Introduction

To solve this hackathon challenge, we conducted six distinct experiments with different variations.

#	SCORE	FILENAME	SUBMISSION DATE	SIZE (BYTES)	STATUS	✔	
1	---	test_gradientgurus2.zip	10/17/2024 16:15:43	1600539	Failed		+
2	0.2275538275	test_gradientgurus3.zip	10/18/2024 03:09:48	1602233	Finished		+
3	0.4111304856	test_gradientgurus4.zip	10/20/2024 09:39:14	1388337	Finished		+
4	0.4135402677	test_gradientgurus5.zip	10/21/2024 09:56:35	1404698	Finished		+
5	0.4111304856	test_gradientgurus4.zip	10/21/2024 10:46:40	30016	Finished		+
6	0.2275538275	test_gradientgurus3.zip	10/21/2024 10:46:44	30607	Finished		+
7	0.4386453234	test_gradientgurus6.zip	10/21/2024 14:28:32	33620	Finished		+
8	0.4414395876	test_gradientgurus7.zip	10/22/2024 05:41:53	38103	Finished		+
9	0.4380875536	test_gradientgurus8.zip	10/23/2024 05:48:39	45737	Finished		+
10	0.4195945068	test_llama_94k_2epoch.zip	10/24/2024 07:22:35	52862	Finished		+
11	5.4152e-06	test_openai_1epoch.zip	10/24/2024 11:47:14	54836	Finished		+
12	0.4432103713	test_openai_1epoch_attempt2.zip	10/24/2024 11:59:13	53602	Finished		+
13	0.4090943552	test_llama_150k_2epoch.zip	10/24/2024 18:19:43	56515	Finished		+
14	---	test_openai_without_price.zip	10/26/2024 03:33:51	56694	Failed		+
15	0.0	test_openai_without_price_attempt2.zip	10/26/2024 03:41:33	60909	Finished		+
16	0.4274195295	test_openai_without_price_attempt3.zip	10/26/2024 03:48:53	61232	Finished		+
17	---	test_openai_with_price_2epoch.zip	10/27/2024 05:58:17	250200	Failed		+
18	0.4527411948	test_openai_with_price_2epoch_attempt2.zip	10/27/2024 11:03:02	68165	Finished		+
19	0.2987804878	test_rag_attempt1.zip	10/27/2024 19:18:00	67704	Finished		+
20	0.4823300697	test_openai_with_price_2epoch_postprocessed.zip	10/28/2024 09:55:14	72527	Finished	✔	+
21	0.4514090456	test_gradientgurus7_postprocessed.zip	10/28/2024 10:30:40	74654	Finished		+

Here, we'll present three of these experiments that showed promising results or hold potential for further exploration to enhance the solution. The experiments are in chronological order.

