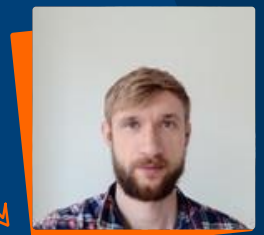
idealo

Hybrid Search at idealo

MICES 2026



Gennady Shabanov

Machine Learning Engineer

Atakan Filgöz

Machine Learning Engineer

Where We Started

idealo

Talks that shaped our journey

The Digitec Galaxus Vector Search Journey
Abel Camacho Guardian · Joel Widmer
MICES 2025

Vectorizing Consumer Electronic Goods
Ruchi Juneja · Johannes Peter
MICES 2024

How Semantic Search Projects Fail
Roman Grebennikov
MICES 2024

Agenda

idealo

- Introducing idealo
- Motivation
- Vector Search - Finetuning
- Hybrid Search Strategies
- Model Serving and Challenges
- Key Takeaways

About idealo

Germany’s 4th largest ecommerce website

50.000+ shops

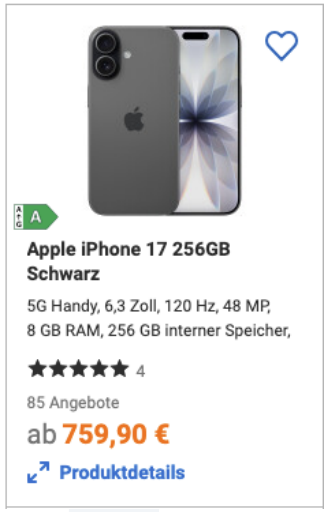
Active in 6 countries (DE, AT, ES, IT, FR, UK)

8M+ products
500M+ offers

18M+ app downloads

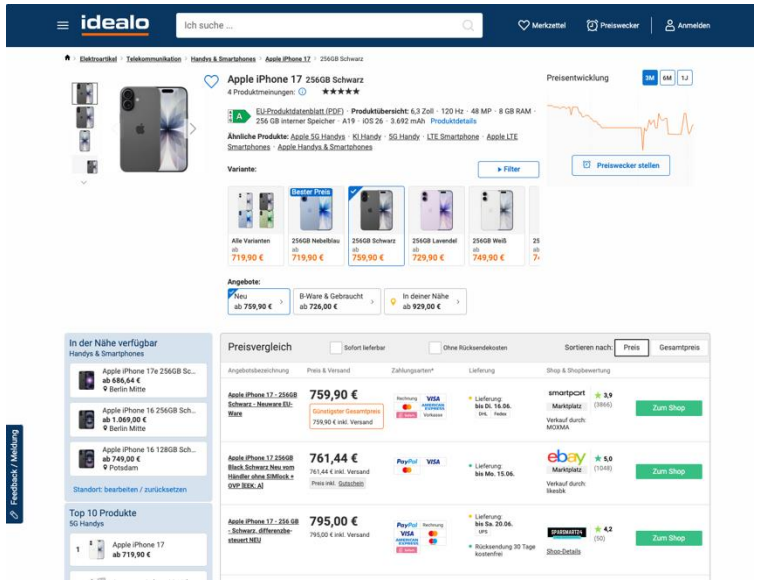
80M+ visits per month

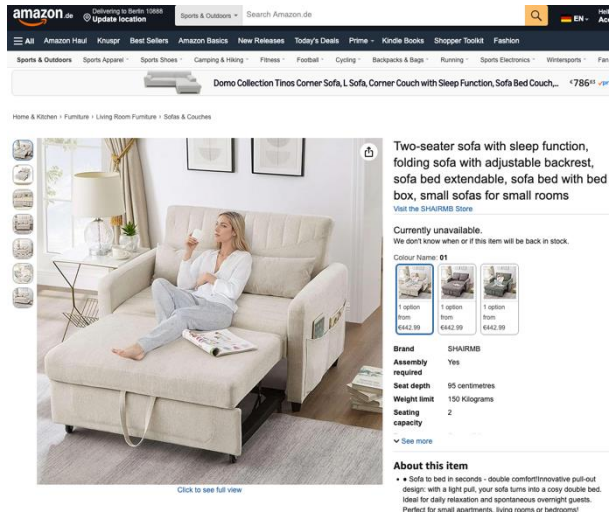
idealo's Catalog



8M+ Products

500M+ Single Offers





Keyword Search @ idealo

- Customized Lucene
- Manually tuned boosts
- Historical global item clicks and query-item clicks affect the boosts

	Bild	Details
Produkt id:203247542 Score :173010 Clickouts (total):4359 Clickouts (query): 5359 Discount:0%		Apple iPhone 15 128GB Schwarz ⊞ Kategorien : Elektroartikel > Telekommunikation > Handys & Smartphones Product-Types : LTE Smartphone, 5G Handy, Phablet
Produkt id:203235721 Score :164960 Clickouts (total):11660 Clickouts (query): 398 Discount:0%		Apple iPhone 15 ⊞ Kategorien : Elektroartikel > Telekommunikation > Handys & Smartphones Product-Types : LTE Smartphone, 5G Handy, Phablet
Produkt id:203247309 Score :163070 Clickouts (total):1307 Clickouts (query): 819 Discount:0%		Apple iPhone 15 256GB Schwarz ⊞ Kategorien : Elektroartikel > Telekommunikation > Handys & Smartphones Product-Types : LTE Smartphone, 5G Handy, Phablet

Keyword Search Pros / Cons

idealo

✓ Strengths

- **Exact matches**
  Model numbers, brands, EANs
- **Fast & low latency**
  Mature, highly optimized retrieval
- **Transparent**
  Easy to inspect and debug

