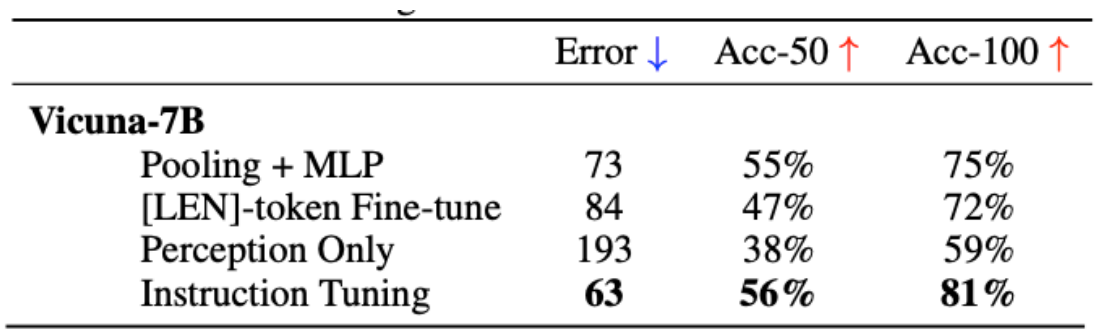


Table 2: Performance of response length perception via Perception in Advance across different LLMs.

	Perception in Advance (PiA)			Perception Only (PO)			
	Error(w) ↓	Acc-50 ↑	Acc-100 ↑	Error(w) ↓	Acc-50 ↑	Acc-100 ↑	Failure ↓
GPT-4	22	80%	100%	100	28%	55%	0%
ChatGPT	51	77%	90%	89	55%	68%	2%
Claude	37	64%	96%	63	52%	84%	0%
Bard	70	44%	72%	130	28%	50%	28%
HugginChat-30B	77	52%	72%	113	56%	72%	12%
Vicuna-13B	94	49%	73%	92	55%	75%	0%
Vicuna-7B	123	40%	65%	122	40%	65%	0%

