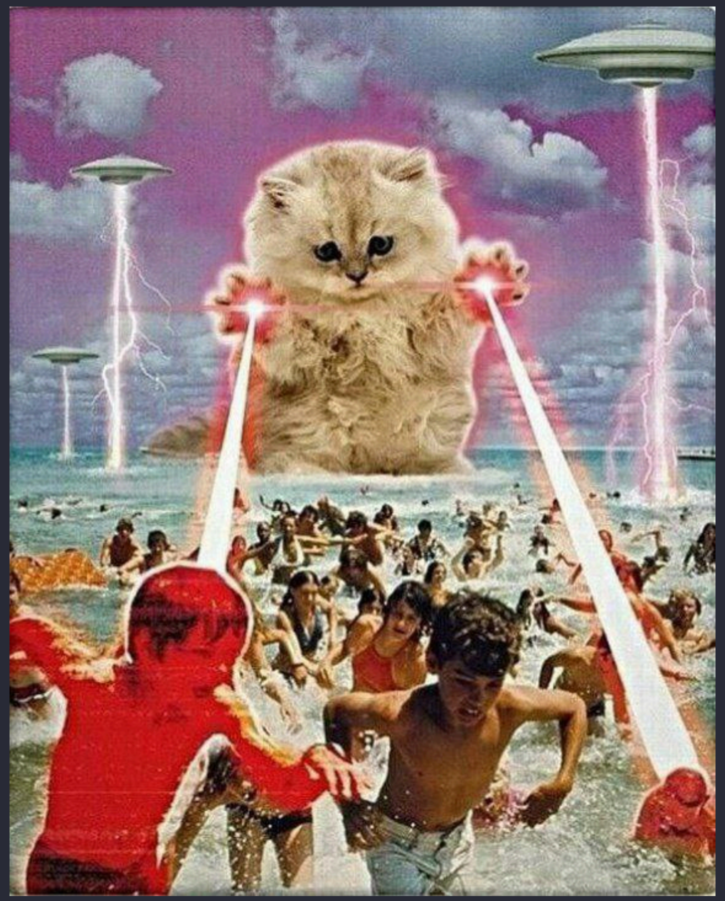
How semantic search projects



FAIL

Roman Grebennikov | Delivery Hero SE | MICES 2024

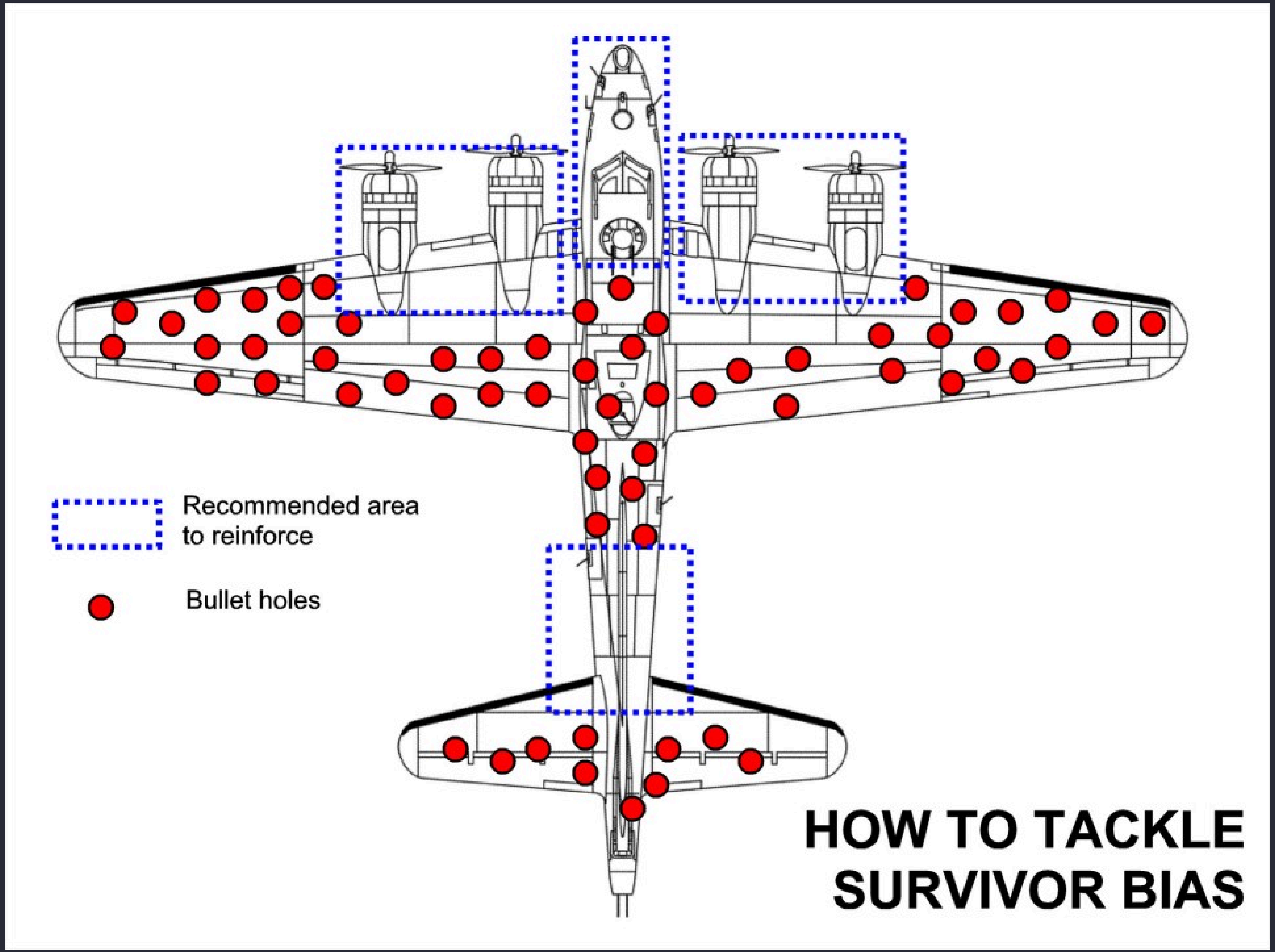
whoami

- PhD in CS, quant trading, credit scoring
- **Findify**: e-commerce search, personalization
- **Delivery Hero**: food search, LLMs
- **Opensource**: Metarank, Nixiesearch

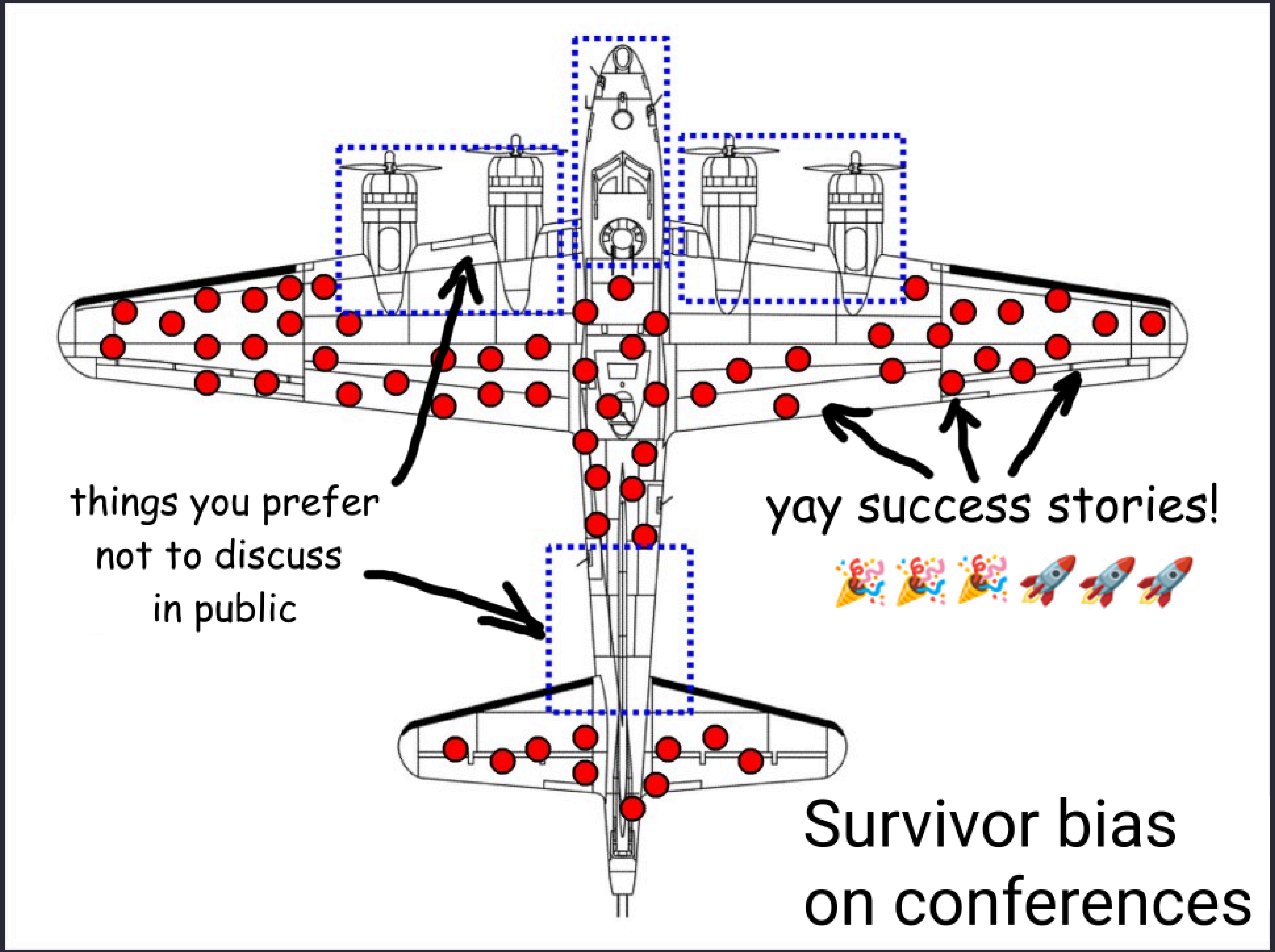
Delivery Hero

- Last-mile food & groceries delivery
- **70** countries, **20** languages
- **1M** restaurants & local vendors

