

Captions Are Worth a Thousand Words

Enhancing Product Retrieval with
Pretrained Image-to-Text Models

Jason Tang¹, Garrin McGoldrick², Marie Al Ghossein²,
Ching-Wei Chen²

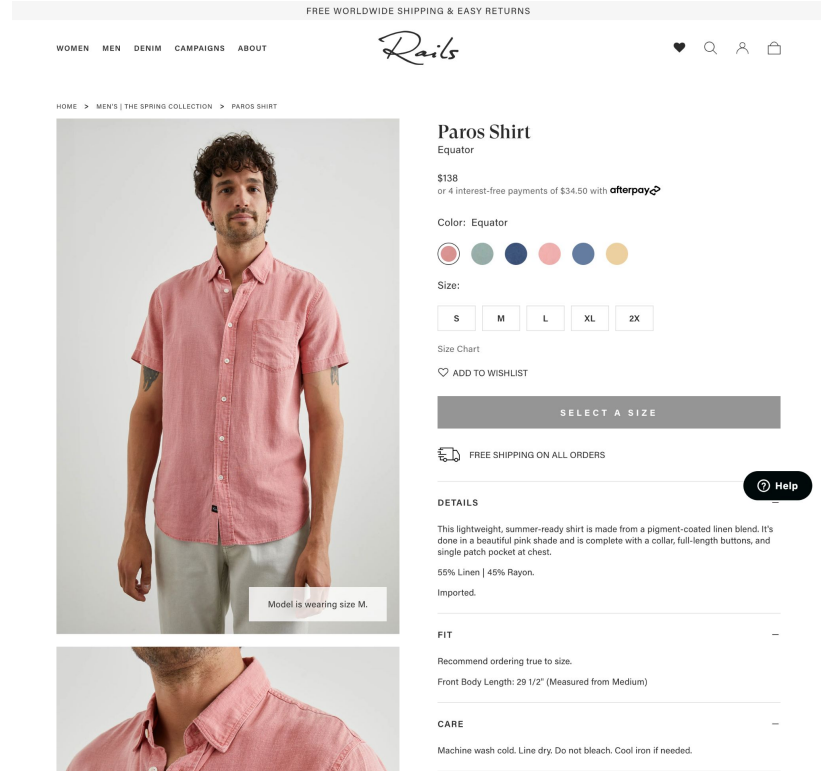
¹ University of Toronto, ² Crossing Minds

Presented at ISIR-eCom @ WSDM 2024
Mar 8th, 2024, Mérida, Mexico



Metadata, Search and Discovery

- Metadata is essential to helping customers find and discover products in a catalog.
- Descriptive product metadata like Name, Color, Size, Description, etc, are heavily used in Search as well as Recommendations.



FREE WORLDWIDE SHIPPING & EASY RETURNS

WOMEN MEN DENIM CAMPAIGNS ABOUT

Rails

HOME > MEN'S | THE SPRING COLLECTION > PAROS SHIRT

Paros Shirt
Equator

\$138
or 4 interest-free payments of \$34.50 with **afterpay**

Color: Equator

Size:

S M L XL 2X

Size Chart

ADD TO WISHLIST

SELECT A SIZE

FREE SHIPPING ON ALL ORDERS

Help

DETAILS

This lightweight, summer-ready shirt is made from a pigment-coated linen blend. It's done in a beautiful pink shade and is complete with a collar, full-length buttons, and single patch pocket at chest.

55% Linen | 45% Rayon.

Imported.

