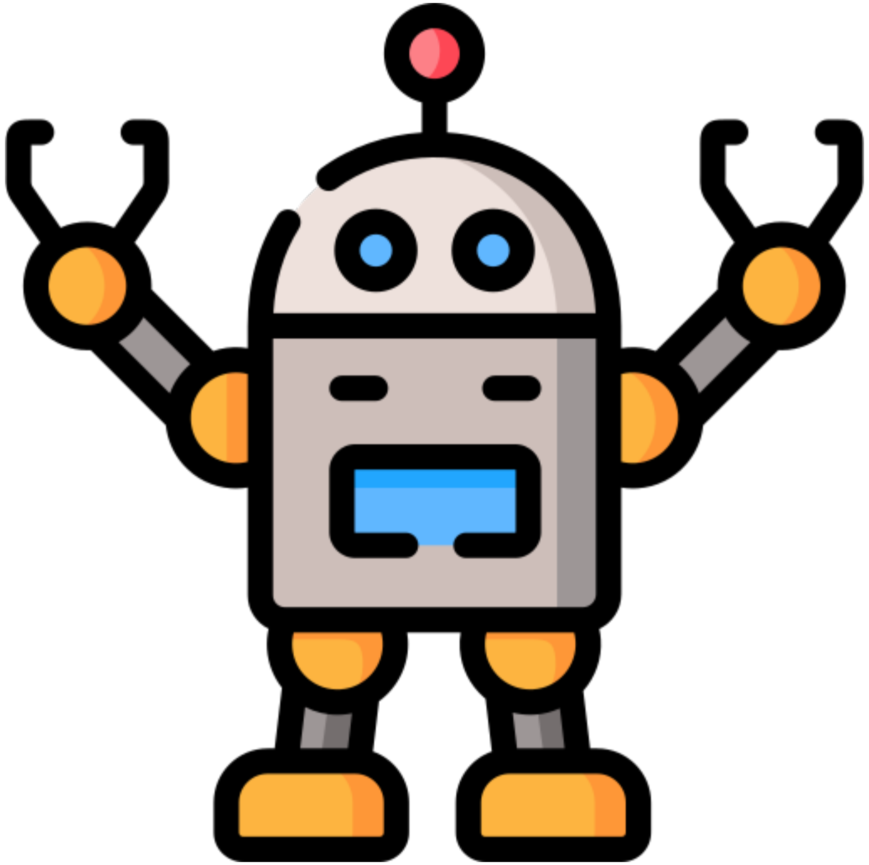
LLMs as Inexpensive Raters

Setting up offline search evaluation infrastructure on a budget



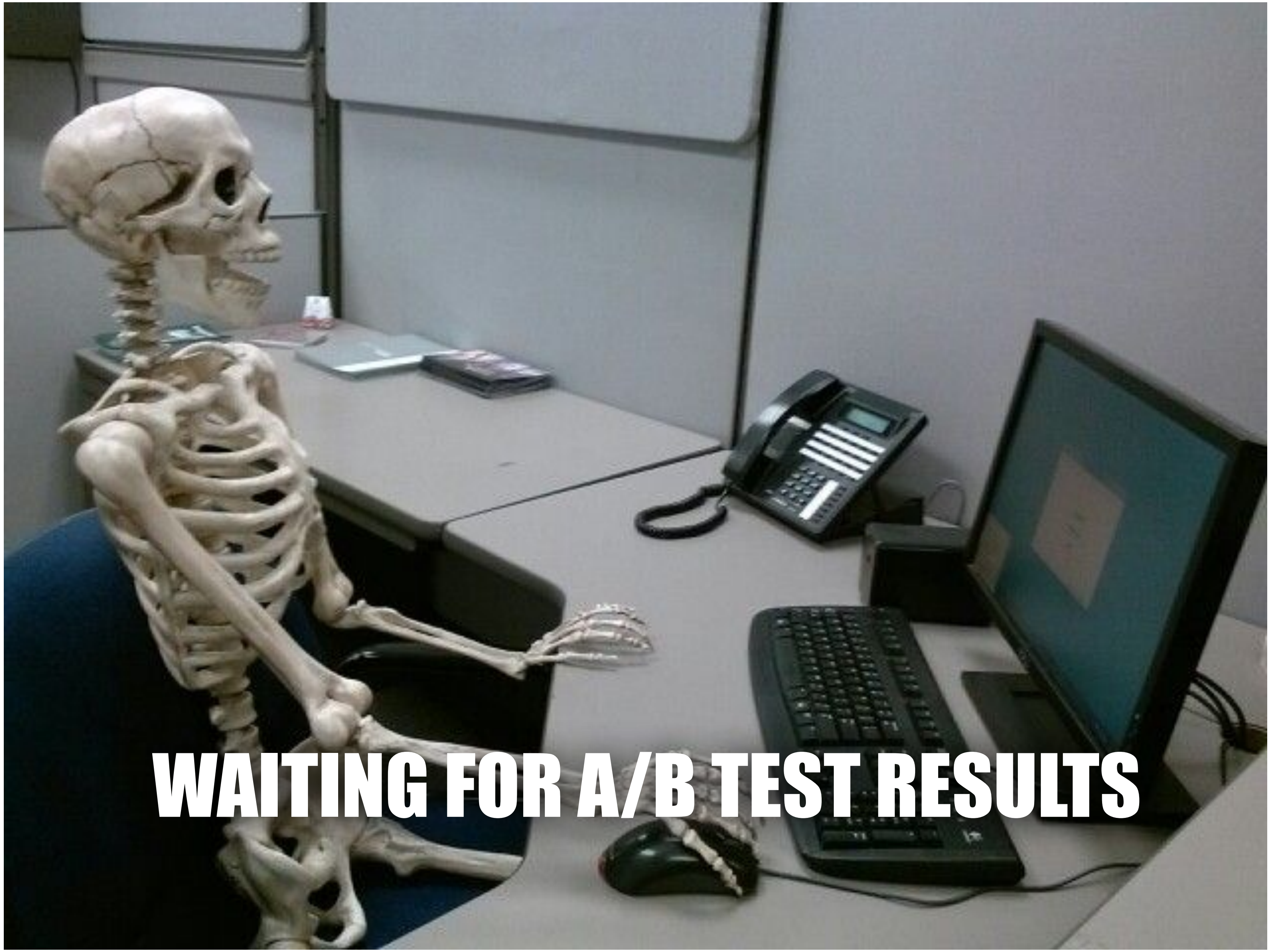
HSE

Mikayla Joy Webster

Search Specialist & Backend Dev @ Home Shopping Europe

E-commerce Marketplace

- 7 years as a backend dev
- 3 years in search


WAITING FOR A/B TEST RESULTS

Test Pool