✗ Weaknesses

- **Spelling mistakes**
  Typos break exact matching
- **Model number variants**
  Same product named differently across shops
- **Semantics**
  Can't handle attributes, synonyms, or intent

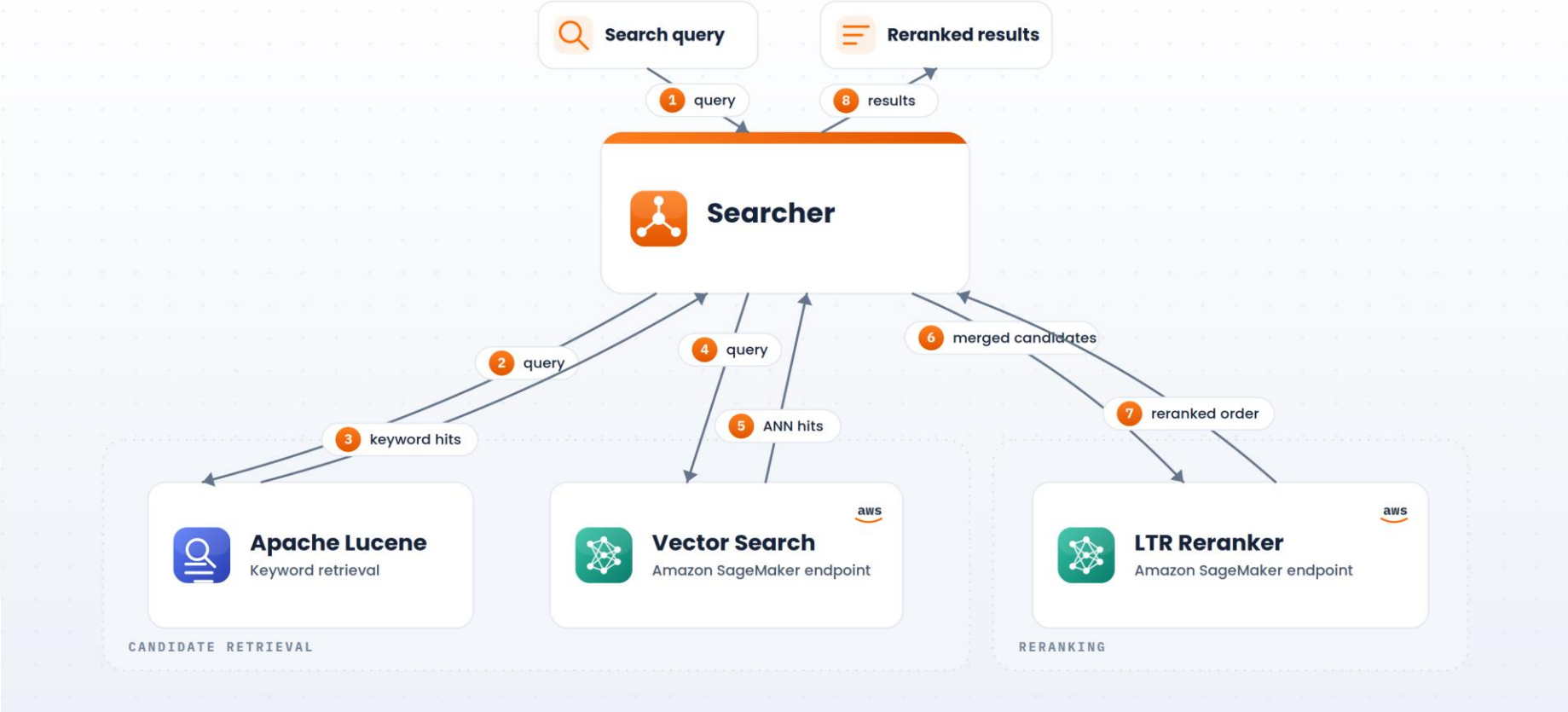
From Keyword Search to Hybrid Search

idealo

- The failures of keyword search are not random:
  - Spelling
  - Semantics (synonyms, attributes, intent)
  - Language
- Keyword-side fixes (rules, synonyms, boosts) only go so far → they don't scale to the long tail
- Data-driven fixes (boosts, click signals) need traffic to tune — head queries have it, the long tail doesn't.
- So why not both?

High Level Overview

idealo



Vector Search – The Approach

idealo

Lots of decisions to be made:

Relevance Labels

Define what makes an item relevant

Pre-trained Models

Start from MTEB Leaderboard

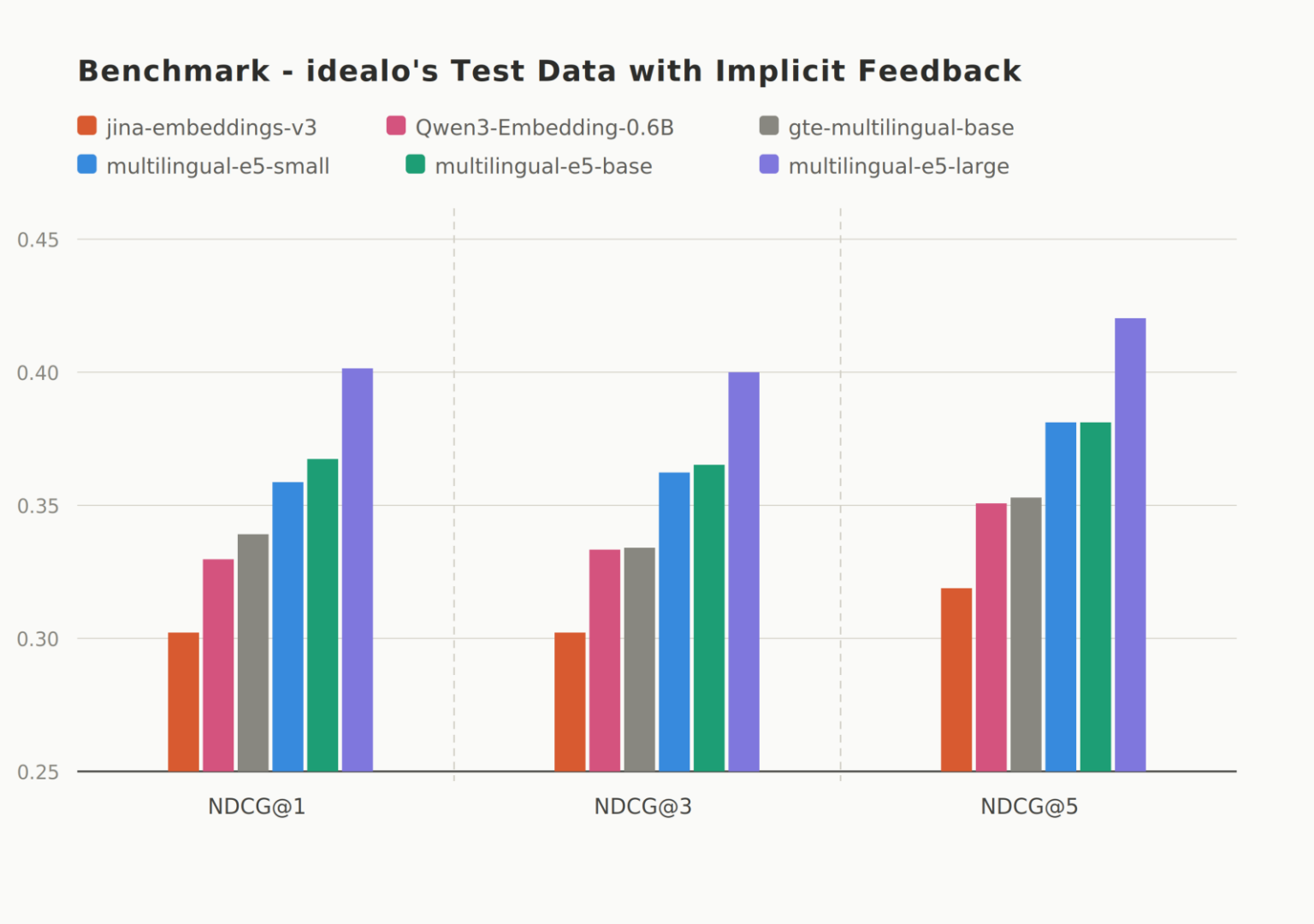
BEIR dataset

Loss Functions

Fine-tune BERT embeddings with contrastive learning

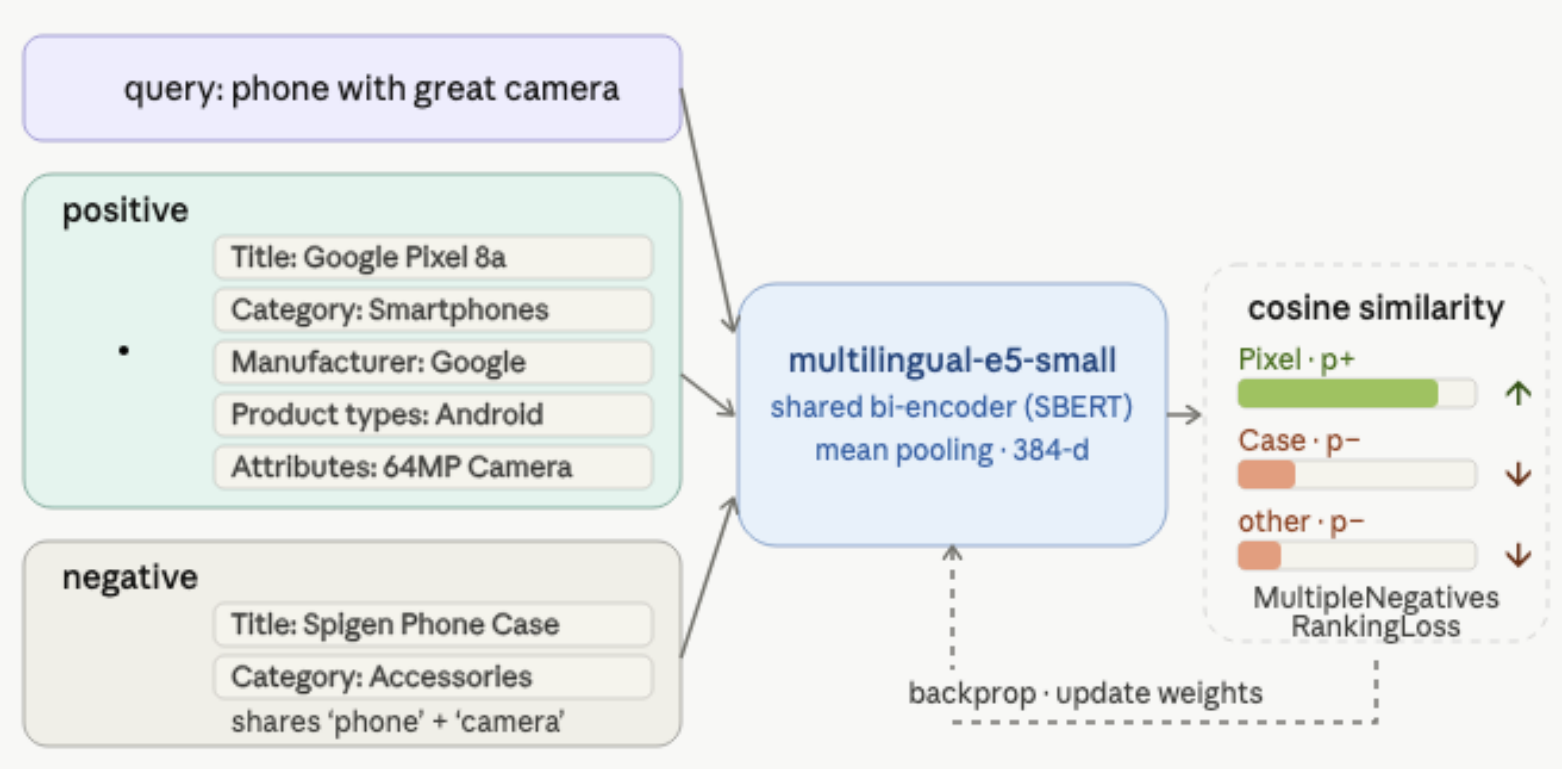
Pre-trained Models

idealo



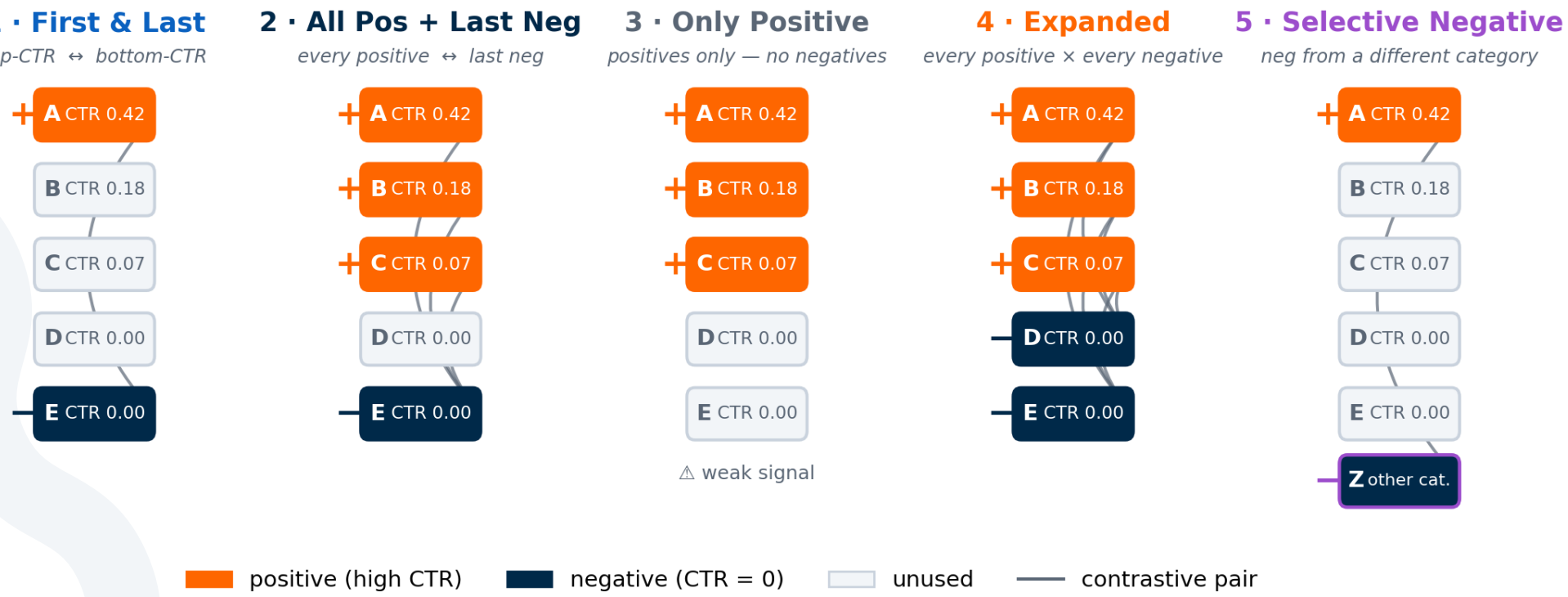
sBERT Finetuning Overview

idealo



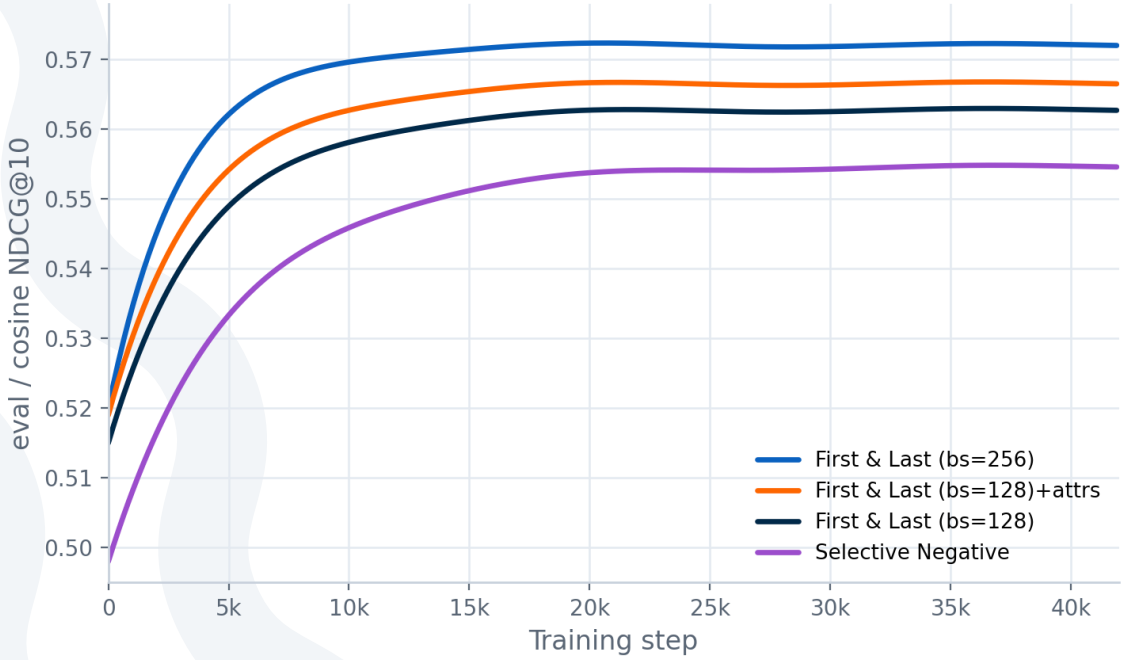
Contrastive Pair Strategies

Model: intfloat/multilingual-e5-small **Loss:** MultipleNegativesRankingLoss **Epochs:** 20

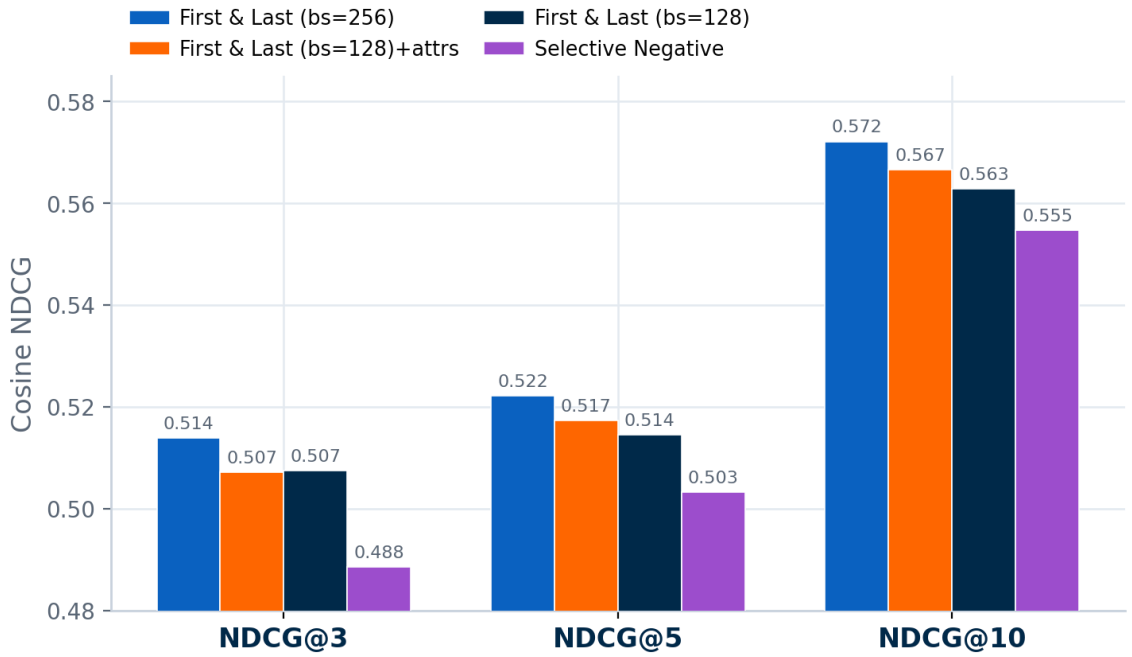


All strategies pair items from the same search result list — **they differ only in which items become positives vs. negatives.**