Survivor bias



Survivor bias on conferences

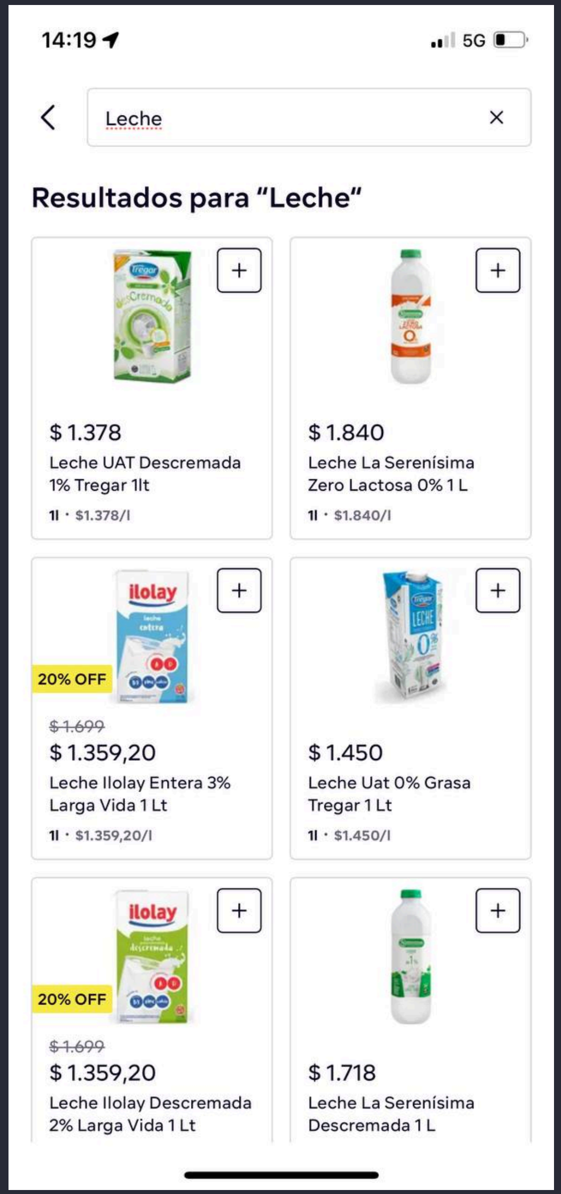


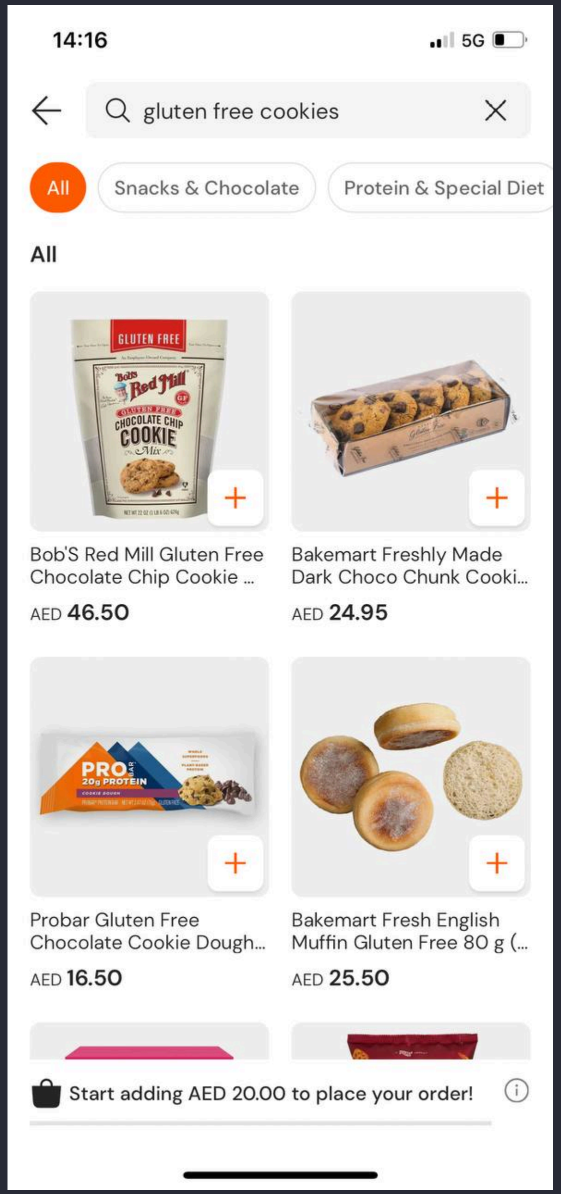
Agenda

🚀

- Do embeddings matter?
- Relevance tuning with semantic search
- Multilingual search
- Semantic search halting problem

Product search in Q-Commerce

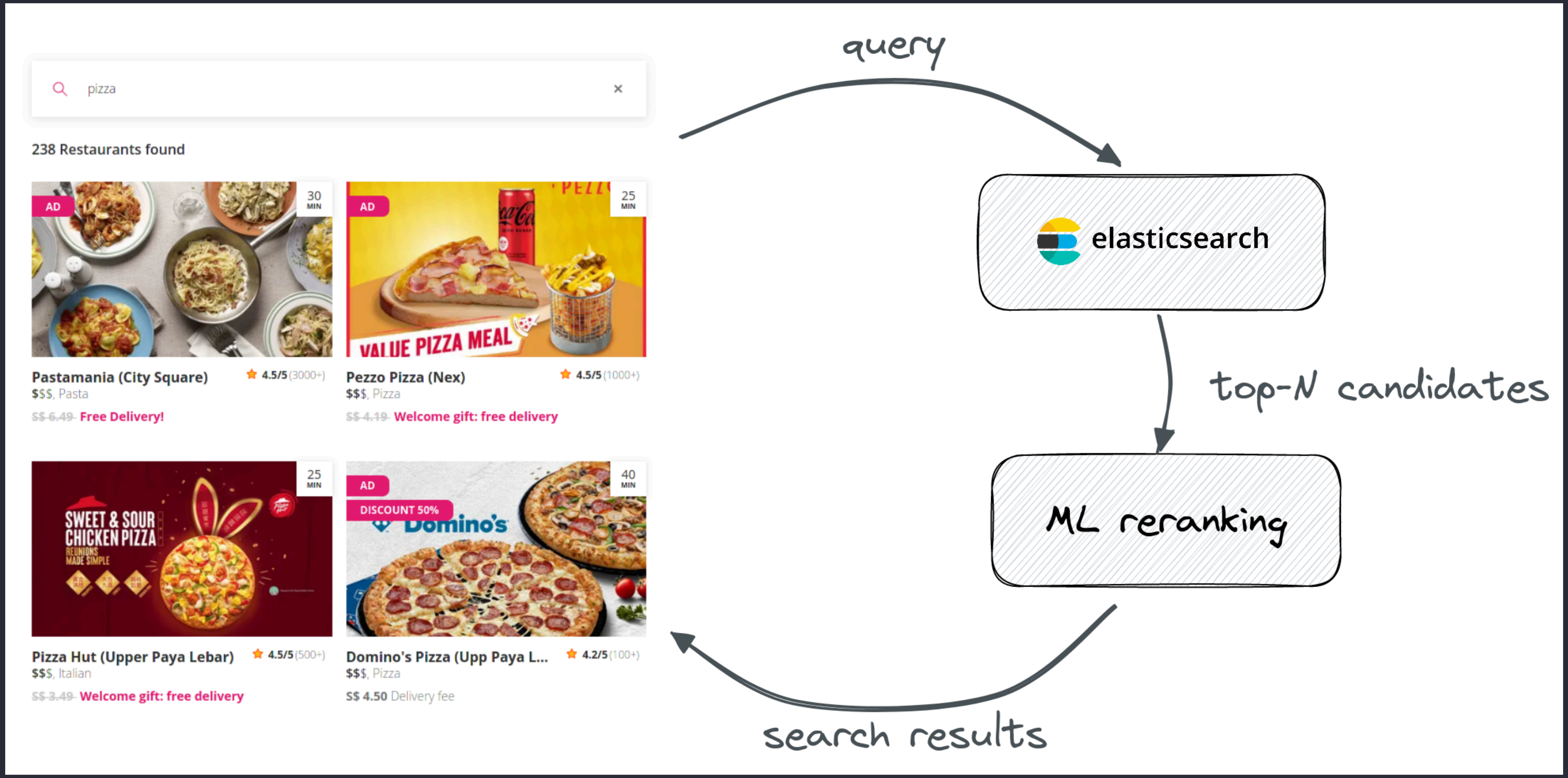




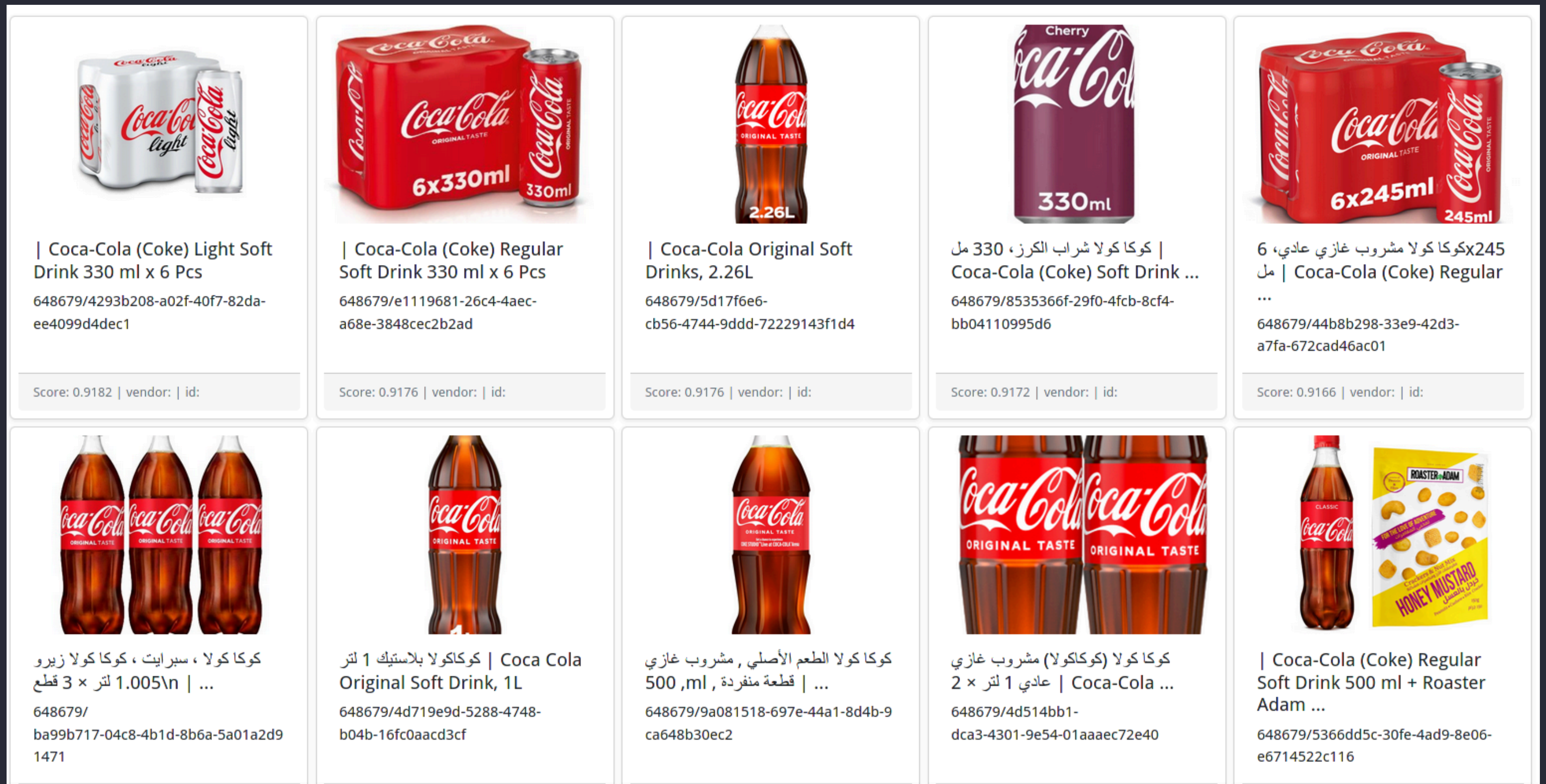
- Large inventory: ~20M items
- Diverse multi-token requests
- ~10% (**OMG!**) zero results rate