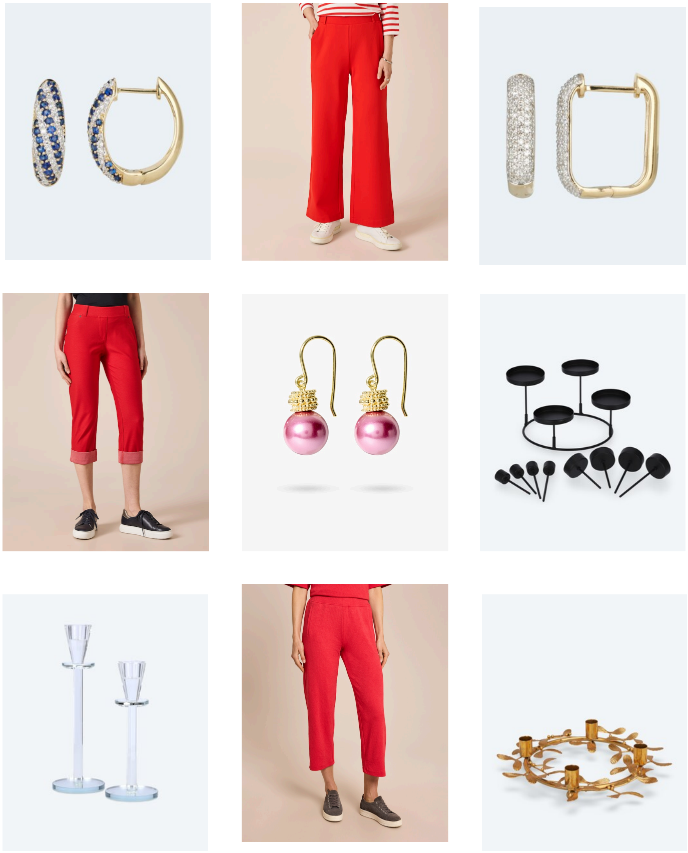
Fixed Query Set

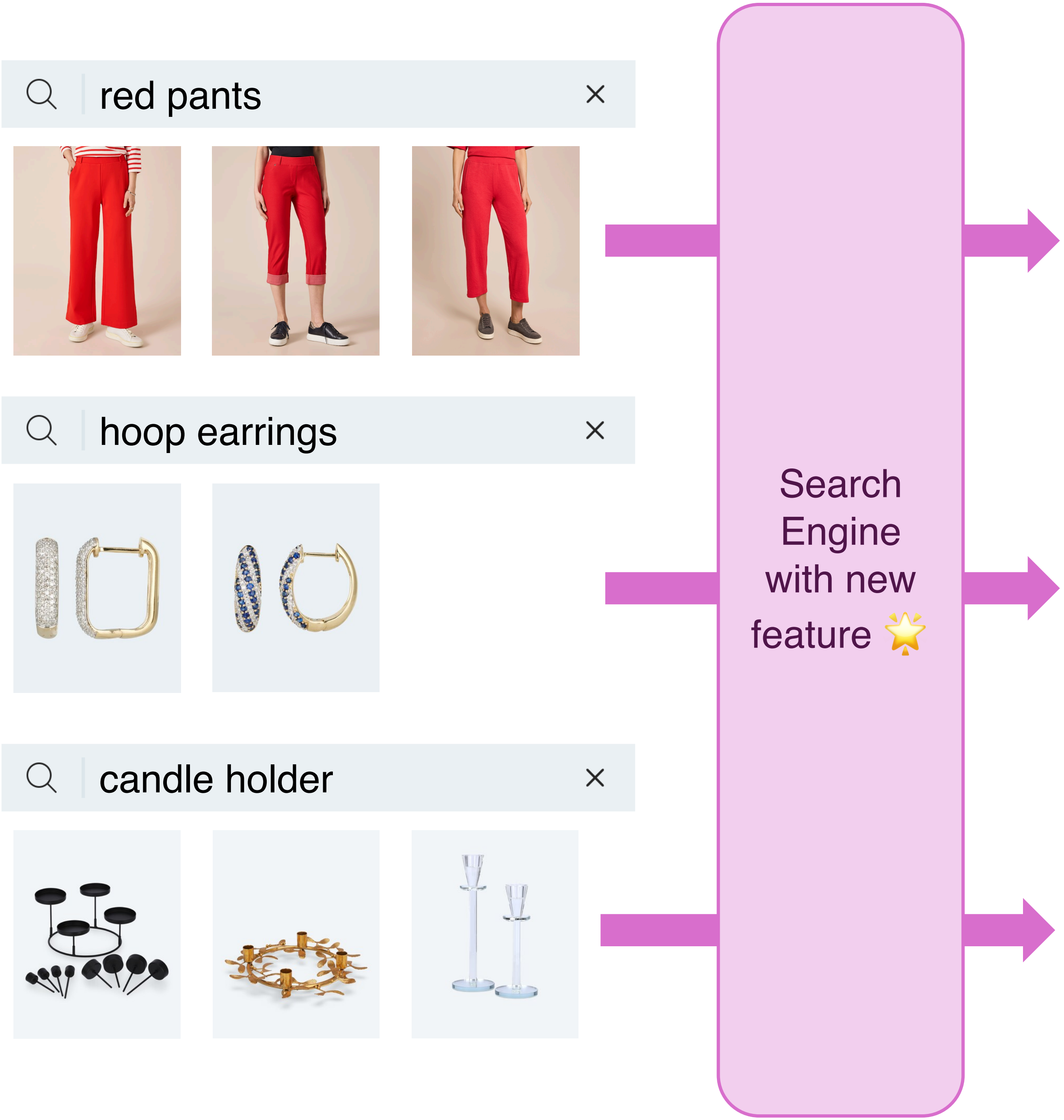
red pants

hoop earrings

candle holder

Fixed Product Set



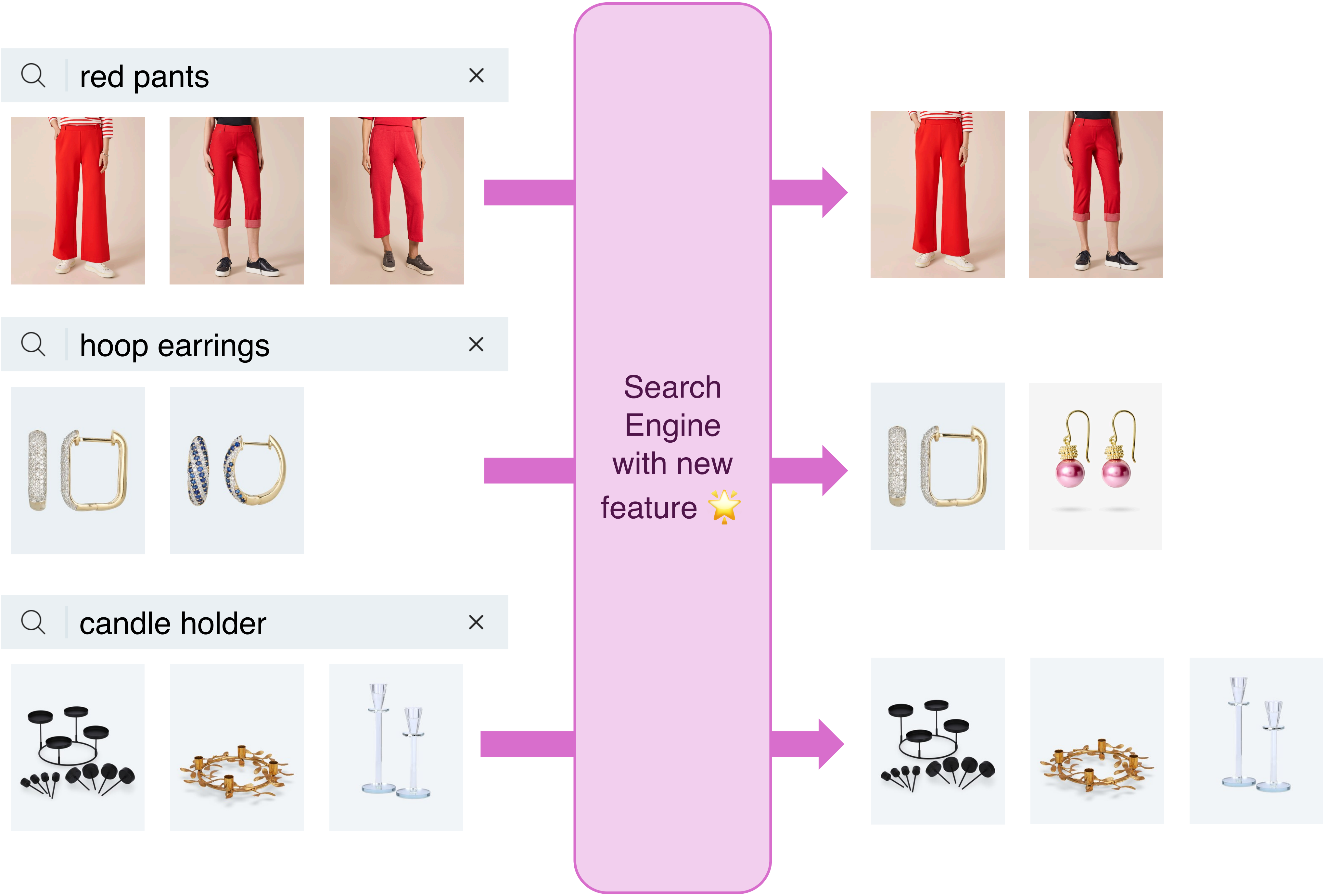


red pants

hoop earrings

candle holder

Search Engine with new feature 🌟

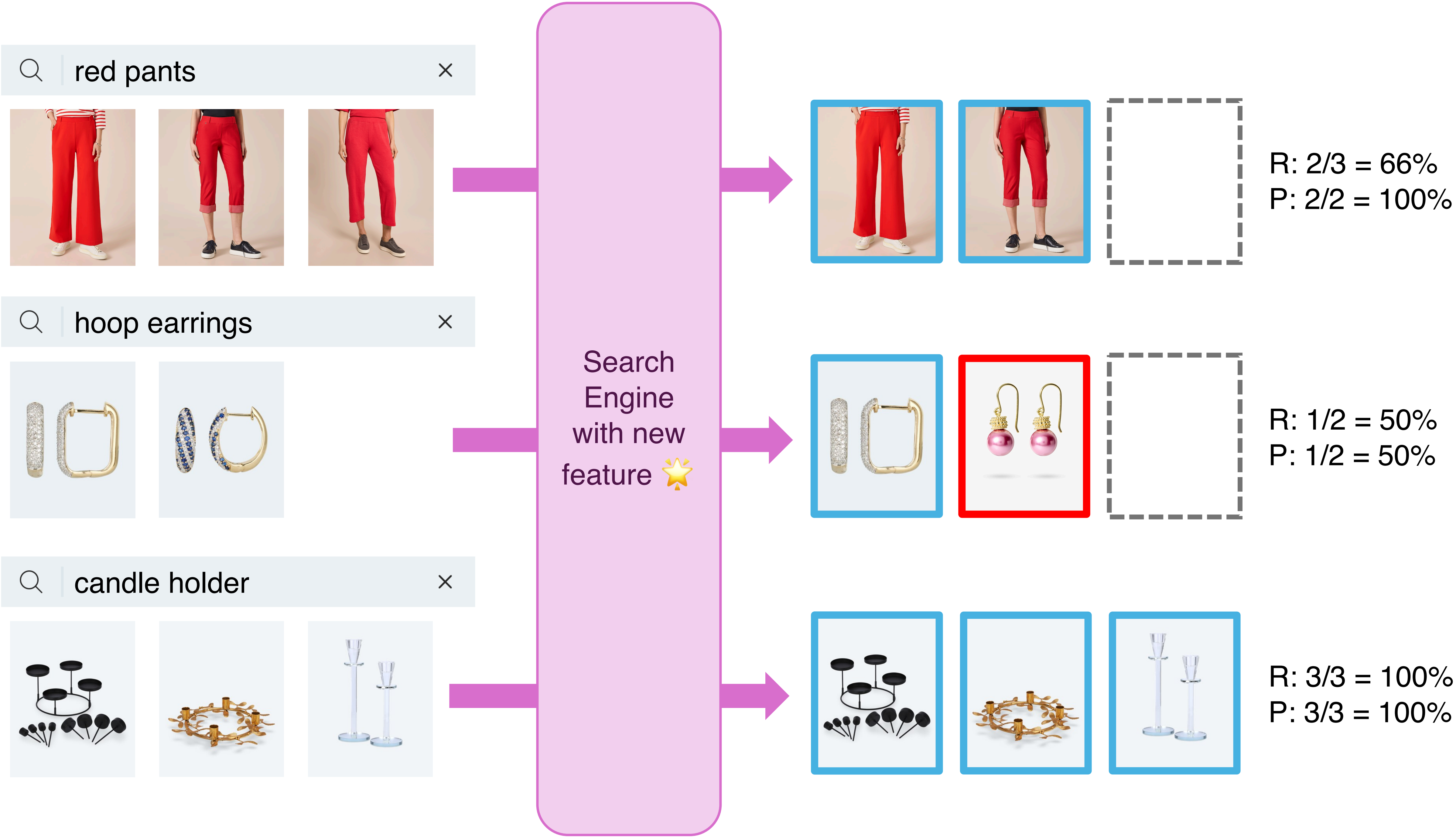


red pants

hoop earrings

candle holder

Search Engine with new feature 🌟



