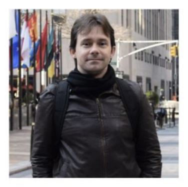
Winning Amazon KDD Cup'24

1st Place - Team NVIDIA


Ahmet Erdem 🇹🇷


Benedikt Schifferrer 🇩🇪


Chris Deotte 🇺🇸

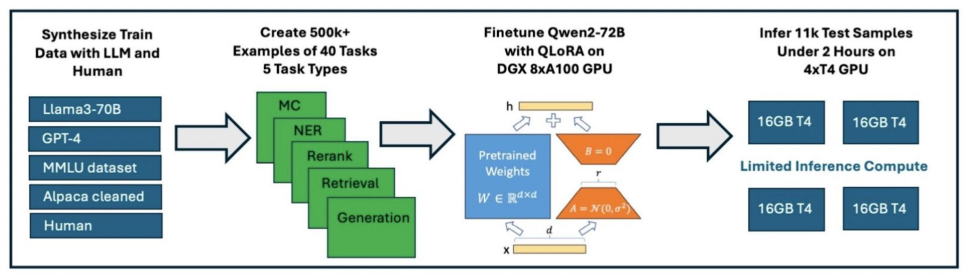

Gilberto Titericz 🇧🇷


Ivan Sorokin 🇫🇮


Simon Jegou 🇫🇷

Solution Summary

Fine-tuning Qwen2-72B



	Track 1	Track 2	Track 3	Track 4	Track 5
Team NVIDIA	**83.3**	**79.1**	**74.6**	**76.1**	**78.8**
2nd place	82.5 (-0.8)	78.4 (-0.7)	73.3 (-1.3)	73.5 (-2.6)	78.2 (-0.6)
3rd place	82.4 (-0.9)	78.1 (-1.0)	72.8 (-1.8)	71.5 (-4.6)	77.3 (-1.5)

amazon KDD Cup 2024

Multi-Task Online Shopping Challenge for LLMs

31,000 $ 10,500

Summary:

- Evaluating Large Language Models as helpful assistance in ecommerce domain
- Test Dataset (ShopBench) contained 20,000 questions covering 57 diverse tasks, representing 5 task types (e.g. Multiple Choice) and organized in 4 tracks
- Code Competition: No access to test dataset and solutions are executed on hosted infrastructure with specific compute and time constraints.
- No Private Test Dataset

Input	The product 'Hanes Men's Beefy-T T-Shirt, Heavyweight Cotton Tee, 1 Or 2 Pack, Big & Tall' appears on an e-commerce website. What type of fabric is used in it? 0. spandex, polyester 1. cotton 2. microfiber 3. It cannot be inferred. Answer:
Answer	1

Challenges

- **No Training Dataset**: Only 96 example questions were shared with the participants
- **Hidden tasks**: The 96 questions represent only 18 of 57 tasks. The model requires to generalize to other tasks
- **Time and compute constraints**: Solutions have to run in a specific timelimit on 4x NVIDIA T4 GPUs with 16GB memory

Track	Time (min)
1	70
2	20
3	30
4	20
5	140

Training Dataset

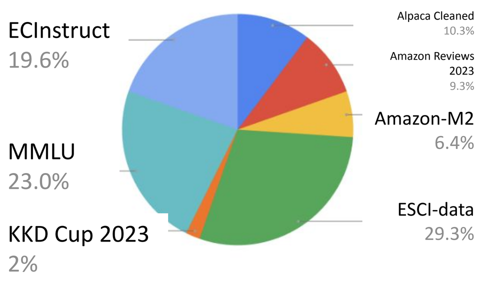
Training Datasets

Input Sources

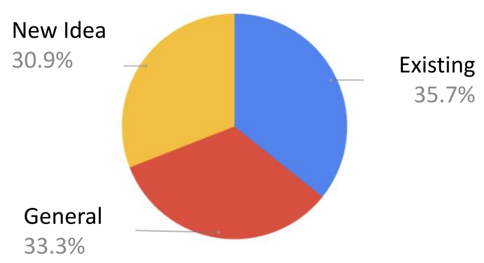
- **Amazon-M2**
  - A multi-lingual Amazon session dataset with rich meta-data used for KDD Cup 2023.
- **Amazon Reviews 2023**
  - A large scale Amazon Review Dataset with rich features and over 500M reviews across 33 categories.
- **NingLab/ECInstruct**
  - Instruction dataset covers 116,528 samples from 10 real and widely performed e-commerce tasks of 4 categories.
- **ESCI-data**
  - Shopping Queries dataset provides a list of up to 40 potentially relevant results, together with ESCI relevance judgements (Exact, Substitute, Complement, Irrelevant) indicating the relevance of the product to the query.
- **MMLU**
  - Massive multitask test consisting of 16k multiple-choice questions
  - and auxiliary 100k multiple-choice training questions from ARC, MC_TEST, OBQA, RACE, etc.
- **Alpaca-Cleaned**
  - Cleaned version of the original Alpaca Dataset released by Stanford.

