

# Find Me the **Better** Product!



Jenson Chang  
Chukwunonso Ebele-Muolokwu  
Catherine Meng  
Jingyuan Wang



## The Team



**Varada Kolhatkar (Mentor)**

Associate Professor of Teaching  
University of British Columbia



**Jenson Chang**

Master of Data Science  
University of British Columbia



**Chukwunonso  
Ebele-Muolokwu**

Master of Data Science  
University of British Columbia



**Catherine Meng**

Master of Data Science  
University of British Columbia



**Jingyuan Wang**

Master of Data Science  
University of British Columbia



## Who is our Capstone Partner?

- Founded in 2022 by two UBC graduates



- Currently focused on personal finance products that deliver rewards to users





## Partner's Needs

- Interested in expanding into e-commerce products

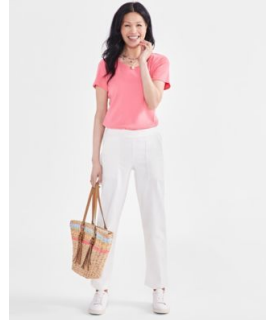


Style & Co Women's High Rise Cropped Pull-On Leggings, Created for Macy's - New Uniform Blu

Price: \$29.99





Liverpool Los Angeles Womens Frayed Tweed Jacket Cap Sleeve Shirt Stride Cropped Wide Leg Jeans




Style & Co Women's Mid-Rise Pull-On Dobby Straight-Leg Jeans, Created for Macy's - Natural




## Partner's Current Search Approach

CREDIT CARDS



**BMO CashBack® Mastercard® for Students**  
5% cash back for the first 3 months, up to \$2,500 in spend



**BMO AIR MILES® Mastercard®\* for Students**  
Get 800 AIR MILES Bonus Miles



## Limitation of Current Approach



- Keyword based searching only matches **exact words**, not the **meanings** and struggles with **paraphrased** user search queries.



## Pain Points

- User search queries are very **hard to predict** and are very **diverse**.

### Basic Queries

- Iphone
- Laptop
- Shoes

### Attribute Queries

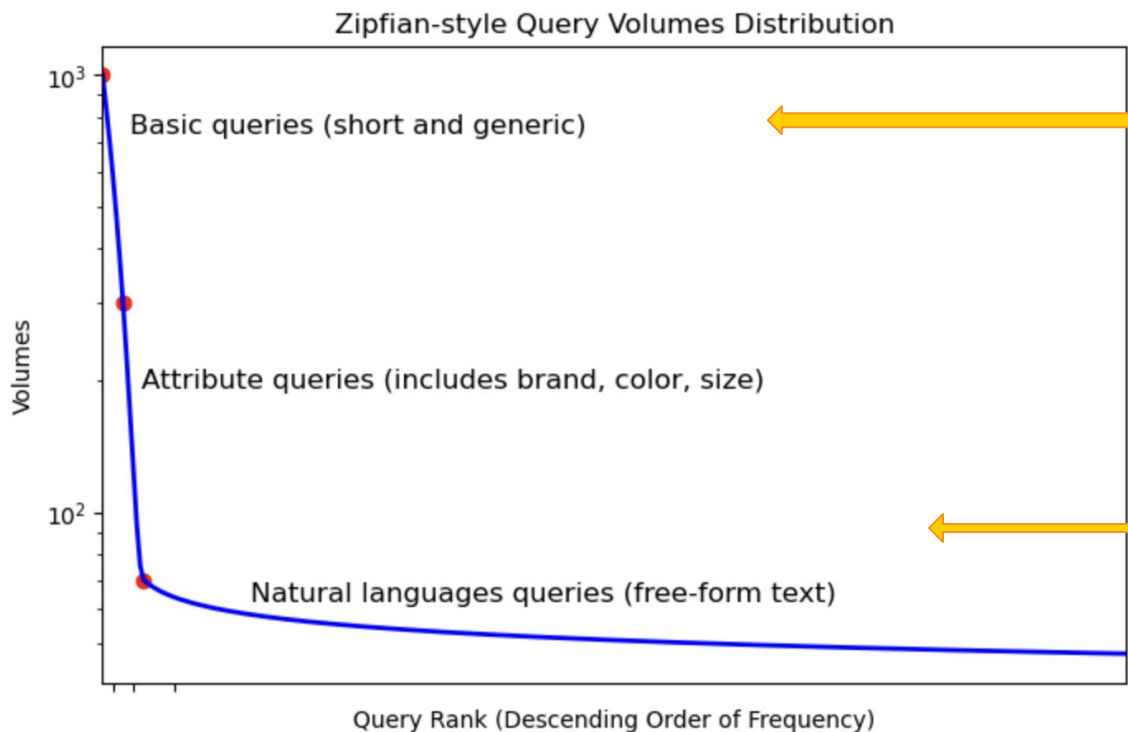
- Iphone 16 in white
- Laptop 15 inches
- Nike Shoes Size 9

### Natural Language Queries

- **Iphone** 16 in white with student discount
- Laptop with 1 year warranty
- Nike Shoes Size 9 in wet condition for women



## Most of the Query Diversity Lies in the Long Tail



Common and very frequent

👍 existing approach (e.g. TFIDF)

Rare and much less frequent

👎 existing approach (e.g. TFIDF)





## Project Goals

- Build a **fast** and **scalable multimodal** search engine that lets users search using text or images to find the **most relevant** products.



