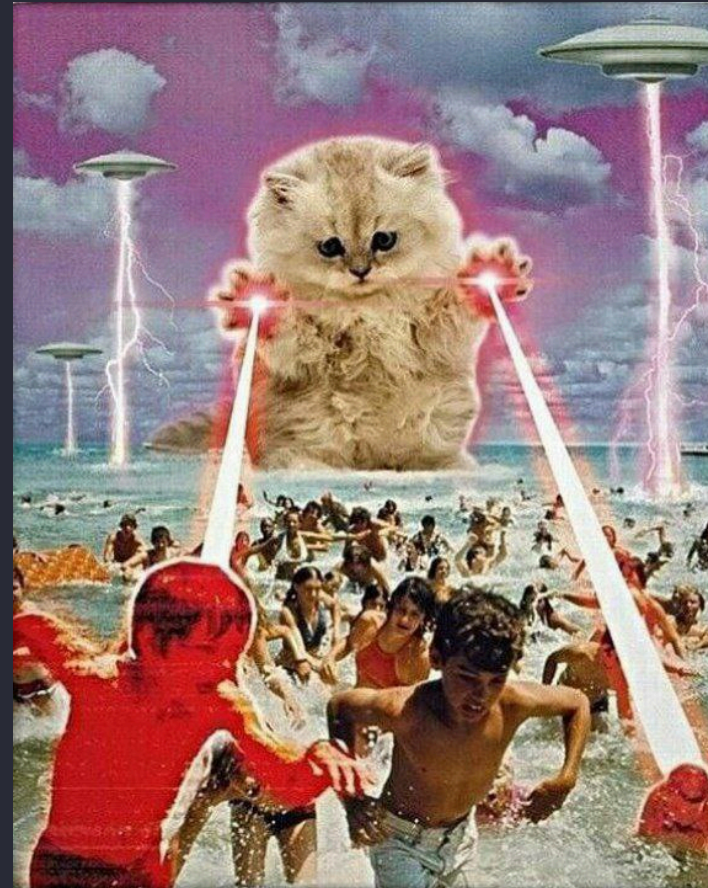


How semantic search projects



FAIL

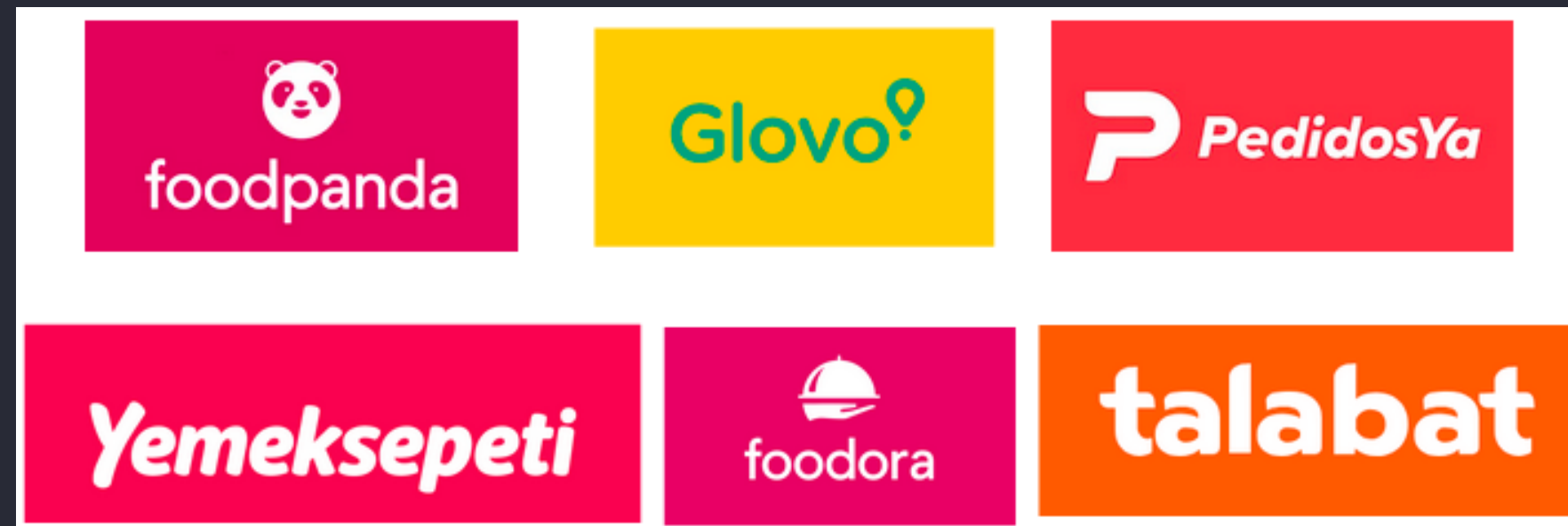
whoami



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

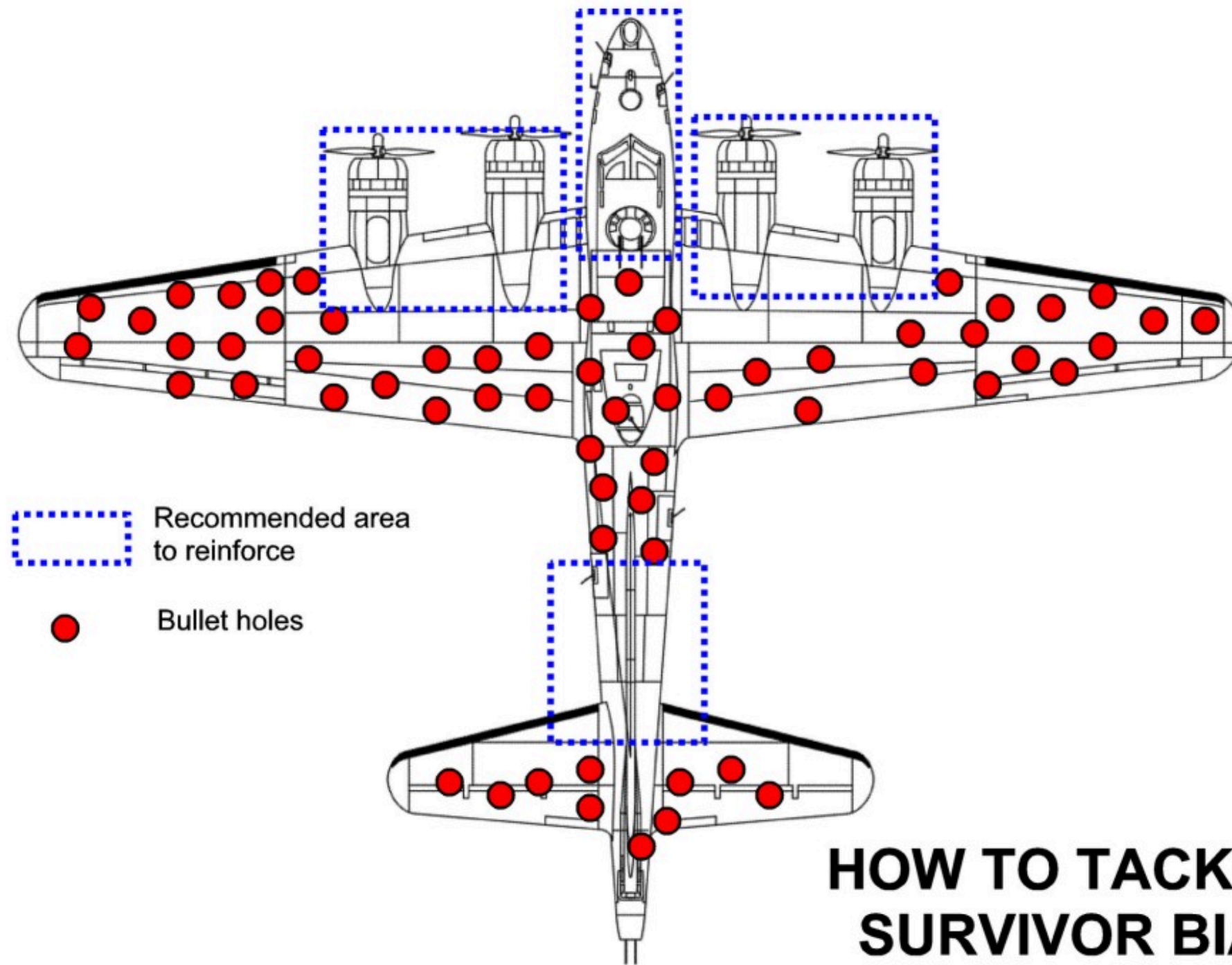


Delivery Hero



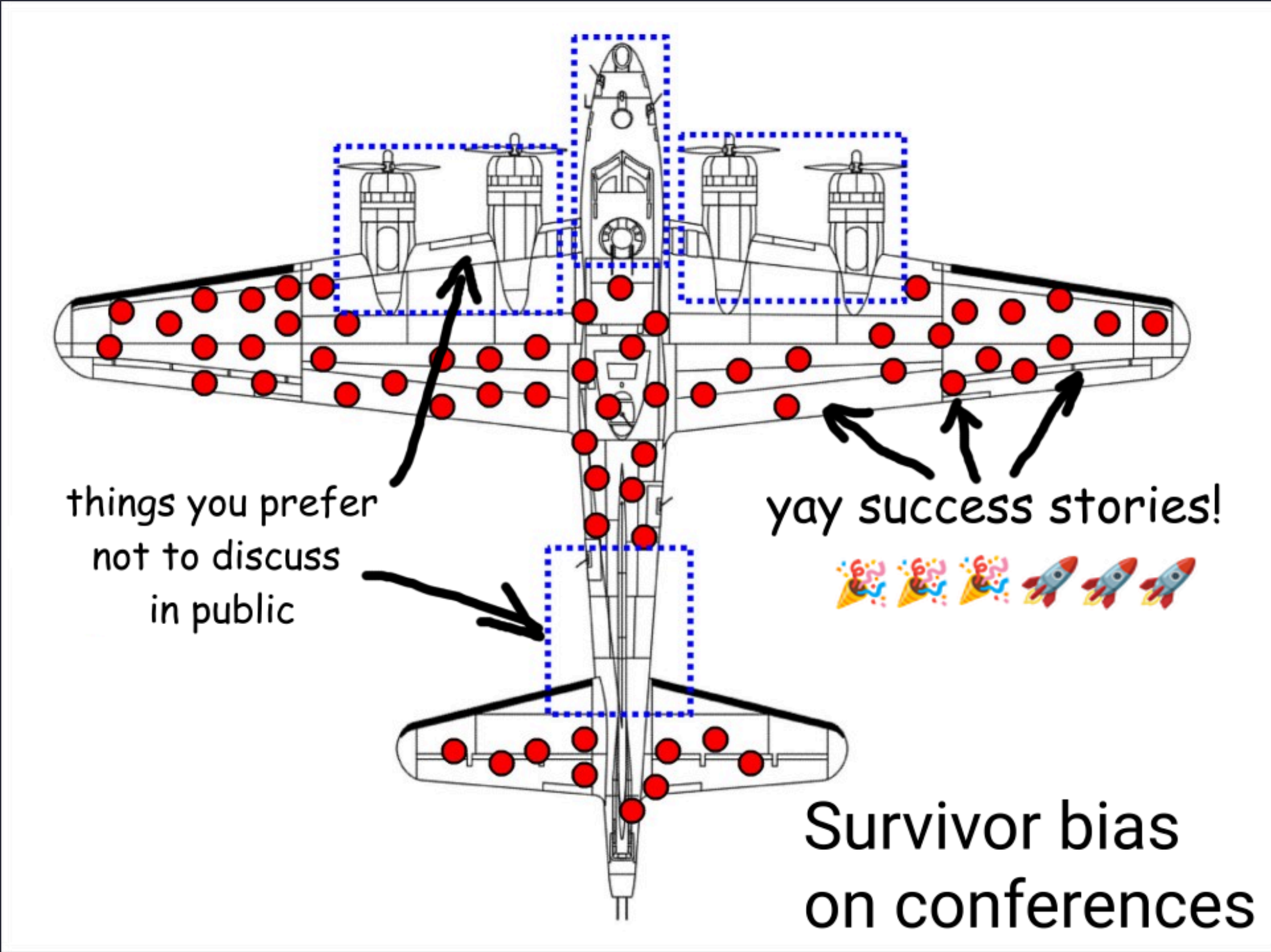