Top-Performing Strategies — Results



NDCG@10 convergence over training steps



Final cosine NDCG@3 / @5 / @10

Key takeaway: "First & Last" with batch size 256 wins consistently, and adding product attributes lifts scores further. **Only-positive pairs perform worst — hard negatives are crucial.**

False Negatives

Not every query has a true negative

position	type	item	image
1(0)	PRODUCT	Dermapharm Ketozolin 2 % Shampoo (120 ml) id:4324339	
2(0)	PRODUCT	Dermapharm Ketozolin 2 % Shampoo id:5650469	
3(0)	PRODUCT	Dermapharm Ketozolin 2 % Shampoo (60 ml) id:4170000	
4(0)	PRODUCT	Dermapharm Ketozolin 2 % Shampoo (2 x 120 ml id:203304826	
5(0)	PRODUCT	Dermapharm Ketozolin 2 % Shampoo (3x120ml) id:209046325	
6(0)	OFFER	KETOZOLIN 2% Shampoo 2x 120ml Sparset plus id:dcb27a4561678890a7ce47411e5d87c5	

Search results for "Ketozolin Shampoo" — every item is relevant

LLMs – Negative Filtering

idealo

Recap: hard negatives come from search results — we take the least-relevant item with 0 CTR as the negative sample.

⚠ The problem

- 0 CTR ≠ irrelevant — an item can be relevant but simply not clicked.
- When every result is relevant, the lowest-CTR item becomes a false negative.
- Training on false negatives teaches the model the wrong signal.

