Demo Session: Price Discovery

Overview

- **What has been completed**

- **Current Progress**

- **Next Steps**

Architecture

Optimization

Problem: R2 score of all the regression models built on Amazon products (~10000 data points) is lower than 0.6

Optimization:

- Try big dataset

  Stylish product dataset on Kaggle (~56000 data points)

- Using large CLIP model
 Switch from *vit-basic* model to *vit-large* model

Result

	Vit_basic (1024 features, 1 hour)			Vit_large (1536 features, 56 hours)		
	Xgboost	Random Forest	LightGBM	Xgboost	Random Forest	**LightGBM**
R2	0.66	0.15	0.73	0.69	0.72	**0.74**
MAPE				0.4	0.33	**0.27**

MAPE: Mean absolute percentage error

Price Range

upper =
LGBMRegressor(objective =
'quantile', alpha = **0.95**)

upper2 =
LGBMRegressor(objective
= 'quantile', alpha = **0.80**)
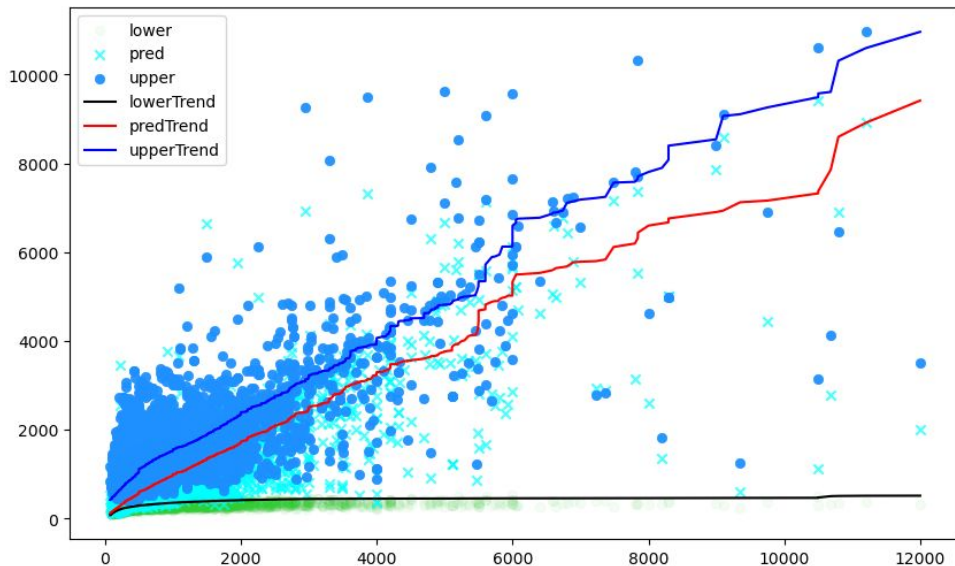
$ 762.8 95% $1238 5% $294

(50%) 80% $1001 20% $539
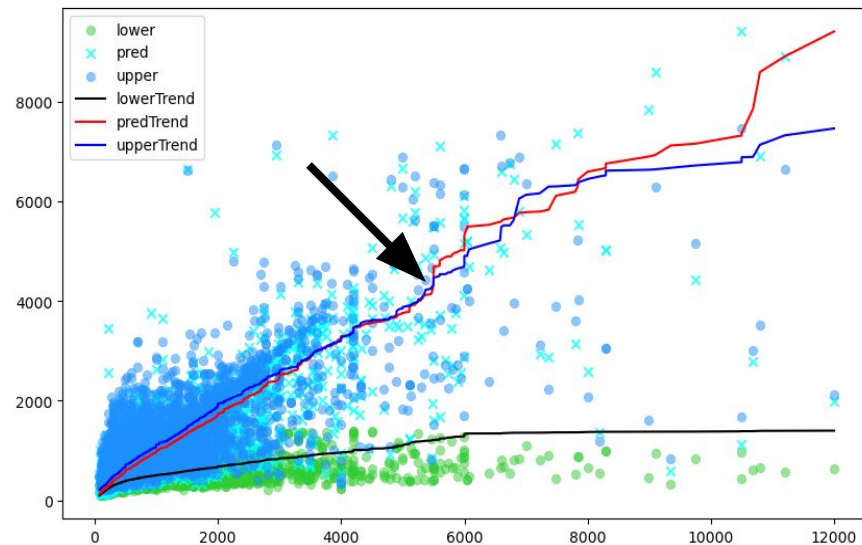
Drawback

- It took a long time to extract image information from products.