red pants

hoop earrings

candle holder

Search Engine with new feature ⭐

R: 2/3 = 66%
P: 2/2 = 100%

R: 1/2 = 50%
P: 1/2 = 50%

R: 3/3 = 100%
P: 3/3 = 100%

Retrieve, Annotate, Evaluate, Repeat: Leveraging Multimodal LLMs for Large-Scale Product Retrieval Evaluation

Kasra Hosseini, Thomas Kober, Josip Krapac, Roland Vollgraf, Weiwei Cheng, Ana Peleteiro Ramallo

Zalando SE, Berlin, Germany (18th Sep, 2024)

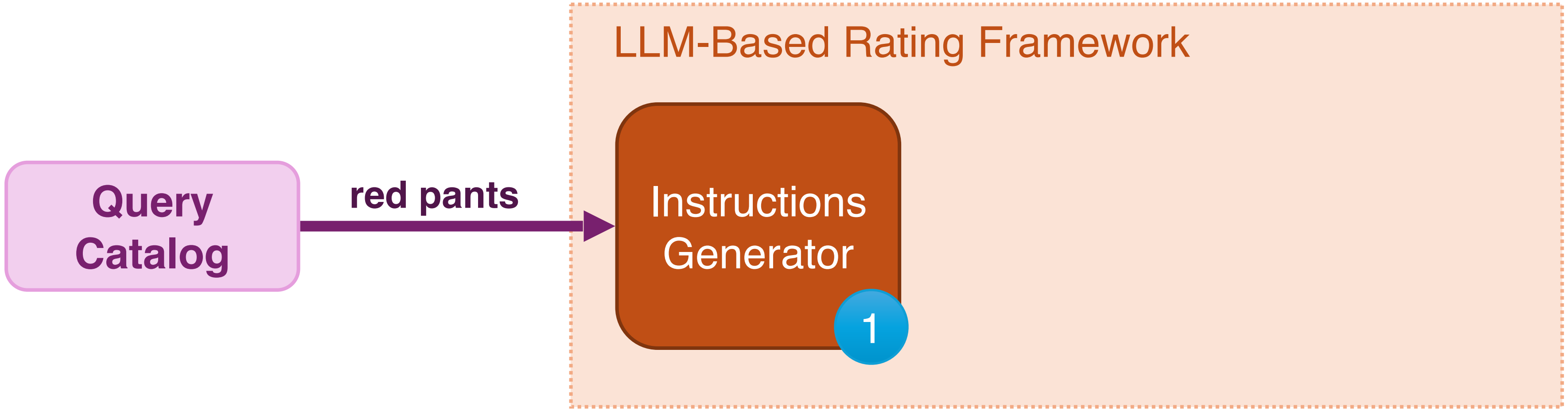
LLM-Based Rating Framework

https://arxiv.org/pdf/2409.11860

Retrieve, Annotate, Evaluate, Repeat: Leveraging Multimodal LLMs for Large-Scale Product Retrieval Evaluation

Kasra Hosseini, Thomas Kober, Josip Krapac, Roland Vollgraf, Weiwei Cheng, Ana Peleteiro Ramallo

Zalando SE, Berlin, Germany (18th Sep, 2024)