✓ The solution: filter negatives with an LLM

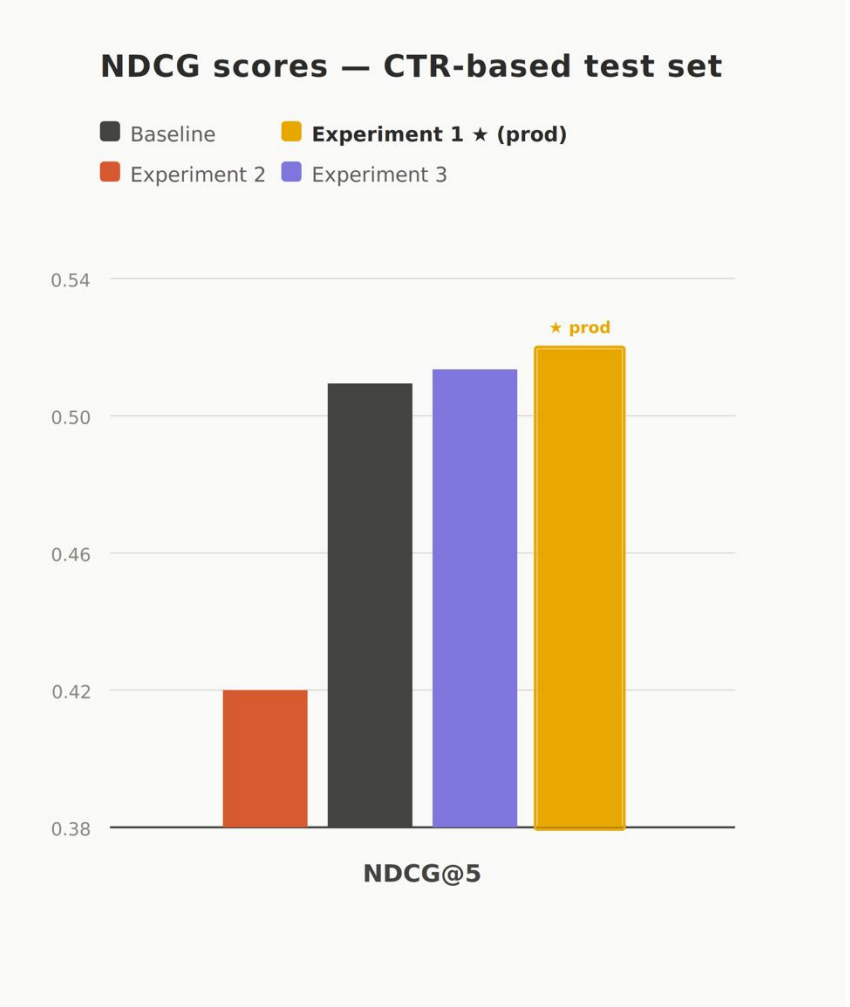
- Before training, an LLM judges whether each candidate negative is truly irrelevant to the query.
- Genuine negatives are kept; false negatives (still relevant) are discarded.
- Handles the non-trivial cases simple CTR rules get wrong.

Example → "Ketozolin Shampoo": every result is the same relevant product in different sizes — yet position 6 (0 CTR) would be picked as the negative.

* Haystack US 2022-René Kriegler,OSC-Modelling implicit user feedback for optimising e-commerce search

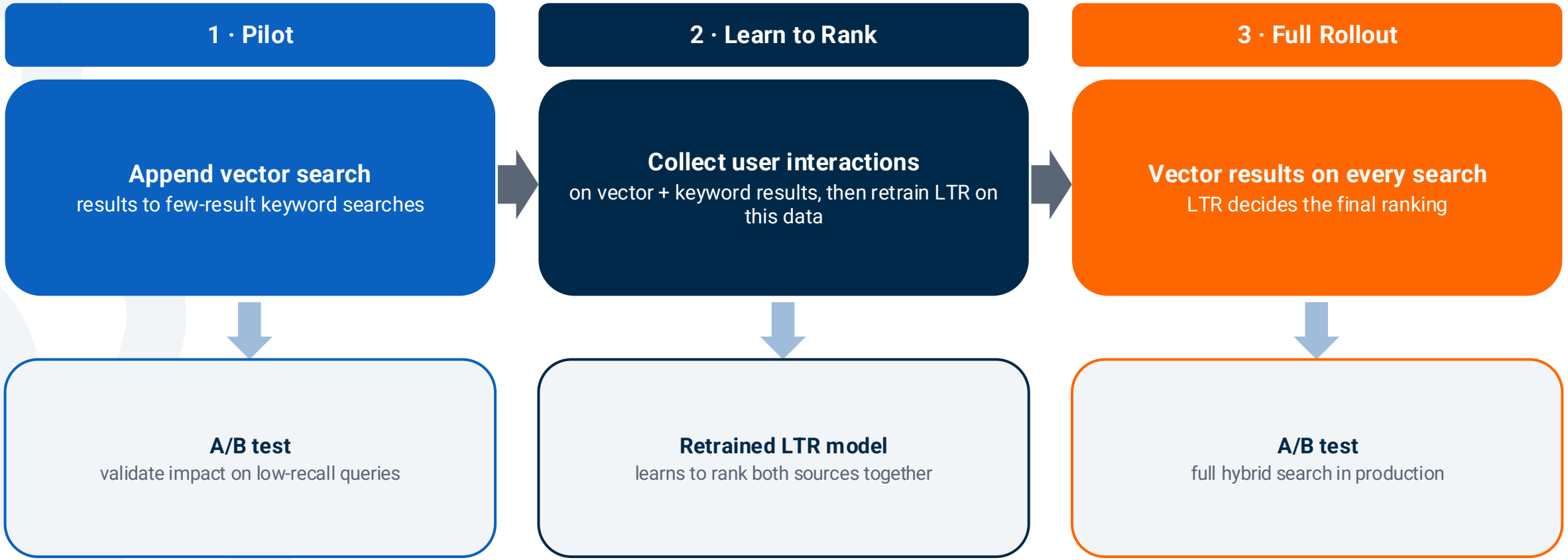
LLMs in Action

- **Baseline (Prev. Best): CTR-based 1+, 1-**
- **Exp1: LLM-Filtered Negs**
  - 25% of query-neg pairs cleaned, random negs instead
- **Exp2: LLM-Generated Single Negative**
  - Hard negatives are generated using LLMs, no CTR-based negs
- **Exp3: Multiple LLM Negatives with Filtered Negatives**
  - CTR-based negatives + 5 LLM generated negatives



