stackoverfit

DATATHON by RISTEK

Integrasi Named Entity Recognition dan Trend-Based Query Expansion sebagai Upaya Peningkatan Relevansi Search Engine

Presented By:

Hafid Sasayuda Ambardi | UGM

Muhammad Dafa Wisnu Galih | UGM

Joseph Greffen Komala | UGM

APA ISU DI E-COMMERCE?

keluhan ahli industri

Keep-up-to-date

Kurangnya Insight

Jadi Dukun

Multimodal ↗

Pencarian product menggunakan berbagai input (e.g. teks & gambar)

Penelitian Harus Dekat Dengan Dunia Asli

Melihat Sebelum Membaca

Relevan, Akurat, Semantik

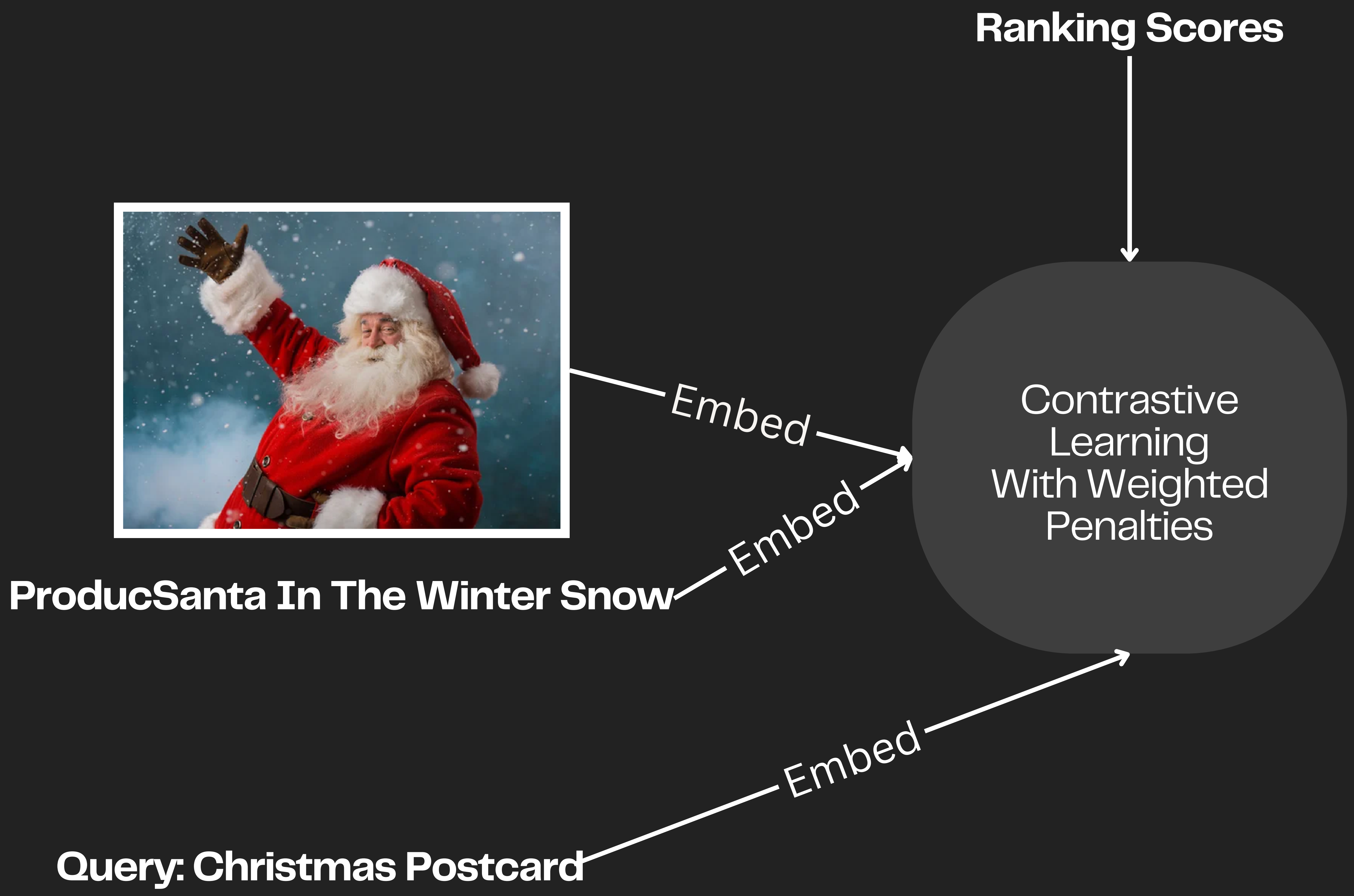
GCL Model

Generalized Contrastive Learning

Multi Input

Weighted Penalty

Fine Tuning with Search Engine Data



Marqo-GS-10M Dataset

IMAGE + TEXT

100
RESULTS / QUERY

Labeled with Scores

15.000+
QUERIES

1.000.000+
PRODUCTS

Implementasi

GCL EMBEDDINGS

Image Embedding **Text Embedding**

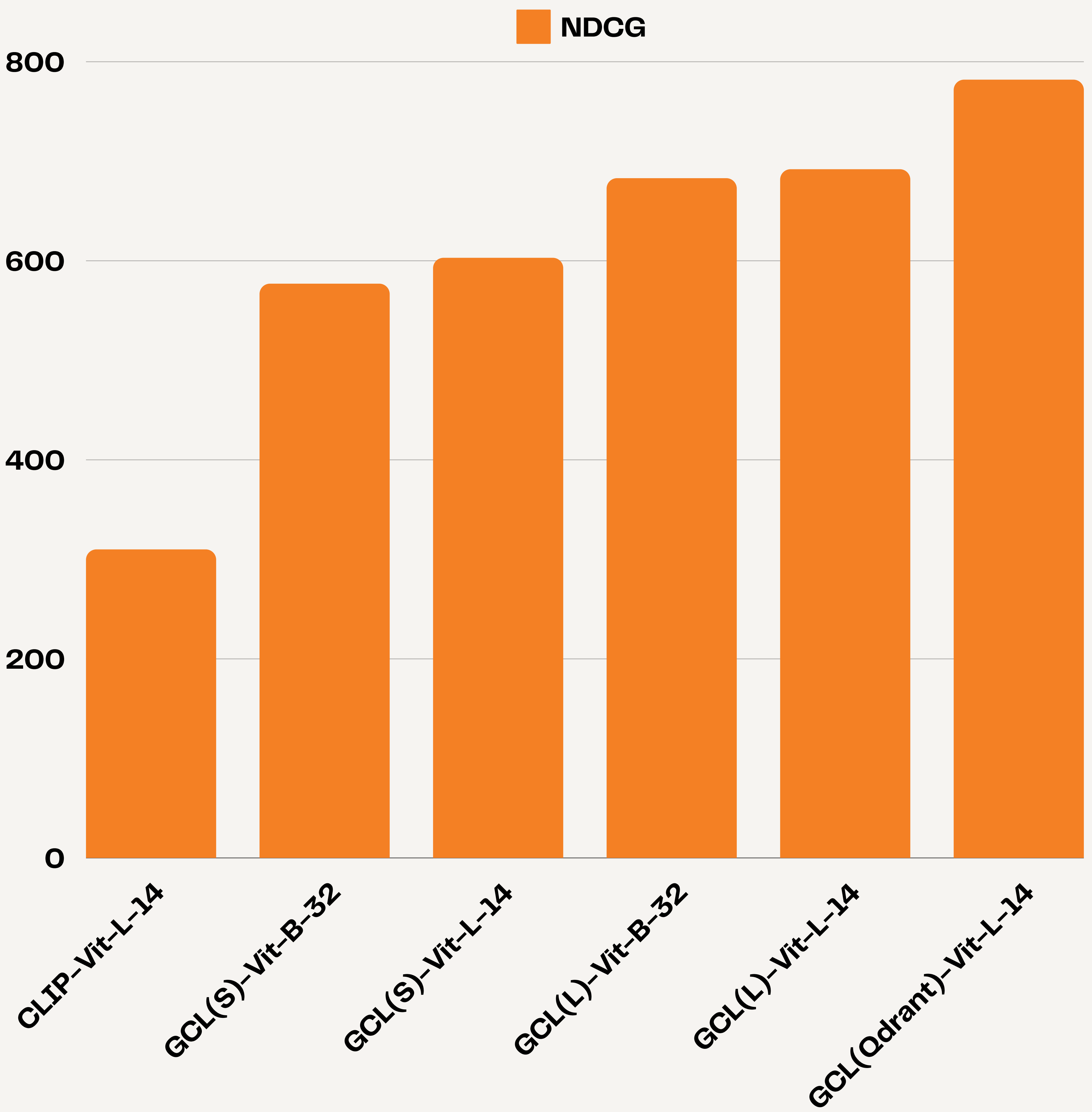
[[4, 5, 2, 4, 5, …, 9, 4, 2], [4, 5, 2, 4, 5, …, 9, 4, 2]]