- Last-mile food & groceries delivery
- **70** countries, **20** languages
- **1M** restaurants & local vendors

Survivor bias



**HOW TO TACKLE
SURVIVOR BIAS**

Survivor bias on conferences

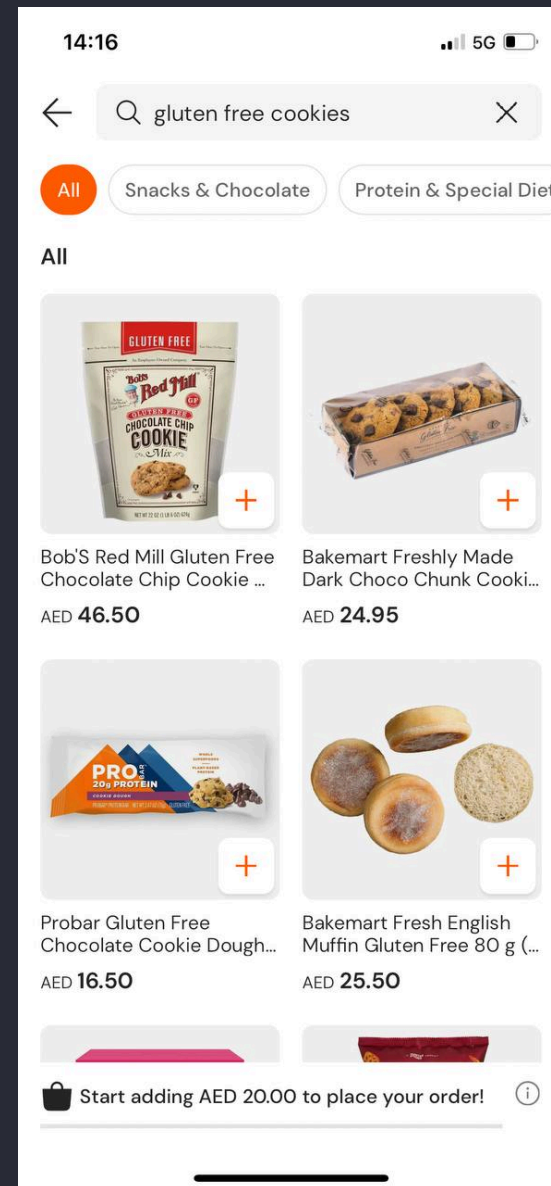
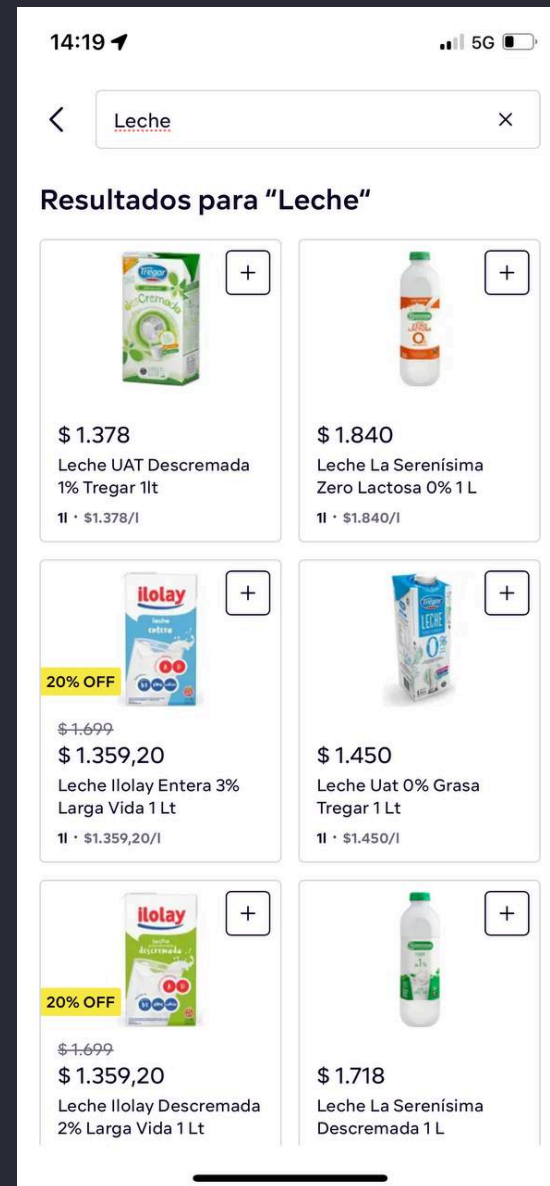