### Must Have

- Support for nature language queries
- Fast response time (< 5s)
- Reusable API endpoints



### Should Have

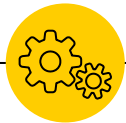
- Reproducible data pipeline



### Nice to Have

- Web interface
- Use larger dataset
- Evaluation plan

# Data & Pipeline





## Product Images and Metadata

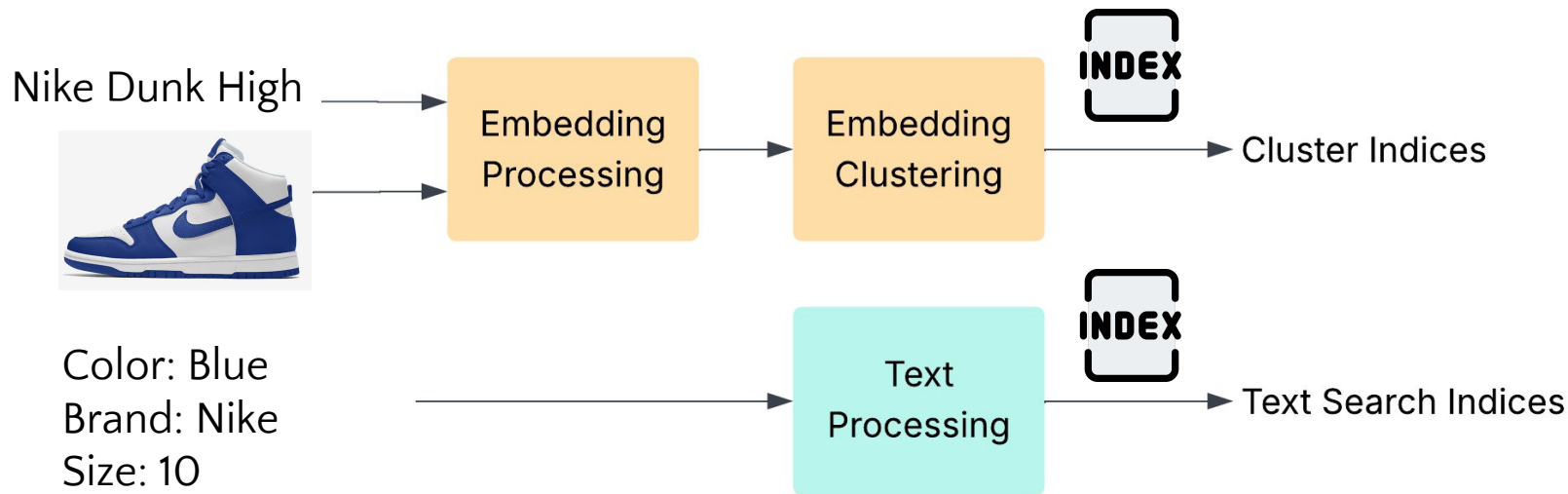
- 1 million low resolution jpeg images of products
- Text file containing product name and product attributes



Field	Example Value
Pid	127.2.DFF8DD86A0648144.3F8EC3F904474916.194145454245
Name	Alfani Men's Mercerized Polo Shirt, Created for Macy's - Neo Navy
Description	Land a preppy look with this long-sleeve polo shirt from Alfani.