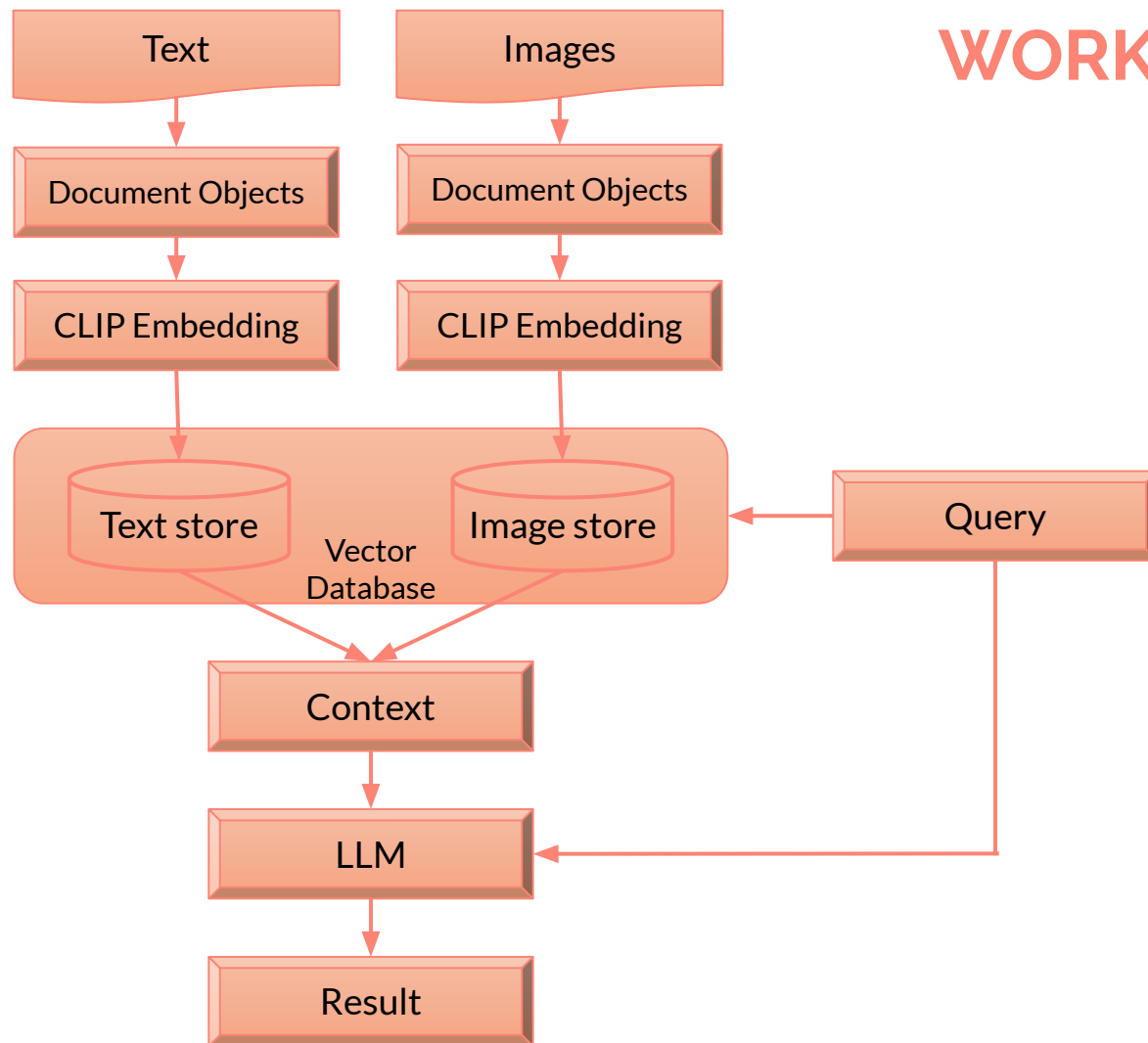 Using GPU

- If you want to get the price range in 20 to 80 percentile, upper price became lower than predict price when price is higher than $6000.

  Try RAG (retrieval-augmented generation)

- Performance is not decent.