FIT

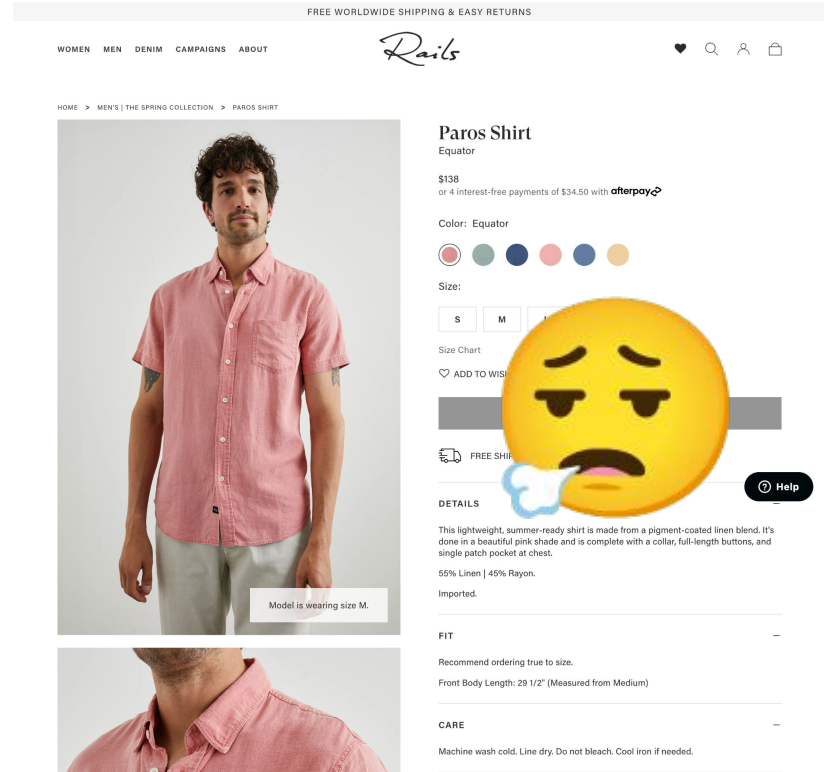
Recommend ordering true to size.
Front Body Length: 29 1/2" (Measured from Medium)

CARE

Machine wash cold. Line dry. Do not bleach. Cool iron if needed.

Metadata, Search and Discovery

- But metadata is costly to manually curate, and many eCommerce sites have varying coverage, quality, and granularity of these tags and descriptors.
- But almost all eCommerce sites will have high quality product images, and paragraph-level product descriptions.



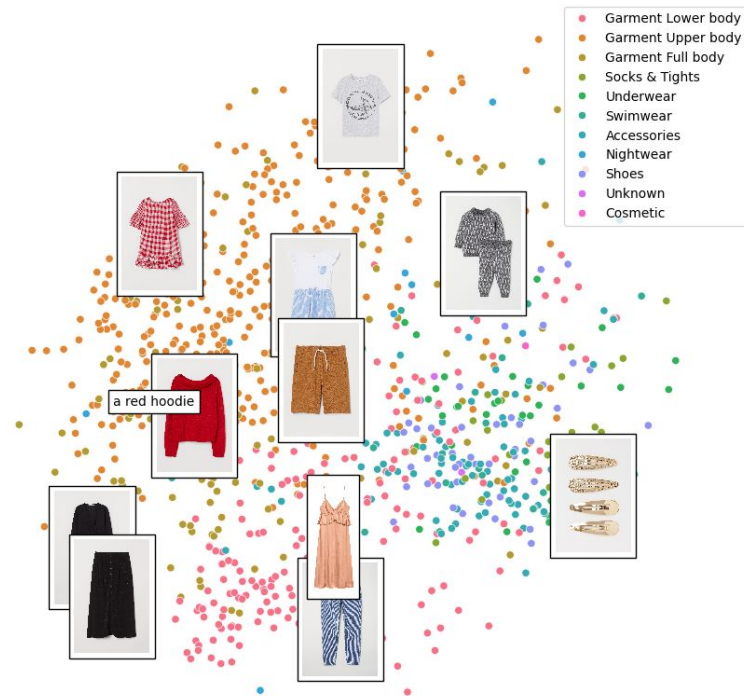
Related Work

CLIP: Image Tagging

- Discriminative image-to-text tagging
- Pre-training task: Predict (image, text) pairs that occur across a batch