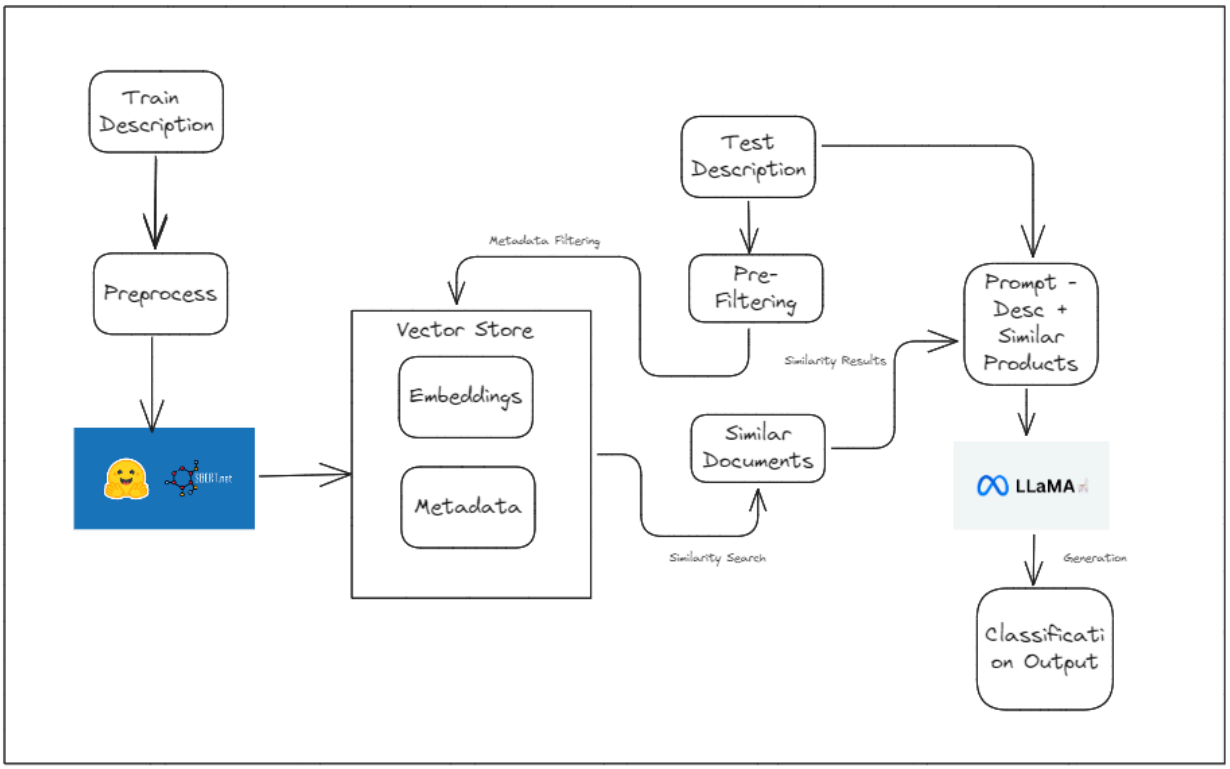
Environments

Following are the three environments used for this competition.

1. Google Colab Pro - A100/T4 GPUs
2. Run Pod A40 GPU - 48GB VRAM ($0.45 per hour)
3. OpenAI - Fine Tuning and Batch API

Experiment - 1 - Retrieval Augmented Generation

Architecture

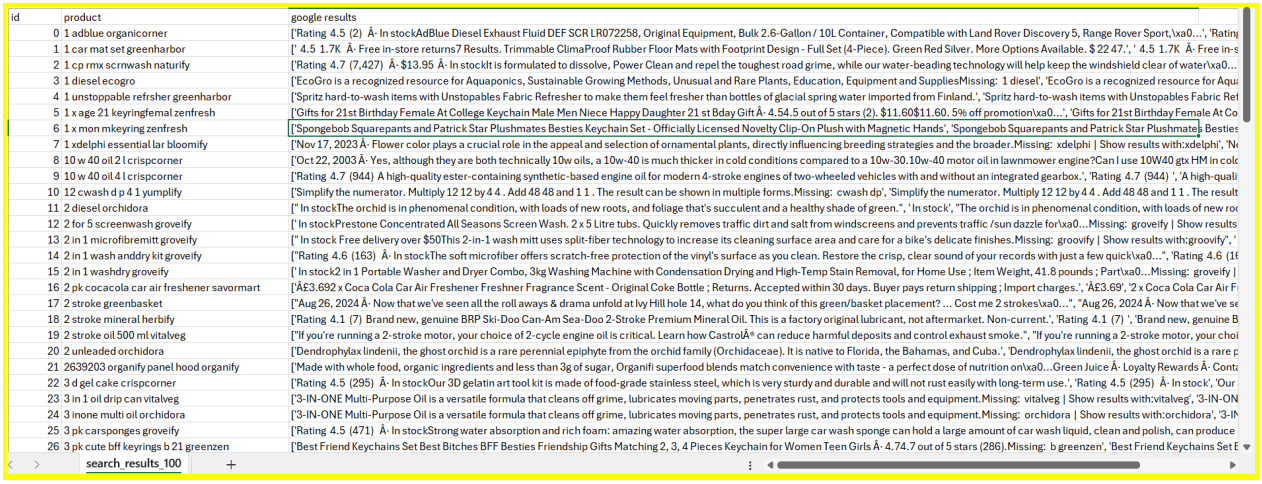


Given an appropriate context, Large Language Models (LLMs) excel at classification tasks.

We found that LLMs could accurately classify products when given more detailed product information. Adding Google search results as context enabled the models to better identify categories such as Supergroup, Group, Module, and Brand.