Agenda



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

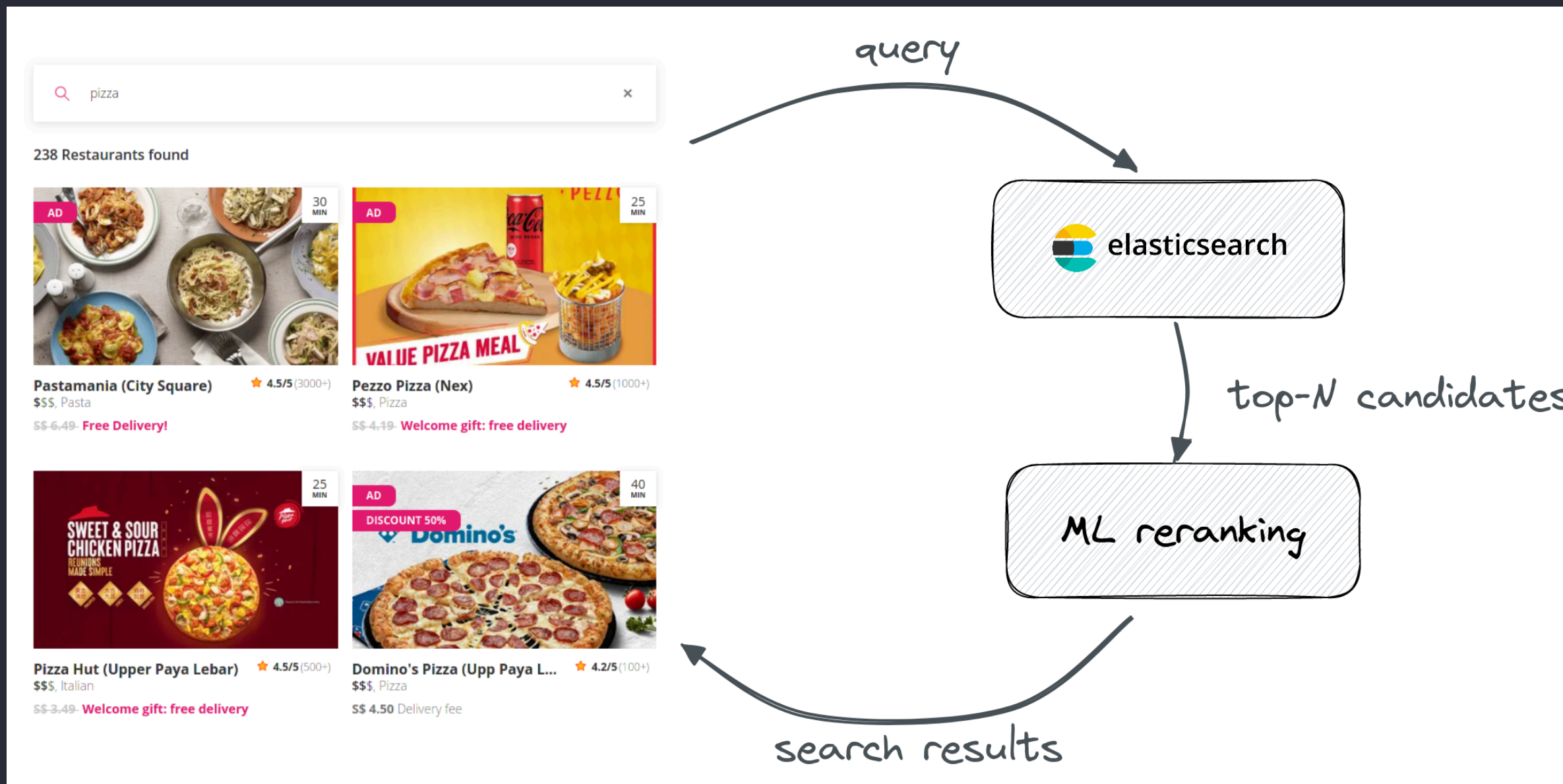
Product search in Q-Commerce



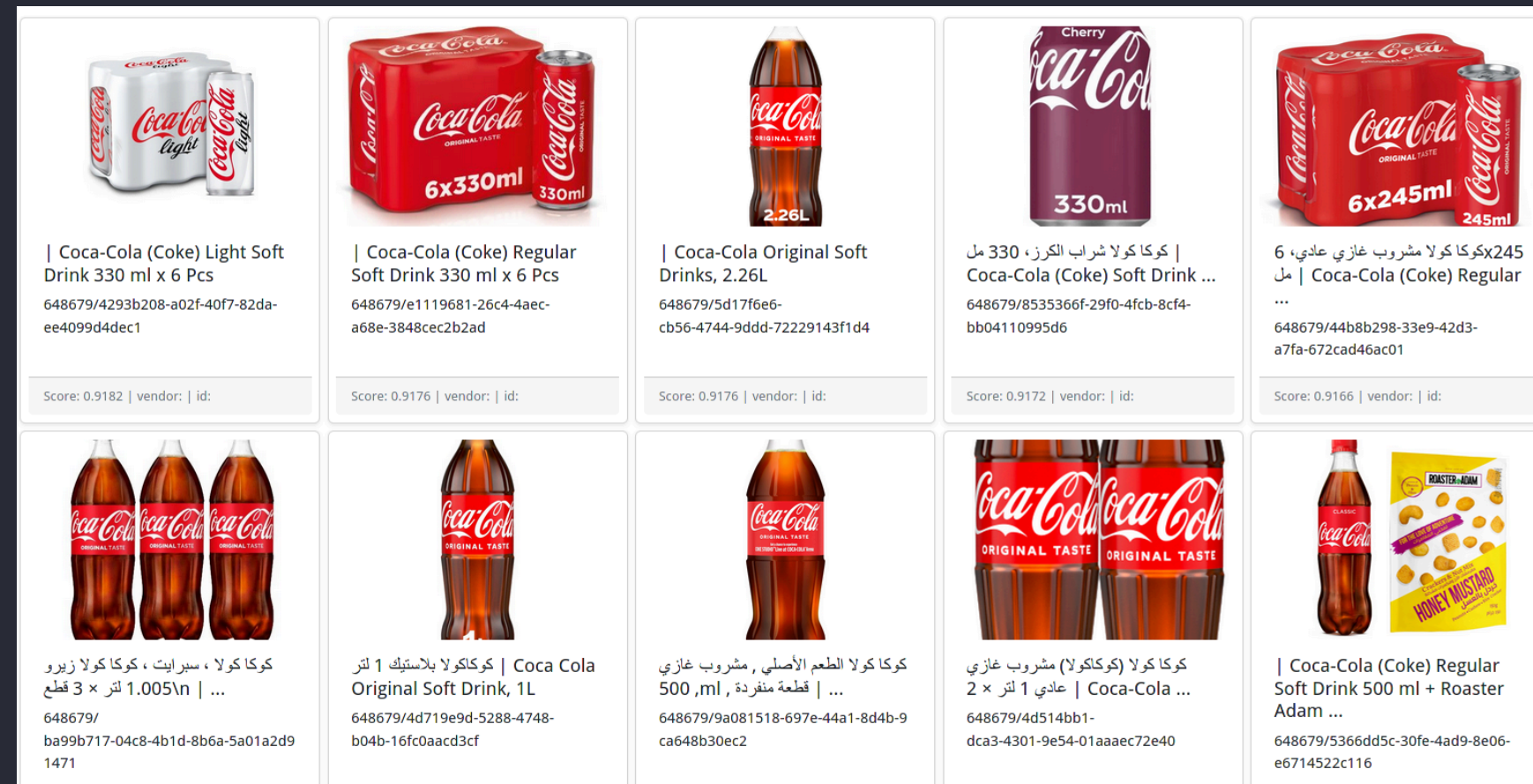
- Large inventory: ~20M items
- Diverse multi-token requests
- ~10% (**OMG!**) zero results rate

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

Index

TB_AE
▼











Query

tomato

Retrieval











Semantic baseline
▼

Search

 <p>CHOPPED TOMATOES</p>				
<p>Pomi Chopped Tomatoes 500 g</p> <p>87eed9eb-0126-48eb-a68d-837228337cb9</p> <p style="font-size: 0.8em; color: #666;">Score: 0.9171 vendor: 648679 id: 87eed9eb-0126-48eb-a68d-837228337cb9</p>	<p>Pure Harvest Local Yoom Tomatoes 250g</p> <p>a1b19ed7-7bae-4c28-a16b-dab5abfeec4a</p> <p style="font-size: 0.8em; color: #666;">Score: 0.9167 vendor: 648679 id: a1b19ed7-7bae-4c28-a16b-dab5abfeec4a</p>	<p>Mtr Tomato Rice 250G</p> <p>579ad4fd-30d8-45aa-9564-ca7c7ae20cea</p> <p style="font-size: 0.8em; color: #666;">Score: 0.9166 vendor: 648679 id: 579ad4fd-30d8-45aa-9564-ca7c7ae20cea</p>	<p>Tomato Roma GCC 1kg</p> <p>de25fba6-a8b7-4875-a1b2-24f8dcc34b19</p> <p style="font-size: 0.8em; color: #666;">Score: 0.9166 vendor: 648679 id: de25fba6-a8b7-4875-a1b2-24f8dcc34b19</p>	<p>Bioitalia Organic Peeled Tomatoes 400 g</p> <p>96dc4050-bb1e-4914-87f0-a52fbf0be93b</p> <p style="font-size: 0.8em; color: #666;">Score: 0.9163 vendor: 648679 id: 96dc4050-bb1e-4914-87f0-a52fbf0be93b</p>
				
<p>Bioitalia Organic Rustic Diced Tomatoes 420 g</p> <p>a5bbd7da-645e-4a9a-8d7d-745daa7dd8da</p>	<p>Bioitalia Organic Chopped Tomatoes 400 g</p> <p>2ffd222f-f7f2-4c80-9eac-a9ef3aad6b12</p>	<p>Petti Il Delicato Tetra Brik Square Con Tappo Passata Extra ...</p> <p>a86a9768-cb11-41e1-9b86-b875c72e064f</p>	<p>Petti 100% Italian Peeled Plum Tomatoes, 400g</p> <p>8313b304-d645-4d9d-9542-0220bd7620b4</p>	<p>Pomi Tomato Paste Tube 200 g</p> <p>8c2aea21-bfaf-4846-8a2e-9706c5755b2c</p>