FashionCLIP

- CLIP fine-tuned on fashion dataset from Farfetch



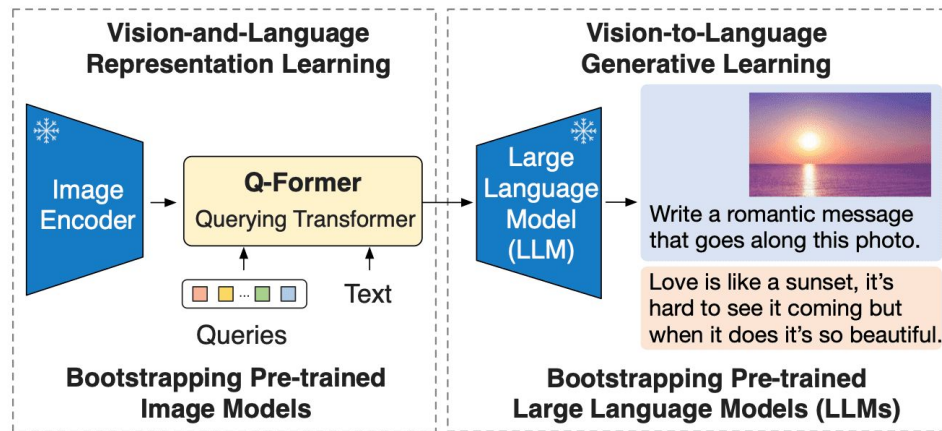
Related Work

BLIP-2: Image captioning

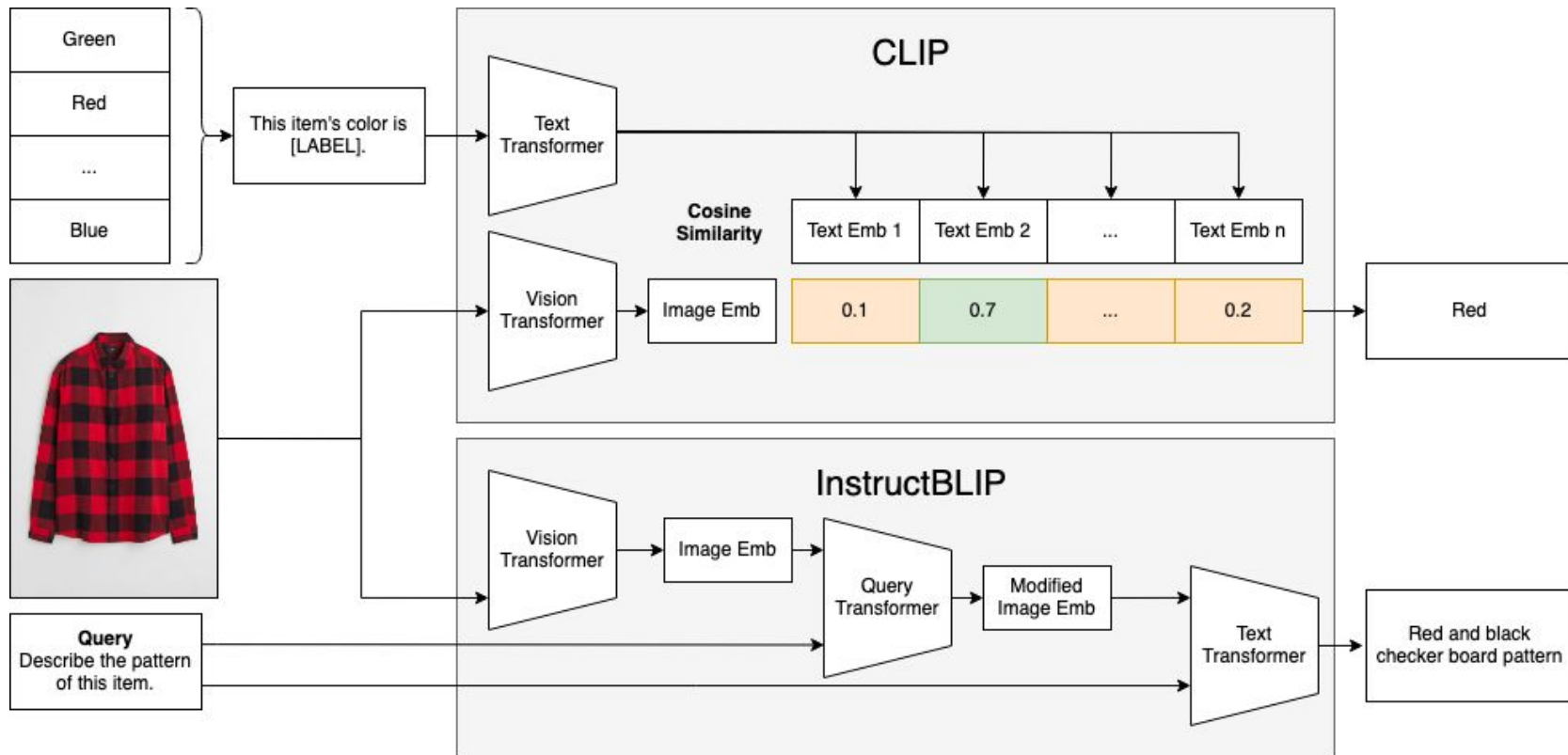
- Generative image-to-text captioning
- Pre-trained jointly with multiple vision-language objectives