Better Results with Context:

When providing only similarity search results and the product description to the LLM, it yields moderate accuracy. However, when an elaborated description from Web Search is also included with the similarity results and product description, the LLAMA's accuracy improves significantly.



id	product	google results
0	1 adblue organicorner	['Rating 4.5 (2) Â· In stockAdBlue Diesel Exhaust Fluid DEF SCR LR072258, Original Equipment, Bulk 2.6-Gallon / 10L Container, Compatible with Land Rover Discovery 5, Range Rover Sport,\xa0…', 'Ratin
1	1 car mat set greenharbor	[' 4.5 1.7K Â· Free in-store returns7 Results. Trimmable ClimaProof Rubber Floor Mats with Footprint Design - Full Set (4-Piece). Green Red Silver. More Options Available. $ 22 47. ', ' 4.5 1.7K Â· Free in-s
2	1 cp rmx scrnwash naturify	['Rating 4.7 (7,427) Â· $13.95 Â· In stockIt is formulated to dissolve, Power Clean and repel the toughest road grime, while our water-beading technology will help keep the windshield clear of water\xa0…
3	1 diesel ecogro	['EcoGro is a recognized resource for Aquaponics, Sustainable Growing Methods, Unusual and Rare Plants, Education, Equipment and SuppliesMissing: 1 diesel', 'EcoGro is a recognized resource for Aqua
4	1 unstoppable refrsher greenharbor	['Spritz hard-to-wash items with Unstopables Fabric Refresher to make them feel fresher than bottles of glacial spring water imported from Finland.', 'Spritz hard-to-wash items with Unstopables Fabric Ref
5	1 x age 21 keyringfemal zenfresh	['Gifts for 21st Birthday Female At College Keychain Male Men Niece Happy Daughter 21 st Bday Gift Â· 4.54.5 out of 5 stars (2). $11.60$11.60. 5% off promotion\xa0…', 'Gifts for 21st Birthday Female At Co
6	1 x mon mkeyring zenfresh	['Spongebob Squarepants and Patrick Star Plushmates Besties Keychain Set - Officially Licensed Novelty Clip-On Plush with Magnetic Hands', 'Spongebob Squarepants and Patrick Star Plushmates Besties
7	1 xdelphi essential lar bloomify	['Nov 17, 2023 Â· Flower color plays a crucial role in the appeal and selection of ornamental plants, directly influencing breeding strategies and the broader.Missing: xdelphi | Show results with:xdelphi', 'Nc
8	10 w 40 oil 2 l crispcorner	['Oct 22, 2003 Â· Yes, although they are both technically 10w oils, a 10w-40 is much thicker in cold conditions compared to a 10w-30.10w-40 motor oil in lawnmower engine?Can I use 10W40 gtx HM in cold
9	10 w 40 oil 4 l crispcorner	['Rating 4.7 (944) A high-quality ester-containing synthetic-based engine oil for modern 4-stroke engines of two-wheeled vehicles with and without an integrated gearbox.', 'Rating 4.7 (944) ', 'A high-quali
10	12 cwash d p 4 1 yumplify	['Simplify the numerator. Multiply 12 12 by 4 4 . Add 48 48 and 1 1 . The result can be shown in multiple forms.Missing: cwash dp', 'Simplify the numerator. Multiply 12 12 by 4 4 . Add 48 48 and 1 1 . The result
11	2 diesel orchidora	[" In stockThe orchid is in phenomenal condition, with loads of new roots, and foliage that's succulent and a healthy shade of green.", ' In stock', "The orchid is in phenomenal condition, with loads of new roo
12	2 for 5 screenwash groveify	[' In stockPrestone Concentrated All Seasons Screen Wash. 2 x 5 Litre tubs. Quickly removes traffic dirt and salt from windscreens and prevents traffic /sun dazzle for\xa0…Missing: groveify | Show results