Introducing Hybrid Search

idealo

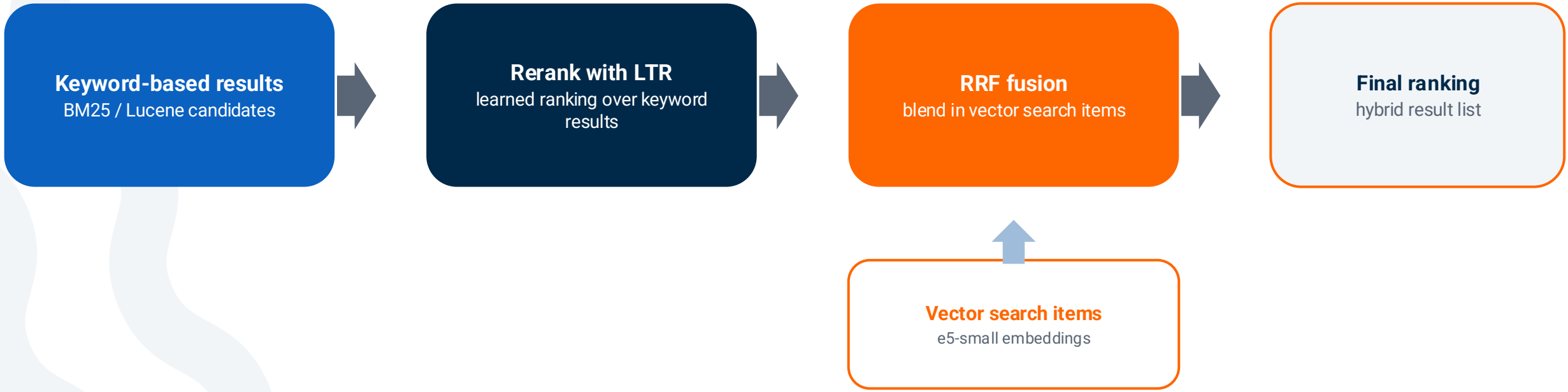


Each stage is validated by A/B testing before the next — a gradual, data-driven rollout of hybrid search.

Alternative: Rerank-then-Fuse

idealo

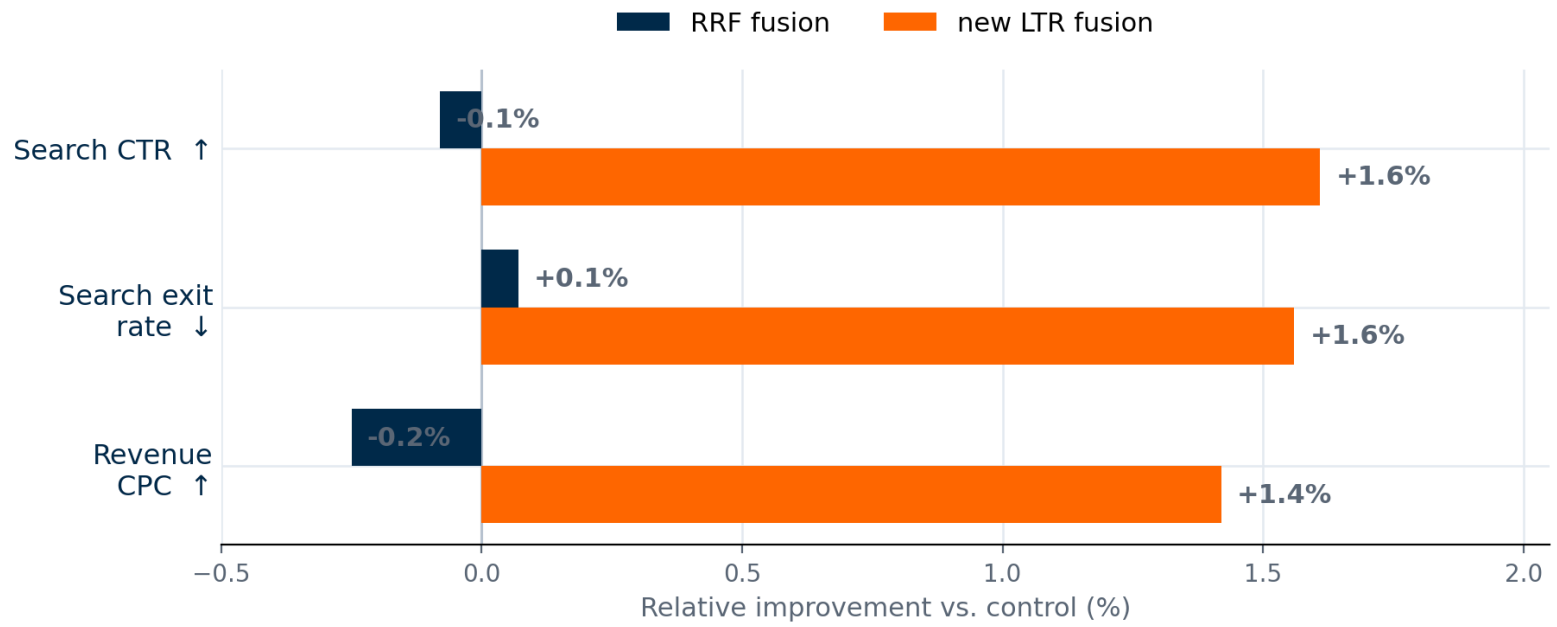
Instead of letting LTR rank both sources jointly, here we first rerank keyword results, then fuse vector items via RRF.