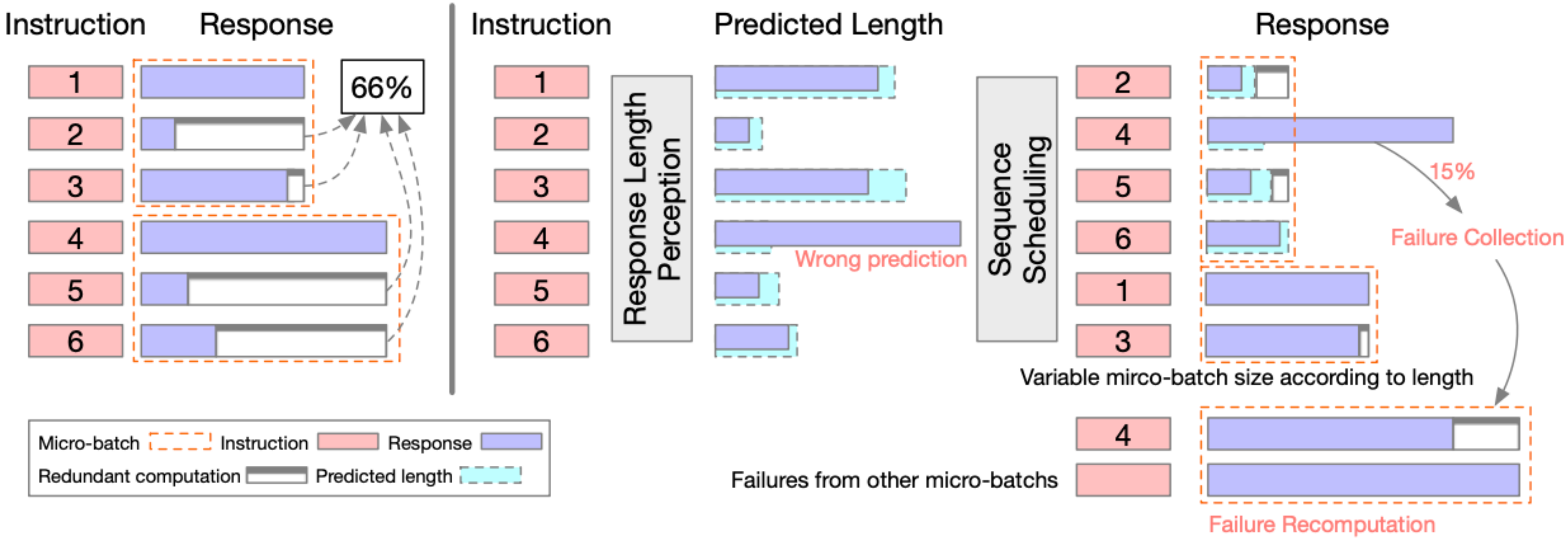
Response Length Perception

For smaller models, we use instruction tunning to improve this ability.



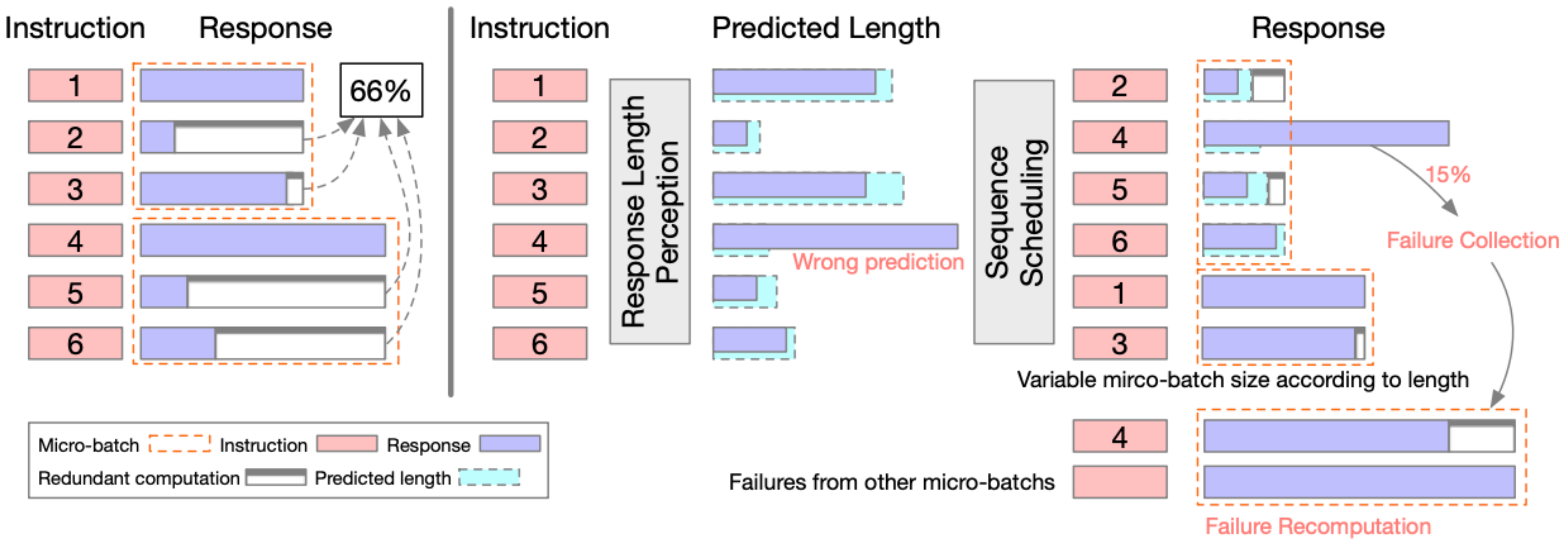
	Error ↓	Acc-50 ↑	Acc-100 ↑
Vicuna-7B			
Pooling + MLP	73	55%	75%
[LEN]-token Fine-tune	84	47%	72%
Perception Only	193	38%	59%
Instruction Tuning	**63**	**56%**	**81%**