13	2 in 1 microfibremitt groveify	[" In stock Free delivery over $50This 2-in-1 wash mitt uses split-fiber technology to increase its cleaning surface area and care for a bike's delicate finishes.Missing: groovify | Show results with:groovify", '
14	2 in 1 wash anddry kit groveify	["Rating 4.6 (163) Â· In stockThe soft microfiber offers scratch-free protection of the vinyl's surface as you clean. Restore the crisp, clear sound of your records with just a few quick\xa0…", 'Rating 4.6 (16
15	2 in 1 washdry groveify	[' In stock2 in 1 Portable Washer and Dryer Combo, 3kg Washing Machine with Condensation Drying and High-Temp Stain Removal, for Home Use ; Item Weight, 41.8 pounds ; Part\xa0…Missing: groveify |
16	2 pk cocacola car air freshener savormart	['Â£3.692 x Coca Cola Car Air Freshener Freshner Fragrance Scent - Original Coke Bottle ; Returns. Accepted within 30 days. Buyer pays return shipping ; Import charges.', 'Â£3.69', '2 x Coca Cola Car Air F
17	2 stroke greenbasket	["Aug 26, 2024 Â· Now that we've seen all the roll aways & drama unfold at Ivy Hill hole 14, what do you think of this green/basket placement? ... Cost me 2 strokes\xa0…", "Aug 26, 2024 Â· Now that we've se
18	2 stroke mineral herbify	['Rating 4.1 (7) Brand new, genuine BRP Ski-Doo Can-Am Sea-Doo 2-Stroke Premium Mineral Oil. This is a factory original lubricant, not aftermarket. Non-current.', 'Rating 4.1 (7) ', 'Brand new, genuine B
19	2 stroke oil 500 ml vitalveg	["If you're running a 2-stroke motor, your choice of 2-cycle engine oil is critical. Learn how CastrolÂ® can reduce harmful deposits and control exhaust smoke.", "If you're running a 2-stroke motor, your choi
20	2 unleaded orchidora	['Dendrophylax lindenii, the ghost orchid is a rare perennial epiphyte from the orchid family (Orchidaceae). It is native to Florida, the Bahamas, and Cuba.', 'Dendrophylax lindenii, the ghost orchid is a rare p
21	2639203 organify panel hood organify	['Made with whole food, organic ingredients and less than 3g of sugar, Organifi superfood blends match convenience with taste - a perfect dose of nutrition on\xa0…Green Juice Â· Loyalty Rewards Â· Conta
22	3 d gel cake crispcorner	['Rating 4.5 (295) Â· In stockOur 3D gelatin art tool kit is made of food-grade stainless steel, which is very sturdy and durable and will not rust easily with long-term use.', 'Rating 4.5 (295) Â· In stock', 'Our
23	3 in 1 oil drip can vitalveg	['3-IN-ONE Multi-Purpose Oil is a versatile formula that cleans off grime, lubricates moving parts, penetrates rust, and protects tools and equipment.Missing: vitalveg | Show results with:vitalveg', '3-IN-ON
24	3 inone multi oil orchidora	['3-IN-ONE Multi-Purpose Oil is a versatile formula that cleans off grime, lubricates moving parts, penetrates rust, and protects tools and equipment.Missing: orchidora | Show results with:orchidora', '3-IN
25	3 pk carsponges groveify	['Rating 4.5 (471) Â· In stockStrong water absorption and rich foam: amazing water absorption, the super large car wash sponge can hold a large amount of car wash liquid, clean and polish, can produce
26	3 pk cute bff keyrings b 21 greenzen	['Best Friend Keychains Set Best Bitches BFF Besties Friendship Gifts Matching 2, 3, 4 Pieces Keychain for Women Teen Girls Â· 4.74.7 out of 5 stars (286).Missing: b greenzen', 'Best Friend Keychains Set B