Customer intent

Index: TB_AE | Query: tomato | Retrieval: Semantic baseline | Search

 <p>CHOPPED TOMATOES</p> <p>Pomi Chopped Tomatoes 500 g</p> <p>87eed9eb-0126-48eb-a68d-837228337cb9</p> <p>Score: 0.9171 vendor: 648679 id: 87eed9eb-0126-48eb-a68d-837228337cb9</p>	 <p>Pure Harvest Local Yoom Tomatoes 250g</p> <p>a1b19ed7-7bae-4c28-a16b-dab5abfeec4a</p> <p>Score: 0.9167 vendor: 648679 id: a1b19ed7-7bae-4c28-a16b-dab5abfeec4a</p>	 <p>MTR Tasty Delights READY TO EAT</p> <p>Tomato Rice</p> <p>Mtr Tomato Rice 250g</p> <p>579ad4fd-30d8-45aa-9564-ca7c7ae20cea</p> <p>Score: 0.9166 vendor: 648679 id: 579ad4fd-30d8-45aa-9564-ca7c7ae20cea</p>	 <p>Tomato Roma Grapes</p> <p>de25fba6-a8b7-4875-a1b2-24f8dcc34b19</p> <p>Score: 0.9166 vendor: 648679 id: de25fba6-a8b7-4875-a1b2-24f8dcc34b19</p>	 <p>Bioitalia pomodori pelati biologici organici peeled tomatoes</p> <p>Peeled Sauce 400</p> <p>Bioitalia Organic Peeled Tomatoes 400g</p> <p>96dc4050-bb1e-4914-87f0-a52fbf0be93b</p> <p>Score: 0.9163 vendor: 648679 id: 96dc4050-bb1e-4914-87f0-a52fbf0be93b</p>
 <p>Diced Tomatoes 420</p> <p>Bioitalia Organic Rustic Diced Tomatoes 420 g</p> <p>a5bbd7da-645e-4a9a-8d7d-745daa7dd8da</p>	 <p>Chopped Tomatoes 400</p> <p>Bioitalia Organic Chopped Tomatoes 400 g</p> <p>2ffd222f-f7f2-4c80-9eac-a9ef3aad612</p>	 <p>Petti IL DELICATO</p> <p>PASSATA EXTRAFINE</p> <p>Petti Il Delicato Tetra Brik Square Con Tappo Passata Extra ...</p> <p>a86a9768-cb11-41e1-9b86-b875c72e064f</p>	 <p>Petti 1925</p> <p>WHOLE PEELED TOMATOES</p> <p>POMODORI PELATI</p> <p>100% TUSCAN TOMATO</p> <p>400g</p> <p>Petti 100% Italian Peeled Plum Tomatoes, 400g</p> <p>8313b304-d645-4d9d-9542-0220bd7620b4</p>	 <p>Pomi TOMATO PASTE</p> <p>Tomato Paste Tube 200g</p> <p>8c2aea21-bfaf-4846-8a2e-9706c5755b2c</p>

40% conv rate

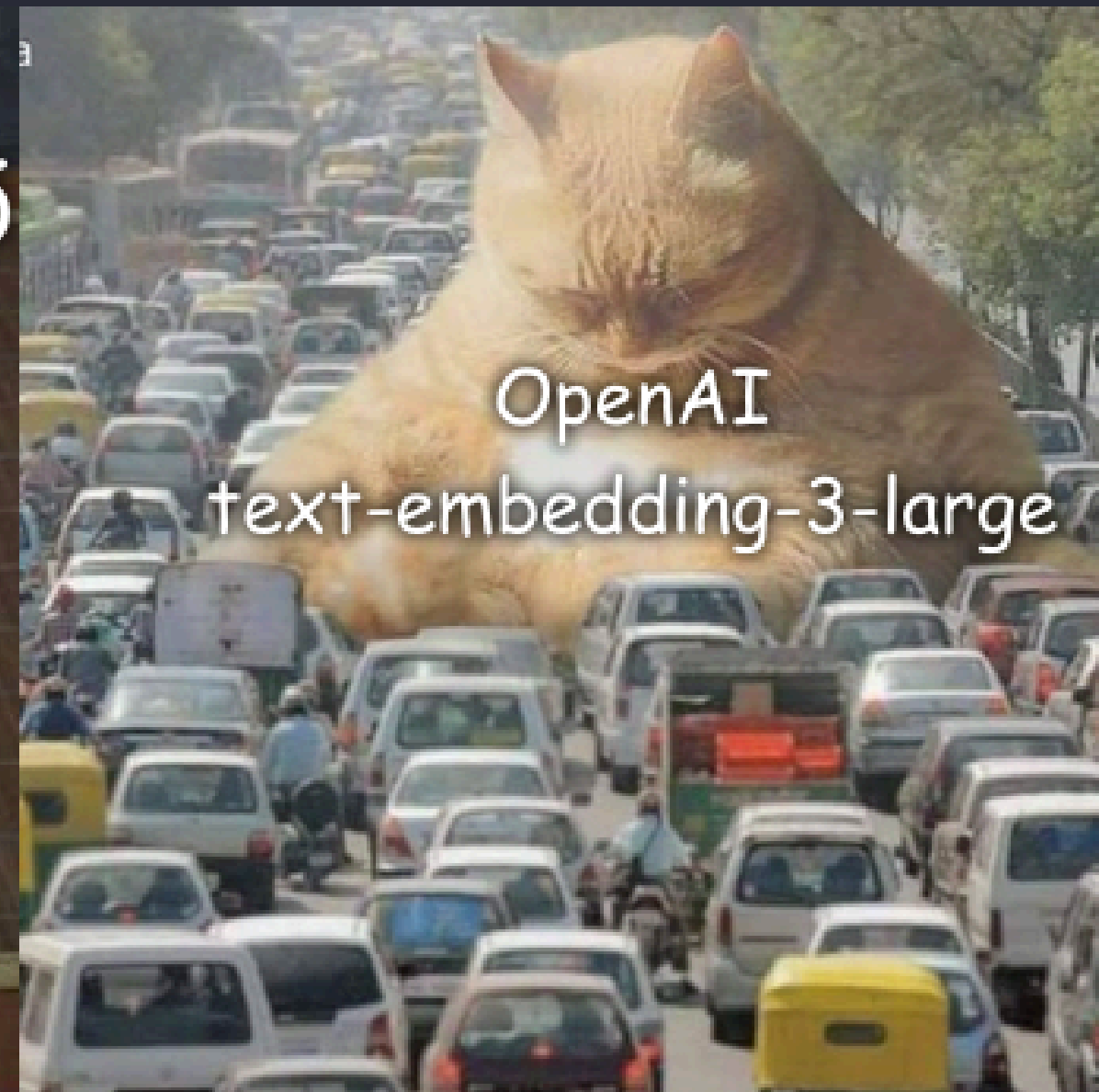
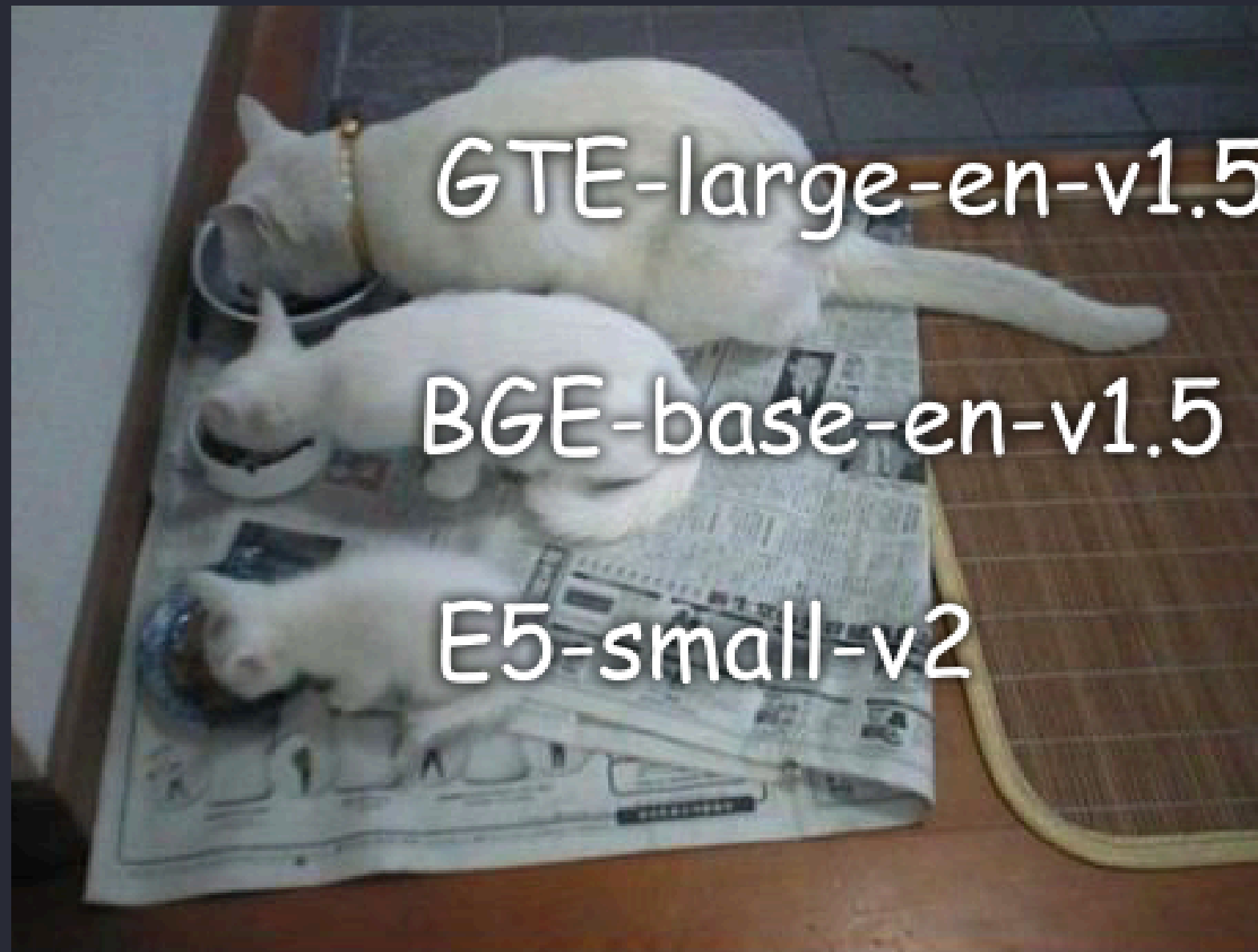
0% conv rate

made a mistake,
but...
where?