InstructBLIP

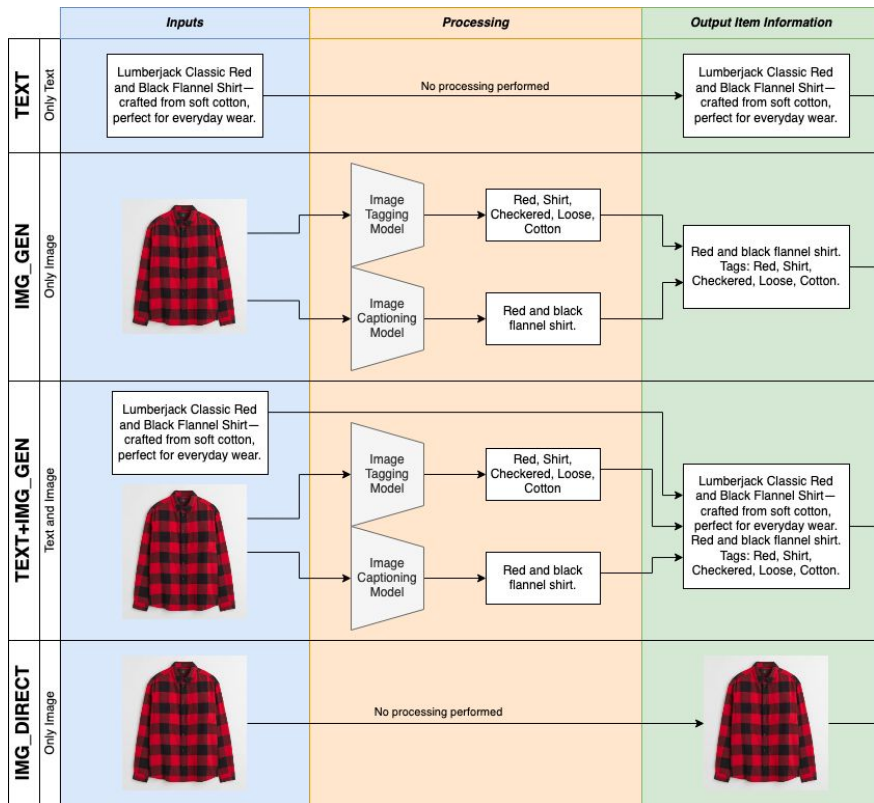
- Instruction tuning format for further pre-training



Proposed Method



Proposed Method



Proposed Method

- Query preprocessing to address misspelled queries and those containing non-English terms
- Done with ChatGPT-3.5, e.g.,

"Extract at least 5 related tags or usage keywords from queries. Output in English as a comma separated list."