search_results_100

However, cost became a limiting factor. Search APIs providing extensive information are not free, and conducting 180k searches with Serp APIs significantly raised the project's expenses. To keep the hackathon project within a $100 budget, we opted for a cost-effective alternative.

Instead of incorporating additional data as context, we maximized the use of available data. Our approach involved similarity search to identify the top-K similar products based on their descriptions, with Retailer and Price data used as filters.

Our process involved the following steps:

1. Indexing all training descriptions in FAISS.
2. Adding metadata (price and retailer information) to a CSV file.
3. Using Sentence Transformer (sentence-transformers/all-mpnet-base-v2) to retrieve the top 5 similar products.
4. Creating prompts based on queries and similar products, then using Llama (meta-llama/Llama-3.1-8B-Instruct) to determine the product class.

Our choices of Llama, Sentence Transformers, and FAISS were guided by budget considerations.

Inference:

To manage GPU costs, we implemented inference in a two-step process:

1. **Prompt Generation** - This was handled on a CPU machine, where we generated 180k prompts and saved them in JSON format.
2. **Classification using LLM** - We employed Llama with vLLM for the classification task.

Cost Estimation

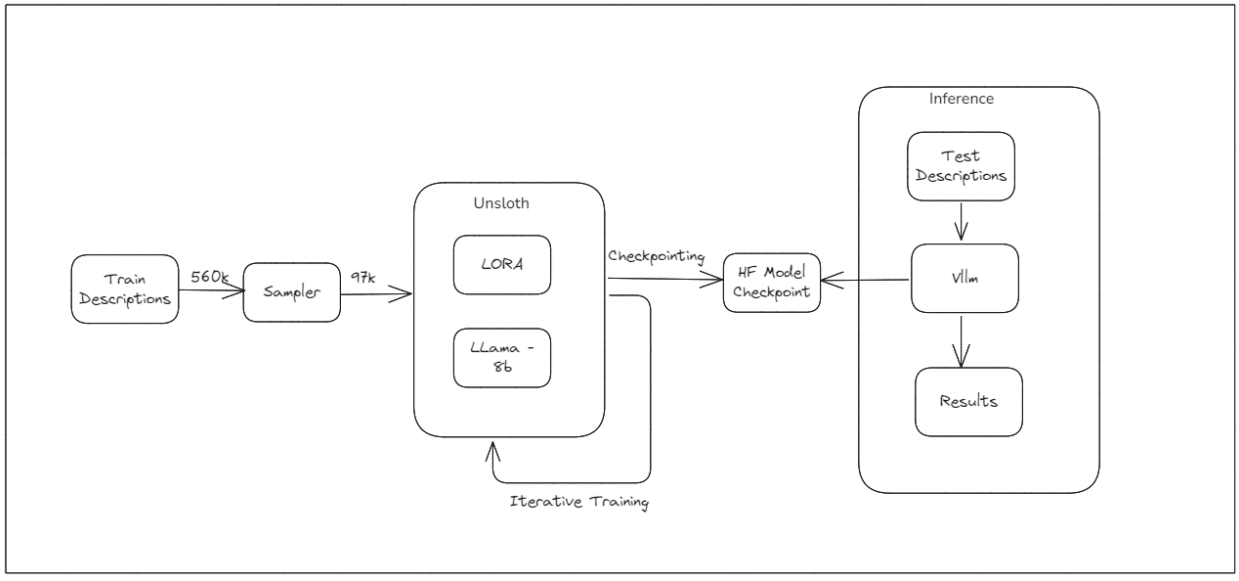
- **Indexing into Vector Index** - Performed on Google Colab - **$0 (No cost)**
- **vLLM Inference for Test Data** - Approximately 4 hours - **~$1.8**

Future Scope

1. Experimenting with different values of *K* (number of matching documents).
2. Enhancing context with additional information from the Serp API.
3. Implementing a re-ranking mechanism for improved results.
4. Refining the filtering process for greater accuracy.
5. Using more advanced embeddings for similarity search.
6. Exploring more powerful Large Language Models, such as GPT-4 or Gemini Pro.

Experiment - 2 - Fine Tuning Open Source LLMs

Architecture



Data Preparation

Due to limited GPU capacity, we couldn't fine-tune the model on the entire dataset. Instead, we used stratified sampling to create a subset of 100k rows (~20% of the data).

For stratified sampling, we combined all classification labels (Supergroup, Group, Module, and Brand) into a single label. This approach aimed to maintain the original class proportions in the training data, ensuring that all classes—including those with only one sample—were represented in the training set.

For training the Llama model, we utilized a standard JSON format. All input fields—such as *description*, *retailer*, and *price*—were included in the training data.