Multi-token, long tail queries 🔴

Retrieve and rerank



Precision vs Recall

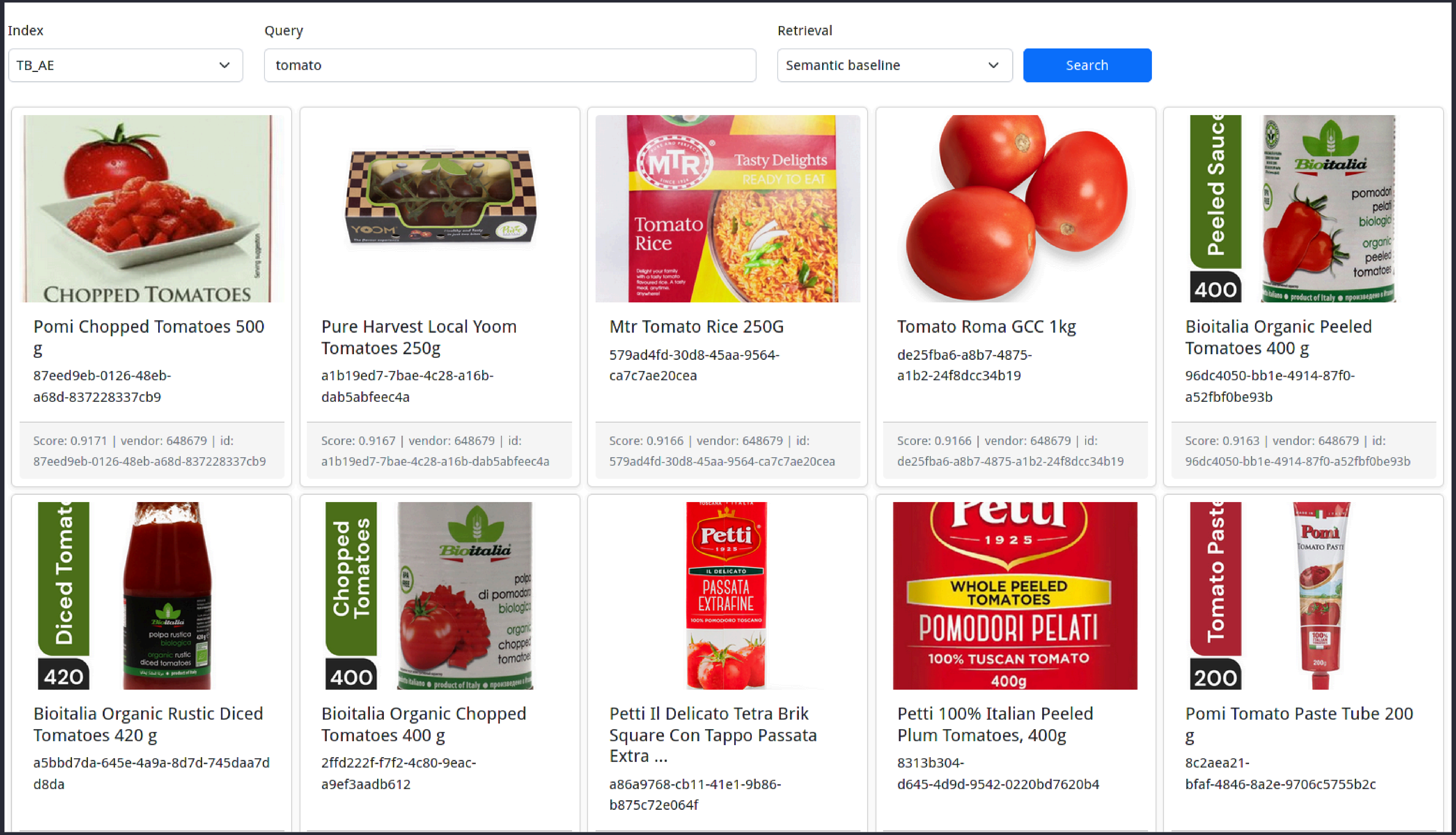


- **coca AND cola AND zero**: zero results
- **coca OR cola OR zero**: matches pepsi
- **coca AND cola AND (zero OR light)**: good luck

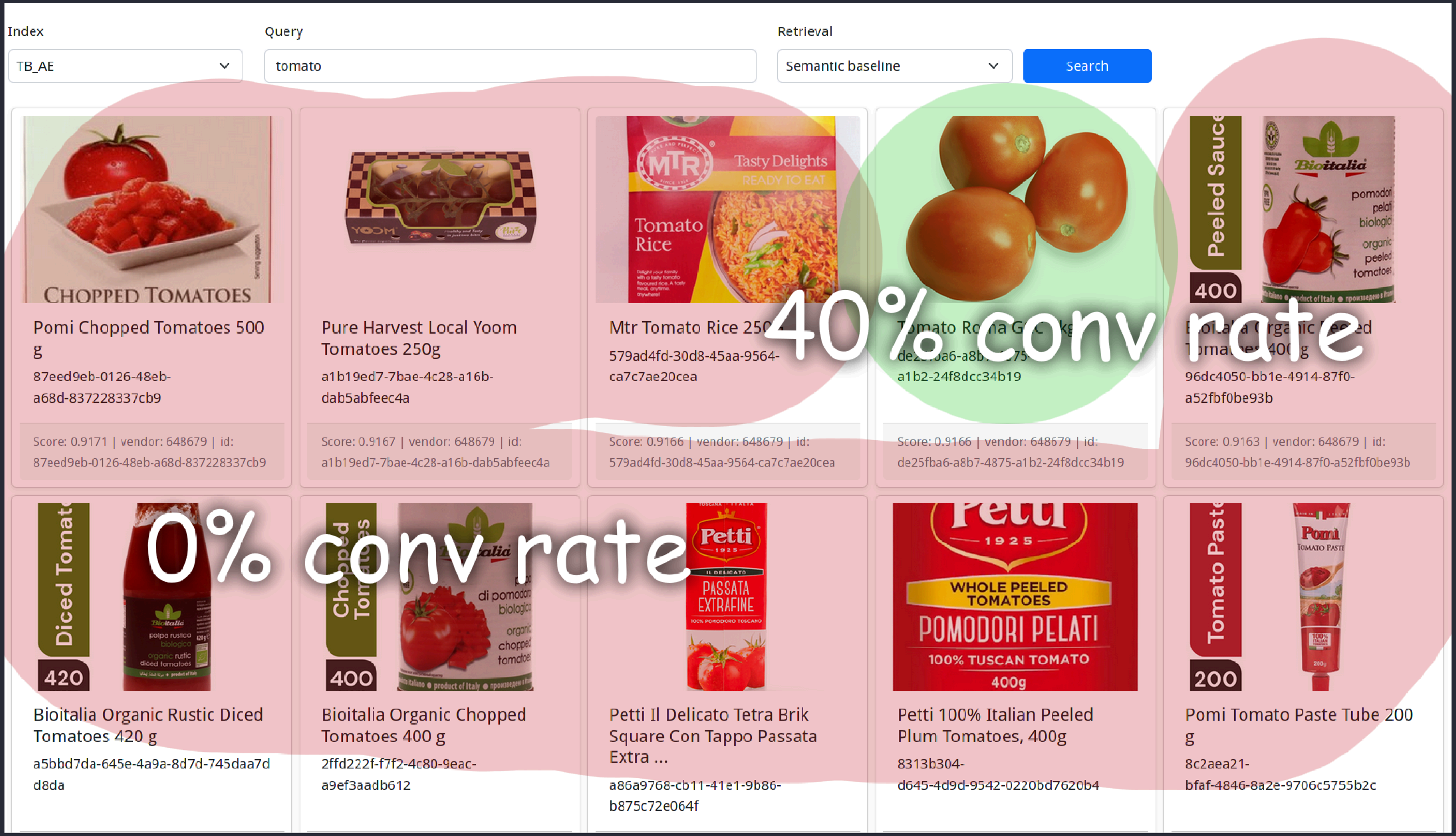
yay semantic search!

- Embed documents with SBERT/OpenAI
- Install a Vector© Search® Database™ 🚀
- ...
- PROFIT