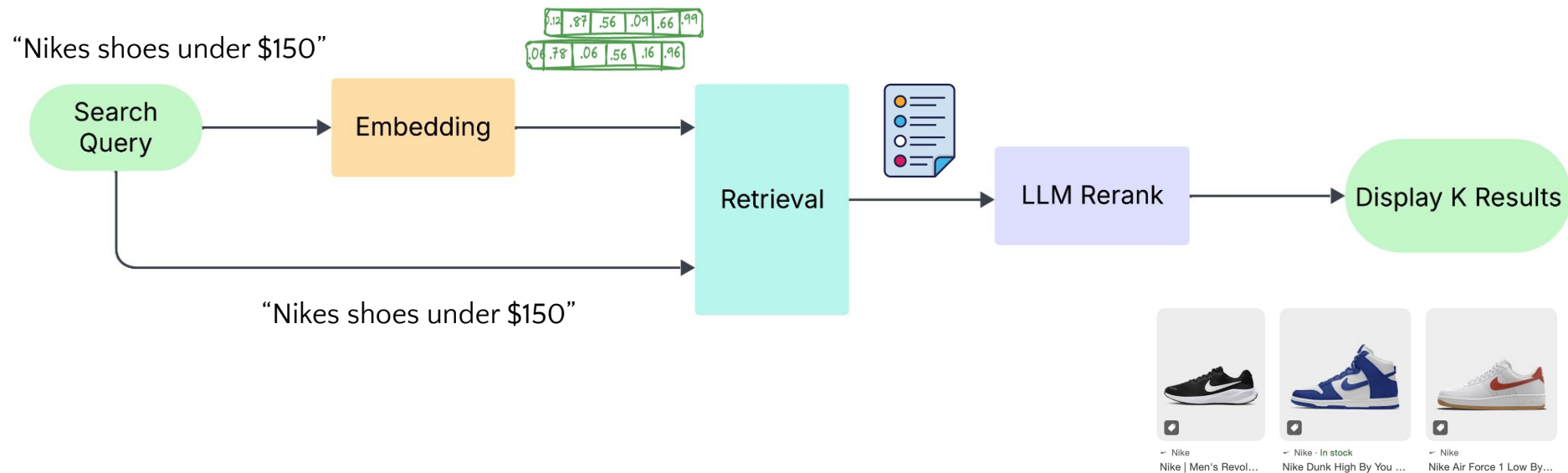


## Indexing 1M Products for Search



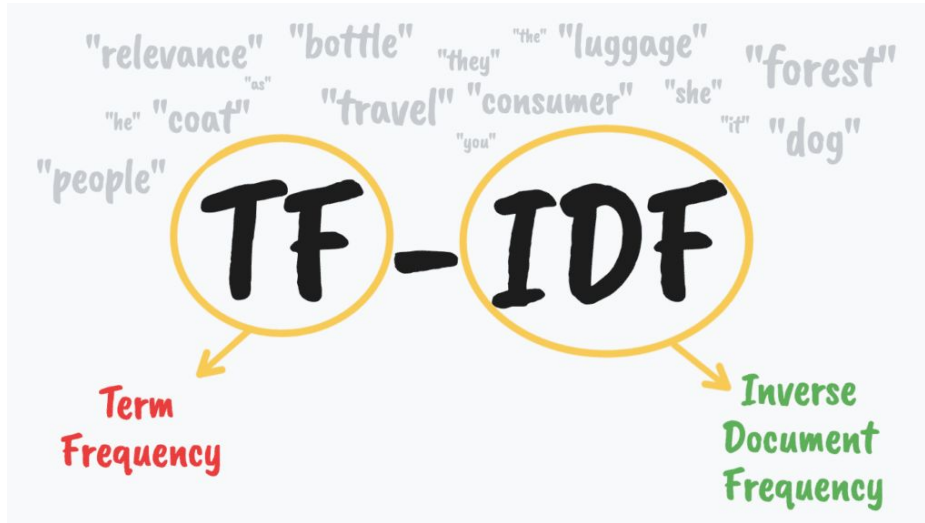


# Search Using Text and Embeddings





## Text Search as Baseline

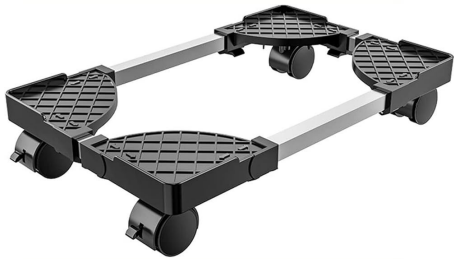


- TF-IDF is applied to product name and metadata
- Products are ranked
  - a. Term weight
  - b. Term frequency
  - c. Length of text



## False Positives with Text Search

Query: “*Desk for office use*”



TEMU Adjustable Stand Locking Casters - Metal, Ventilation & Cooling For And Office Desk Use, In /, For



TEMU Jmhud 5-layer Paper Organizer With Handle, Tray Mesh Desktop File Organizer, Paper Sorter. Desktop Organizer For Office



TEMU Aluminum Alloy Rotatable Cell Phone And Tablet Stand - Adjustable Desk And Bedside Mobile Holder, Universal



## Improve Pipeline with Embeddings

### Search Query

“Desk for office use”



CLIP +  
MiniLM  
Embed

.06 .78 .06 .56 .16 .96

Retrieve top 20 most  
relevant products



### Database of Products



CLIP +  
MiniLM  
Embed

.06 .78 .06 .56 .16 .96  
.012 .87 .56 .09 .66 .99  
.012 .87 .56 .09 .66 .99  
.012 .87 .56 .09 .66 .99





## Improved But Needs Further Optimization

Query: “*Desk for office use*”



**Gouun L-Shaped Office Desk with Storage Drawers and Keyboard Tray - Black**



**TEMU Adjustable Stand Locking Casters - Metal, Ventilation & Cooling For And Office**



**Tribesigns 78.7-Inch Executive Desk, Large Computer Office Desk Workstation, Modern**



## LLM to Rerank Results by Relevance

1.



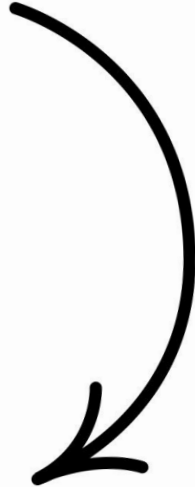
2.



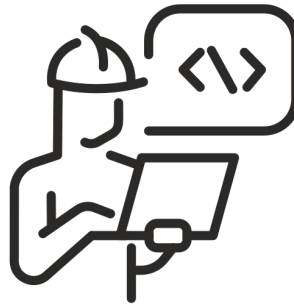
3.



4.