(Multi Vector)

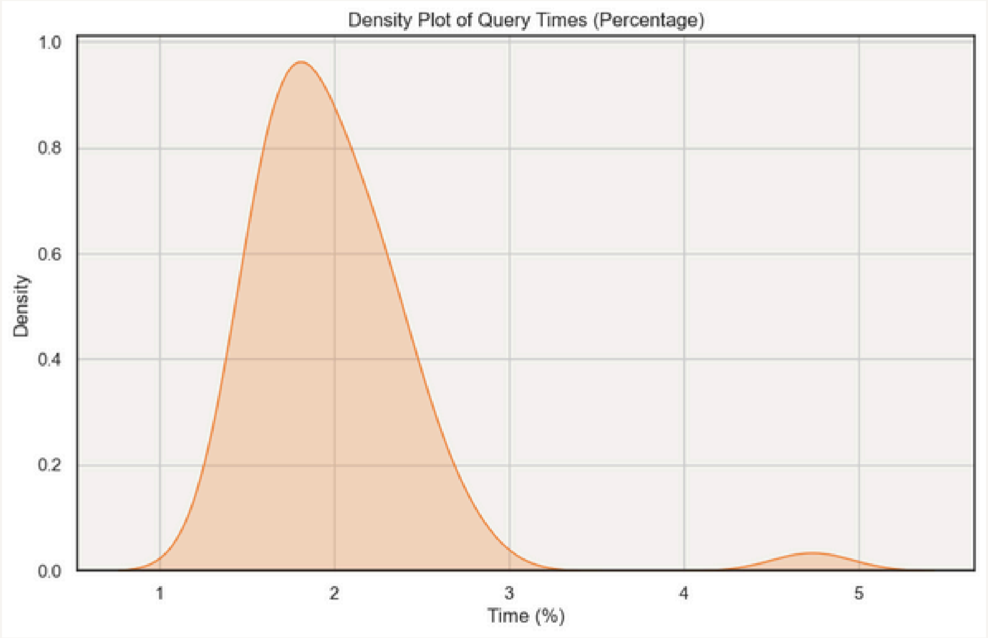
MAX SIM

RESULTS

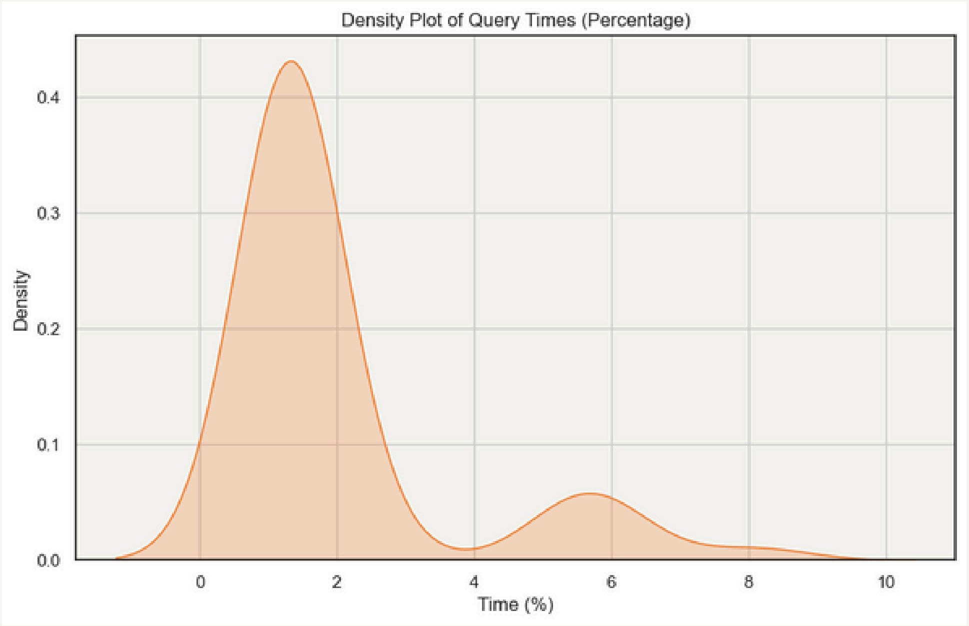
Evaluation



34 MS



11 MS



UDAH?

"A natural language processing (NLP) method that extracts and **categorizes** important information in text known as named entities."

named entity ↗ recognition (ner)

intuitively,

embedding vektor
yang dapat dipahami
oleh manusia.

model, data, and fine tuning ↘

Model

QueryNER by bltlab

- Model baseline adalah bert-base-uncased.
- 108,919,331 parameters.
- Di-fine tuning dengan data beranotasi manual.

Data

Query History

- Dianotasi oleh model GPT-4o.
- 13,592 baris dengan total 140,192 tokens.
- Split 80:20 (train:test)

Finetuning

Trainer API from Huggingface 🤗

- Untuk meningkatkan performa model terhadap data spesifik yang diberikan.

finetuning result ↘

metrics	baseline	finetuned
loss	5.972	**0.425**
accuracy	0.307	**0.907**
f1 score	0.312	**0.907**

300%

peningkatan
performa model

NER in action! ↘

	Affordable	chic	gowns	for	formal	scene
Baseline Model	modifier	creator	core_product_type	O	modifier	modifier
Fine tuned Model	price	modifier	core_product_type	O	occasion	occasion

Class: core_product_type, modifier, creator, product_name, department, occasion, content, UoM, color, time, price, shape, condition, product_number, quantity, origin, O

terus,
buat apa?