Sequence Scheduling



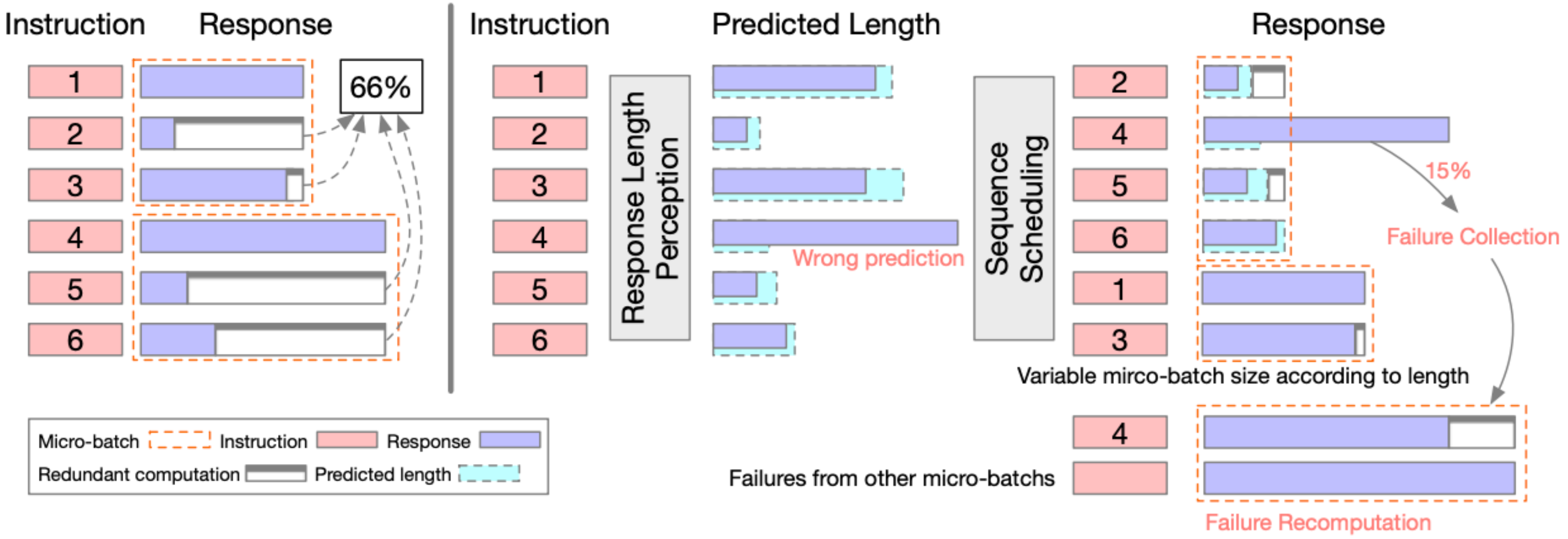
Our method predicts the length of responses first and groups the ones with similar lengths into batches. Two techniques are used to further improve the performance: Failure Collection and Re-computation (FCR), and Variable Batch Size (VBS)