Customer intent

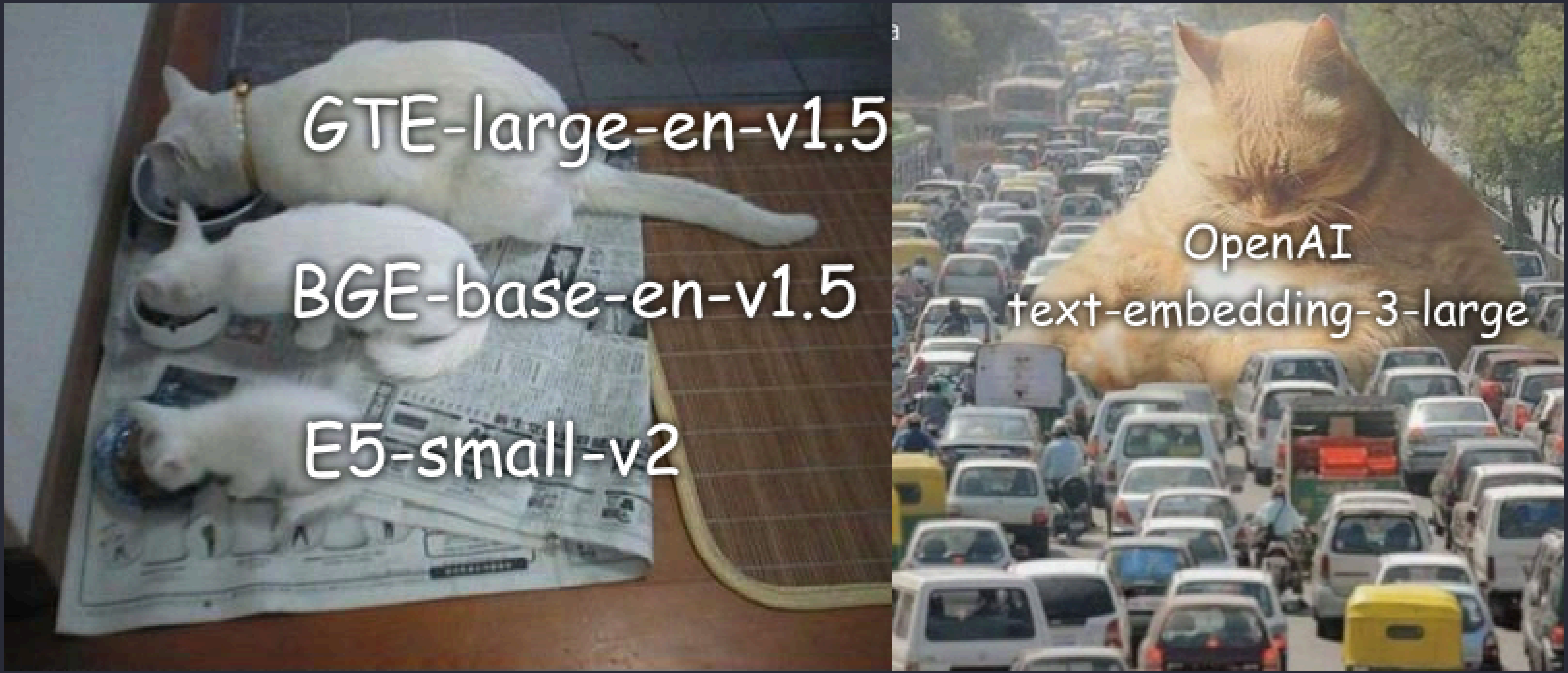


Customer intent



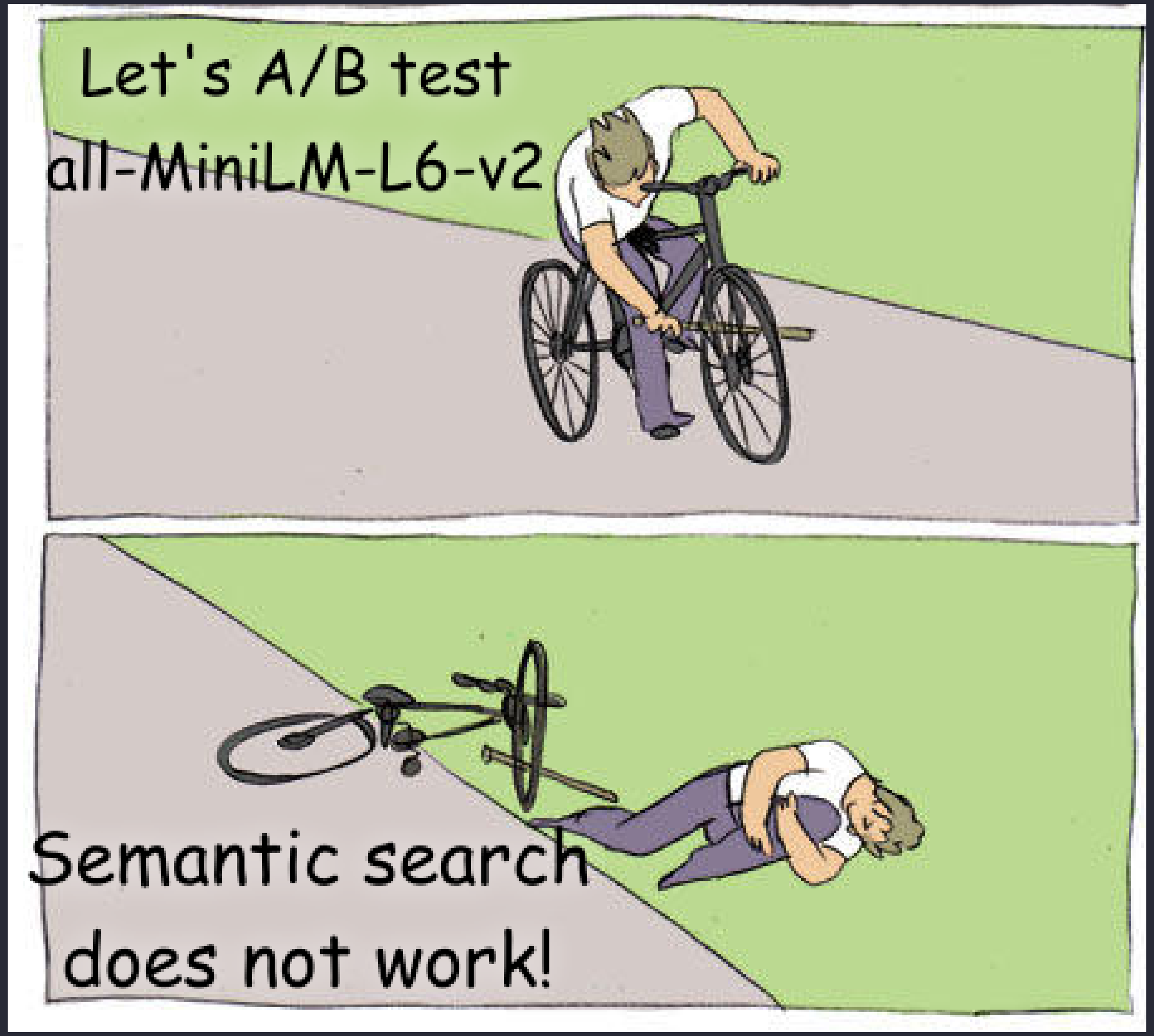
- **Relevance is subjective**: depends on intent
- **Embedding model**: no idea about your audience

Bigger models?



demo

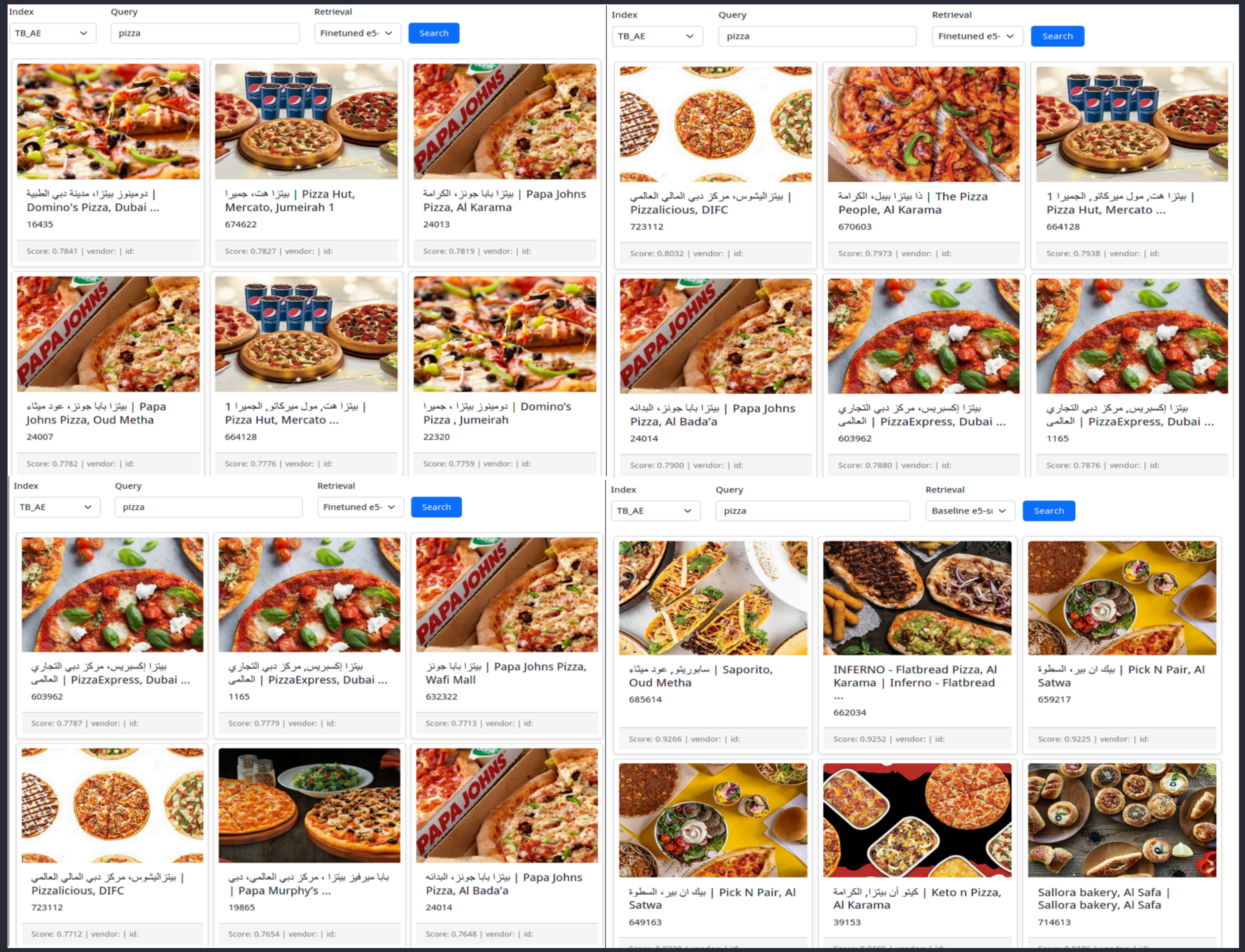
Does size matter?



- Big models: more into small details
- Still no idea about customer intent :(

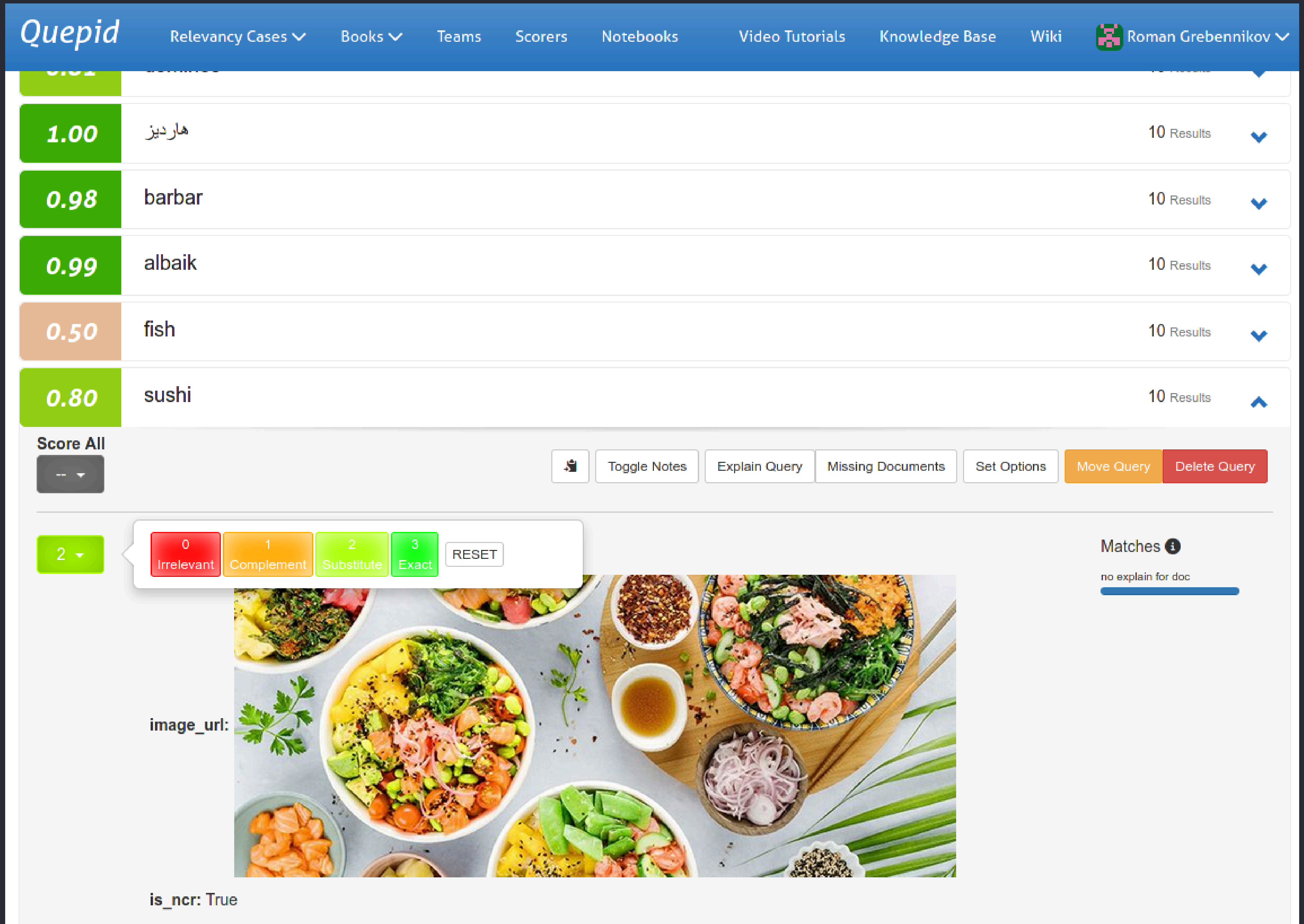
Semantic search relevance tuning

- **Lexical search**: relevance labels, tinker with retrieval
- **Semantic search**: relevance labels, tinker with retrieval