Tren

What

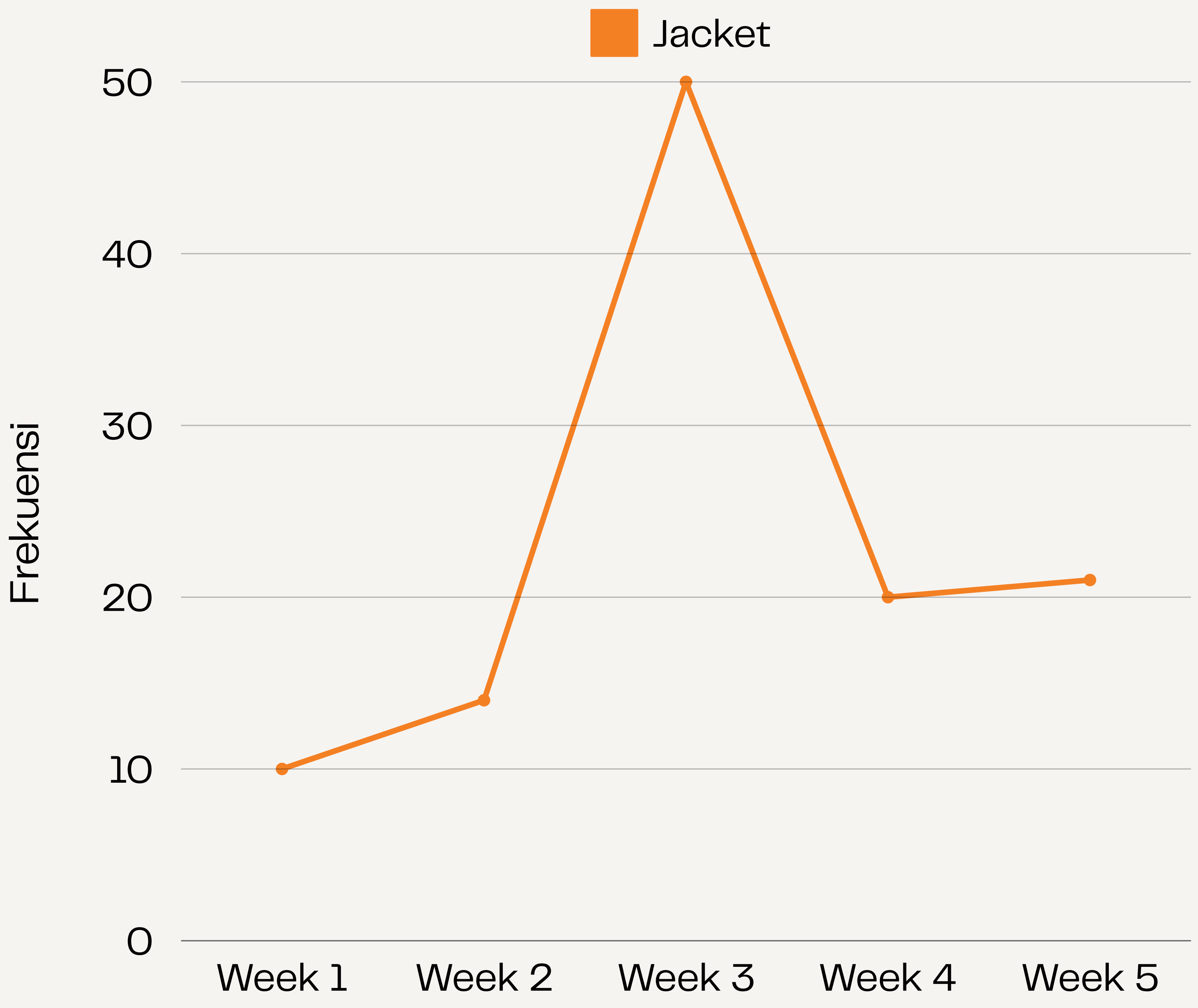
Kondisi saat banyak orang berinteraksi pada kategori barang yang sama

Named Entity Recognition (NER)

Tren

How

Frekuensi pencarian dalam suatu rentang waktu



EMANG CUKUP?!?!

Dataset :

1

Social Media
(e.g., X(Twitter))

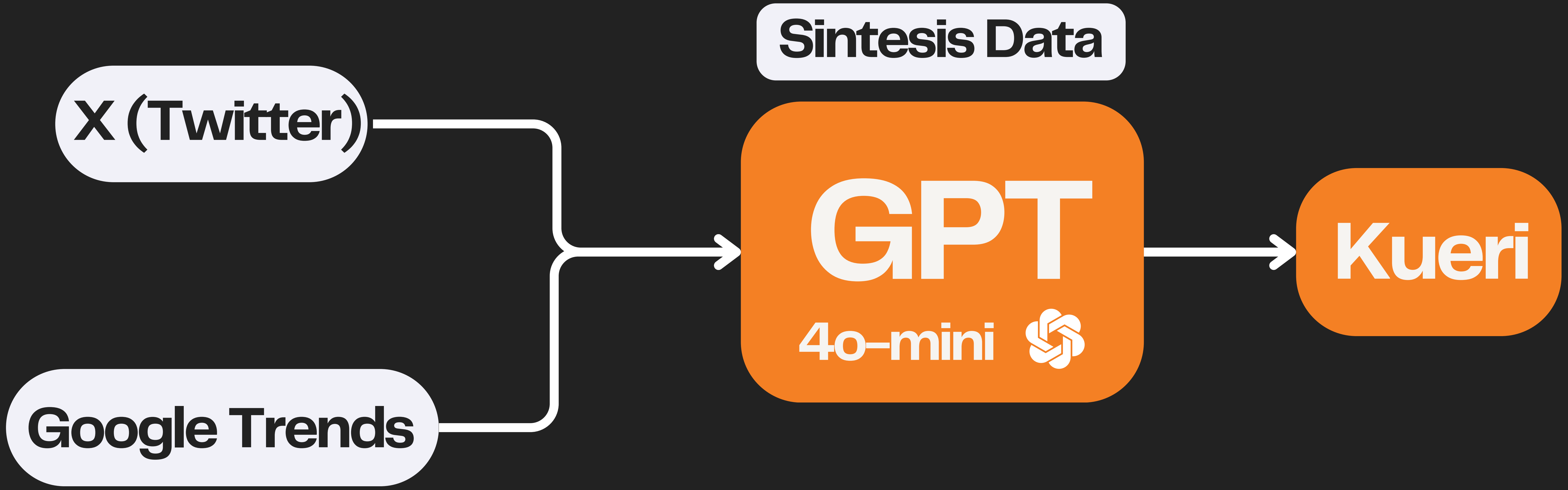
2

Web Search
(e.g., Google Trend)