Training Datasets - 39 Diverse Datasets with total of ~500,000 Samples

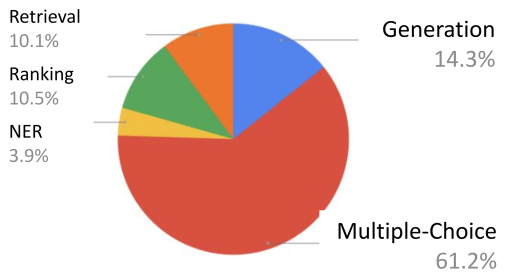
Distribution by Source Dataset



Distribution by Task Cluster



Distribution by Task Type

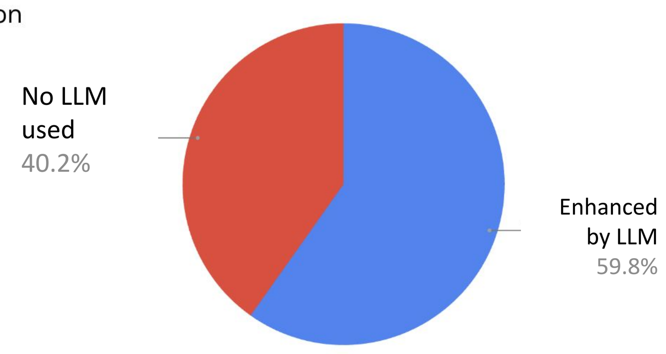


- We build 39 different datasets based on 7 public available datasets as an input, resulting in total 500,000 samples
- Around 30% of the samples were based on own ideas
- Majority of samples were multiple choice questions (61%) followed by generation (14%)

Training Datasets

Synthetic data generation

1) Prompt LLM to construct the task from the multiple seed data
   - (a) combine product attributes, target entity, instruction
   - (b) combine user query, product list, instruction
   - (c) combine question, documents, instruction
2) Enrich the seed data with missing details
   - (a) extract entities from product description
   - (b) identify the product type or category
3) Generate instructions with different wordings
   - (a) replace existing instruction with new wordings



Training Datasets - A Few Own Examples

Example 1	The product 'American Flag Patch, US Military Patches Independence Day Tactical Patch Waterproof Non-Fading Flag Patches for Backpacks Caps Clothes.' is available on an online shopping website. **Which of the following reviews was written for this product:** 0. <Random Review> 1. looks great holding it's color in the hot sun. quality material no rips or frays from windy conditions here. very satisfied. 2. <Random Review> 3. <Random Review> **Output: 1**
Example 2	The product 'Shine Whitening - Zero Peroxide Teeth Whitening System - No Sensitivity' has multiple product reviews. Given the following numbered list of 5 reviews, **please rank the reviews according their helpfulness to a user.** The most helpful review should appear first and the least helpful review should be last. Review List: <List of Review> You should output a permutation of 1 to 5. There should be a comma separating two numbers. Each review and its number should appear only once in the output. Only respond with the ranking results. Do not say any word or explanations. Output: **Output: 2, 1, 4, 5, 3**
Example 3	A user is searching for the product 'ZEN Bundles Zen Pipe Cleaners Hard Bristle, 132 Count (Pack of 3)'. Given the following numbered list of 4 queries, **please rank the queries according their relevance with the product**. Query List: 1. straight bong 2. brown pipe cleaners 3. pipe softy bits 4. chillum pipe You should output a permutation of 1 to 4. There should be a comma separating two numbers. Each query and its number should appear only once in the output. Only respond with the ranking results. Do not say any word or explanations. Output: **Output: 3,2,1,4**

Methods

LLM comparison without fine-tuning

- LLMs without fine-tuning provide great results out of the box
- During Phase 1, we focused on prompt engineering and model selection
- **Qwen2-72B** without fine-tuning would score 9th overall and 4th place on Track 5 at the end of the competition
- At the end of Phase 1, initial experiments demonstrated the potential benefits of fine-tuning

Model	Track 1	Track 2	Track 3	Track 4	Track 5
Bagel-34B-v0.5	0.701	**0.661**	0.634	0.587	0.683
Smaug-72B	0.718		0.656	0.648	0.698
LLaMa3-70B	0.781	0.653	0.666	0.624	0.718
Qwen2-72B	**0.798**	0.641	**0.719**	**0.692**	**0.749**

Wise-ft

- We used wise-ft to deal with the **distribution shift** between our training set and the ShopBench dataset.
- Wise-ft linearly interpolates between the base model and the fine-tuned model
- Wise-ft brought gains **+1.5** of **+1.3** and **+0.8** on Tracks 1, 3 and 5.