First step is still the same

You cannot improve search if you cannot measure it

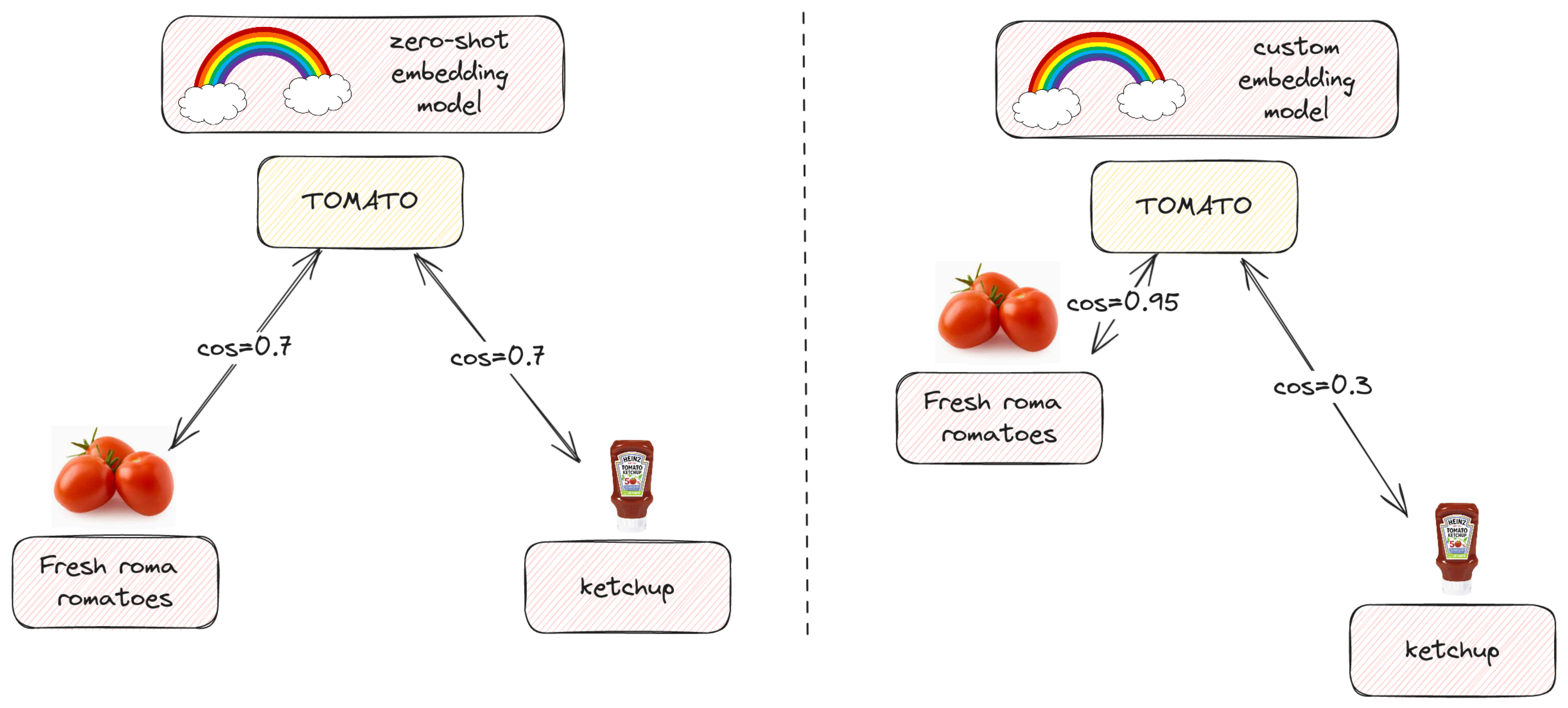


[0] - github.com/o19s/quepid

Relevance tuning?

- **Lexical search**: boosts, synonyms, queries
- **Semantic search**: fine-tuning

Fine-tuning



- Relevant docs: make them closer to query
- Irrelevant docs: make them further from query

What is positive and negative?

- 1 click, 3 impressions = 33% CTR?

Mixing clicks and confidence

$$CVR\_b = \frac{c * CVR + conversions}{c + impressions}$$

- **Bayes correction**: mix prior and posterior
- **Low confidence**: strong shift to avg
- **High confidence**: almost no shift to avg

[1]: Haystack US22: R.Kriegler, Modelling implicit user feedback for optimising e-commerce search

Bayes corrected CVR as label

	conv	impressions	CVR	bCVR, c=10	bCVR, c=50
 Pomi Chopped Tomatoes 500 g	1	3	0.33	0.230	0.207
 Pomi Tomato Paste Tube 200 g	0	2	0.0	0.166	0.192
 Tomato Roma GCC 1kg	40	100	0.4	0.381	0.333
 Mtr Tomato Rice 250G	0	300	0.0	0.006	0.028

avg conv rate = 0.2

Bayes corrected CVR as label

	conv	impressions	CVR	bCVR, c=10	bCVR, c=50	
 Pomi Chopped Tomatoes 500 g	1	3	0.33	0.230	0.207	LOW CONFIDENCE :(
 Pomi Tomato Paste Tube 200 g	0	2	0.0	0.166	0.192	
 Tomato Roma GCC 1kg	40	100	0.4	0.381	0.333	POSITIVE!
 Mtr Tomato Rice 250G	0	300	0.0	0.006	0.028	NEGATIVE!

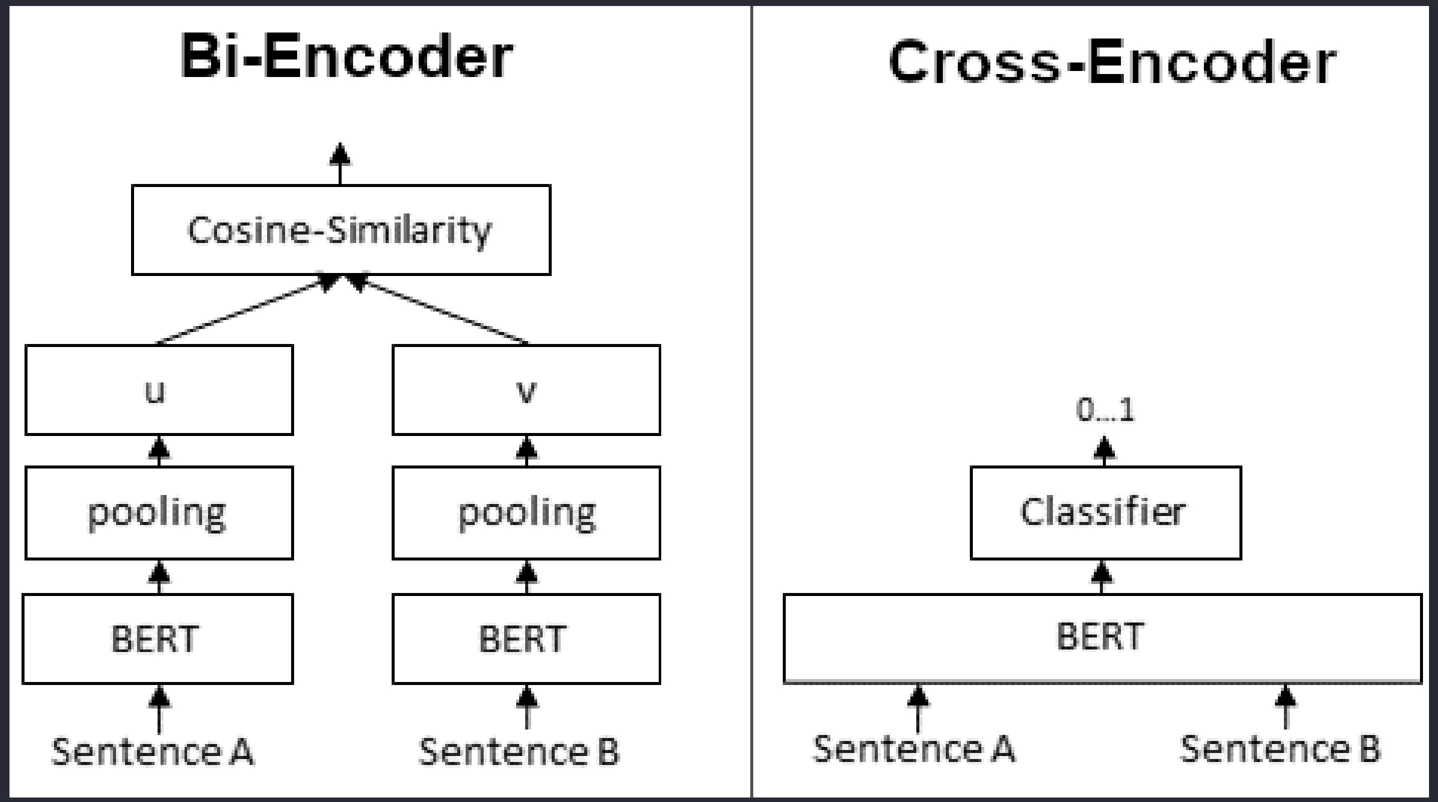
avg conv rate = 0.2

demo

Implicit labels are noisy

- **High confidence, avg CVR**: oops
- **Long tail queries**: not enough data to reach confidence
- **Bias towards existing ranking**