1. **Semantic similarity** to the query intent
2. Direct **keyword** matches
3. **Brand** name mentions
4. **Price** comparison





## Relevant Results are Ranked Higher

Query: “*Desk for office use*”



**Gouun L-Shaped Office Desk with Storage Drawers and Keyboard Tray - Black**



**Tribesigns 78.7-Inch Executive Desk, Large Computer Office Desk Workstation, Modern**



**Tribesigns 78" L-Shaped Executive Desk, Large Office Desk with Drawers and Lateral**

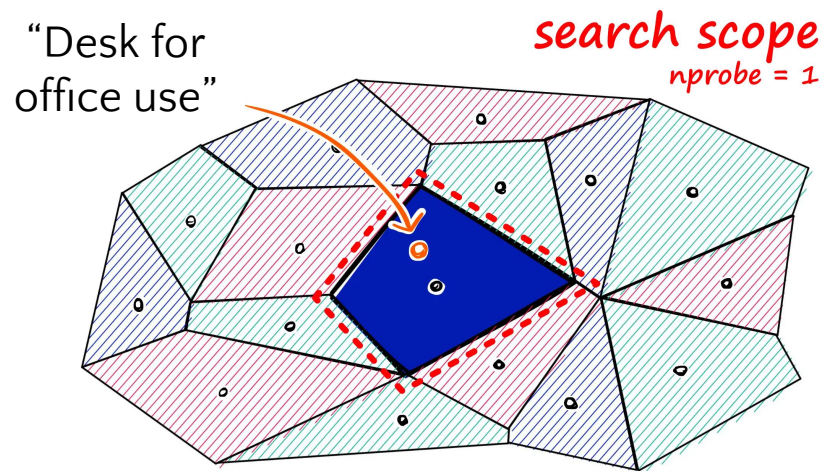
This is great, but can we search  
through 1M products in  
reasonable amount of time?





## Clustered Search

- FAISS – Facebook AI Similarity Search
- Speeds up search time but only searches a subset of the product catalog



# Evaluation





## Evaluating Performance

---

- No evaluation metrics from partner
- Synthesized benchmark dataset of 300 queries

### **Recall@20**

*Is the target product in the top 20 results?*

### **Precision@20**

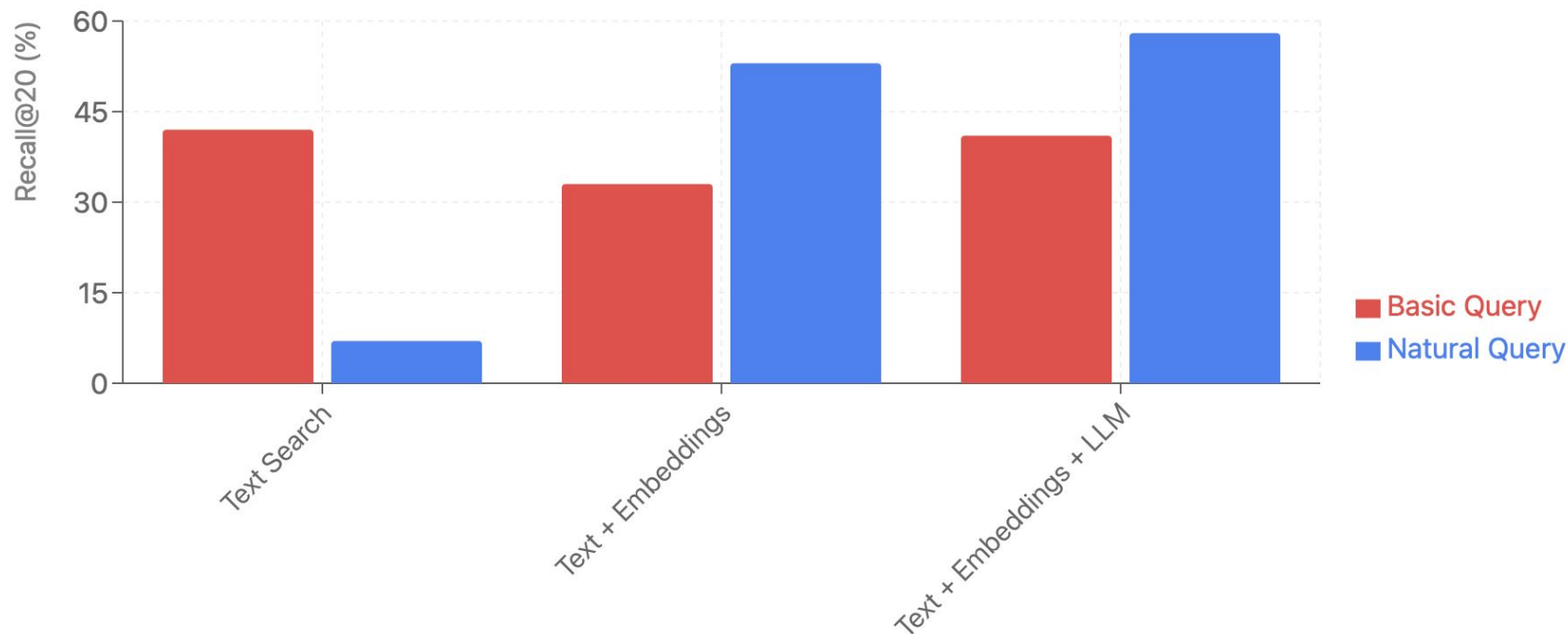
*How many relevant products are in the top 20 results?*