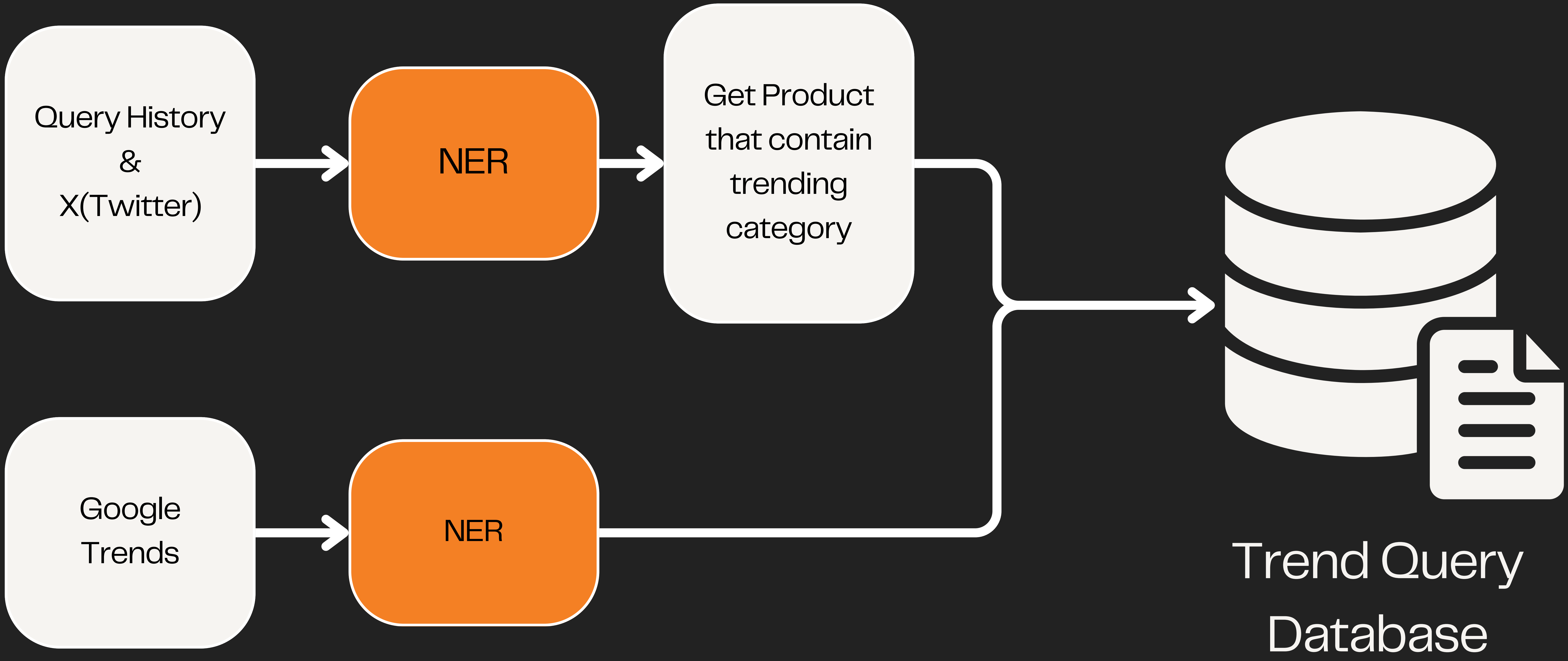
3

Query History

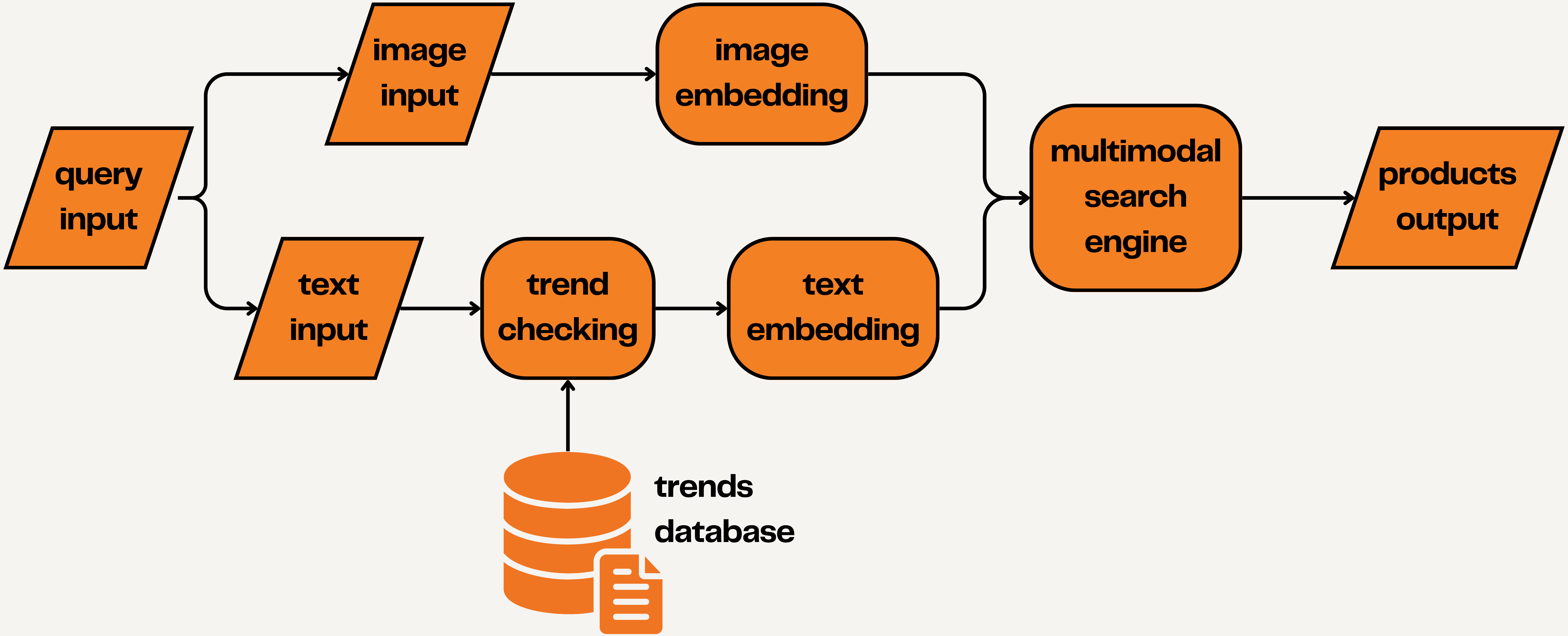
Sintesis Data dengan LLM



Implementasi NER & Trend

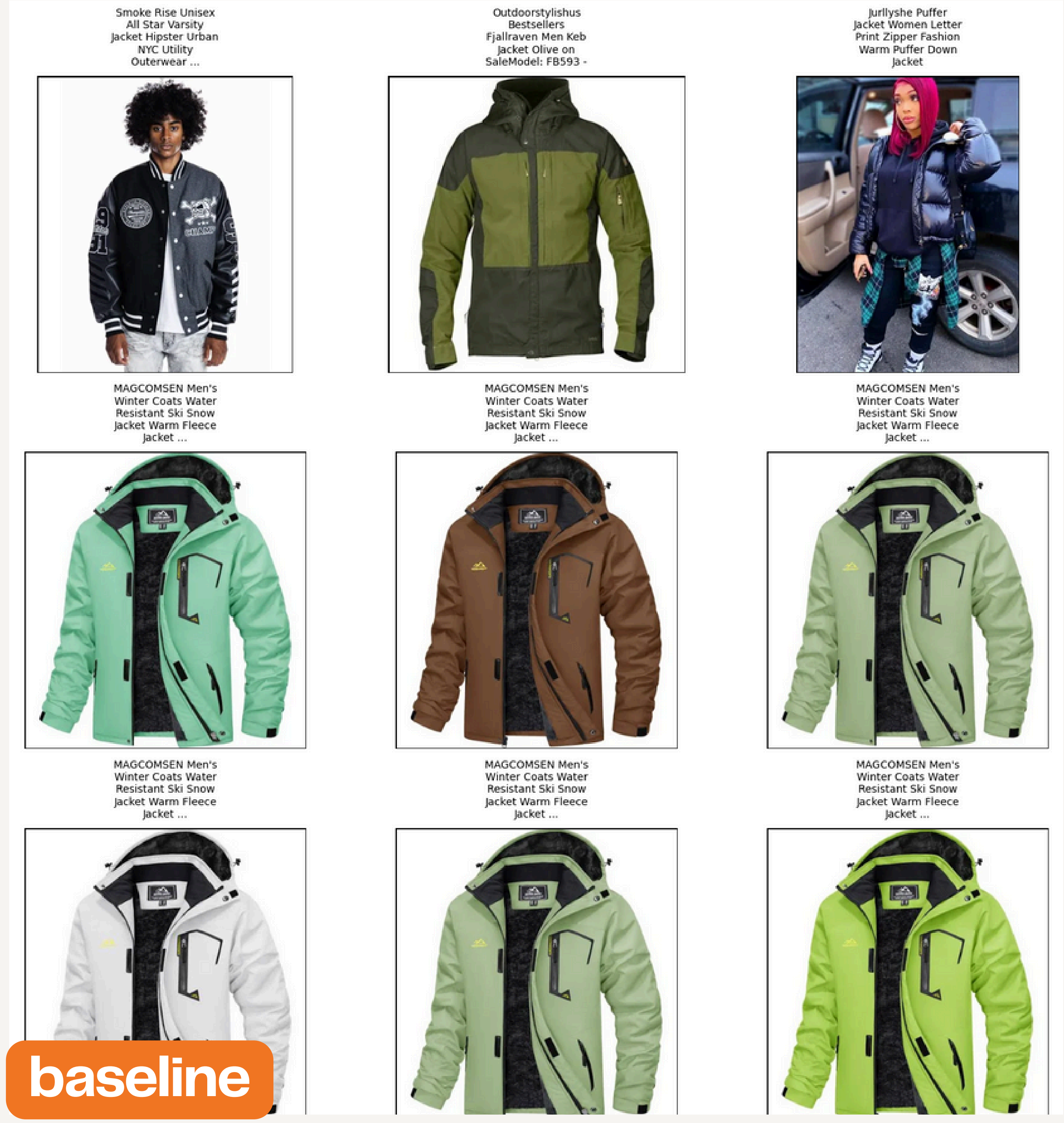


overall flowchart ↘



hasilnya?