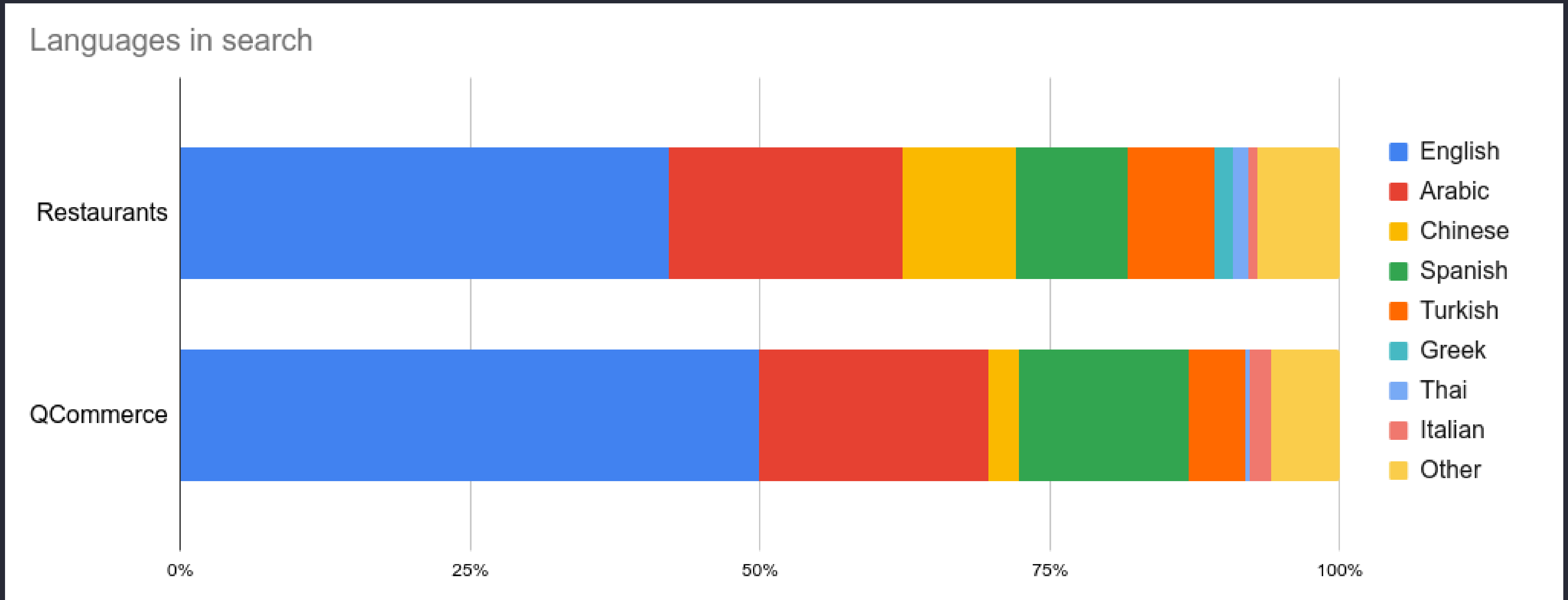
Future plans



- **LLM relabeling**: use explicit labels to fine-tune Cross-Encoder
- **Llama3 CE**: much faster convergence on small data
- **Distillation**: train embeddings on re-labeled dataset

[1] - sbert.net: Cross-Encoders

Non-English search



Problem: all MTEB leaderboard models are English

Multilingual search

Guess the amount of non-english train samples:

2 Training Methodology

	# Sampled
Wikipedia	150M
mC4	160M
Multilingual CC News	160M
NLLB	160M
Reddit	160M
S2ORC	50M
Stackexchange	50M
xP3	80M
Misc. SBERT Data	10M
Total	∼1B

Table 1: Data mixture for contrastive pre-training.

	# Sampled
MS-MARCO Passage	500k
MS-MARCO Document	70k
NQ, TriviaQA, SQuAD	220k
NLI	275k
ELI5	100k
NLLB	100k
DuReader Retrieval	86k
Fever	70k
HotpotQA	70k
Quora Duplicate Questions	15k
Mr. TyDi	50k
MIRACL	40k
Total	∼1.6M

Table 2: Data mixture for supervised fine-tuning.

[1] L.Wang et al. Multilingual E5 Text Embeddings: A Technical Report

Out of domain

Dataset	BM25	Multilingual E5 (small)
MIRACL:sw	0.4243	0.6755
MIRACL:yo	0.6831	0.4187
BEIR:trec-covid	0.6823	0.7139

in domain

out of domain

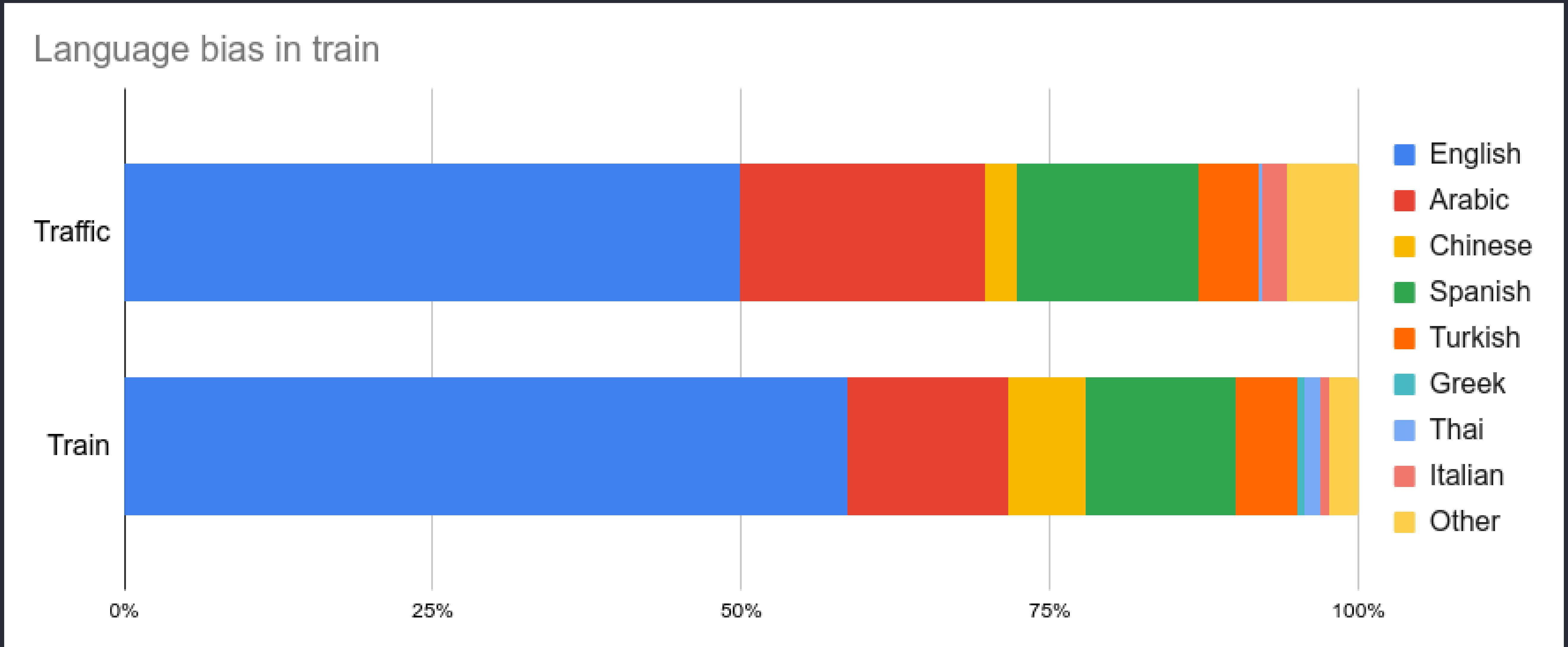
Retrieval effectiveness for BM25 and E5 small (NDCG@10)

Food & groceries search - out of domain 😭

[1] - J.Bergum: Vespa Blog - Simplify Search with Multilingual Embedding Models

Fine-tuning on implicit data 💀

Confidence based labels = more bias to English



Hack: up-sample non-English training data (and get more noise!)

demo

Mixed language data

				ndcg			
base	negs	format	batch	1	3	5	10
multilingual-e5-base	2	local	512	0.7858	0.6771	0.5751	0.6521
multilingual-e5-base	2	local+category	512	0.7837	0.6719	0.5687	0.6478
multilingual-e5-base	2	local+category+master	512	0.7901	0.6836	0.5798	0.6574
multilingual-e5-base	2	local+master	512	0.7965	0.6901	0.5892	0.6654
multilingual-e5-base	2	local+english	512	0.8136	0.7134	0.6112	0.6845
multilingual-e5-base	2	local+english+category	512	0.8047	0.6988	0.5966	0.6724
multilingual-e5-base	2	local+english+category+master	512	0.8175	0.7097	0.6076	0.6813
multilingual-e5-base	2	local+english+master	512	0.8142	0.7099	0.6069	0.6813

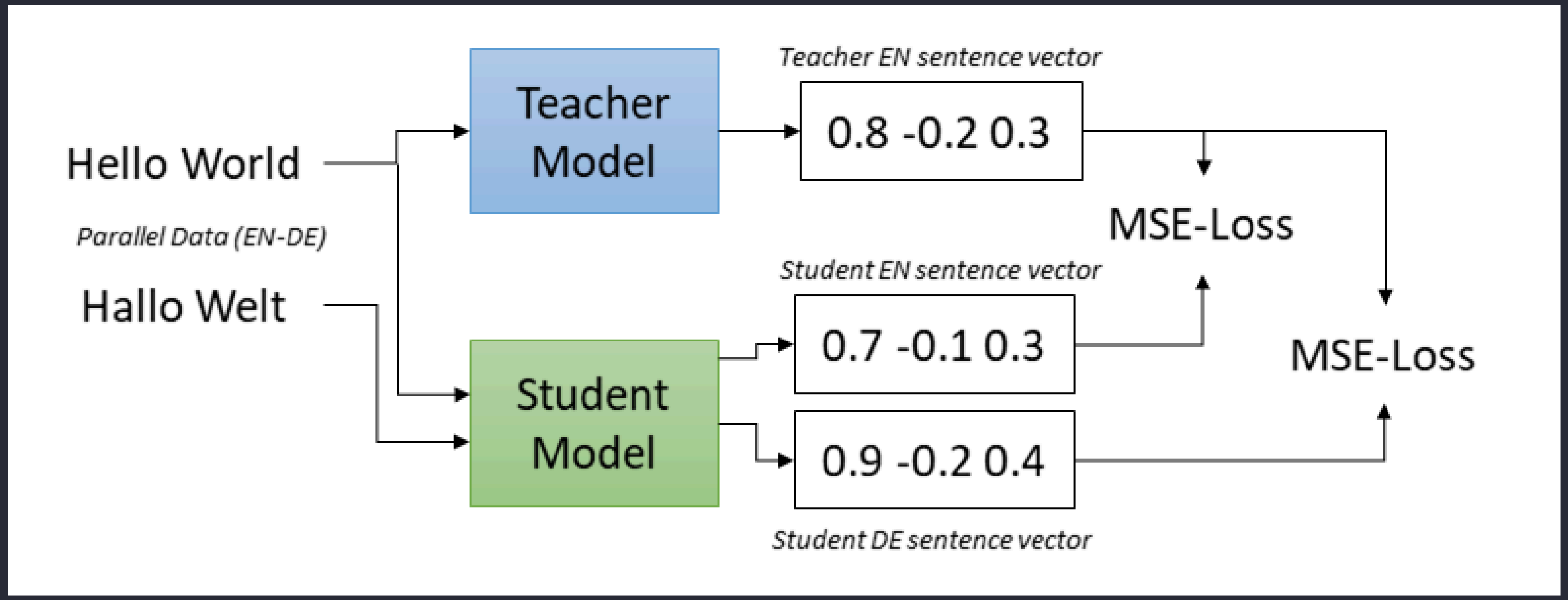
Worked well: mixed language data fine-tuning