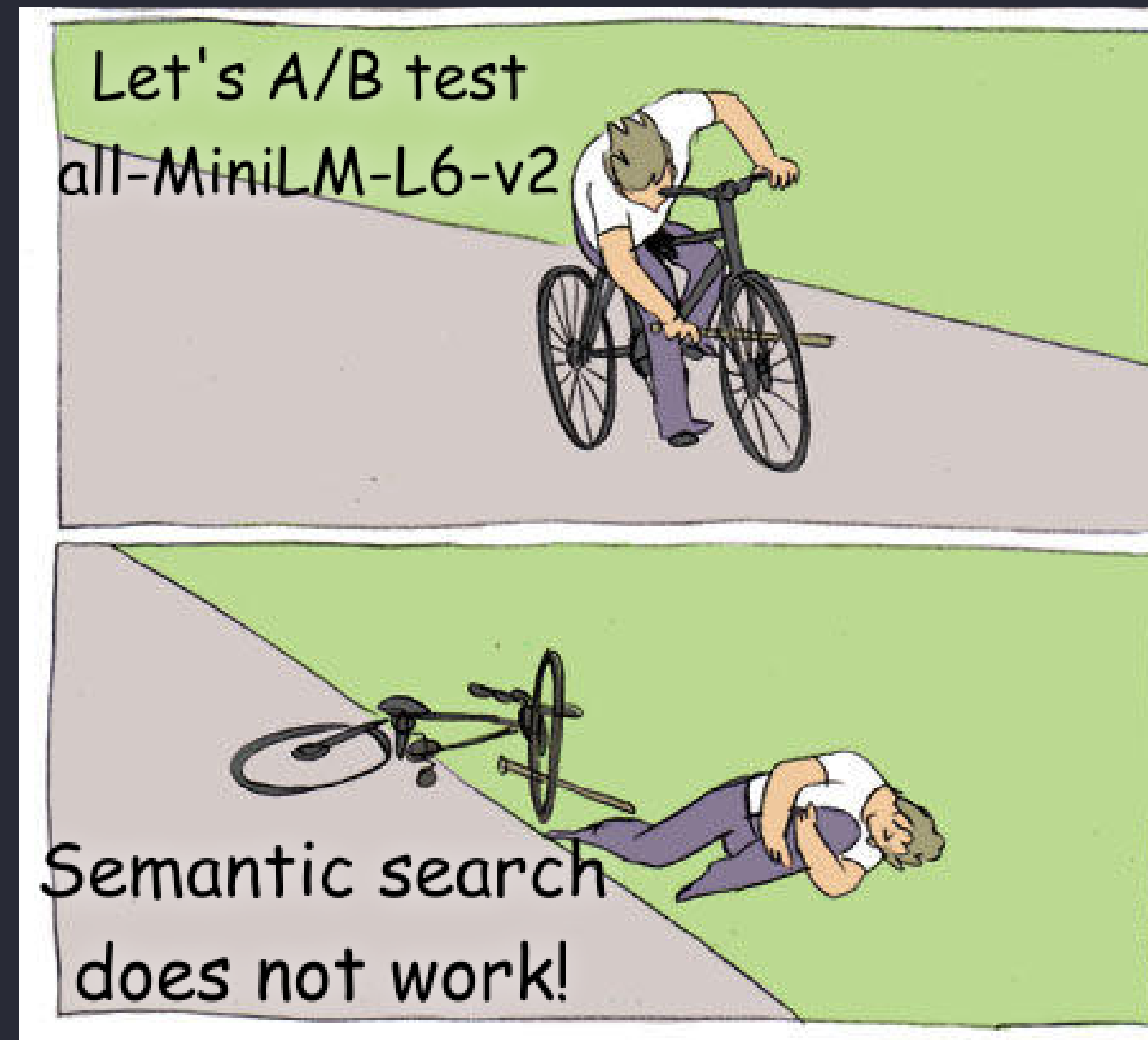
- **Relevance is subjective:** depends on intent
- **Embedding model:** no idea about your audience

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

The image displays four screenshots of search results for the query 'pizza', comparing lexical and semantic search models. Each screenshot shows a grid of search results with images, titles, and scores.

Top Left (Lexical Search - Finetuned e5-): Shows results for 'pizza' with a score of 0.7841. Results include Domino's Pizza, Pizza Hut, and Papa Johns.

Top Right (Lexical Search - Baseline e5-s1): Shows results for 'pizza' with a score of 0.8032. Results include Pizzalicious, The Pizza People, and Pizza Hut.

Bottom Left (Semantic Search - Finetuned e5-): Shows results for 'pizza' with a score of 0.7782. Results include Papa Johns, Pizza Hut, and Domino's.

Bottom Right (Semantic Search - Baseline e5-s1): Shows results for 'pizza' with a score of 0.9266. Results include Saporito, Inferno Flatbread Pizza, and Pick N Pair.

First step is still the same

You cannot improve search if you cannot measure it

The screenshot shows the Quepid search interface. At the top, there is a navigation bar with the Quepid logo and various menu items: Relevancy Cases, Books, Teams, Scorers, Notebooks, Video Tutorials, Knowledge Base, Wiki, and a user profile for Roman Grebennikov. Below the navigation bar, a list of queries is displayed with their respective scores and result counts:

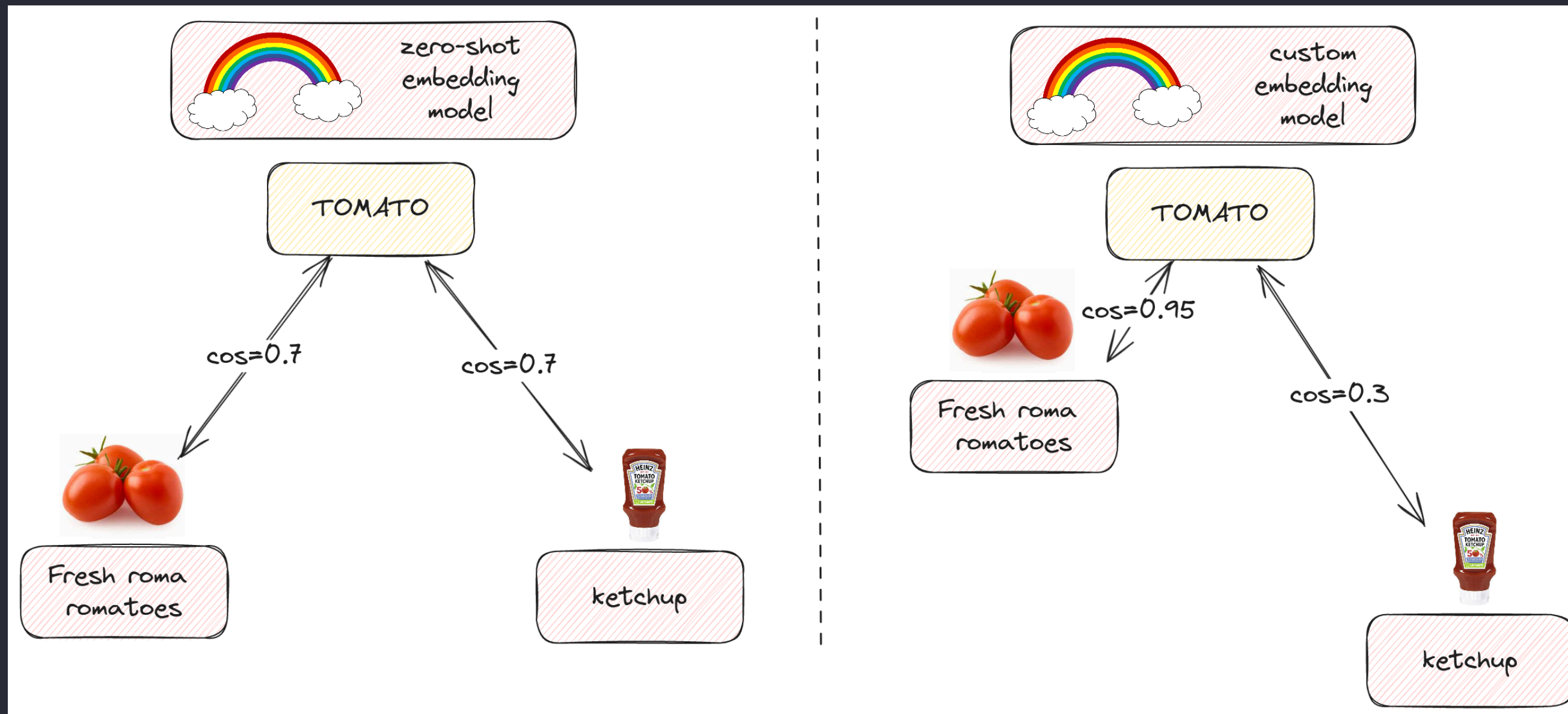
Score	Query	Results
1.00	হারদیز	10 Results
0.98	barbar	10 Results
0.99	albaik	10 Results
0.50	fish	10 Results
0.80	sushi	10 Results

Below the list, there are several action buttons: Score All, Toggle Notes, Explain Query, Missing Documents, Set Options, Move Query, and Delete Query. A detailed view of a match for the query 'sushi' is shown below. It includes a score of 2, a rating dropdown set to 2, and a set of buttons for rating: 0 Irrelevant, 1 Complement, 2 Substitute, 3 Exact, and a RESET button. The match itself is an image of a sushi platter with various types of sushi, including nigiri, maki, and a bowl of ramen. The image is labeled with 'image_url:' and 'is_ncr: True'. To the right of the image, there is a 'Matches' section with a dropdown menu set to 'no explain for doc'.