query: jackets



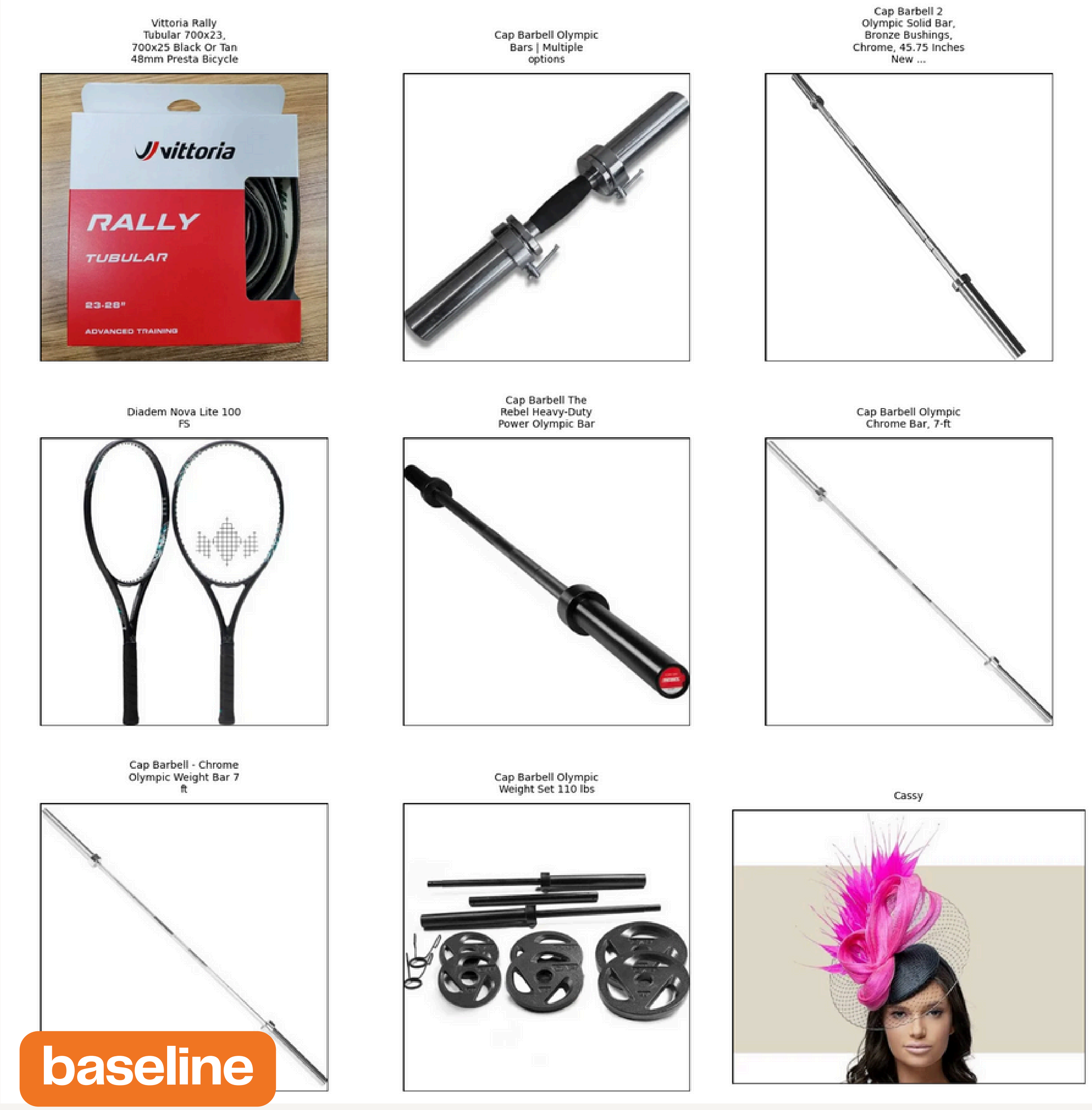
baseline

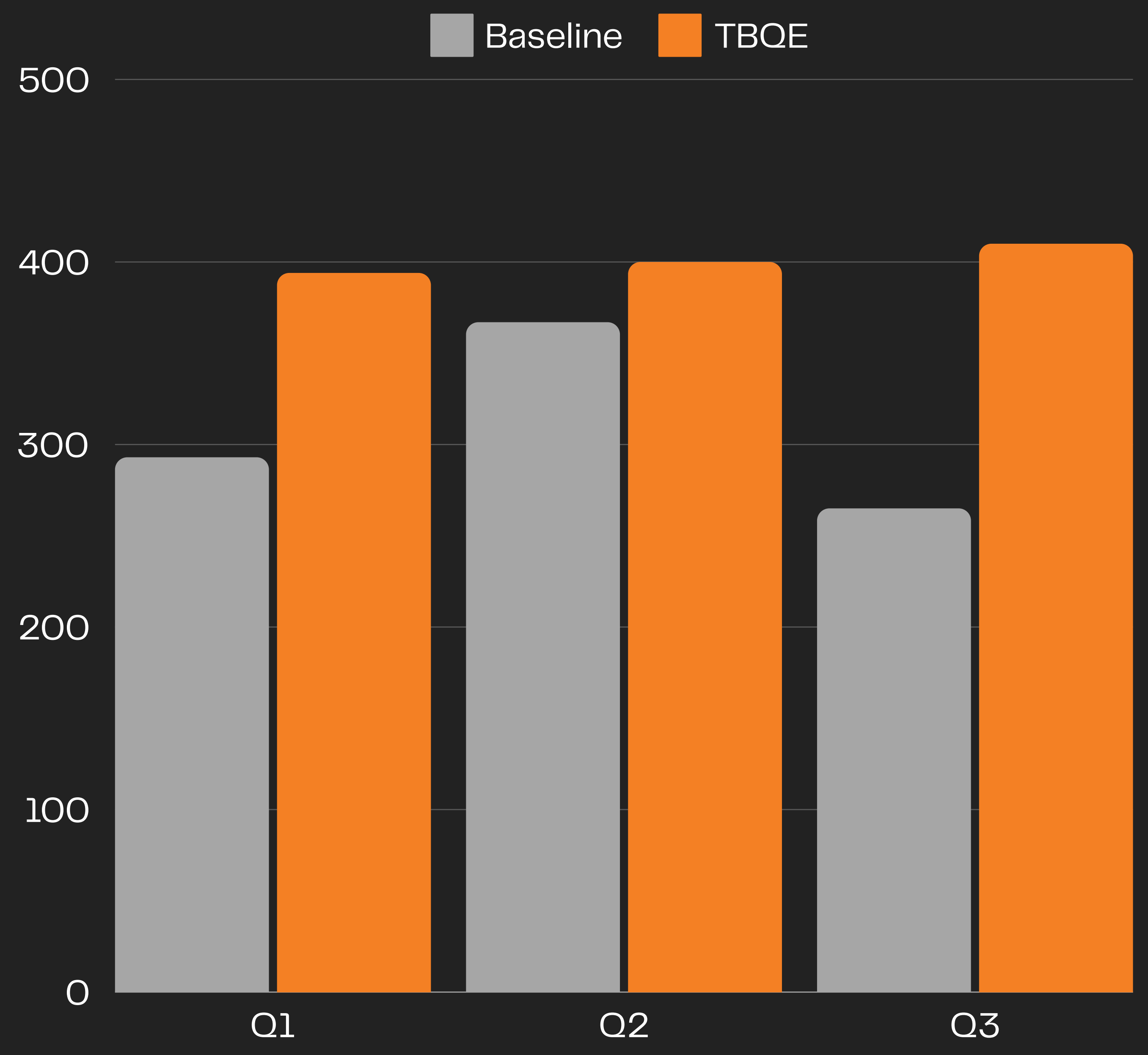
TBQE

trend: winter

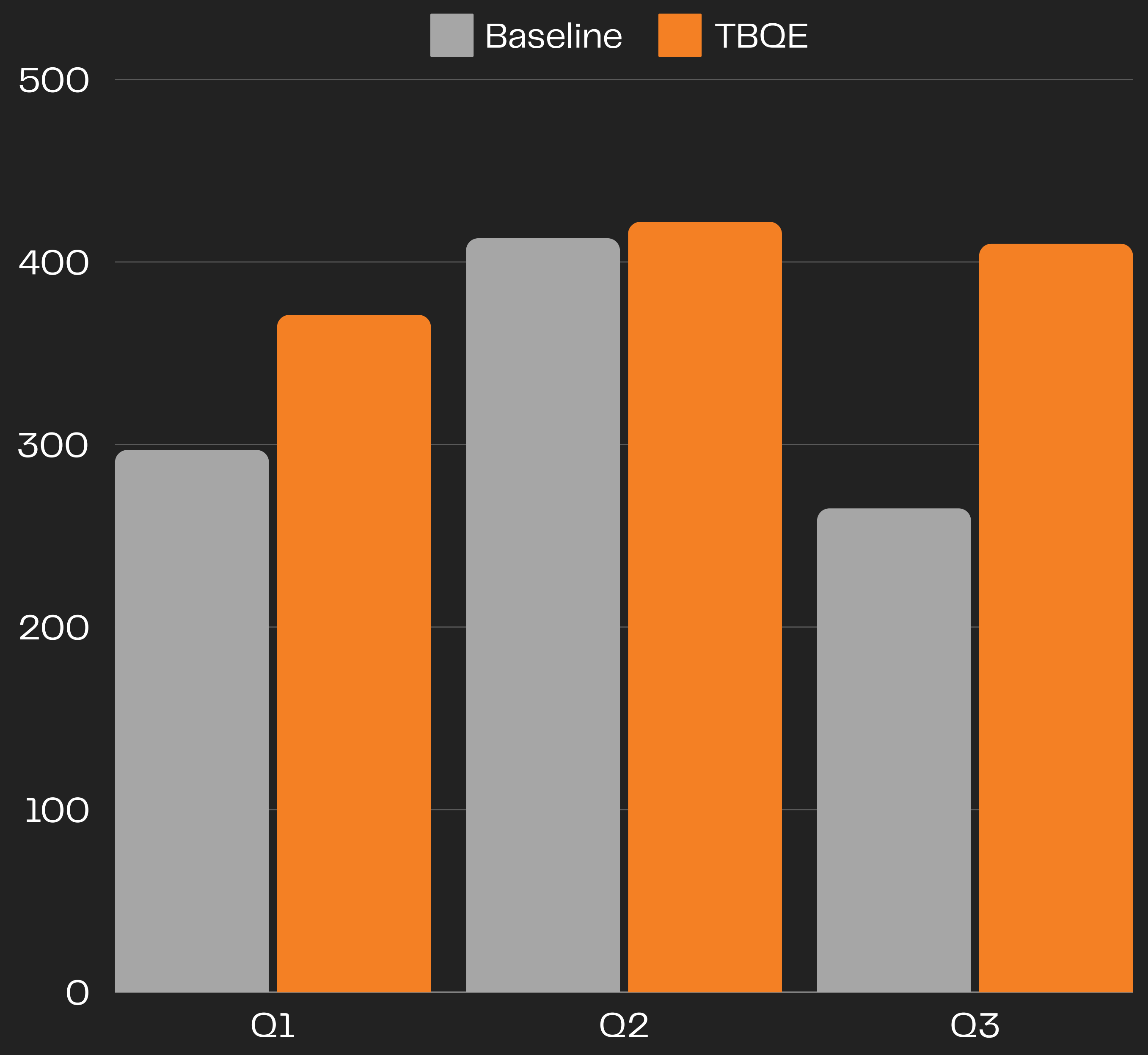
hasilnya? (2) ↘

query: olympics





Skor Trend



Skor Relevansi

TBQE = Trend-Based Query Expansion