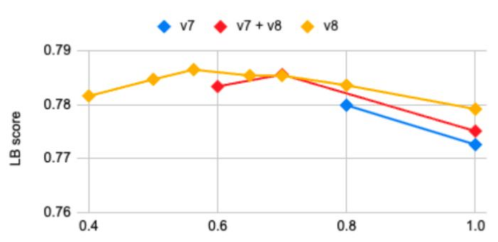
$$W_{wise} = (1-\alpha) * W_{base} + \alpha * W_{ft} \quad + \quad W_{ft} = W_{base} + W_A \cdot W_B \quad \longrightarrow \quad W_{wise} = W_{base} + \alpha * W_A \cdot W_B$$

~ strong on any distribution (zero-shot)

strong on the training set distribution

LoRA fine-tuning

We just need to rescale the LoRA weights by $\sqrt{\alpha}$



Accuracy gain on Track 5 using Wise-ft

Following KDD Cup, this method has been implemented in PEFT:

```
from peft.helpers import rescale_adapter_scale

with rescale_adapter_scale(model, alpha):
    outputs = model(**inputs)
```

Logits processor

- During phase 1, we used logits processors to constrain the LLM generation process
  - For MC: 1 token among [0, 1, 2, 3, 4, 5]
  - For retrieval and ranking: numbers separated by commas
  - For NER: increase the logits of prompt tokens by a constant value
  - For generation: no constraints
- During phase 2, fine-tuning reduced the need for logits processors but we kept them

Fine-Tuning Qwen2

- We fine-tuned Qwen2-72B-Instruct with **QLoRa** using the axolotl library
- Fine-tuning ran on **8x A100** GPUs each with 80 GB GPU memory for 24 hours
- Loss is calculated on the answer tokens using SFT. Hypothesis: more complex methods such as RLHF is not required as answers contain very few tokens
- System prompt contains the task type: `"You are a helpful online shopping assistant. Your task is {task_type}."`. During inference, simple heuristics are used to determine the task type.

Hyperparameter	Value
Optimizer	AdamW
LR Scheduler	cosine
Learning Rate (LR)	0.0002
Weight Decay	0.01
Warm Up Steps	10
Micro Batch Size	1
Gradient Accumulation	4
QLoRa R	64
QLoRa Alpha	32
QLoRa Dropout	0.05
QLoRa Linear	TRUE
Quantization	4-bit

Model	Track 1	Track 2	Track 3	Track 4	Track 5
Qwen2 72B Base Model	0.798	0.641	0.719	0.692	0.749
Qwen2 72B Fine-Tuned	0.816 (+1.8)	0.787 (+14.6)	0.729 (+1.0)	0.758 (+6.6)	0.779 (+5.0)

Iterative fine-tuning

- We fine-tuned our models a second time on slightly different datasets and obtain a boost of **+0.2** to **+0.4**
- This second round of fine-tuning is much faster: 3-8h compared to 24h
- Goal is to explore different dataset blends

	Model	Track 1	Track 2	Track 3	Track 4	Track 5
Iteration 1	Dataset	v8	v7	v8	v7	v8
	Weight	0.56	1	0.56	1	0.56
	LB score	0.831	0.787	0.742	0.758	0.787
Iteration 2	Dataset	v9b	v7b	v9b	v7b	v9b
	Weight	0.75	0.5	0.25	0.5	0.25
	LB score	0.833 (+0.2)	0.791 (+0.4)	0.746 (+0.4)	0.761 (+0.3)	0.788 (+0.1)

Quantization & vLLM

Quantization

- 4xT4 = 64GB of memory → too few for144GB of weights in bfloat16
- We merged the LoRA adapter into Qwen-72B weighted and quantized them to int4 using AWQ → 37GB
- We used the 96 QA pairs for calibration, it took ~1 hour on a single A100 GPU
- GPTQ-Int4 gave very similar results

vLLM

- Before quantization, we padded MLP weights with 128 zeros to allow tensor-parallelism in vLLM on 4 GPUs