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?

	conv	impressions	CVR	
 CHOPPED TOMATOES Pomi Chopped Tomatoes 500 g	1	3	0.33	🤔
 Tomato Paste 200 Pomi Tomato Paste Tube 200 g	0	2	0.0	🤔
 Tomato Roma GCC 1kg	40	100	0.4	
 Mtr Tomato Rice 250G	0	300	0.0	

avg conv rate = 0.2

- 1 click, 3 impressions = 33% CTR?





Mixing clicks and confidence

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

- **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
 CHOPPED TOMATOES Pomi Chopped Tomatoes 500 g	1	3	0.33	0.230	0.207
 Tomato Paste 200 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	
 <p>CHOPPED TOMATOES Pomi Chopped Tomatoes 500 g</p>	1	3	0.33	0.230	0.207	LOW CONFIDENCE :(
 <p>Tomato Paste 200 Pomi Tomato Paste Tube 200 g</p>	0	2	0.0	0.166	0.192	
 <p>Tomato Roma GCC 1kg</p>	40	100	0.4	0.381	0.333	POSITIVE!
 <p>Mtr Tomato Rice 250G</p>	0	300	0.0	0.006	0.028	NEGATIVE!

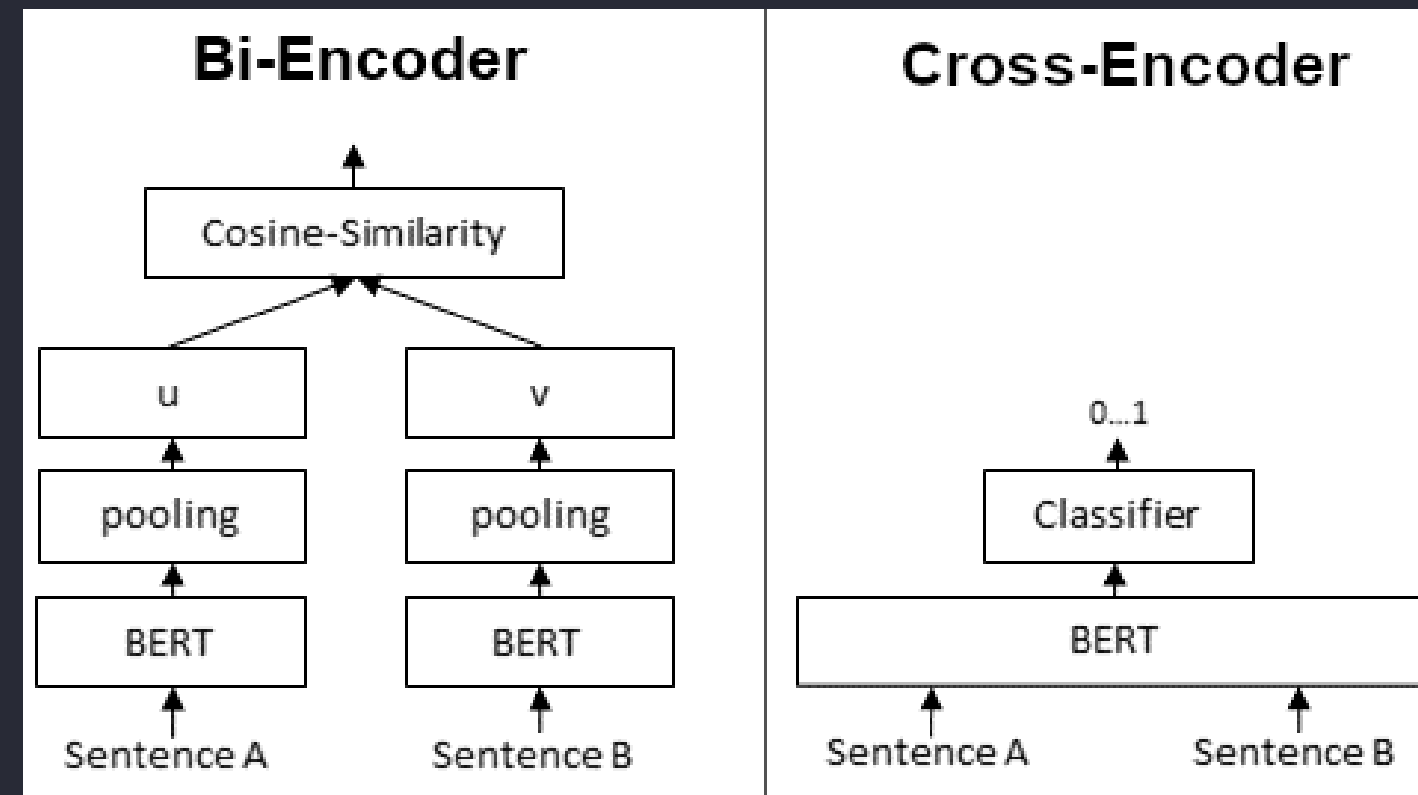
avg conv rate = 0.2

demo

Implicit labels are noisy

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

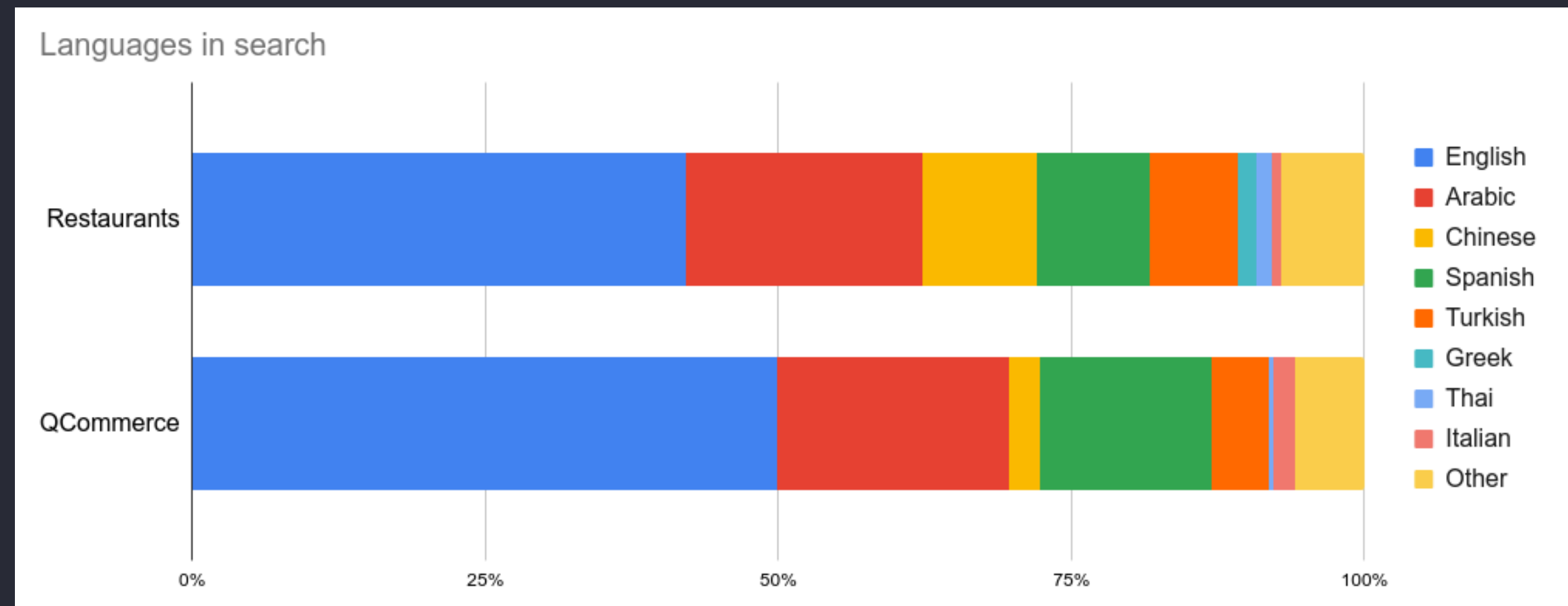
Future plans



- **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.