Steps Taken for implementing RAG

- Created the text and images database
- Stored 100000 instances in both text and images
- Passing the product title/ description to the text database and getting top 5 similar results
- Passing the image to the image database and getting top 5 similar results
- Consolidating the results obtained from the two databases and adding the same in the prompt
- Passing the prompt to the LLM
- LLM gives the result in the form of product range

Components for Iteration 1

Vector Database -   ChromaDB

Text Embeddings: all-mpnet-base-v2

Image Embeddings: Open-clip / Clip
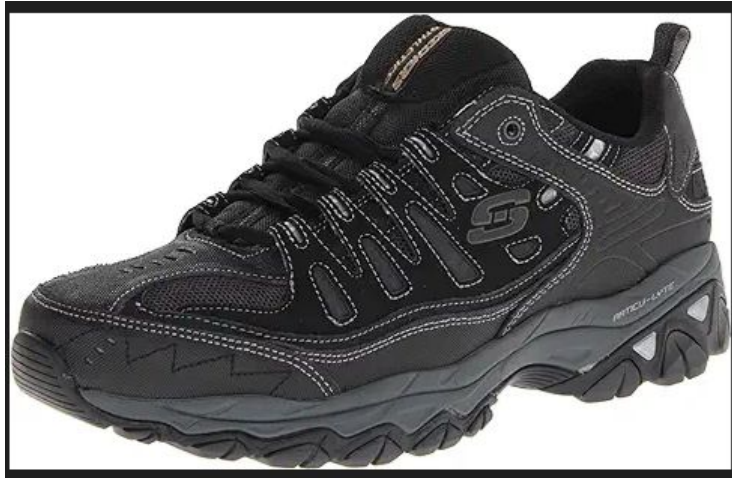
Framework for passing prompt: Langchain

LLM: Microsoft Phi-2

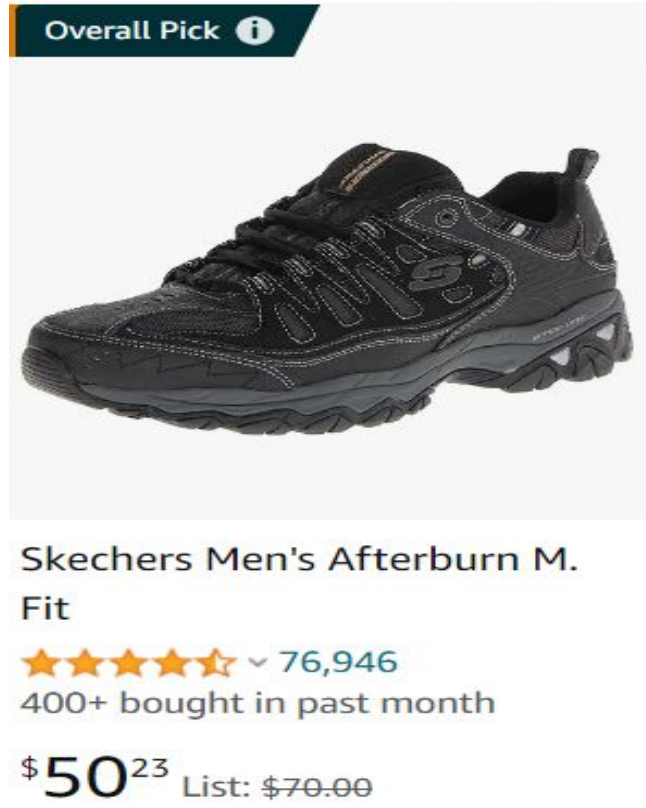# Results of Iteration 1

Passed Text: Skechers Men's Afterburn M. Fit

Passed Image: Price Range given:

$99.50-$249.00

Actual product on Amazon

Skechers Men's Afterburn M. Fit

★★★★⯪ ∨ 76,946

400+ bought in past month

$50²³ List: $70.00

Components for Iteration 2

Vector Database -   ChromaDB

Text Embeddings: Open-Clip / Clip

Image Embeddings: Open-clip / Clip

Framework for passing prompt: Langchain

LLM: Microsoft Phi-2

Results of Iteration 2

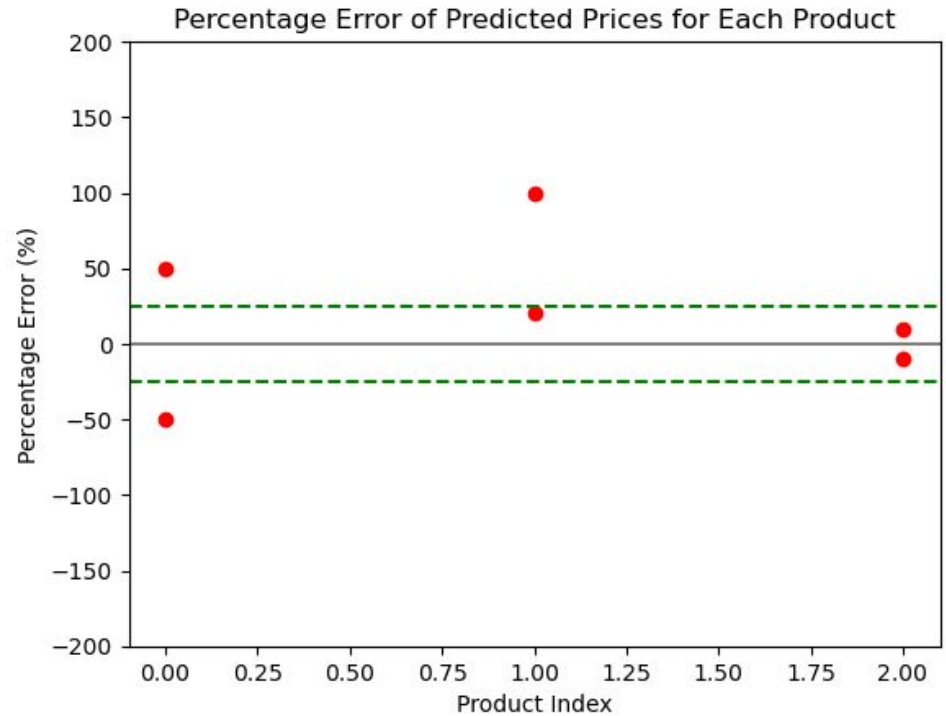Passed Text: Skechers Men's Afterburn M. Fit

Passed Image:

Price Range given:

$60-$80

Evaluation

Original price	Predicted price range
$20	$10 - $30
$50	$60 - $100
$100	$90 - $110

Percentage Error of Predicted Prices for Each Product
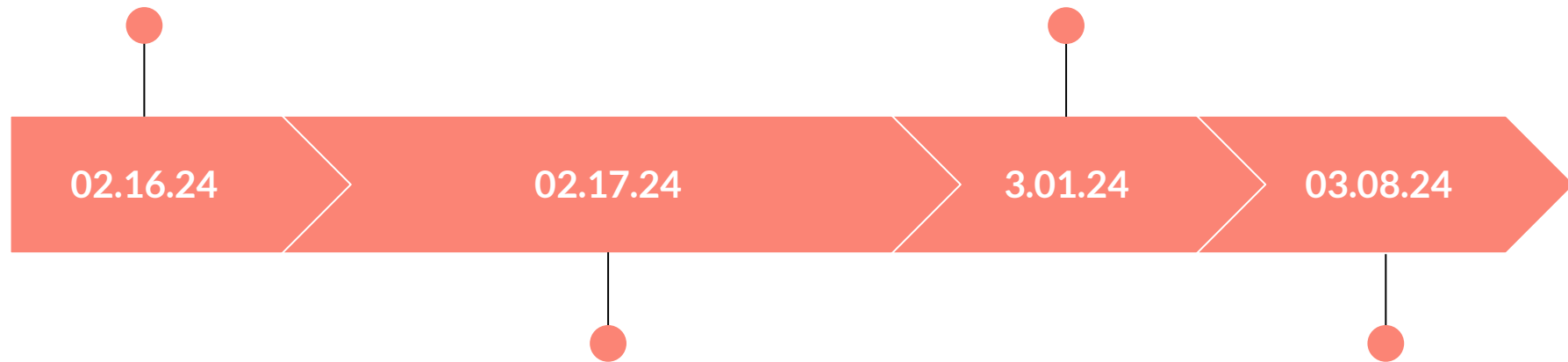
Plan for improvement

- Improving the prompts.

- Migrating the Vector Database from local machine to Cloud.

- Improving the data by exploring different embeddings from different multimodal llms or looking into Knowledge graph database.

- Enabling / Fine tuning the LLM to be aware of events like Black Friday sale, etc. eg. using agents to search on current market situation.

- Improving on the decided evaluation metric that makes sure the range given is suitable and that the price is in the range.

- Experimenting with different LLM's.

Next Steps

Questions?