User Query	Processed Query
!awnmower tires without rims	lawnmower, tires, without rims
#20 paper bags without handle	paper bags, without handle, packaging, eco-friendly, retail
paws	animal, pets, claws, dogs, cats
apple iphone 11 pro unlocked	apple, iPhone, 11 pro, unlocked
자전거트레일러	bicycle trailer, bike trailer, cycling trailer, bike cart, bike carrier
眼镜框	eyeglass frames, glasses frames, eyewear, spectacle frames, glasses

Evaluation

- The [Amazon Shopping Queries dataset](#) consists of a list of 1.1M query-product pairs annotated with E/S/C/I (Exact, Substitute, Complement, Irrelevant) labels
- Focus on task 1 of the KDD Cup'22: **Query-Product Ranking** and on the English subset of the ESCI dataset
- Use product identifiers to **scrape images** and select first image to constitute dataset

KDD Cup 2022 Workshop: ESCI Challenge for Improving Product Search

Held in conjunction with [KDD'22](#) Aug 17th, 2022 – Washington D.C., USA

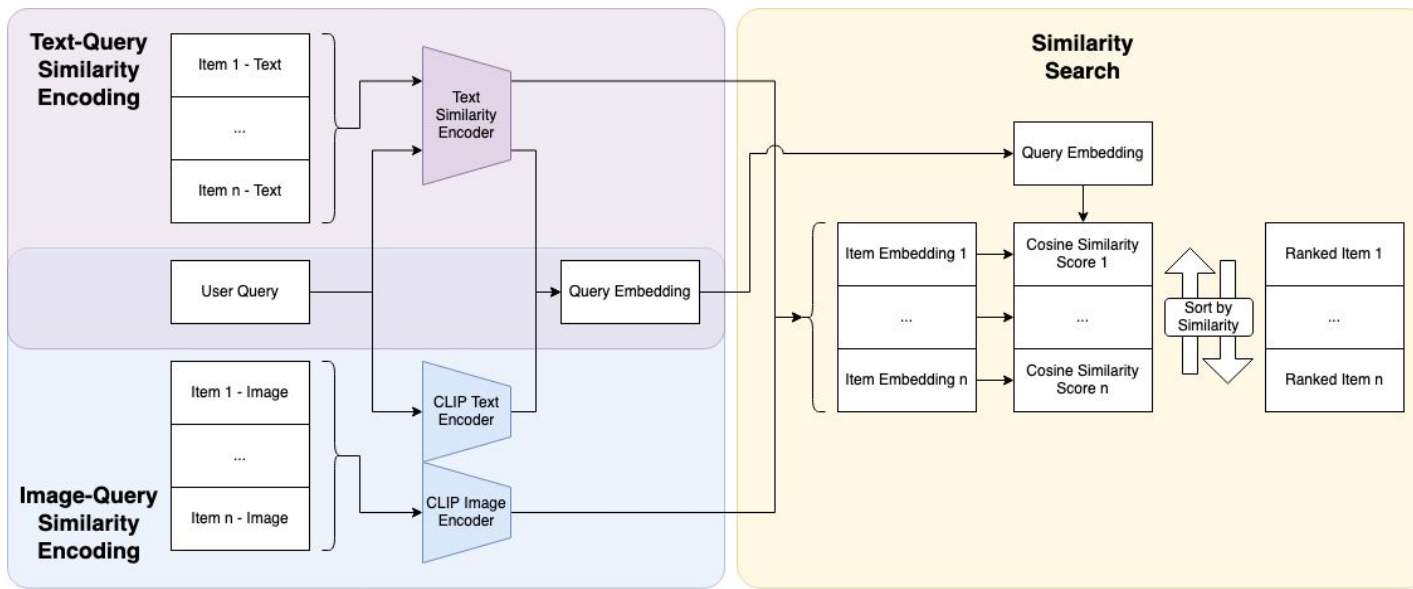


Dataset sampling

- Filter the dataset and consider products that appear in at least 3 queries (out of the 482k products) for efficiency reasons, given the cost of retrieving images
 - Resulting dataset: 21.6k products, 19.8k queries, 4.3 products per query
- Use padding, consisting in adding random products per query with I label, in order to mitigate the facts that:
 - The datasets mainly contain E and S labels per query,
 - Relevant information (e.g., product dimensions and version) might not be discernible from images

