Second Place Overall Solution for Amazon KDD CUP 2024

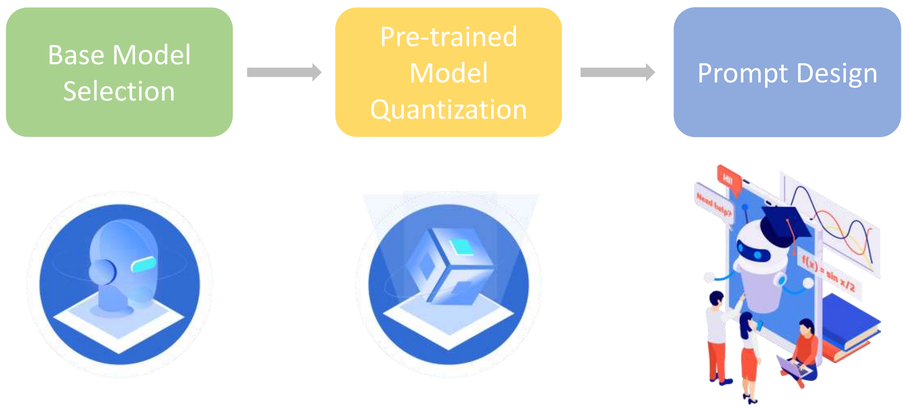
Pengyue Jia[1]*, Jingtong Gao[1]*, Xiaopeng Li[1]*, Zixuan Wang[1]*, Yiyao Jin[1]*, Xiangyu Zhao[1]

[1] School of Data Science, City University of Hong Kong

Code

CityU AML Lab

1. Overview



1.1 Introduction

➢ Challenges

This competition is challenging because of **no training dataprovided**, **multi-task** and **complex online shopping questions**, and **sharp limitations on inference time and GPU memory**.

➢ Our Solution

To tackle the above challenges, we introduce a pipeline containing three parts: **base model selection, pre-trained model quantization, and prompt design**. Our solution across all five tracks adheres to these three steps and demonstrates robust performance.

➢ Stand Out Feature

It is worth noting that there is **no fine-tuning** in our solution, which broadens the usability of our pipeline.

2. Base Model Selection

2.1 Which Base Model?

➢ Following **Chatbot Arena** to select candidate model

➢ Llama, Gemma, Mistral, ChatGLM, Llama3, Qwen, Qwen2…

➢ **Llama3** and **Qwen2** stand out on the **development set**.

Chatbot Arena

Model	Track1	Track2	Track3	Track5
Llama3-70b-awq	0.8013	0.7067	0.7064	0.7545
Qwen2-72b-awq	**0.8166**	**0.7141**	**0.7181**	**0.7685**

2.2 A Small Model or a Quantized Version of Large Models?

➢ **Quantized version of a large model is better.**

Model	Dtype	Temperature	Top_p	Accuracy
Llama3-8B-Instruct	bfloat16	0.0	0.9	0.4824
Llama3-70B-Instruct	AWQ quantization	0.0	0.9	0.7210

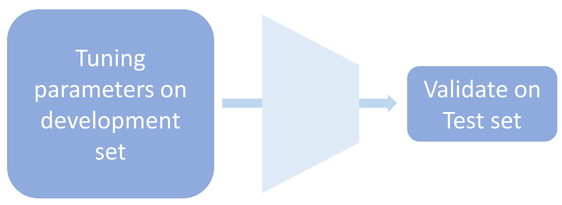
3. Pre-trained Model Quantization

3.1 Our Pipeline

➢ Tuning parameters on development set

➢ Testing on test set

➢ This approach is set to save time



3.2 AWQ Quantization

➢ **Why AWQ?**

➢ Following work [1] using **AWQ quantization on Qwen2**. The paper found that **4 bit quantization** of **Llama3** will only lose **2% performance** comparing to **8B** model and **0.05%** performance comparing to **70B model**.

3.3 Parameter Tuning

➢ Max Data Length: This parameter controls the maximum length of one sample to be considered.

➢ Block Size: Controls the granularity of dividing each block.

➢ n_samples: The number of samples involves in the quantization process.

➢ Dataset Seed

➢ Group Size: Weight metrics will be divided into groups, with each group containing group_size columns.

Different combinations of hyper-parameters have **significantly different performance** on the **development dataset**, with the **best** and **worst** combinations differing by approximately **13%** in accuracy.

Max Data Length	Block_Size	N_Samples	Dataset Seed	Group Size	Accuracy
512	128	128	42	128	0.60864
128	128	128	42	128	0.61321
256	128	128	42	128	0.59657
128	256	128	42	128	**0.67484**
256	256	128	42	128	0.62959
128	512	128	42	128	0.64282
128	256	256	42	128	0.63509

Performance of different parameter combinations on the development set

3. Pre-trained Model Quantization

We select the best-performing set of parameters and experiment on the test set:

Track	Model	Generation	Multi-Choice	NER	Ranking	Retrieval	Overall
1	Qwen2-72b-awq	0.6914	0.8456	0.6764	-	0.8296	0.7970
	Qwen2-72b-awq-s	0.6969	0.8453	0.7096	-	0.8226	0.7985
2	Qwen2-72b-awq	-	0.7619	-	-	0.6165	0.7438
	Qwen2-72b-awq-s	-	0.7540	-	-	0.5977	0.7345
3	Qwen2-72b-awq	0.7259	0.8632	0.8071	-	0.8246	0.8199
	Qwen2-72b-awq-s	0.7236	0.8617	0.8051	-	0.8259	0.8186
5	Qwen2-72b-awq	0.6465	0.7953	0.7449	0.8372	0.8034	0.7657
	Qwen2-72b-awq-s	0.6508	0.7976	0.7474	0.8334	0.8076	0.7685

Better on track 1 and track 5.