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)

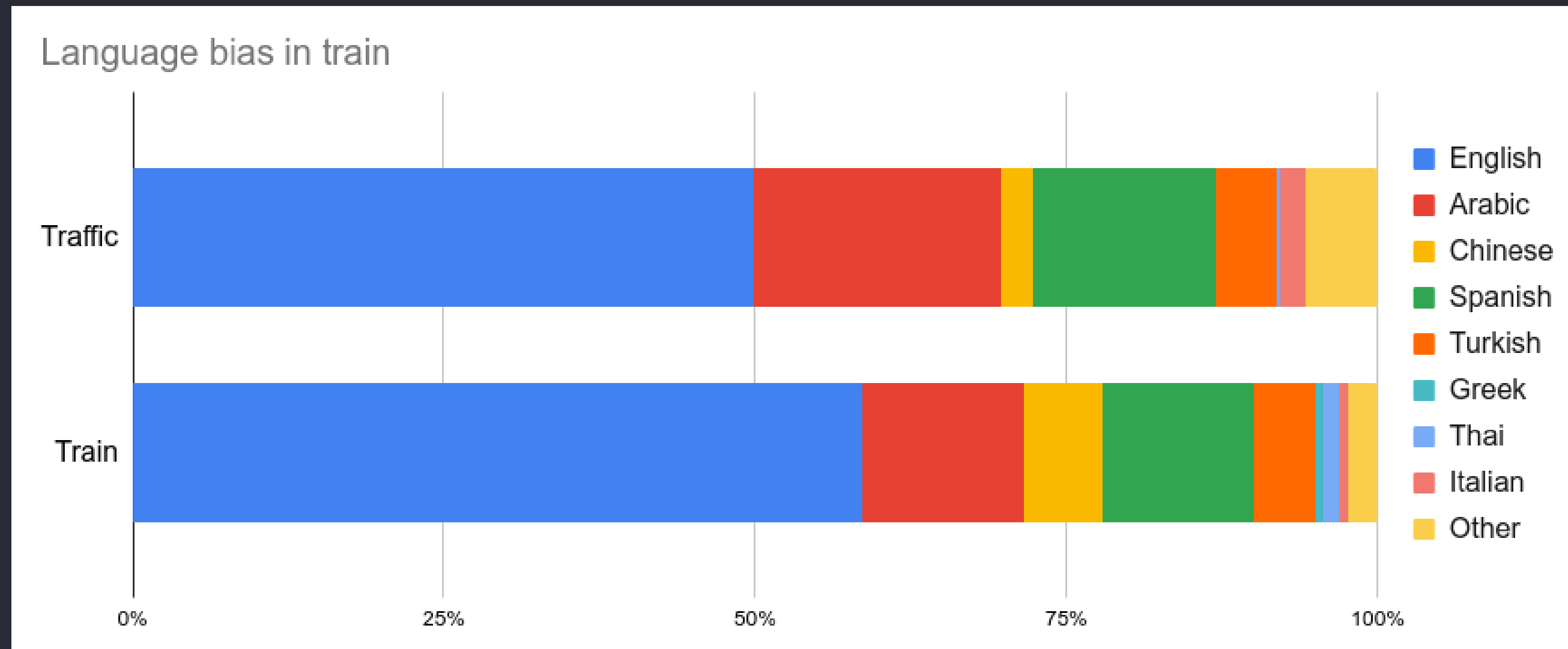
Food & groceries search - out of domain 🤔

[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

base	negs	format	batch	ndcg			
				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

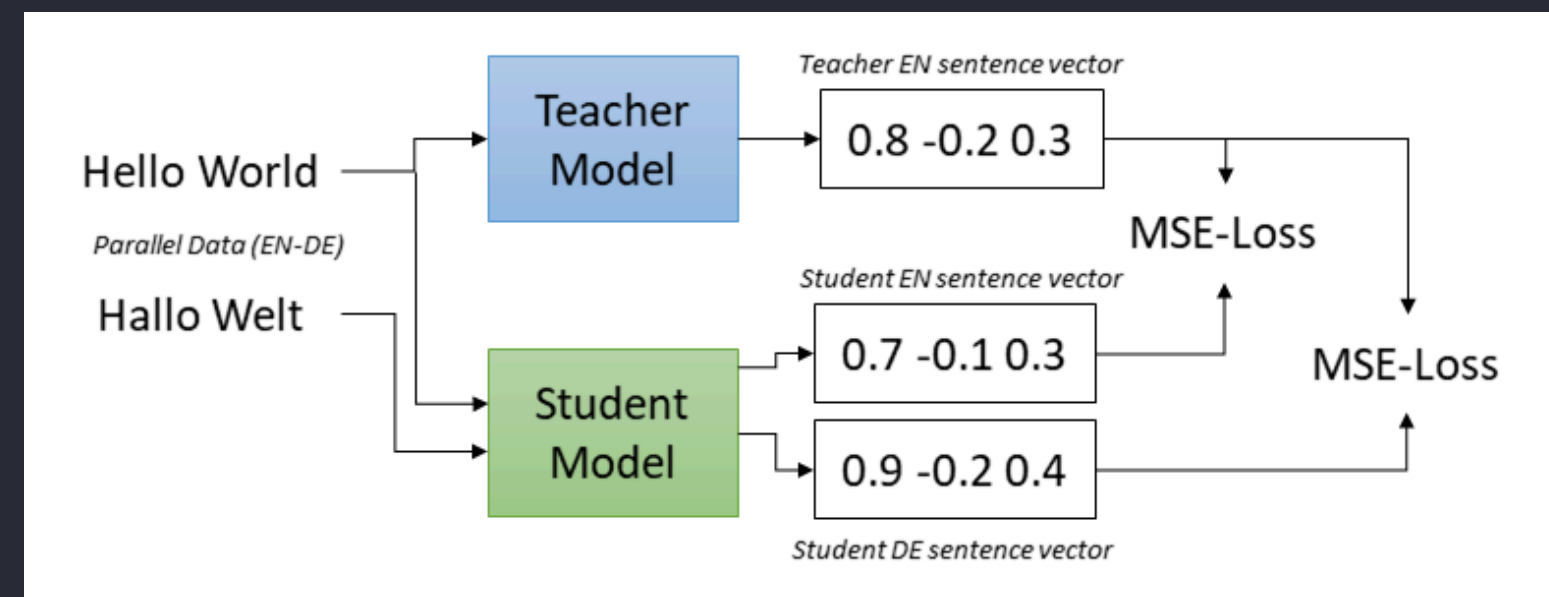
Worked well: mixed language data fine-tuning

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



Semantic search halting problem

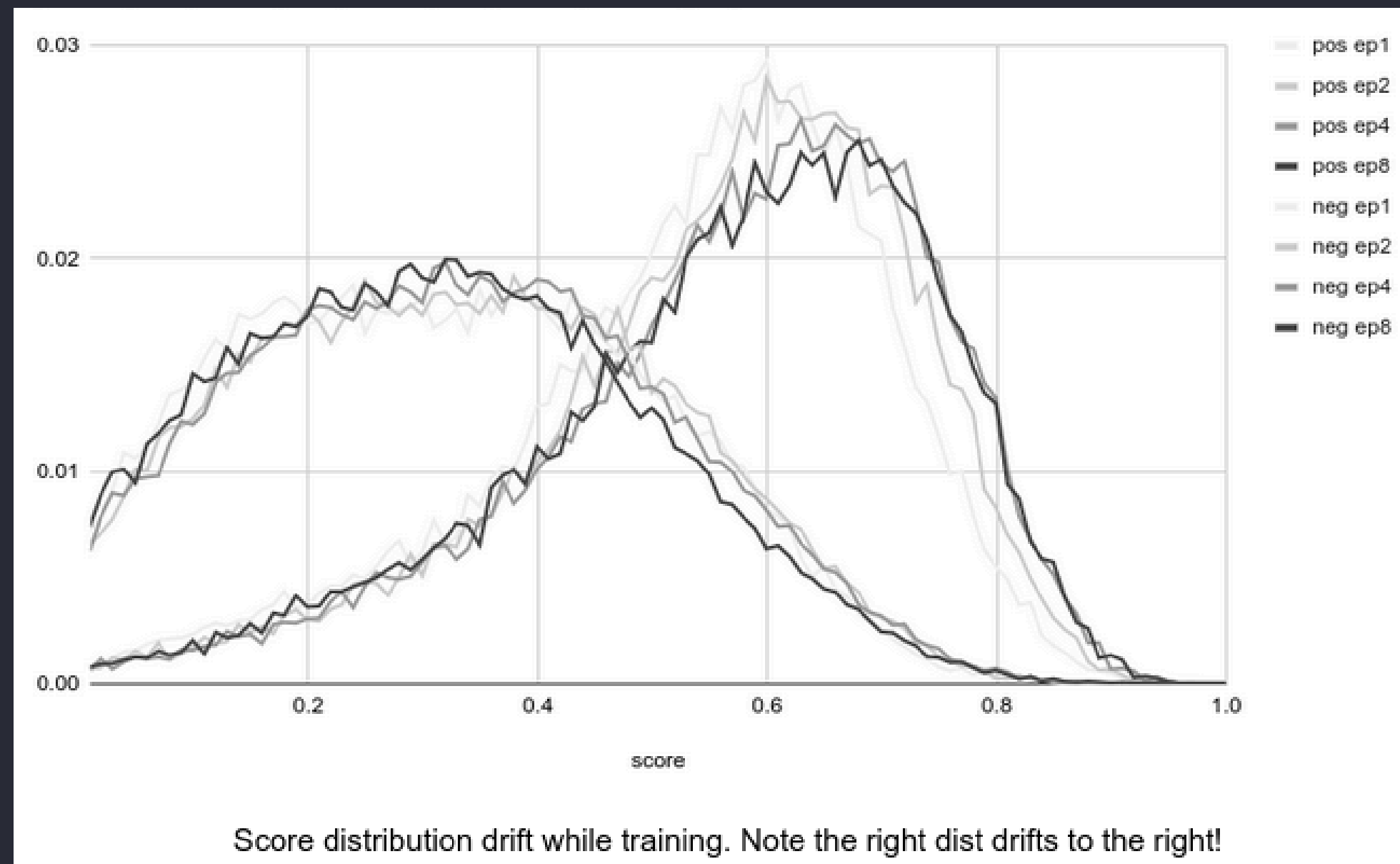
Problem: semantic search always has something found