Similarity Search

- Multi-qa-mpnet-base-dot-v1
- ms-marco-MiniLM-L-2-v2



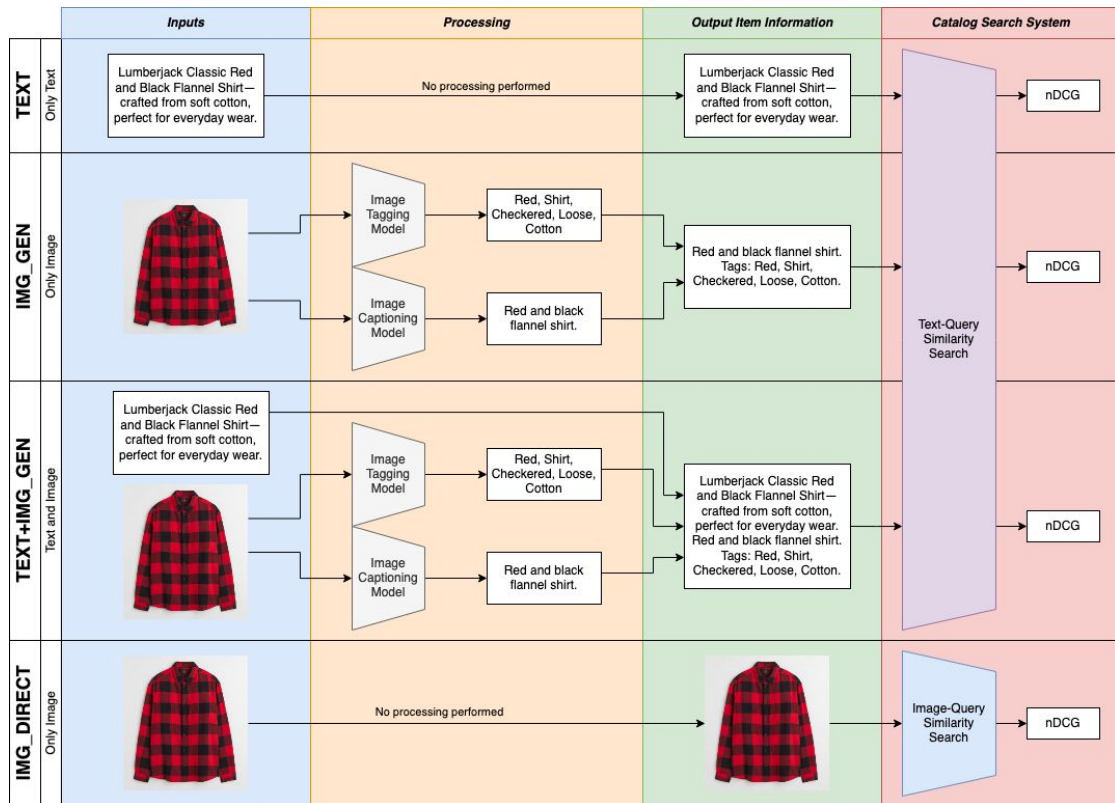
NDCG calculation



Source: [A Practical Guide to NDCG](#)