To maintain consistency, the model's output was also set to JSON format, with keys for *supergroup*, *group*, *module*, and *brand*. Below is a sample of the input and expected output:

Sample Input:

```
{"description": "baby disposable bed mats 10 s", "retailer": "herbgrove",
"price": 2.99}
```

Sample Output:

```
{"supergroup": "baby care", "group": "baby care detail unknown total",
"module": "baby care", "brand": "pure baby"}'}
```

Model Parameters

For the Llama model training, the following hyperparameters were used:

- **Model Configuration**
  - `model_name`: `"unsloth/llama-3-8b-Instruct-bnb-4bit"`

    - `max_seq_length`: 512
    - `dtype`: `torch.bfloat16`
    - `load_in_4bit`: `True`
    - `use_gradient_checkpointing`: `"unsloth"`
    - `attn_implementation`: `"flash_attention_2"`

- **LoRA (Low-Rank Adaptation) Parameters**
    - `r`: 16 (Suggested values: 8, 16, 32, 64, 128)
    - `target_modules`: `["q_proj", "k_proj", "v_proj", "o_proj", "gate_proj", "up_proj", "down_proj"]`
    - `lora_alpha`: 16
    - `lora_dropout`: 0 (Optimized for this value)
- **Training Parameters**
    - `per_device_train_batch_size`: 16
    - `gradient_accumulation_steps`: 4
    - `warmup_steps`: 5
    - `num_train_epochs`: 3
    - `learning_rate`: 2e-4

The model was trained for 3 epochs, with validation accuracy measured using random data points. Unsloth offers an integration to merge the LoRA adapters and save the complete model directly to Hugging Face.

Inference:

For inference, the fine-tuned Llama model was deployed on Runpod's A40 GPU.

We tested inference with both Text Generation Inference (TGI) and vLLM. Using a batch size of 128, TGI inference took around 1.5 hours to process all 180k+ samples.

In contrast, vLLM inference was significantly faster. Thanks to continuous batching support, the total inference time was reduced to approximately 30 minutes, which also contributed to cost savings.

Cost Estimation:

The training process took 1.5 hours per epoch, resulting in a total training time of 4.5 hours for 100k samples, with a cost of $2.25.

For inference:

- Using TGI cost $0.675.
- Using vLLM cost $0.225.

Minimum Requirements:

- To fully reproduce the model, a GPU with at least 48 GB of VRAM is needed.
- For testing on a lower GPU, a minimum of 16GB of VRAM is required, with a batch size of 11.

Model Profiling Report

Training Performance

- **Self CPU Time**: 737.47 ms
- **Self CUDA Time**: 526.07 ms
- **GFLOPs**: 3845.88

The training phase achieved a high GFLOPs count, indicating efficient GPU utilization and substantial compute power.

Inference Performance

- **Inference Time**: 1.78 seconds
- **GFLOPs**: 1702.93

Inference was completed swiftly with reduced GFLOPs, showing the model's efficiency for real-time tasks.

To click here to view profiling result colabfile.

Other Experiments:

Several variations of these experiments were conducted:

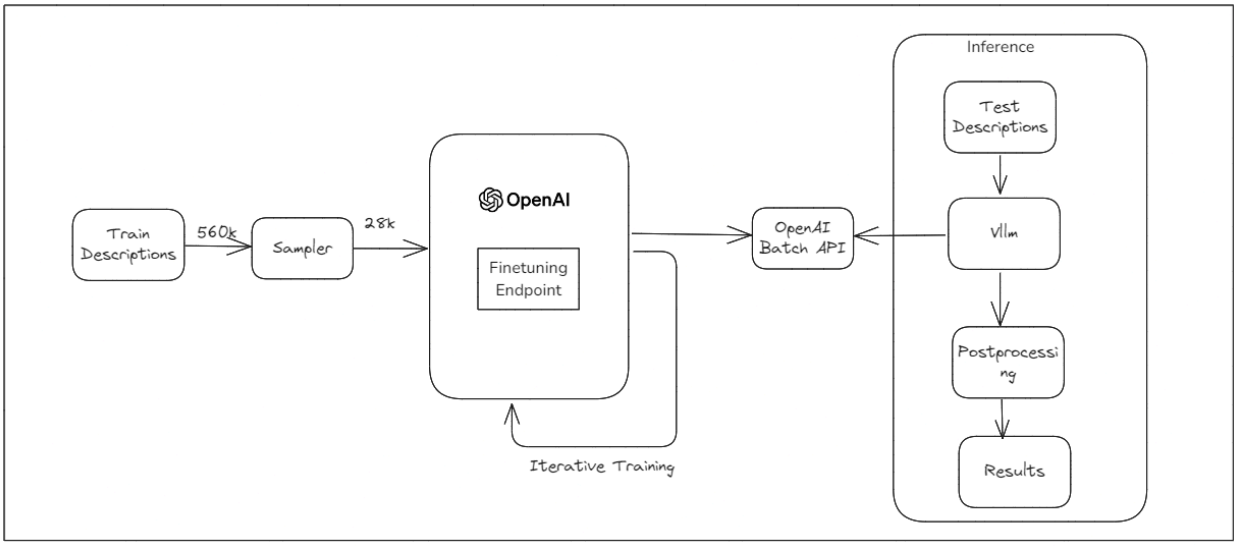
1. **Fine-tuning with the phi-3 model on the entire dataset** - Results indicated that smaller models did not perform as well as Llama-8b.
2. **Fine-tuning with 150k samples** - This experiment examined the impact of removing numbers/prices on overall accuracy. The results confirmed that number representations are essential for model stability.
3. **Grouping the entire training dataset by retailer to create a reduced representation** - This approach provided an option to pretrain and fine-tune the entire dataset effectively.