### **Search Time**

*How long does it take to retrieve search results?*



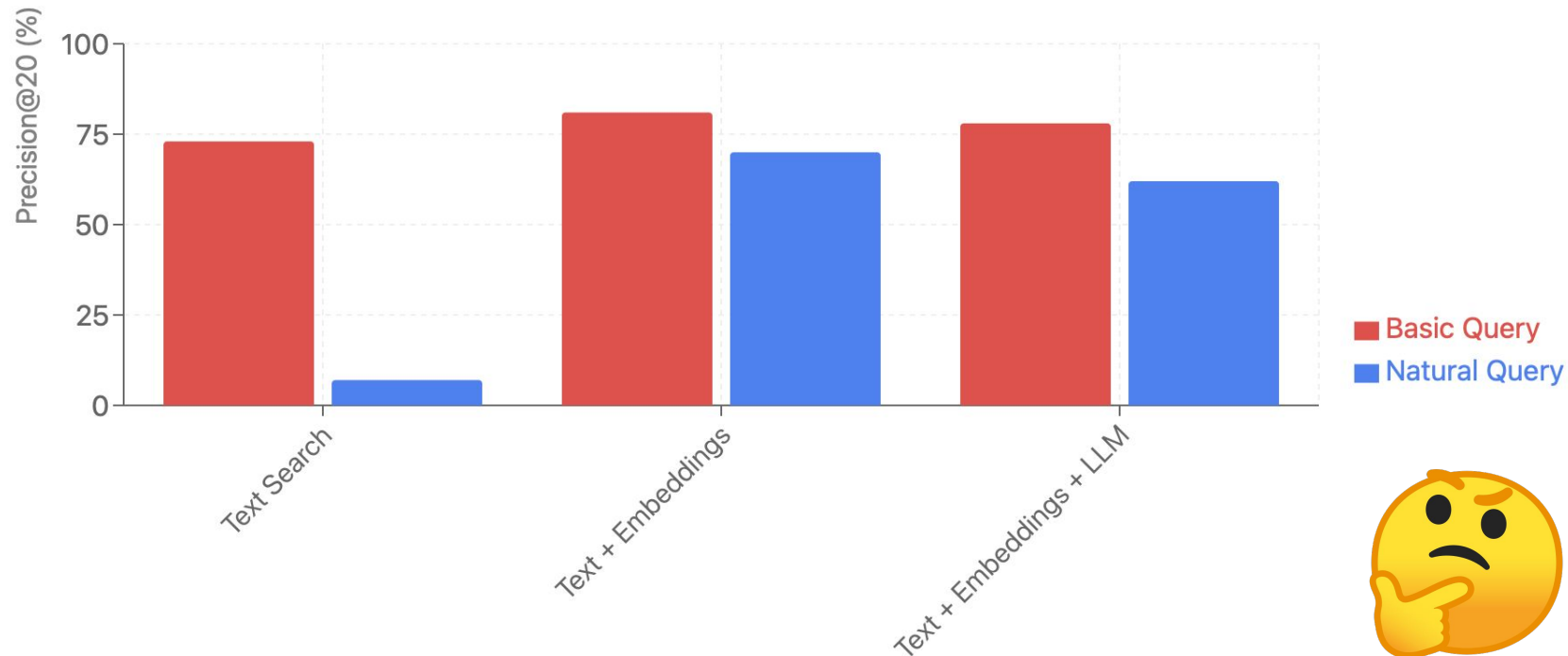
## Recall Improvement on Natural Queries