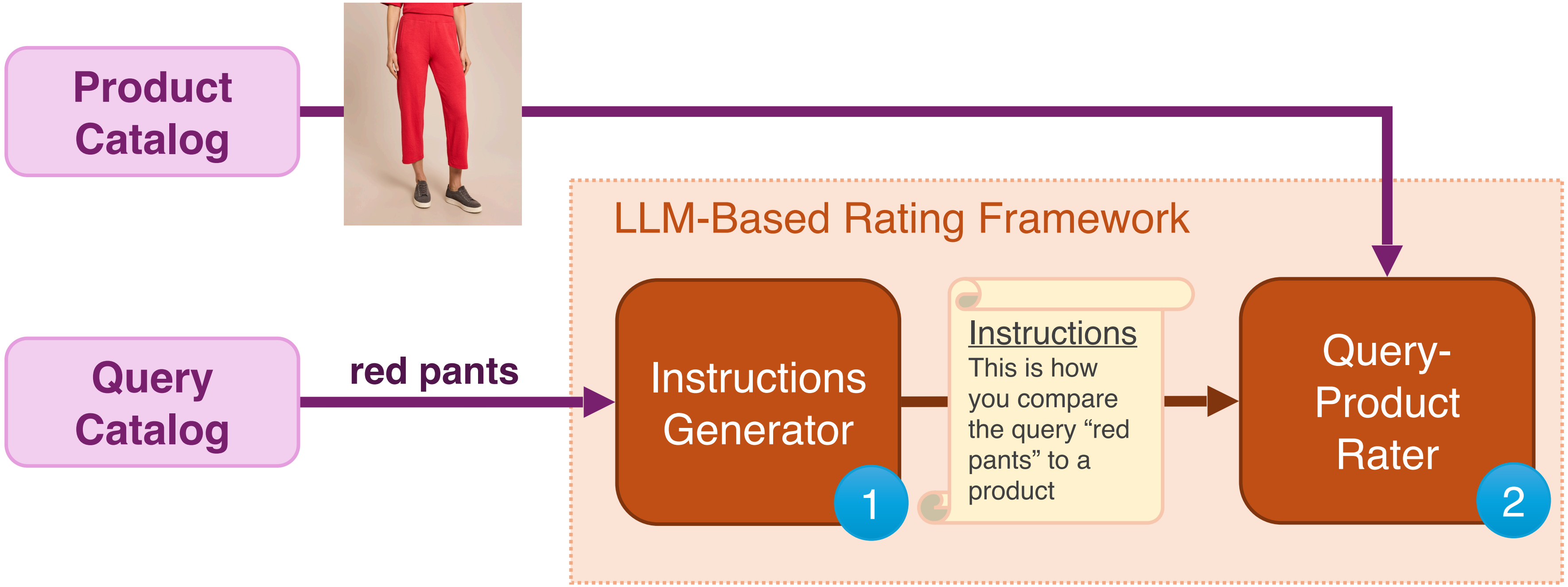


https://arxiv.org/pdf/2409.11860

Retrieve, Annotate, Evaluate, Repeat: Leveraging Multimodal LLMs for Large-Scale Product Retrieval Evaluation

Kasra Hosseini, Thomas Kober, Josip Krapac, Roland Vollgraf, Weiwei Cheng, Ana Peleteiro Ramallo

Zalando SE, Berlin, Germany (18th Sep, 2024)

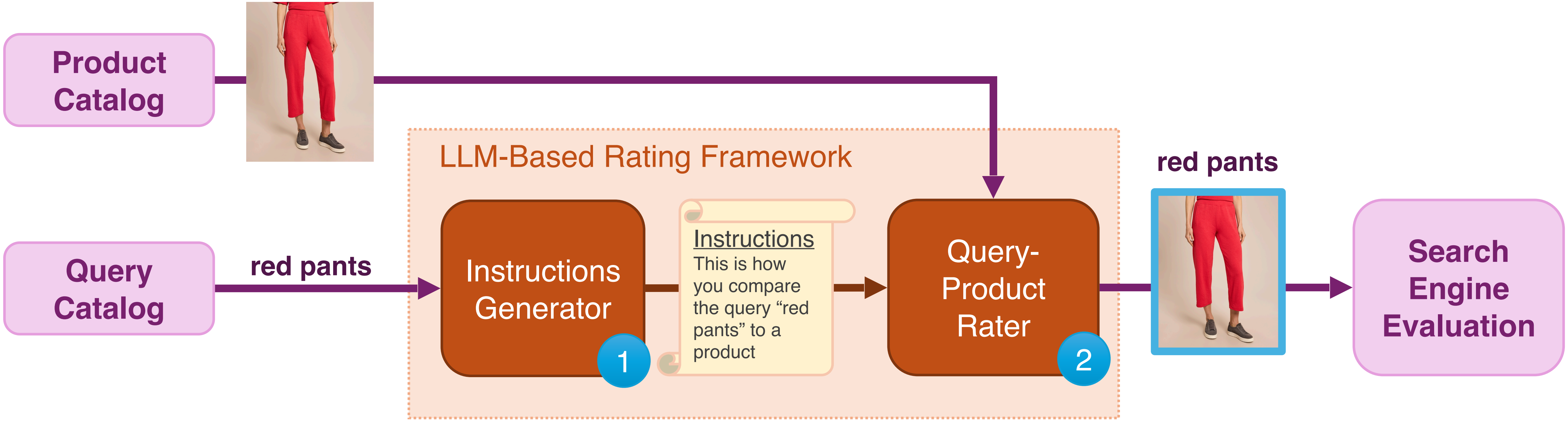


https://arxiv.org/pdf/2409.11860

Retrieve, Annotate, Evaluate, Repeat: Leveraging Multimodal LLMs for Large-Scale Product Retrieval Evaluation

Kasra Hosseini, Thomas Kober, Josip Krapac, Roland Vollgraf, Weiwei Cheng, Ana Peleteiro Ramallo

Zalando SE, Berlin, Germany (18th Sep, 2024)