Future plans

- Experiment #1: machine-translation assisted fine-tuning

```
{
  "query": ["water", "wasser", "水", "ماء", "agua"],
  "positive": ["Oasis Drinking Water", "Oasis Trinkwasser", "綠洲飲用
  "negative": ["Coca-Cola Zero", "可口可樂零"]
}
```

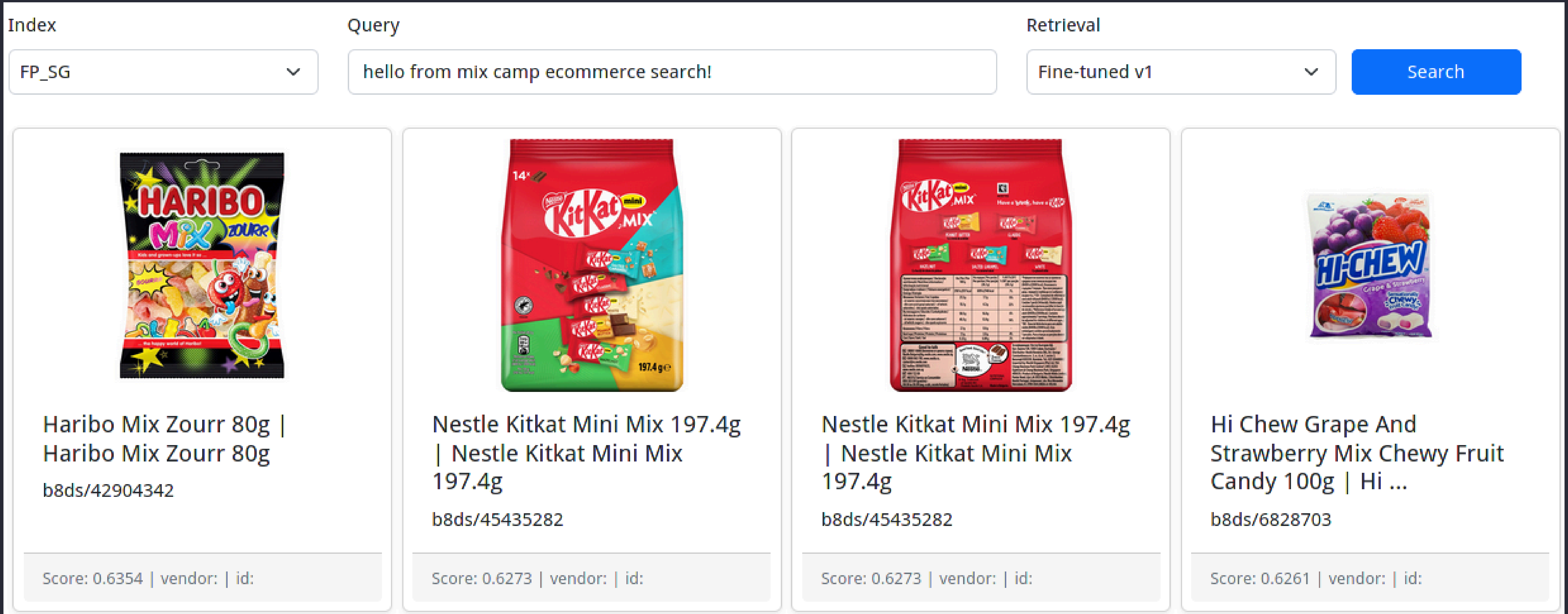
- Experiment #2: distill multi-lingual from English-biased model



[1]: sbert.net: Training Examples » Multilingual Models

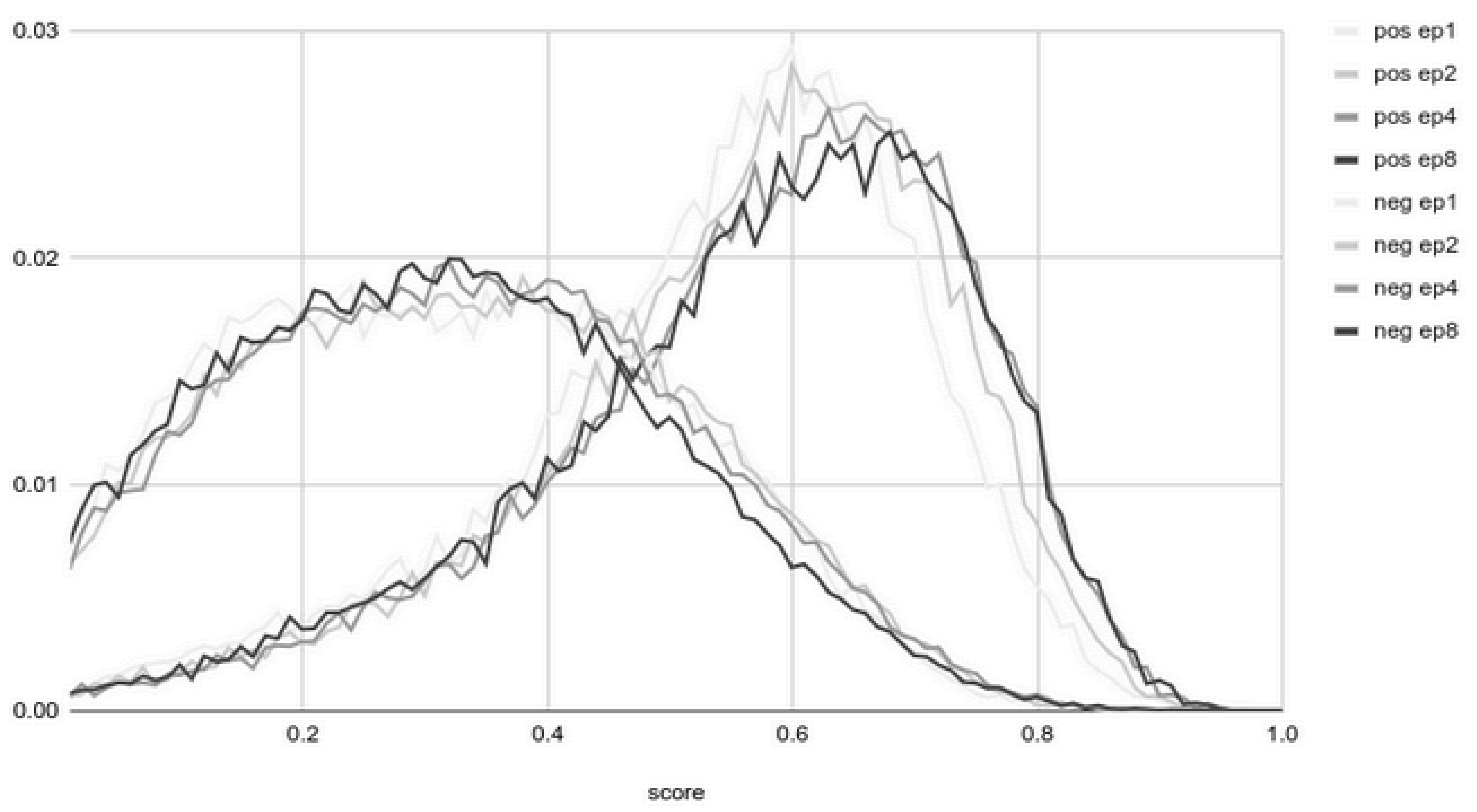
Semantic search halting problem

Problem: semantic search always has something found



demo

Finding a perfect threshold



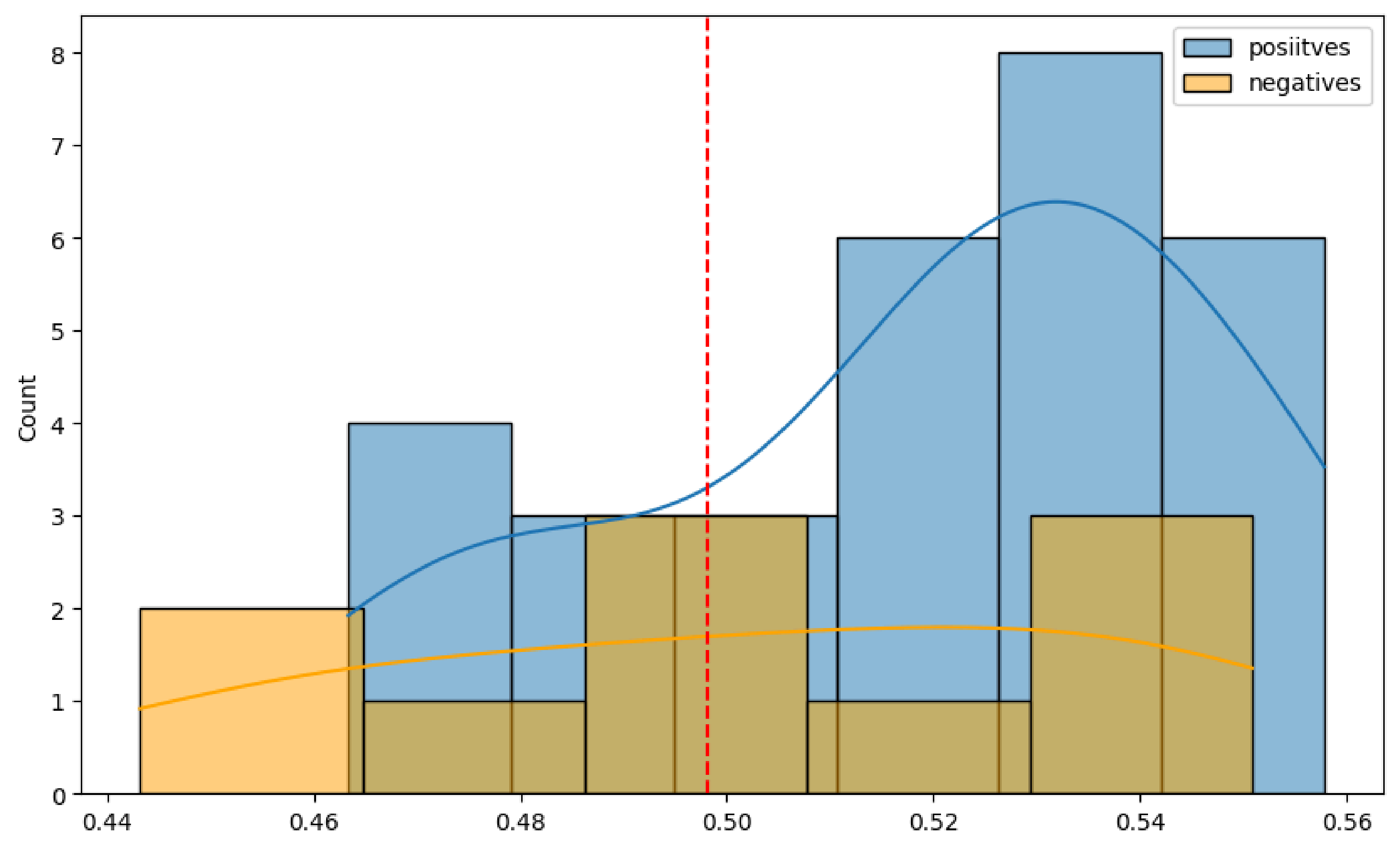
Score distribution drift while training. Note the right dist drifts to the right!