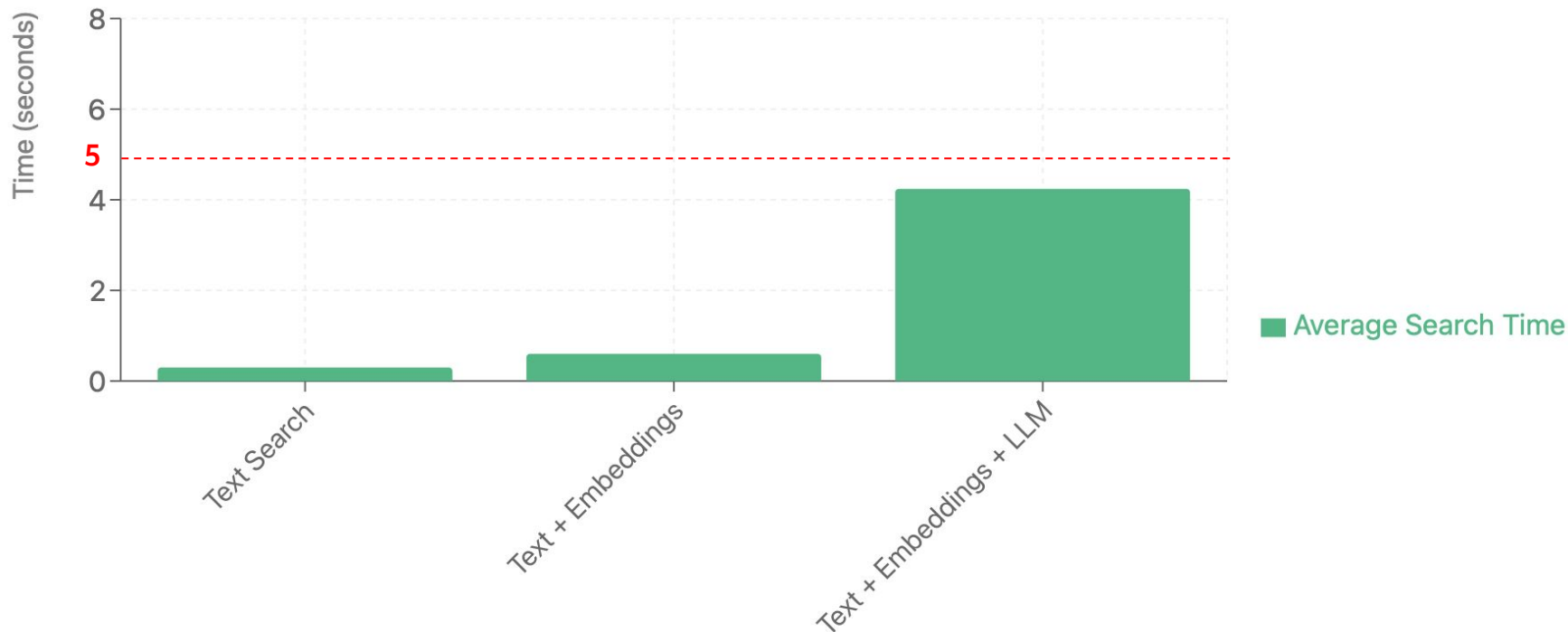


## Precision Results Show Variability





## LLM Adds Additional Search Time





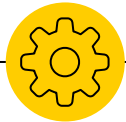
## Evaluation Summary

---

Text Search + Embedding + LLM

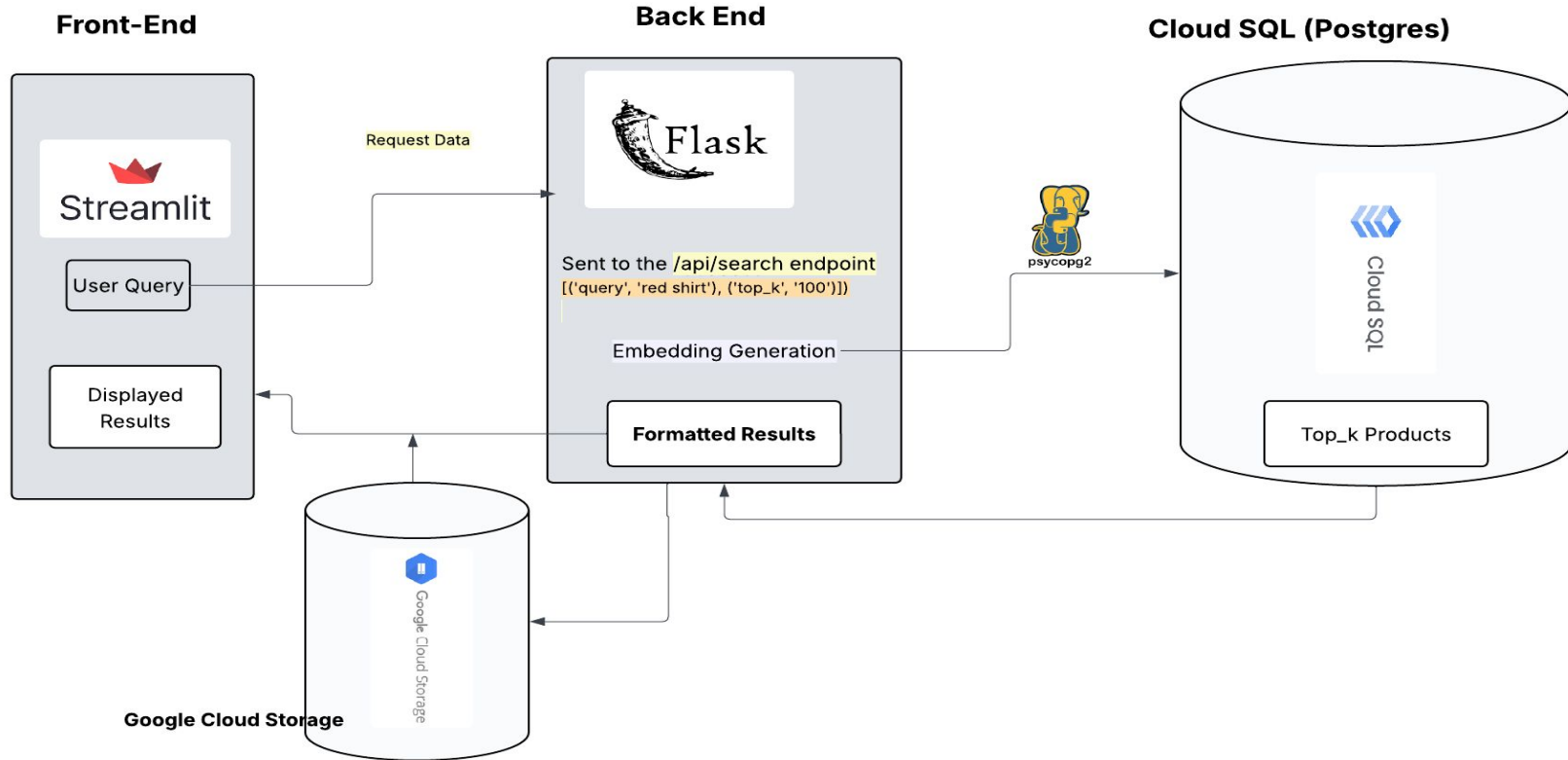
- Recall@20: 0.56
- Precision@20: 0.64
- Average search time: 4.24s