Product Catalog

Query Catalog

red pants

LLM-Based Rating Framework

Instructions Generator

1

Instructions

This is how you compare the query “red pants” to a product

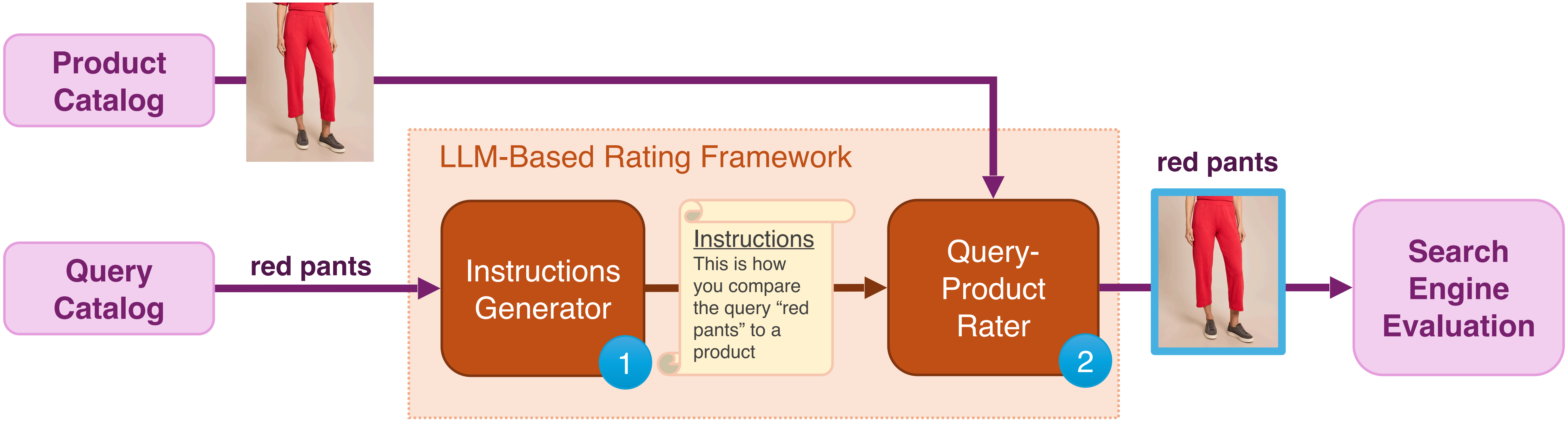
Query-Product Rater

2

red pants

Search Engine Evaluation

https://arxiv.org/pdf/2409.11860



Product Catalog

Query Catalog

red pants

LLM-Based Rating Framework

Instructions Generator 1

Instructions

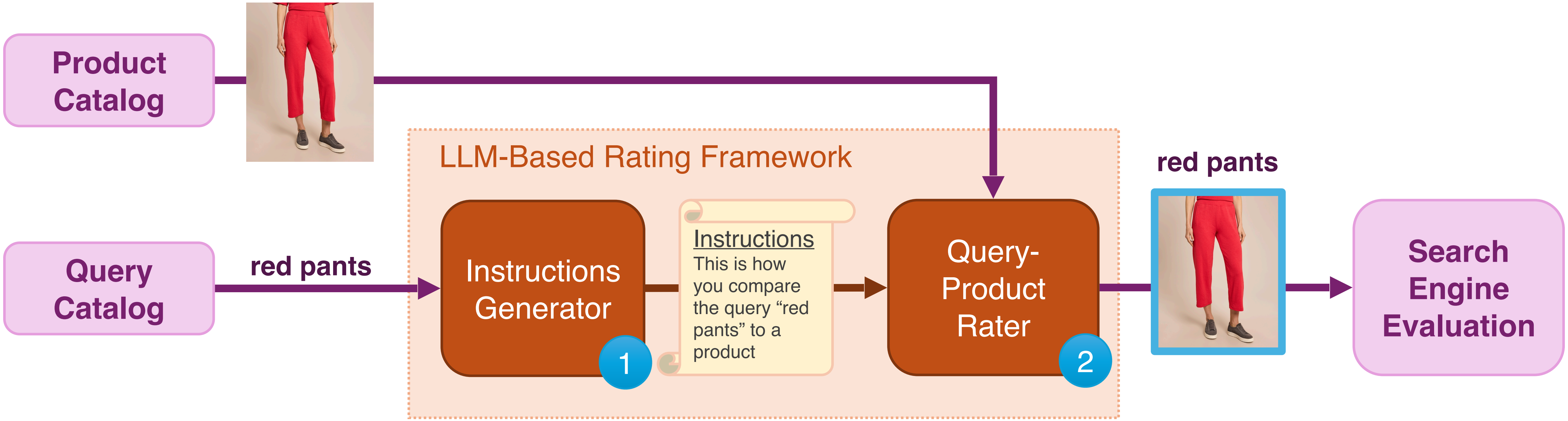
This is how you compare the query “red pants” to a product

Query-Product Rater 2

red pants

Search Engine Evaluation

	Unique Queries (DE)	**Unique Products** (DE)	**Unique Query-Product Pairs** (DE)	**Costs to generate all ratings** (MLLM-multi DE)	**Time to generate all ratings** (MLLM-multi DE)	**Costs to generate all ratings** (humans)	**Time to generate all ratings** (humans)
Zalando 👠	500	8,076	10,000	€72	10 minutes	€15,000	3 weeks