Experiment - 3 - Finetuning GPT4o-mini

In our previous experiments, we observed that models with a higher number of parameters delivered better performance. We aimed to fine-tune larger models like Llama-70b or Qwen-14b, but the cost of fine-tuning these models on a large dataset presented budget challenges.

OpenAI offers the opportunity to train the GPT-4o-mini model at no cost (up to 2,000,000 tokens). To take advantage of this, we chose to fine-tune the GPT-4o-mini model using a reduced version of our dataset.

Architecture



Data Preparation

Since GPT-4o-mini is a chat model, the dataset needed to be structured in a chat-like format.

To optimize the total number of tokens, we employed the following format for the input data:

```
description | retailer | price
```

Additionally, we combined the output labels into a single label formatted as:

```
supergroup__group__module__brand
```

This approach helped streamline the dataset while maintaining clarity in the training process

Example:

```
{
"messages":
 [ {"role": "system", "content": "Your task is to classify the user text"},
{"role": "user", "content": "baby disposable bed mats 10 s | herbgrove |
2.99 ->"},
{"role": "assistant", "content": "baby care__baby care detail unknown
total__baby care__pure baby"}
 ] }
```

When training GPT models, it's essential to be mindful of the volume of data sent to the training process, as charges apply to both input and output tokens.

To manage our budget effectively, we aimed to minimize costs since OpenAI charges the same rate for inferring the fine-tuned model as for the base model. Given the substantial test volume of over 180k samples, the inference costs could accumulate quickly. Therefore, we focused on training the model at no cost to stay within our budget constraints.

We selected **only 28,000 samples,** representing approximately 5% of the overall dataset, and were able to achieve the best accuracy compared to all other models.

Inference

The inference cost for GPT-4o-mini is as follows:

- **Input**: $0.150 per 1M input tokens
- **Output**: $0.600 per 1M output tokens

To reduce the overall inference cost, we utilized OpenAI's Batch API module, which offers a 50% cost reduction. As a result, we were able to run the entire inference for less than $3.

Post Processing

Upon analyzing the results, we noticed inconsistencies in the model output; some inference results were in different cases, and a few were missing the separator (__).

To address these issues, we implemented a post-processing step aimed at correcting such rows using the similarity search module.

We employed Sentence Transformers to refine the classifications for supergroup, group, module, and brand. This enhancement ultimately improved our overall results by 3-4%.

Cost Estimation

Training GPT-4o-mini - Free of cost (<2M tokens).

Estimated training cost until October 31 - $6 ($3 per million tokens)
Estimated training cost after October 31 - $0.6 ($0.3 per million tokens)

After October 31, 2024, the cost for training entire training dataset (540k samples) - $10-$15

Estimated inference cost - ~$3

Future Scope

The model can be trained using the entire dataset (20 million tokens), which could lead to improved overall accuracy, as the current model has not been exposed to the full dataset.

Additionally, the inference cost will remain manageable, as we will continue to leverage the Batch API for efficiency.

Reference

All the code are submitted in the following link - here