3.4 Synthetic Dataset as Quantization Dataset?

We also tried to create Synthetic Dataset as the quantization dataset, but did not improve performance on the development set.

4. Prompt Design

4.1 Instruction

➢ Role definition; Task definition; Tips [2]

An example of the instruction

> You are a very intelligent and helpful online shopping assistant for Amazon who can give reasonable answers or outputs to the online shopping questions. These are your Q&A history, please continue to answer questions concisely like these until the answer or output to the last question is finished. If your answer is excellent, you will get tips.

Prompt	Tips	Multi-choice	Retrieval	Overall
A	0	0.7402	0.5313	0.7141
A	$200	0.7486	0.4624	0.7128
B	**$200**	0.7620	0.6165	0.7438
B	$2000	0.7610	0.5301	0.7322

4.2 Exemplar Selection

➢ **Sub-task Division:**

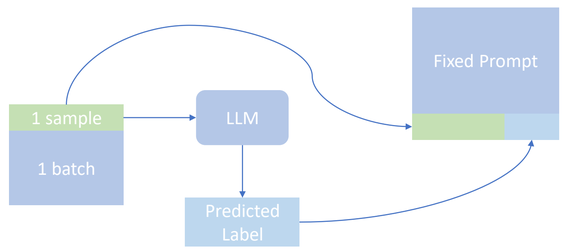
➢ Multi-Choice / Not Multi-Choice

➢ Generation / Multi-Choice / NER / Ranking / Retrieval

➢ **Exemplar:**

➢ Multi-Choice, NER, Ranking -> task-specific exemplars better

➢ Generation, Retrieval -> much general exemplars achieve better performance



➢ **Dynamic Exemplars**

➢ For each batch, we concat the first instance to the fixed prompt

4.3 Hyper-parameters Tuning

➢ Max Context Length: 4096 -> 8192 achieves better performance, but 8192 -> 12288 not.

➢ Max New Tokens: Multi-Choice (1), NER (15), Ranking (15), Retrieval (10), Generation (65) in Track 5.

➢ Temperature: Keeping 0 to ensure reproducibility.

➢ Top_p: Keeping 0.9.

4.4 Output Formatting

➢ Multi-Choice Verification

➢ Whether the output is a valid option, or a default option will be chosen (1)

➢ Non-Multi-Choice Checking

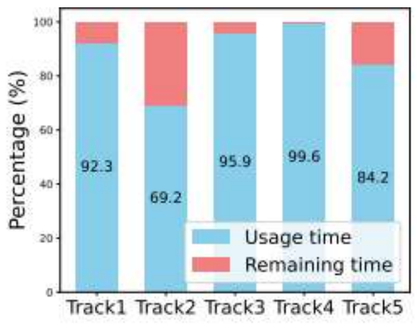
➢ Checking whether the generated text contains any unnecessary information

➢ The word “Question”

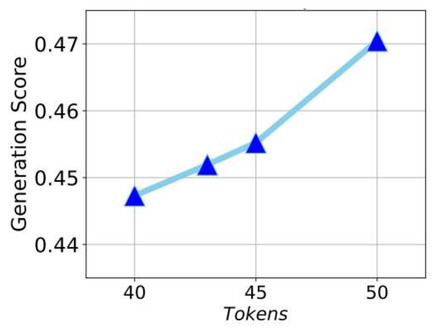
➢ Some unwanted characters

Track	Model	Instruction (Tips)	Sub-task Division	Exemplar	Param Tuning	Generation	Multi-Choice	NER	Ranking	Retreival	Overall
2	Qwen2-72b-awq	No	No	Fixed	No	-	0.7402	-	-	0.5313	0.7141
2	Qwen2-72b-awq	Yes	No	Fixed	No	-	0.7486	-	-	0.4624	0.7128
5	Qwen2-72b-awq	Yes	No	Fixed	No	0.5094	0.7561	0.1236	0.7158	0.6468	0.6763
5	Qwen2-72b-awq	Yes	Multi-Choices	Fixed	No	0.6544	0.7954	0.7430	0.8339	0.8083	0.7679
5	Qwen2-72b-awq-s	Yes	Multi-Choices	Fixed	No	0.6521	0.7977	0.7474	0.8235	0.8045	0.7680
5	Qwen2-72b-awq-s	Yes	All	Fixed	No	0.6509	0.7977	0.7474	0.8334	0.8077	0.7685
5	Qwen2-72b-awq-s	Yes	All	Fixed+Dynamic	No	0.6509	0.8131	0.7474	0.8334	0.8066	0.7776
5	Qwen2-72b-awq-s	Yes	All	Fixed+Dynamic	Yes	0.6627	0.8132	0.8272	0.8360	0.8054	0.7815

5. Efficiency



Time Consuming Statistics



Max New Token Impact on Generation for Track 4

[1] Wei Huang, Xudong Ma, Haotong Qin, Xingyu Zheng, Chengtao Lv, Hong Chen, Jie Luo, Xiaojuan Qi, Xianglong Liu, and Michele Magno. 2024. How good are lowbit quantized llama3 models? an empirical study. arXiv

[2] https://minimaxir.com/2024/02/chatgpt-tips-analysis/