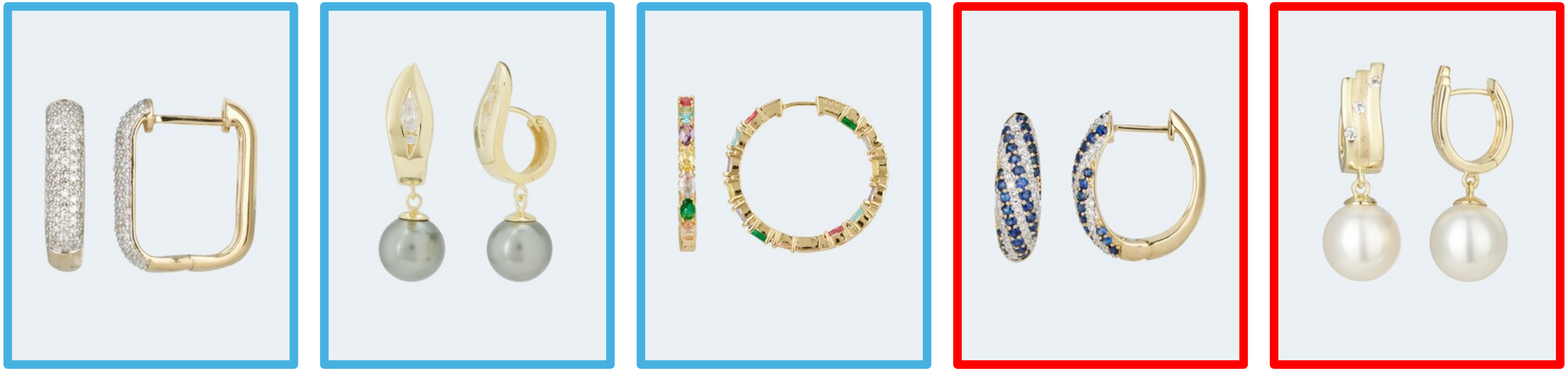
This is how you compare the query “red pants” to a product

	Unique Queries (DE)	**Unique Products** (DE)	**Unique Query-Product Pairs** (DE)	**Costs to generate all ratings** (MLLM-multi DE)	**Time to generate all ratings** (MLLM-multi DE)
Zalando 👠	500	8,076	10,000	€72	10 minutes
Me 😊	87	843	9,515	€70	18 hours

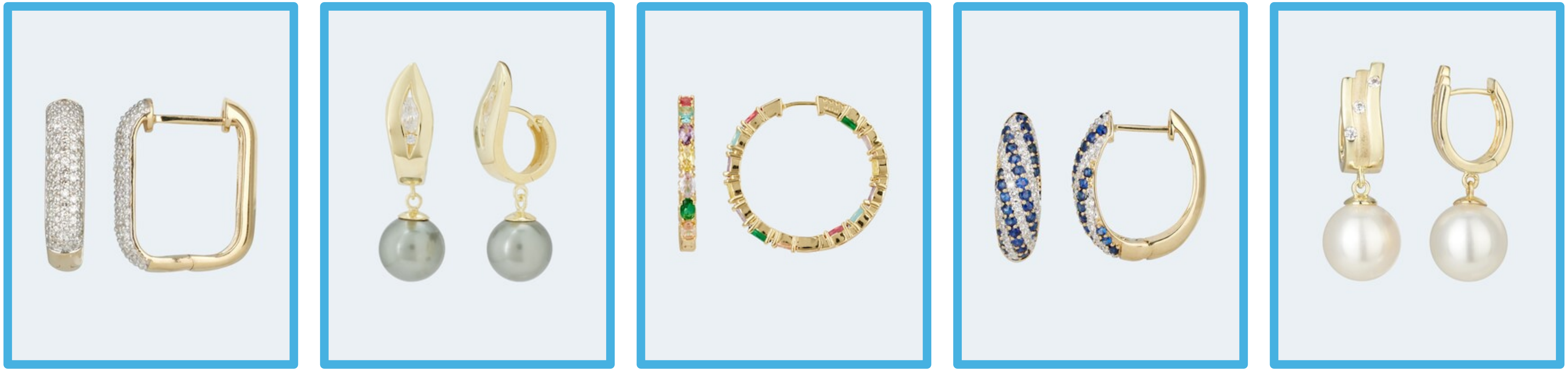
HIGHLIGHTS

Highlight #1: LLMs really don‘t fatigue and humans really do!

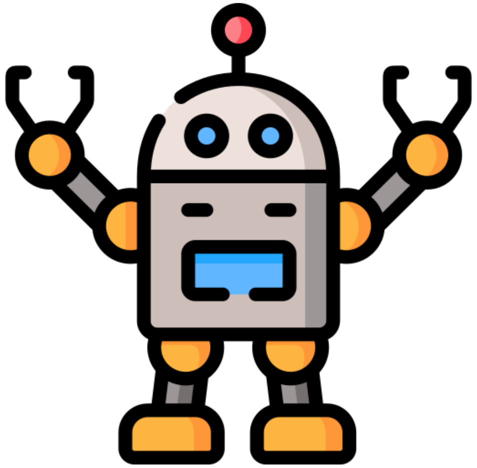
human rater:
“hoop earrings”



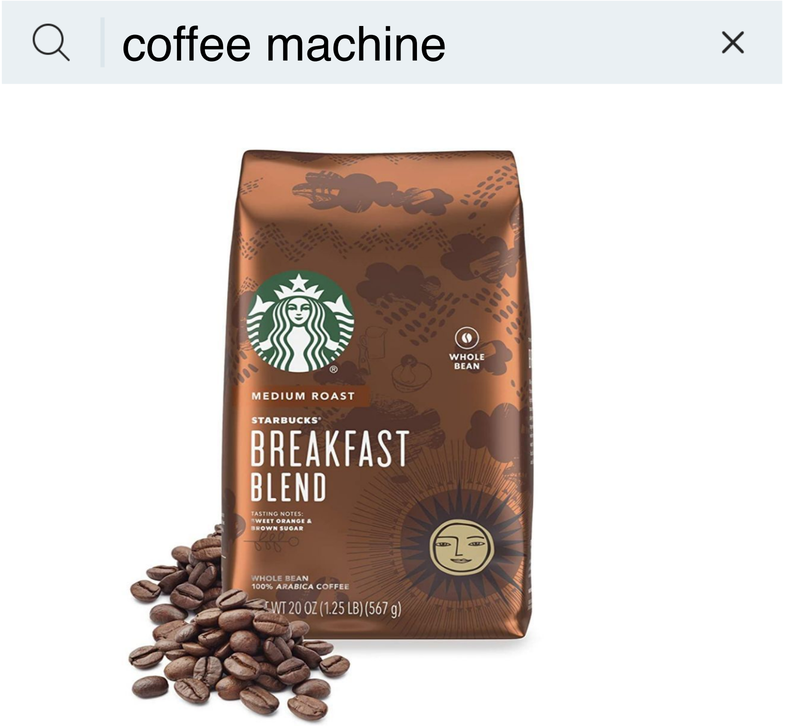
LLM rater:
“hoop earrings”