Sequence Scheduling



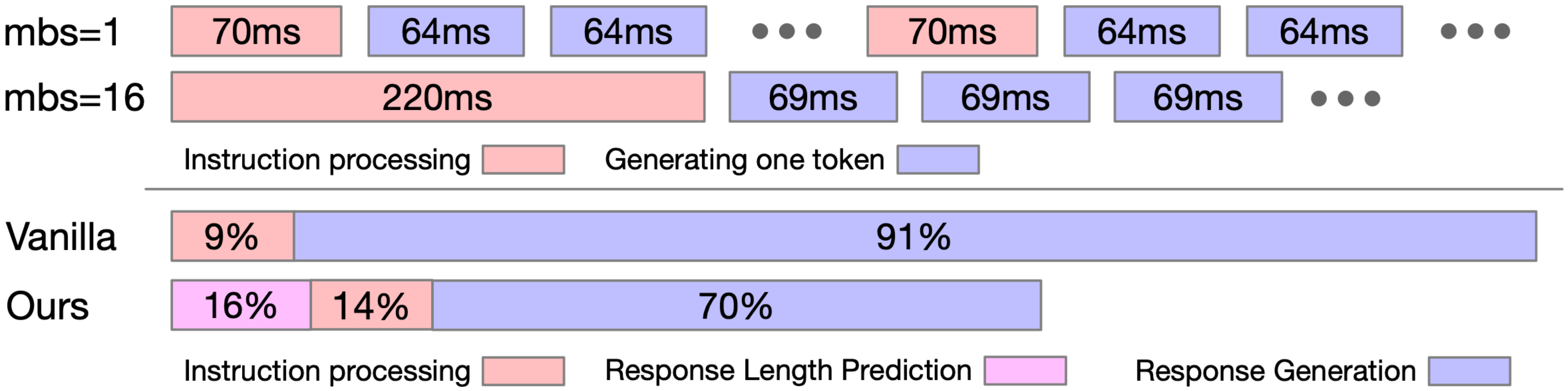
Failure Collection and Re-computation (FCR): If a long response is mistakenly predicted as a short one and included in a batch with predominantly short responses, the overall processing time is affected as the short queries are forced to wait for the completion of the long one. We restrict the number of newly generated tokens to be at most the maximum predicted length within a batch and re-compute the failed ones at the end of the group.

Sequence Scheduling



Variable Batch Size (VBS): Shorter responses require less memory compared to longer ones. We allocate a larger batch size for shorter responses.

Experiment



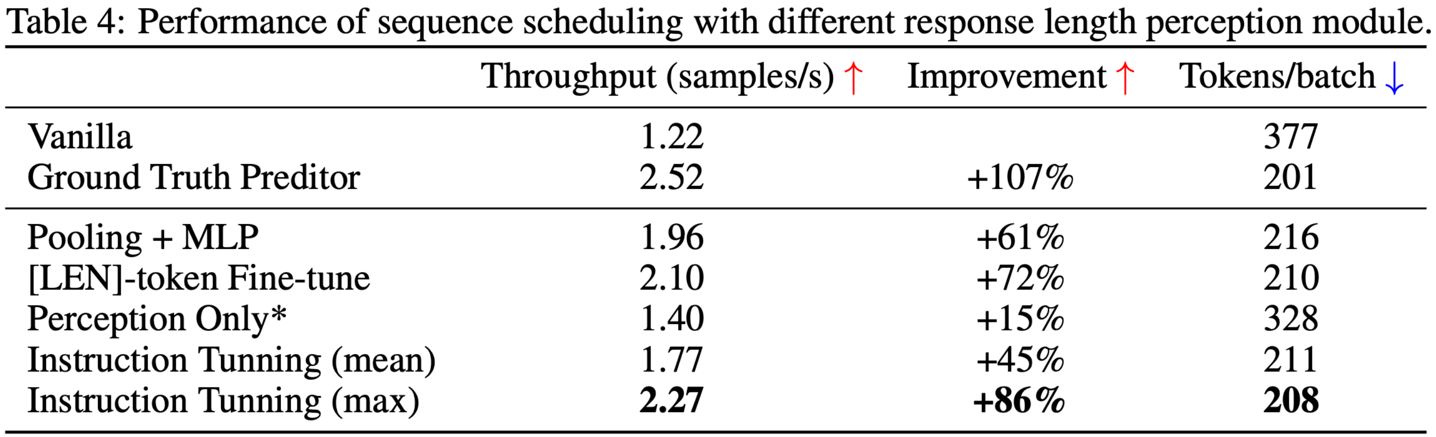


Table 4: Performance of sequence scheduling with different response length perception module.

	Throughput (samples/s) ↑	Improvement ↑	Tokens/batch ↓
Vanilla	1.22		377
Ground Truth Preditor	2.52	+107%	201
Pooling + MLP	1.96	+61%	216
[LEN]-token Fine-tune	2.10	+72%	210
Perception Only*	1.40	+15%	328
Instruction Tunning (mean)	1.77	+45%	211
Instruction Tunning (max)	**2.27**	**+86%**	**208**

Summary

LLMs can **estimate the length** of their responses, which can be used to **group batches for efficient inference**

Thank you