A/B testing

Summary

Multi-Modal

NER → Trend

Isu → SOLVED

stackoverfit

DATATHON
by RISTEK

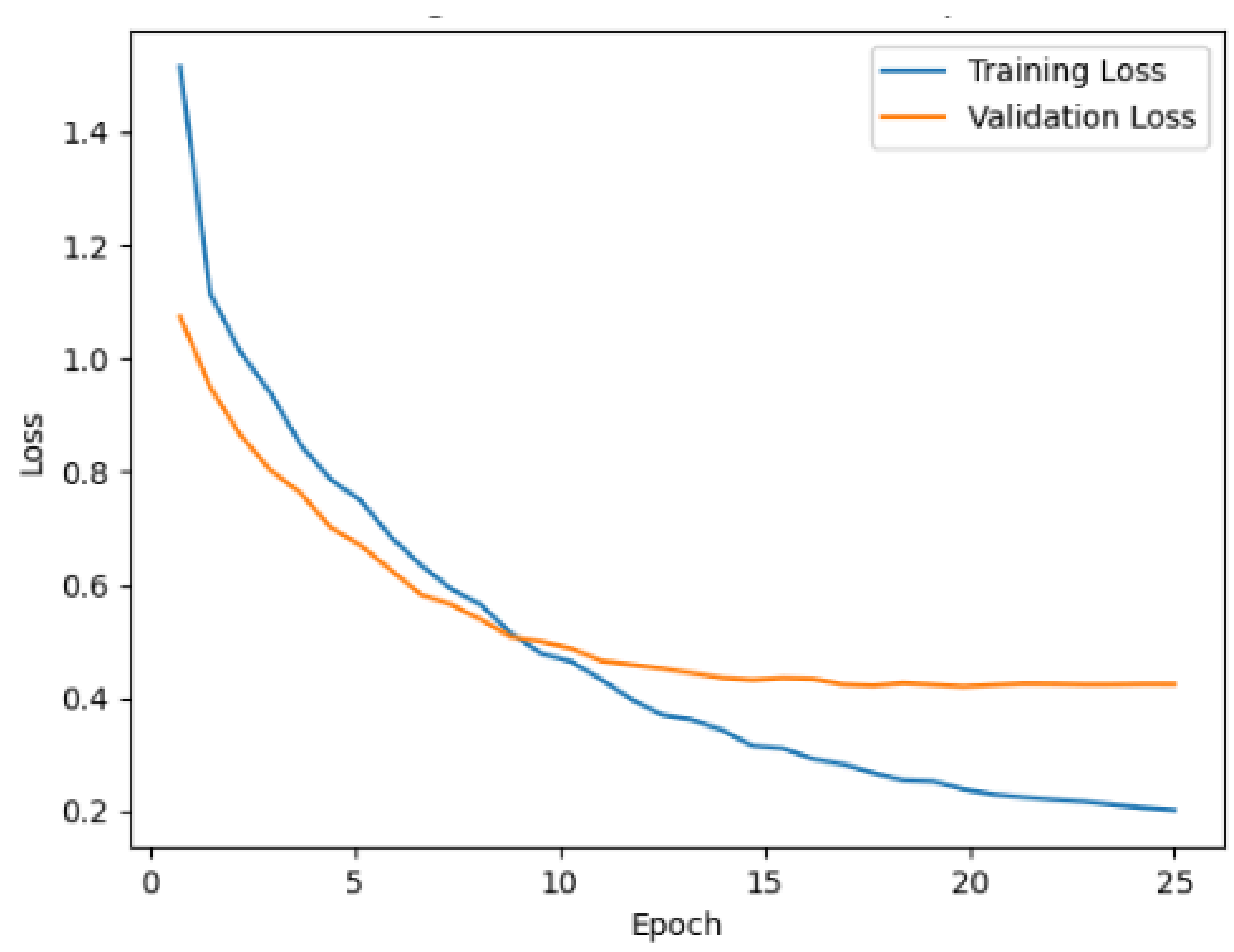
Thank You for Your Attention

QnA Session

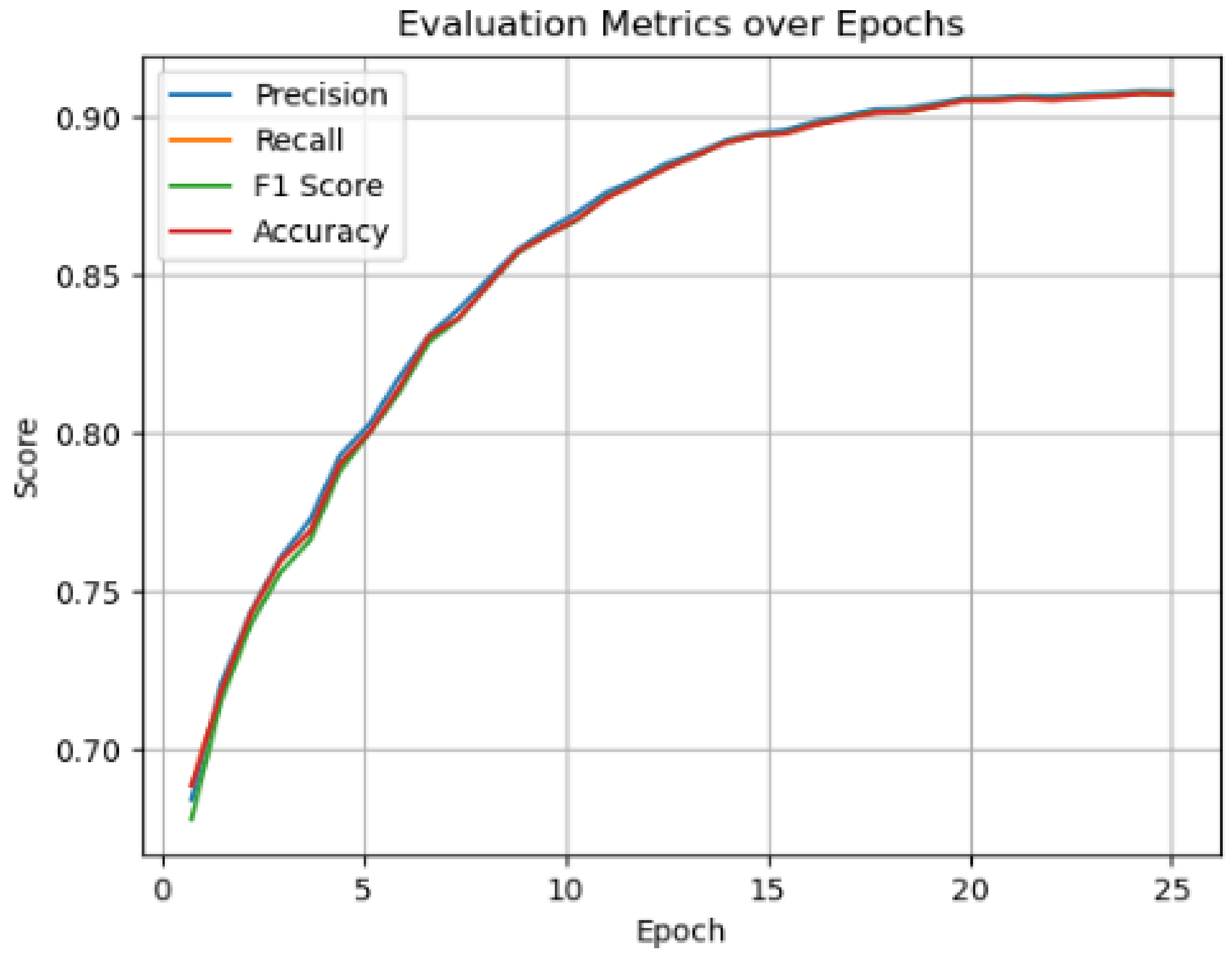
Slides Tambahan (jika diperlukan)