Attempt #1: set 0.7 as threshold => 70% zero results

Attempt #2: similar queries?

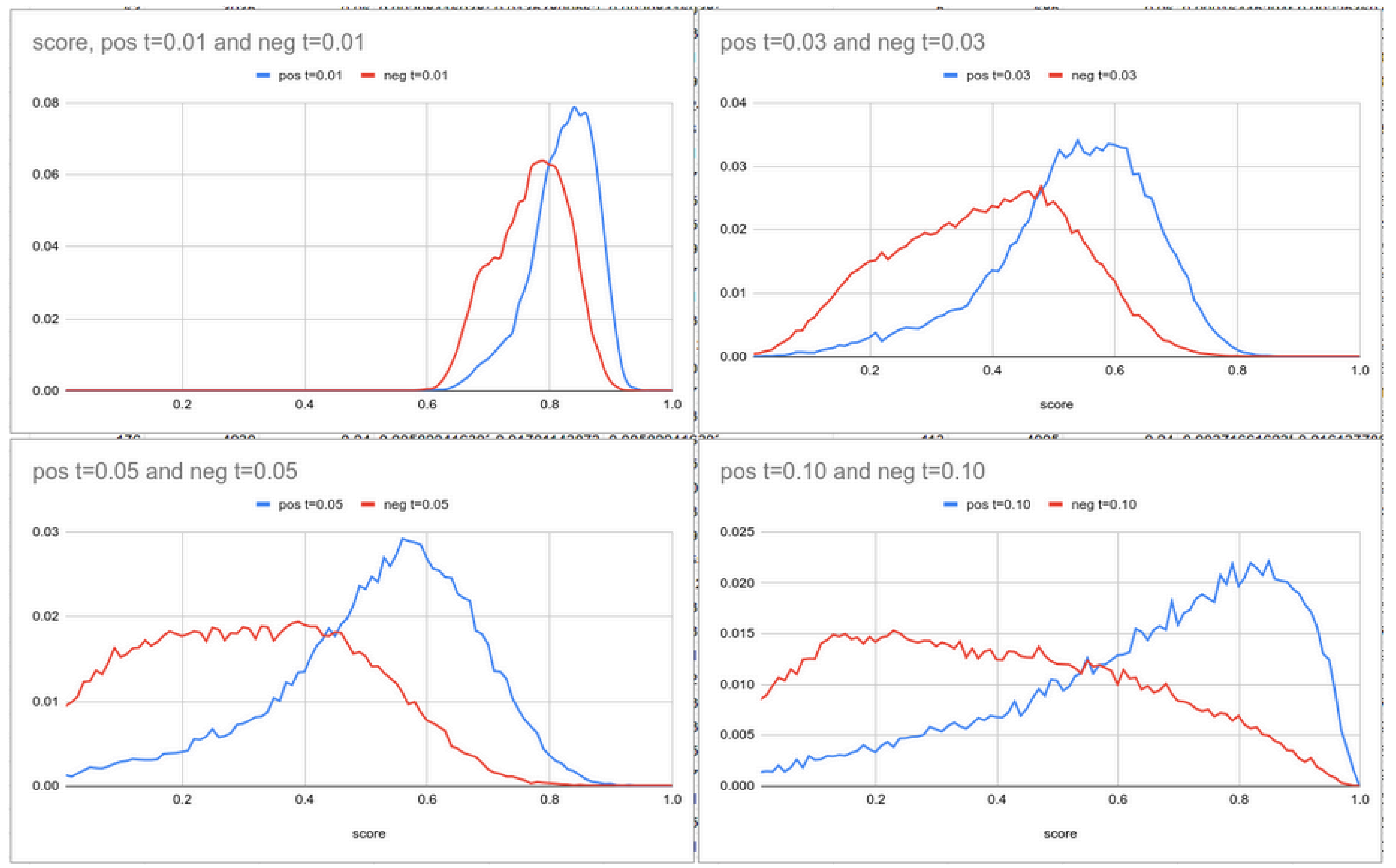
Ecommerce: **a lot of repeated queries!**

- Find a "good enough" threshold for all seen queries
- Threshold of unseen query = avg(**threshold of top-N seen q**)



FAIL: too much noise

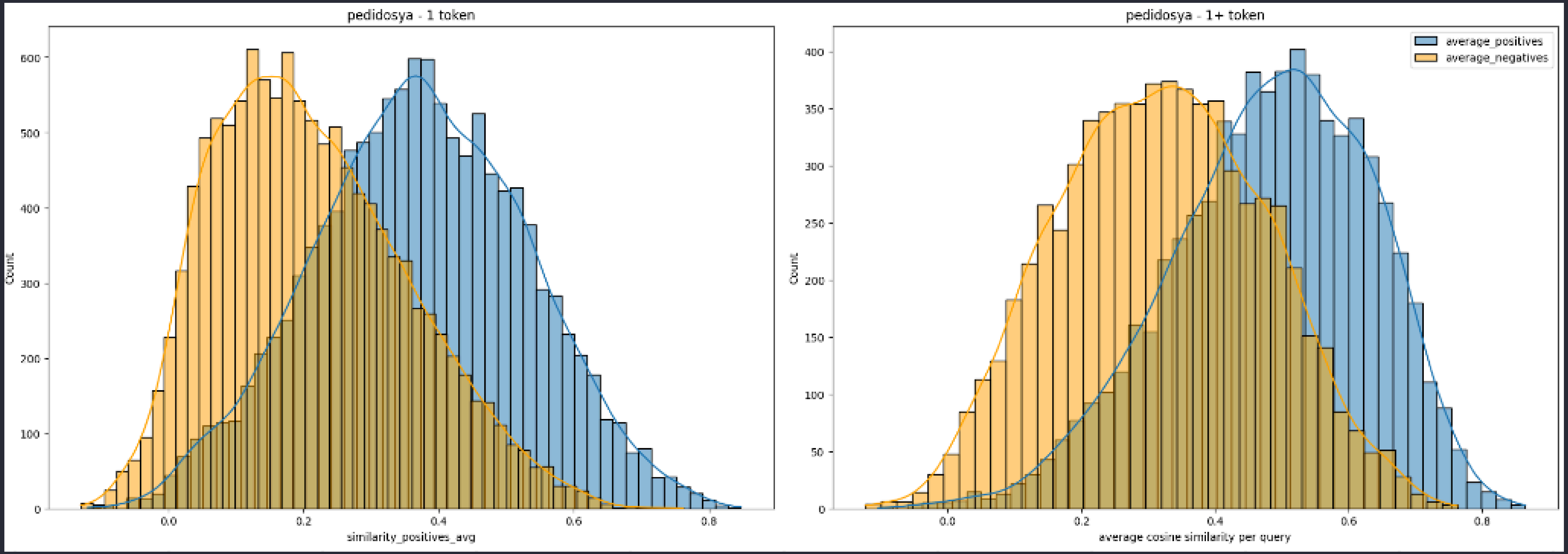
Threshold depends on the model!



Positive/negative score distributions for different temperatures.

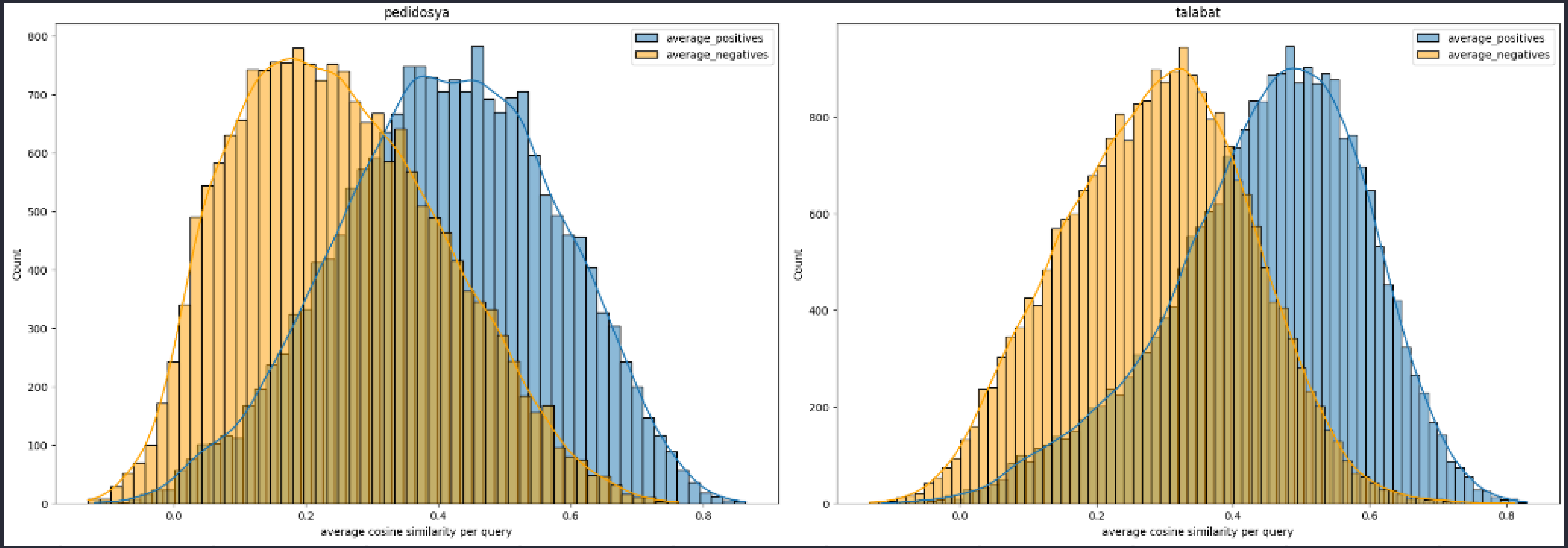
InfoNCE training temperature: model confidence level

Query length and threshold



Longer the query - higher the cosine similarity!

Language and threshold



Left: Spanish, right: Arabic

More training data = more model confidence!

#3: language/token threshold

Pre-computed thresholds:

- Single and multi-token
- Per each brand (and language)
- e5-base-multilingual: temp=0.05, range=0.62..0.70

Does it work?

A/B test: Control vs Hybrid for 2+ tokens

Region	GMV	Orders	Clicks	Click pos	ZRR
SA	+3.9%	+1.6%	+4.2%	-3.4%	N/A
UAE	+0.7%	+0.7%	+2.5%	-2.4%	-40%
APAC	+1%	0%	+1.2%	-1%	-27%
Turkey	+0.6%	+0.4%	+1.14%	0%	N/A
Latam	0%	0%	+0.5%	0%	-12%

Does it work? (yes/no)

- **Depends on baseline**: tough to beat well-built lexical search
- **Focus on recall**: use reranking for precision
- **Should you fine-tune**: yes

Links

- **Linkedin**: linkedin.com/in/romangrebennikov/
- **MTEB Leaderboard**: huggingface.co/spaces/mteb/leaderboard
- **Sentence-transformers v3**: sbert.net/

questions?