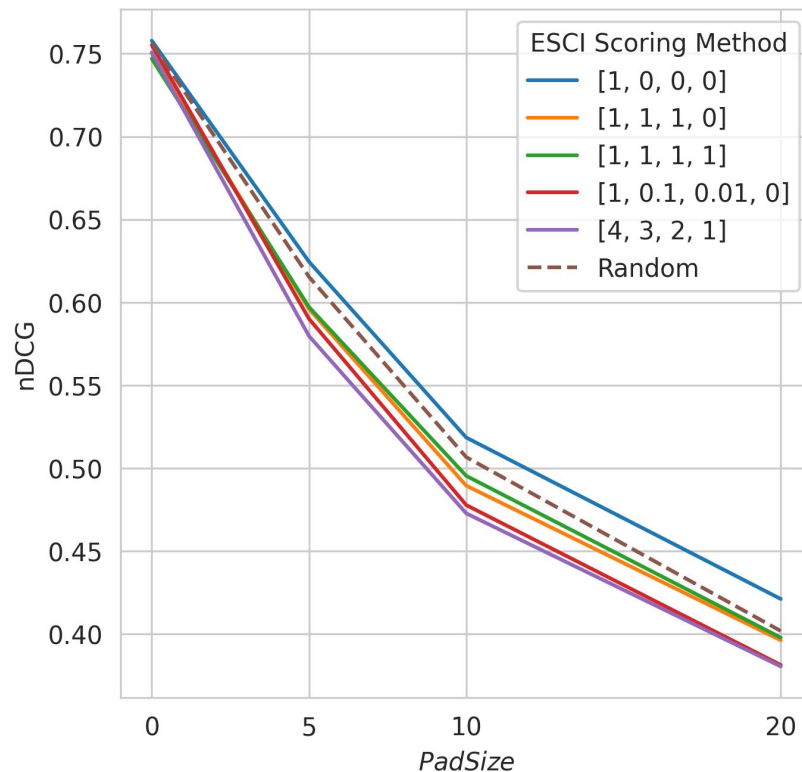
ESCI rating scores from KDD Cup 2022

Rating	<u>Exact</u>	<u>Substitute</u>	<u>Complement</u>	<u>Irrelevant</u>
Weight	1.0	0.1	0.01	0.0