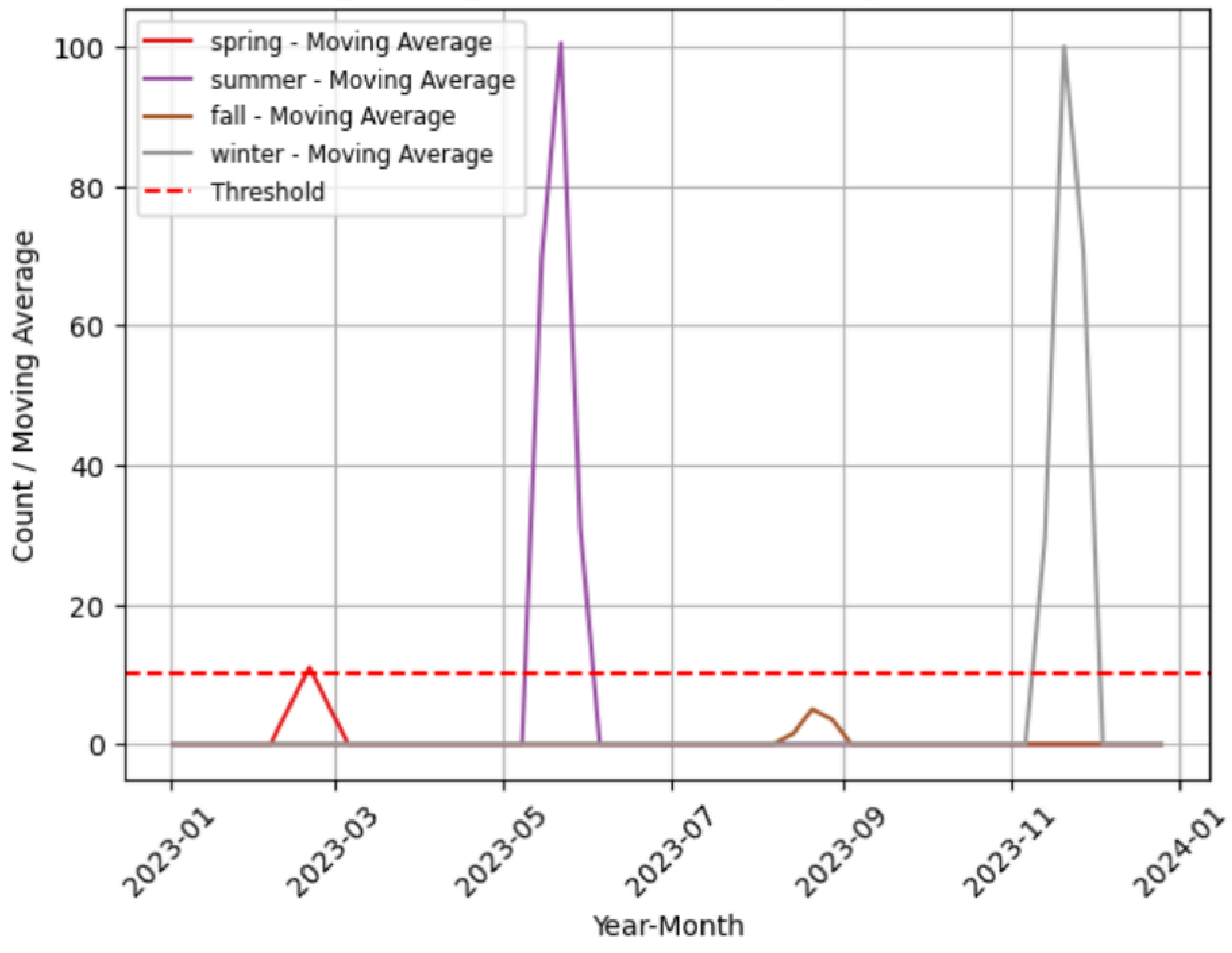
TABEL II. PERFORMA MODEL QUERYNER BASELINE DAN FINE TUNED

Metriks Evaluasi	Model *Baseline* QueryNER	Model *Fine Tuned* QueryNER
Loss	5.972	**0.425**
Accuracy	0.307	**0.907**
Precision	0.394	**0.908**
Recall	0.307	**0.907**
F1 Score	0.312	**0.907**