We tested this approach as well — RRF fusion makes blending the two ranked lists simple, with no joint retraining required.

A/B Test: LTR Fusion vs RRF Fusion

Across every key search metric, new LTR fusion beats RRF fusion.



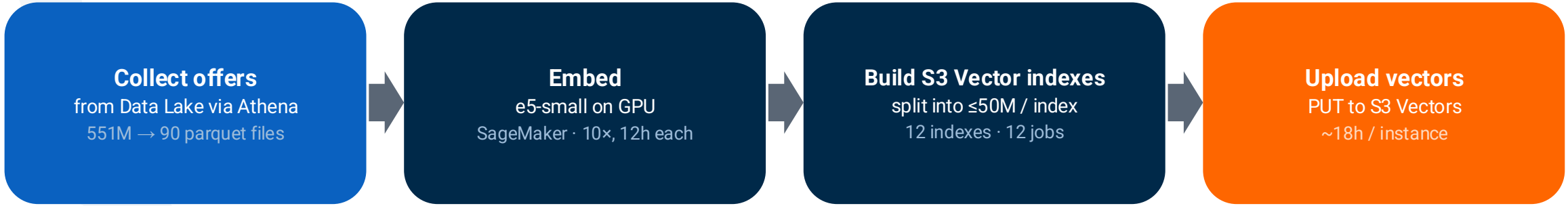
Relative change vs. control — oriented so higher = better.

Key takeaway: new LTR fusion delivers the largest, most consistent gains — higher CTR, more clickouts, lower exit rate and better revenue per click than RRF fusion.

First Approach: AWS S3 Vectors

idealo

A quick-win trial pipeline to index ~556M offer + product embeddings.



Cost per run

- SageMaker processing ≈ €201
- S3 Vectors (storage + PUT) ≈ €273

≈ €474 total · ~556M embeddings · 384-dim

The bottleneck

- Upload capped at 5 req/s, max 500 vectors/batch
- → ~18h per instance just to write vectors
- Embedding generation: 12h × GPU instances
- **Slow queries: ~600–700 ms per lookup**

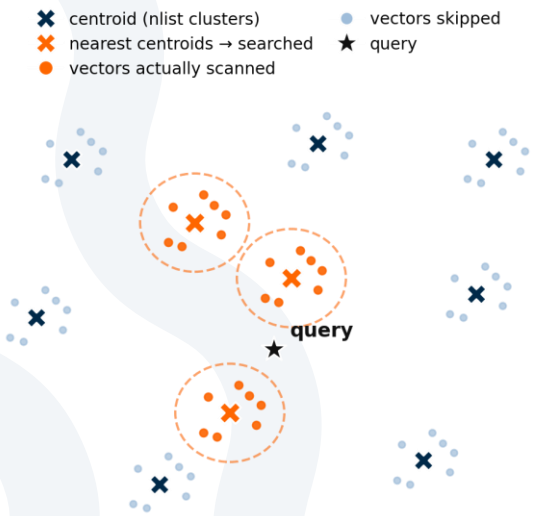
Verdict: fast to stand up, high query latency make S3 Vectors a trial-only setup — not yet a production pipeline.

Production Approach: FAISS IVF-PQ

Two purpose-built indexes — HNSW for product precision, FAISS IVF-PQ for offers at scale.

Products → HNSW index
graph-based, high precision
~8M products

Offers → FAISS IVF-PQ
inverted file + product quantization
~500M offers · 10 shards



How IVF-PQ search works

1. **Build:** k-means clusters all vectors into nlist centroids; each vector is assigned to its nearest centroid.
2. **Query:** find the nearest centroids to the query vector.
3. **Scan:** search only the vectors inside those cells — not the full 500M.
4. **PQ:** product quantization compresses vectors → smaller memory footprint.

→ **Average query latency ≈ 60 ms** (vs ~600–700 ms on S3 Vectors)

FAISS IVF-PQ: Indexing Pipeline

idealo

One extra step vs. a flat index — training the quantizer — but it unlocks fast search at scale.

1 · Embed
create embeddings
for offers + products

→

2 · Train index
run k-means → nlist centroids
12h on 15% of data

→

3 · Add vectors
insert into trained index
500M items · 10 CPU instances · 7h

~1 day
end-to-end pipeline
incl. embedding creation

≈ €500
cost per full run
for ~500M offers

10 shards
queried in parallel
at search time

72 GB
in-memory to serve
combined HNSW + FAISS index

In production today: ~10× faster queries and similar cost to the S3 trial — IVF-PQ scales offers to production without sacrificing latency.

Key Takeaways

idealo

1. **Train on your own data**
   Off-the-shelf zero-shot embeddings don't know your customers' intent — fine-tuning on your data drives relevance.

2. **Hard negatives matter**
   Contrastive pairs need real hard negatives; only-positive training performs worst.

3. **0 CTR ≠ irrelevant**
   Filter false negatives with an LLM — not all low-CTR items are truly bad matches.

4. **Bigger batches, better NDCG**
   “First & Last” pairs at batch size 256 won; adding product attributes helped further.

5. **Recall first, then rerank**
   Optimize retrieval for recall and let LTR handle precision — LTR fusion beat RRF on every key metric.

6. **Match the index to the scale**
   FAISS IVF-PQ serves ~500M offers at ~60 ms; pick the index for your scale and latency budget.

idealo

Questions

Gennady Shabanov

Machine Learning Engineer

Atakan Filgöz

Machine Learning Engineer