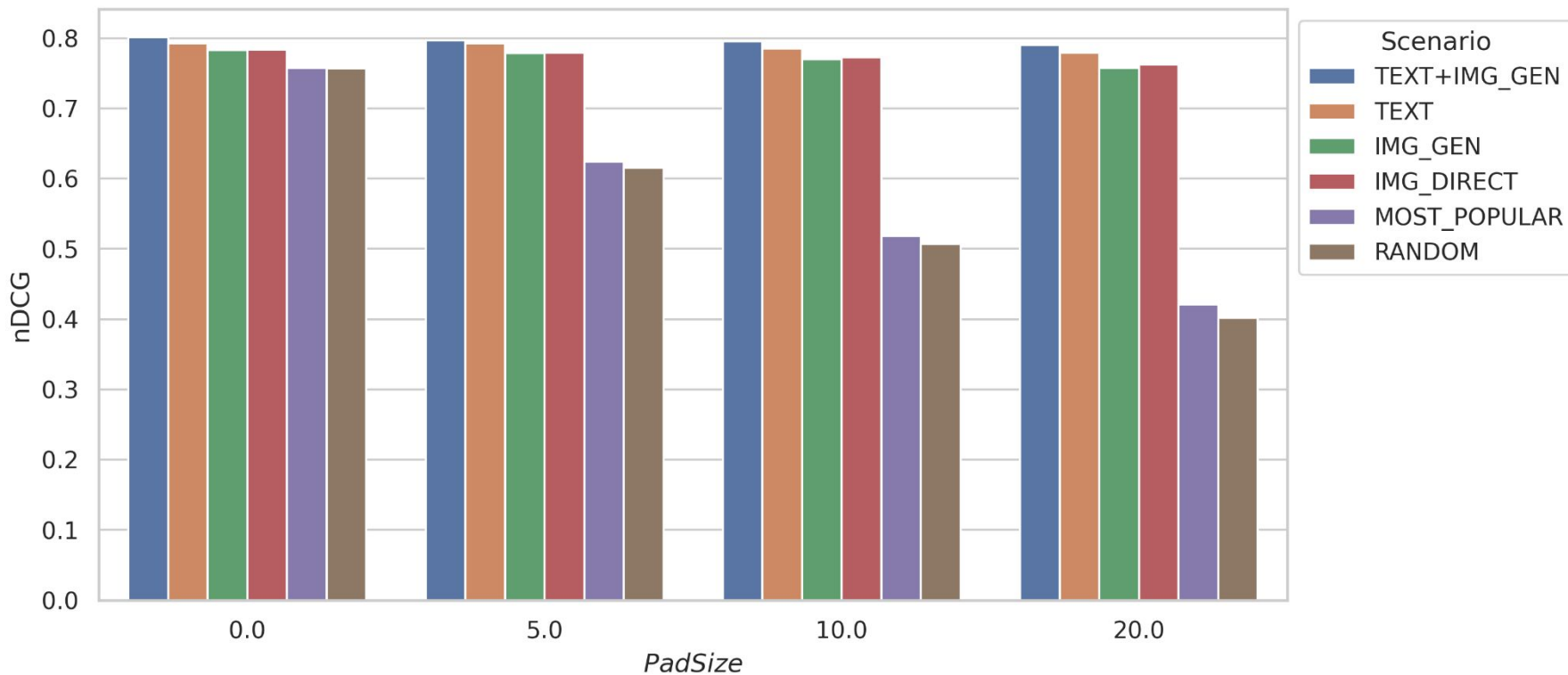


Baselines

- Random
- Most popular,
with popularity computed
according to different scores
attributed to the [E, S, C, I] labels



Results



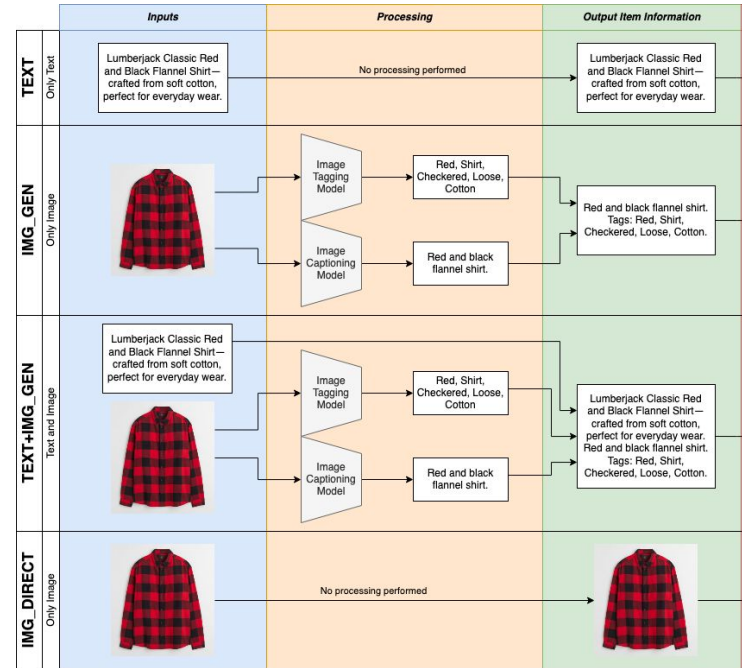
Results

- Query preprocessing showed an increase in performance across the board.

<i>PadSize</i>	Original Query	GPT Preprocessing
0	0.780	0.782
5	0.767	0.774
10	0.756	0.762
20	0.734	0.745

Conclusion

- We demonstrated the ability of multimodal models to generate high-performing image-derived descriptions that enable eCommerce platforms without substantial textual metadata to supplement existing text or images to improve item retrieval performance.



Future Directions

- Explore larger LLMs
- Evaluate against BM25 search engine baseline
- Apply to winning techniques from KDD Cup 2022

Thank You!