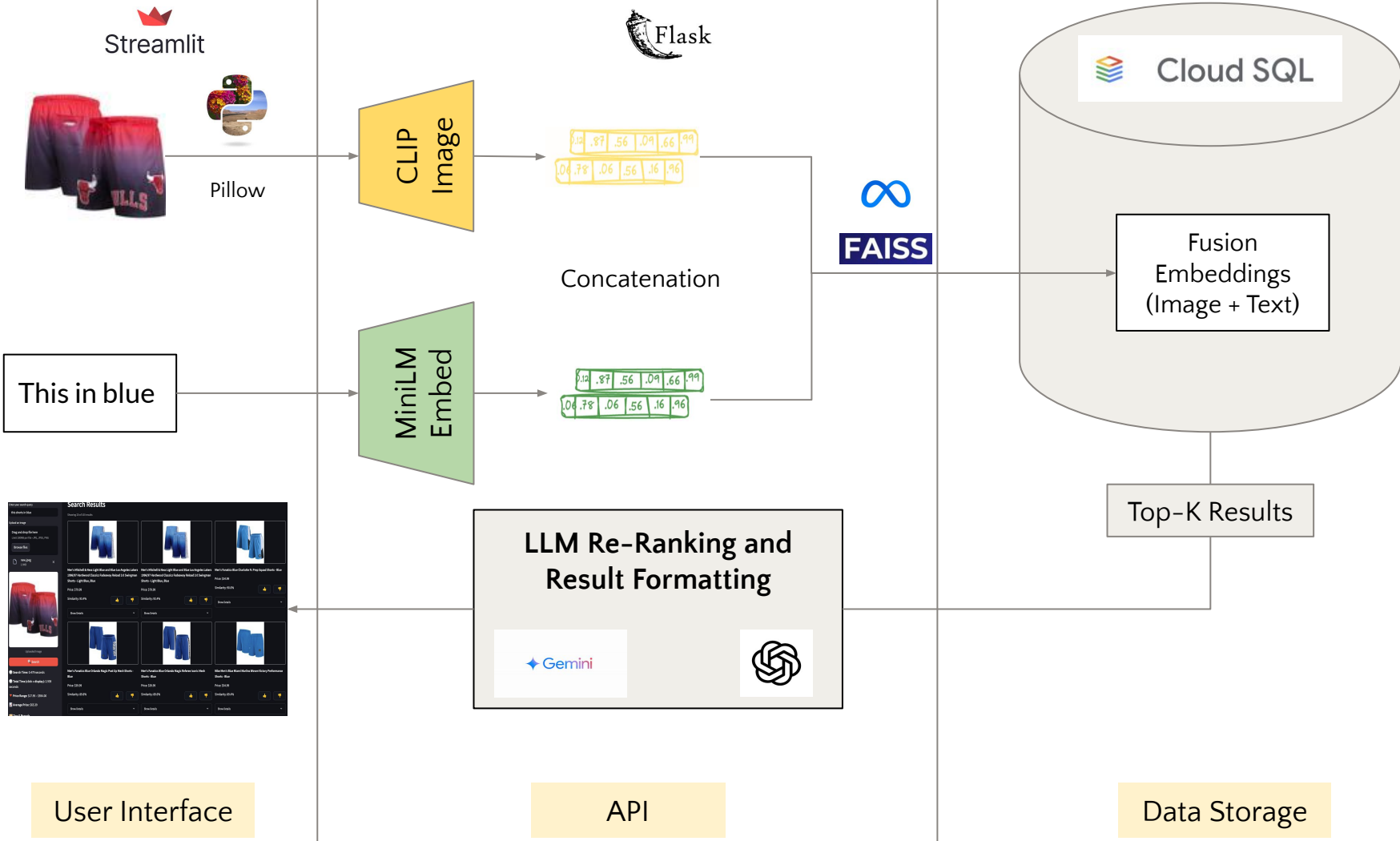
# Workflow





# Workflow Diagram







## Evaluating Performance



### Recall@20

*Is the target product in the top 20 results?*

### Precision@20

*How many relevant products are in the top 20 results?*

### Search Time

*How long does it take to retrieve query results?*



# Tracking Experiments (MLflow)

mlflow 2.22.0

Experiments

Models

Prompts













GitHub

Docs

## Experiments



Search experiments

- ☒ 100k\_pgv\_keyberts  
- ☒ 1M\_faiss\_fusion  
- ☒ 100k\_pgv\_fusion  
- ☐ 100k\_pgv\_combined  
- ☐ 1M\_faiss\_hyperparam  

## Displaying Runs from 3 Experiments

Share

Runs

Evaluation

Experimental

Traces



metrics.rmse < 1 and params.model = "tree"



Time created ▾

State: Active ▾










Datasets ▾

Sort: overall\_recall\_at\_k ▾

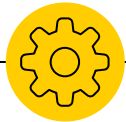
Columns ▾

Group by ▾

Metrics							
<input type="checkbox"/>	Run Name	Created	Duration	attribute_query_	basic_query_	natural_query_	overall_recall_
<input type="checkbox"/>	 fusion_pgv_keyberts_202...	 14 days ago	2.2min	0.74	0.5	0.63	0.62
<input type="checkbox"/>	 pgv_fusion_20250527_15...	 14 days ago	1.6min	0.74	0.5	0.62	0.62
<input type="checkbox"/>	 tfidf_fusion_gpt3.5_turbo_...	 6 days ago	21.2min	0.7	0.41	0.58	0.56
<input type="checkbox"/>	 tfidf_fusion_20250603_20...	 7 days ago	3.0min	0.68	0.33	0.53	0.51
<input type="checkbox"/>	 tfidf_20250603_195853	 7 days ago	1.5min	0.58	0.42	0.07	0.36