Highlight #2: If I made a mistake when setting up the task…

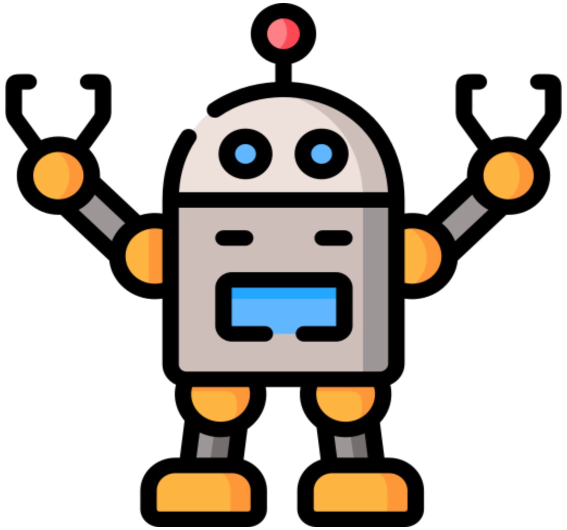


1. Relevant
2. Irrelevant



coffee machine

Highlight #2: … I could redefine the task instructions easily!

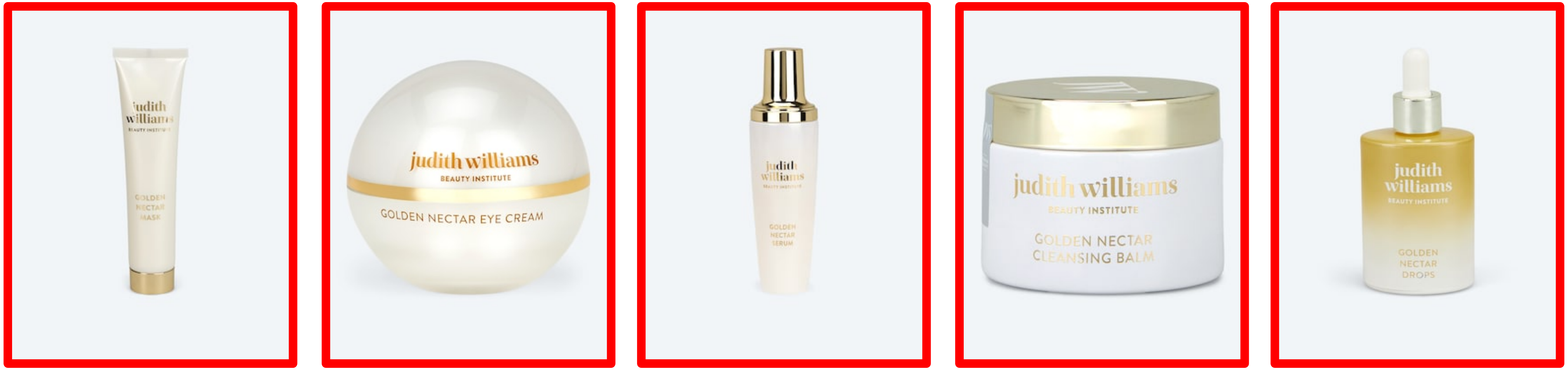


1. ~~Relevant~~
2. ~~Irrelevant~~

1. Relevant
2. Acceptable Substitute
3. Irrelevant

Highlight #3: Retrieving explanations is easy!

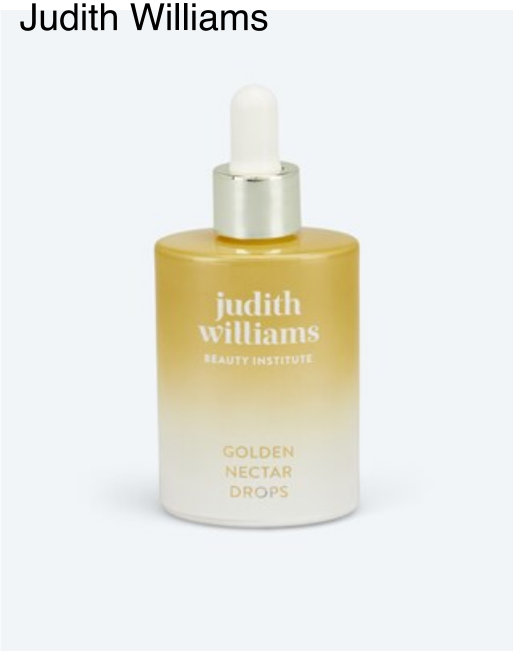
LLM rater: "golden nectar"



Highlight #3: Retrieving explanations is easy!

Golden Nectar Drops

Judith Williams



Query

golden nectar

Judgement Explanation

Relevancy explanation: 1. Must-have: Product is a golden-colored liquid or substance The product image shows a bottle containing a golden-colored liquid, which meets this requirement. 2. Must-have: Product is edible or drinkable The product is clearly a cosmetic item called "Golden Nectar Drops" by Judith Williams Beauty Institute. It appears to be a facial serum or oil, which is not intended for consumption. This does not meet the requirement of being edible or drinkable. While the product's name and color are reminiscent of golden nectar, it is a cosmetic product rather than an edible or drinkable substance. Therefore, it does not fully meet the query requirements. Label: Irrelevant

Judgement Label

irrelevant

Conclusion

IMO worth the investment

- No longer solely dependent on A/B tests
- The time and cost savings are just so huge
- No fatigue
- Better possibilities for iterative task definition improvements
- Can retrieve explanations and reasoning easily

Caveats

- Does not work “out of the box” for domain-specific queries
- Need other infrastructure (like human annotations) to evaluate the LLMs against to ensure quality, so not as cheap as it seems at first

Thank you!
Questions?