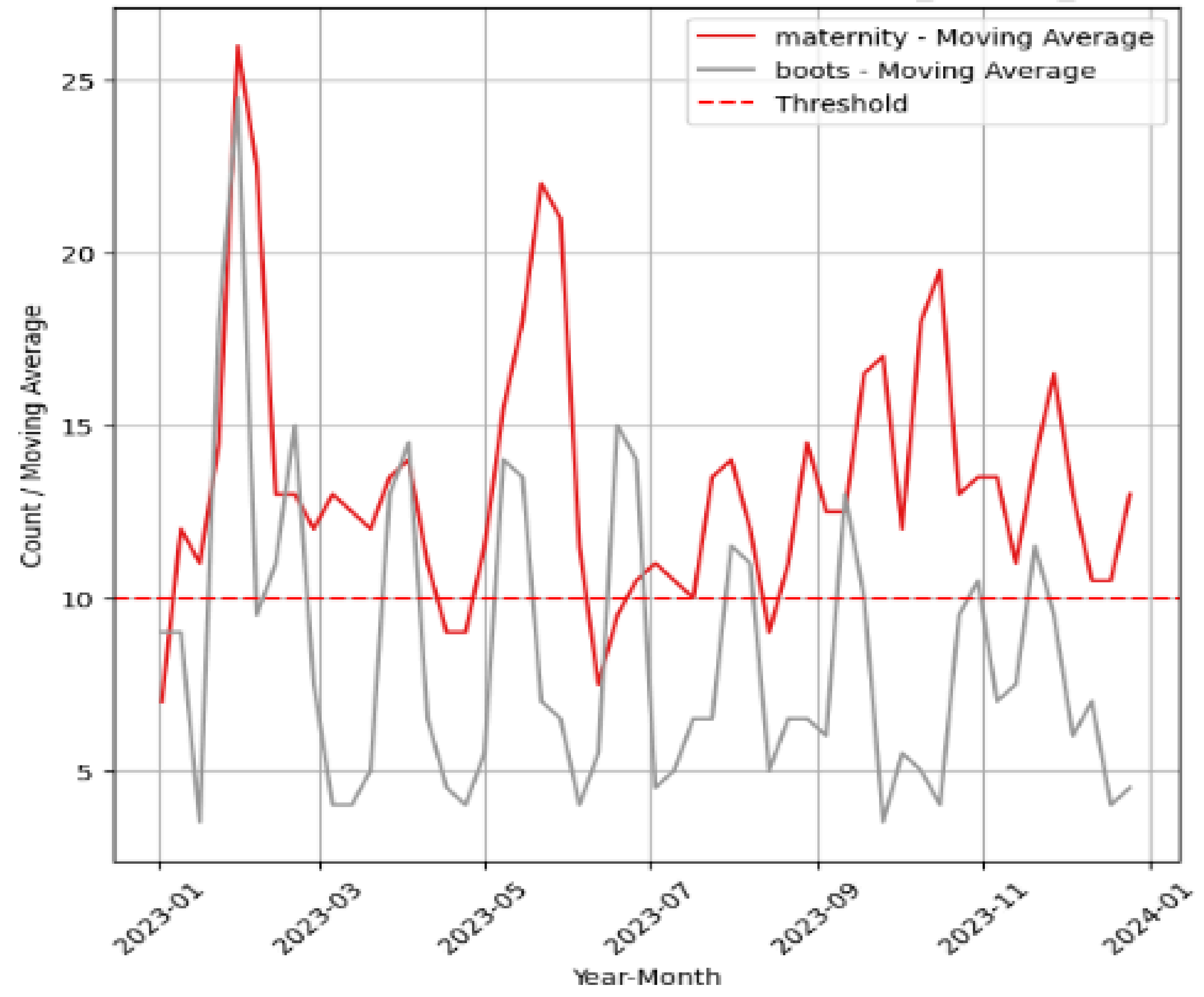


Gambar 8. Diagram Alir *Fine Tuning* Model QueryNER





Gambar 10. Diagram Alir *Fine Tuning* Model QueryNER



$$nDCG_k = \frac{DCG_k}{IDCG_k}$$

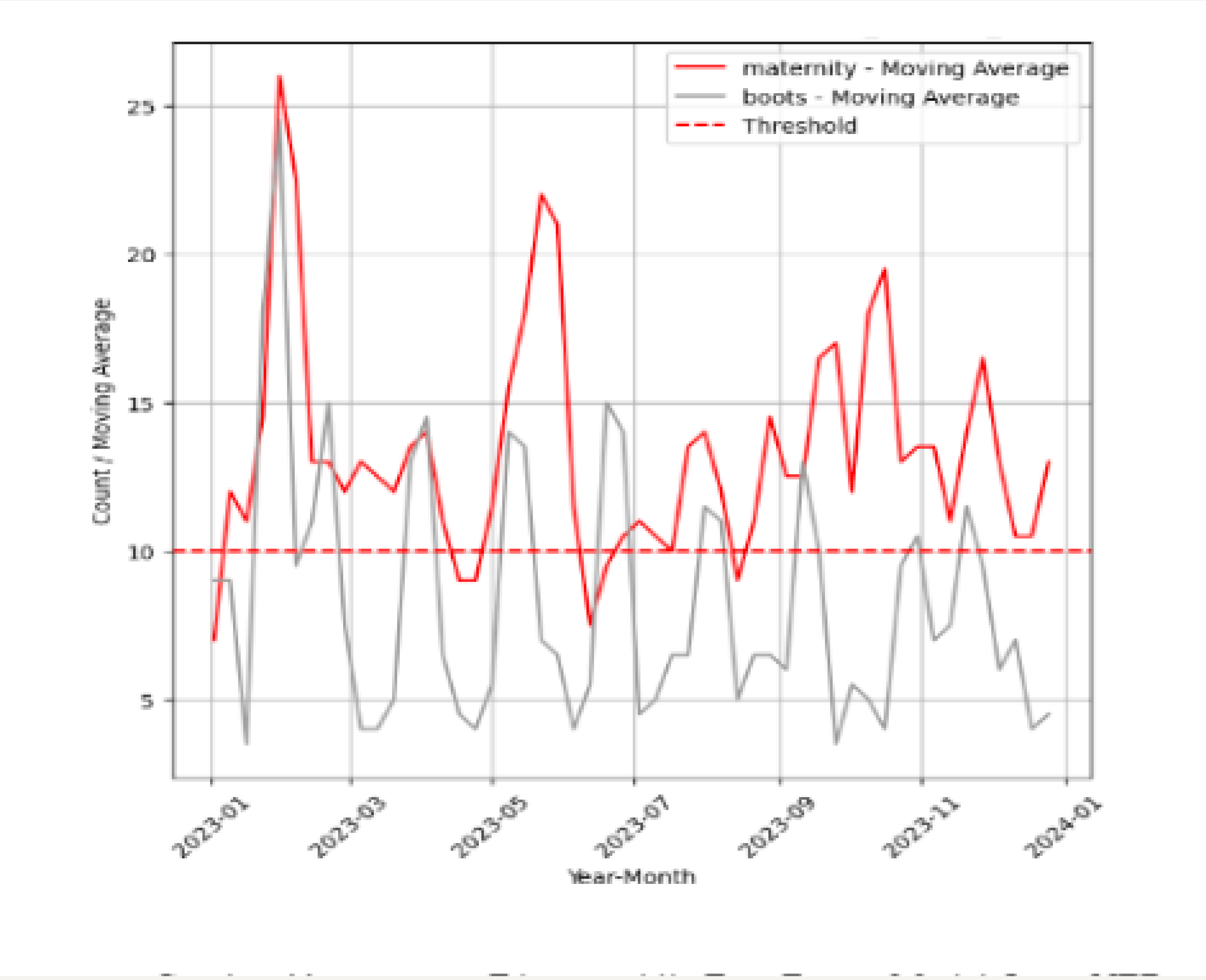
$$DCG_k = \sum_{i=1}^{k} \frac{rel_i}{log_2(i+1)}$$

$$IDCG_k = \sum_{i=1}^{k} \frac{rel_i}{log_2(i+1)}$$

$$Recall = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$F1\, score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$




```
FUNCTION get_trend_overall(data_dict, query_ts, df,
    ↪ threshold):
    CONVERT query_ts to weekly period
    INITIALIZE result_dict as empty list

    FOR EACH category, values IN data_dict:
        FOR EACH value IN values:
            FILTER df WHERE COLUMN category CONTAINS
                ↪ value

            CREATE 'week_ts' column AS weekly period
                ↪ of 'ts' column

            GROUP filtered df BY 'week_ts' AND COUNT
                ↪ occurrences

            CREATE all_weeks AS weekly period range
                ↪ from '2023-01-01' to '2023-12-31'
            REINDEX grouped DataFrame TO INCLUDE
                ↪ all_weeks WITH MISSING VALUES
                ↪ FILLED WITH 0
            RECALCULATE rolling moving average

            GET moving average value FOR
                ↪ query_timestamp

            IF moving average value EXCEEDS threshold:
                ADD value TO result_dict

    RETURN result_dict
```

Kasus Negatif. Selain itu, kami mengumpulkan 1.000 kueri yang tidak berhubungan dengan mode wanita, seperti senjata, kalender, speaker, dan sebagainya, serta memeriksa apakah mesin pencari kami dapat menyaring produk yang memang tidak ada dalam katalog. Dari semua kueri tersebut, 78,30% memiliki nilai kesamaan di bawah 1,6.