# Tools Used





## Tools (Cloud)

- Project goals (Speed)
- Cloud technologies (Cost and scalability)



Google Cloud Platform



## Tech Stack (Cloud)

### Google Cloud SQL

- Managed database hosting
- Separation of compute and storage

### Google Cloud Storage

- Object storage
- Images and indexes

### Google Cloud Run

- Fully managed compute platform by Google
- Allows you to run containerized applications
- Docker, Docker-Compose



## Tech Stack (Summary)

Category	Tools
Data Processing	Pandas, Numpy, Pillow & io
Front & Backend	Streamlit, Flask
Retrieval	Psycopg2, FAISS
AI & ML	Hugging Face, LangChain, OpenAI, Gemini



# Summary





## Project & Solution

**Project goal:** Build a **fast** and **scalable multimodal** search system that captures the **semantic** meaning of user queries.

### Semantic

- **MiniLM** in text embedding
- **LLM** for post retrieval reranking

### Multimodality

- **CLIP** model for both text and image
- **TF-IDF** as baseline

### Efficiency & Scalability

- **FAISS** index for fast similarity search
- **Google Cloud** for scalability



## **Final Product**

---

Reproducible pipelines to build the search engine from the ground up, including:

- Indexing Pipeline
- Inference/Evaluation Pipeline



## Final Product

### Indexing Pipeline (One-time process)

Partner runs: `make train`

Data  
Preprocessing

*data\_clean.py*

Create embeddings by  
Pretrained Models

*generate\_embed.py*

Generate  
Index

*compute\_faiss\_index.py*

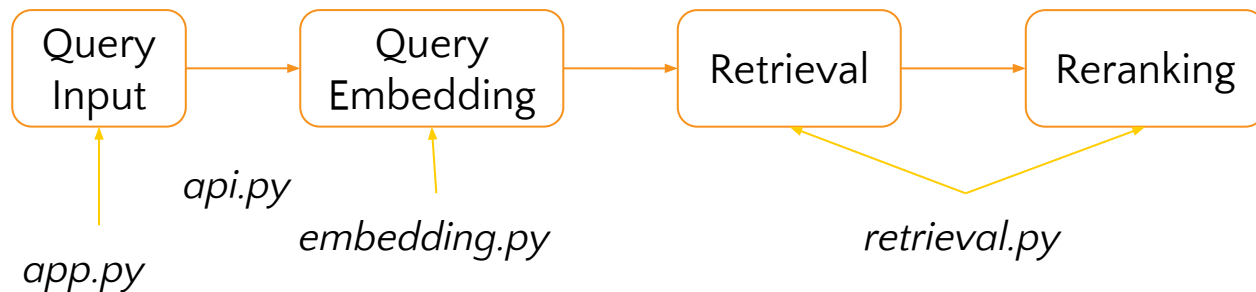




## Final Product

### Inference/ Evaluation Pipeline

Partner runs: `make run`





## Future Explorations

Imperfect relevance evaluation	Standardized guidelines for annotation
LLM only focuses on a subset	Apply adaptive cutoff
No training	Add projection layer (after labeling)
Used only 1M samples	Scale up if more resources permit



## How Did We Do?



### Must Have

- ✓ Support for nature language queries
- ✓ Fast response time (< 5s)
- ✓ Reusable API endpoints



### Should Have

- ✓ Reproducible data pipeline

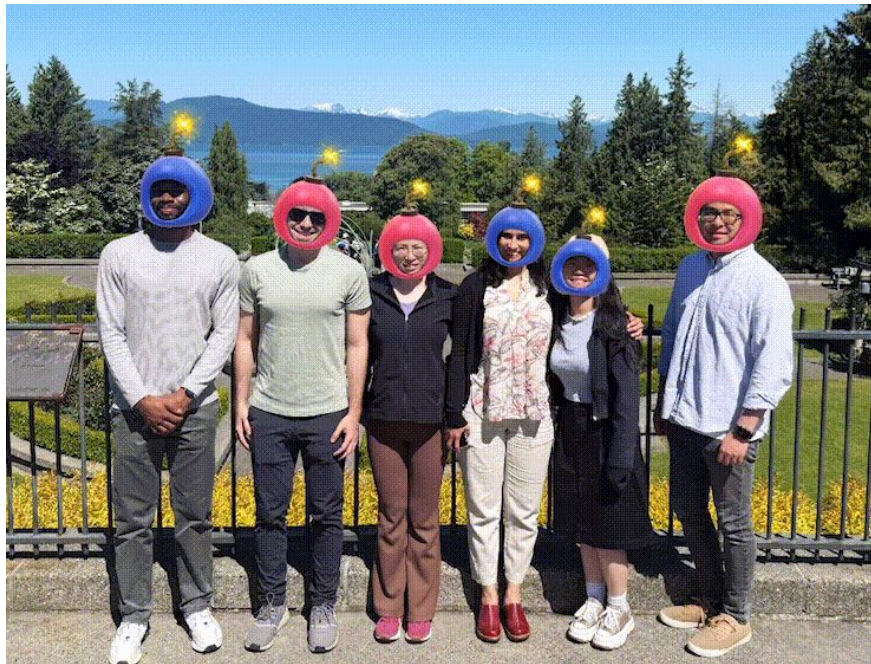


### Nice to Have

- ✓ Web interface
- ✗ Use a larger dataset
- ✓ Evaluation plan



Any Questions?



Try our **demo!**