Index: FP_SG | Query: hello from mix camp ecommerce search! | Retrieval: Fine-tuned v1 | Search

 <p>Haribo Mix Zourr 80g Haribo Mix Zourr 80g b8ds/42904342</p> <p>Score: 0.6354 vendor: id:</p>	 <p>Nestle Kitkat Mini Mix 197.4g Nestle Kitkat Mini Mix 197.4g b8ds/45435282</p> <p>Score: 0.6273 vendor: id:</p>	 <p>Nestle Kitkat Mini Mix 197.4g Nestle Kitkat Mini Mix 197.4g b8ds/45435282</p> <p>Score: 0.6273 vendor: id:</p>	 <p>Hi Chew Grape And Strawberry Mix Chewy Fruit Candy 100g Hi ... b8ds/6828703</p> <p>Score: 0.6261 vendor: id:</p>
--	--	--	--

demo

Finding a perfect threshold

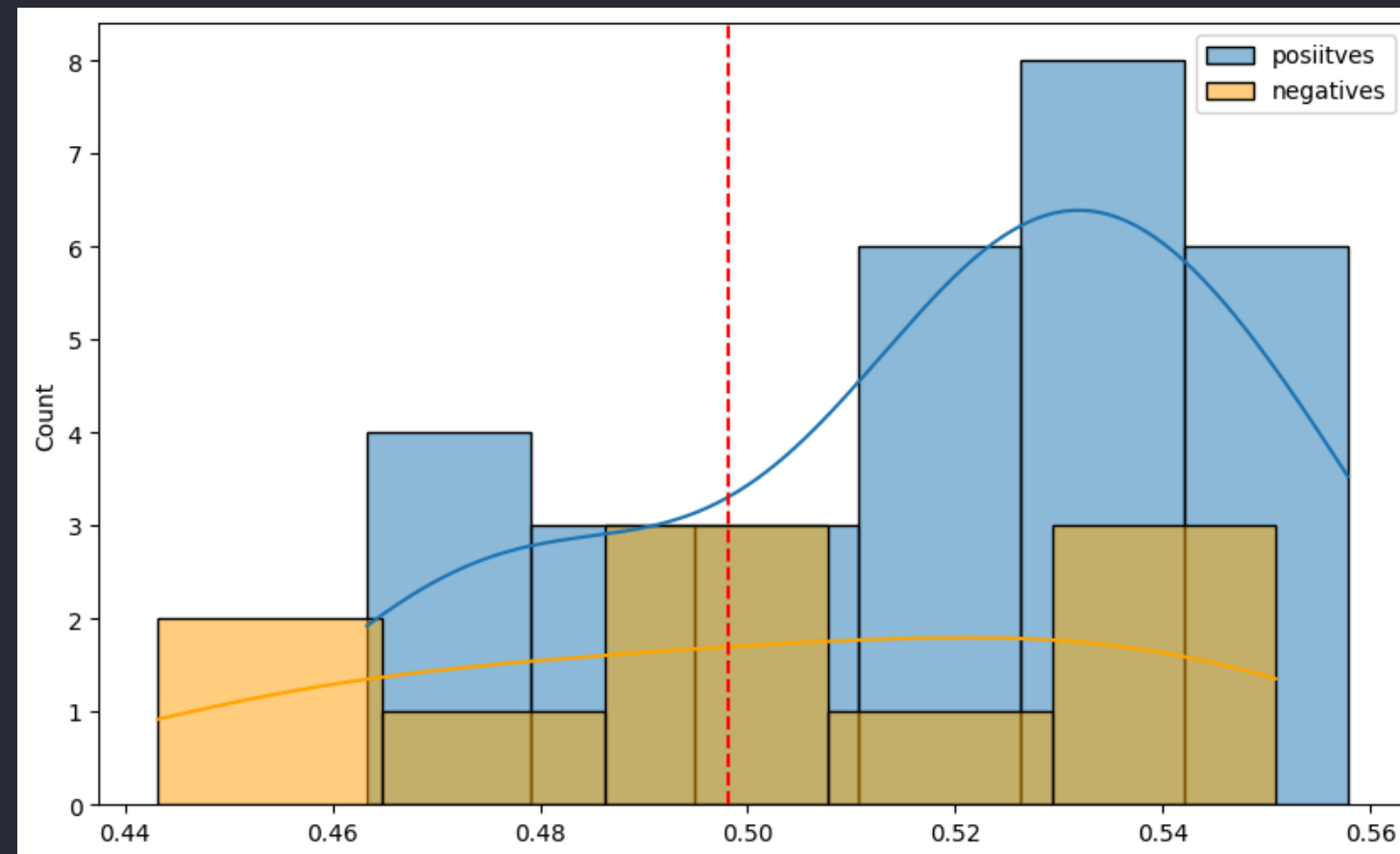


Attempt #1: set 0.7 as threshold => 70% zero results

Attempt #2: similar queries?

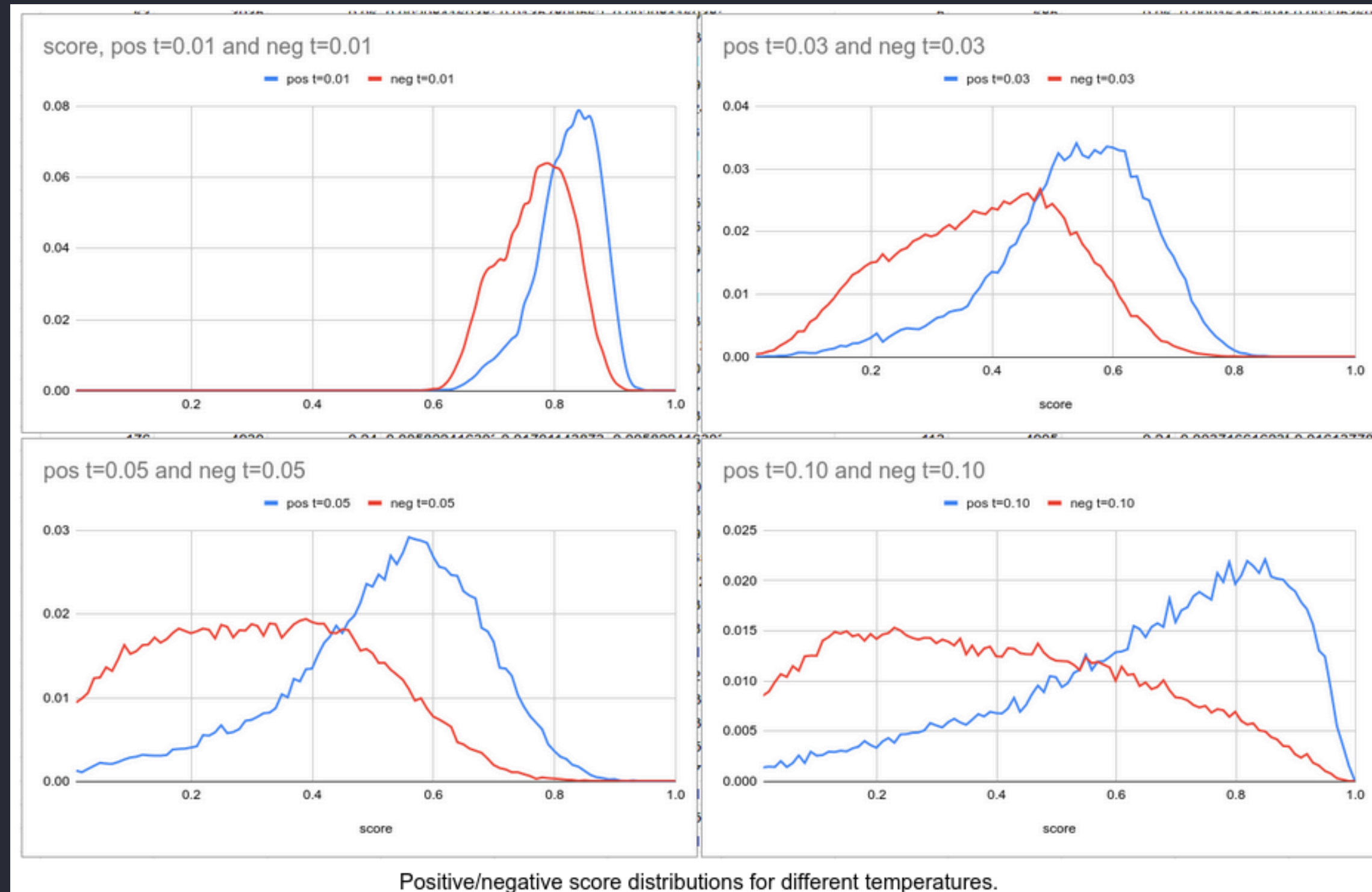
Ecommerce: **a lot of repeated queries!**

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



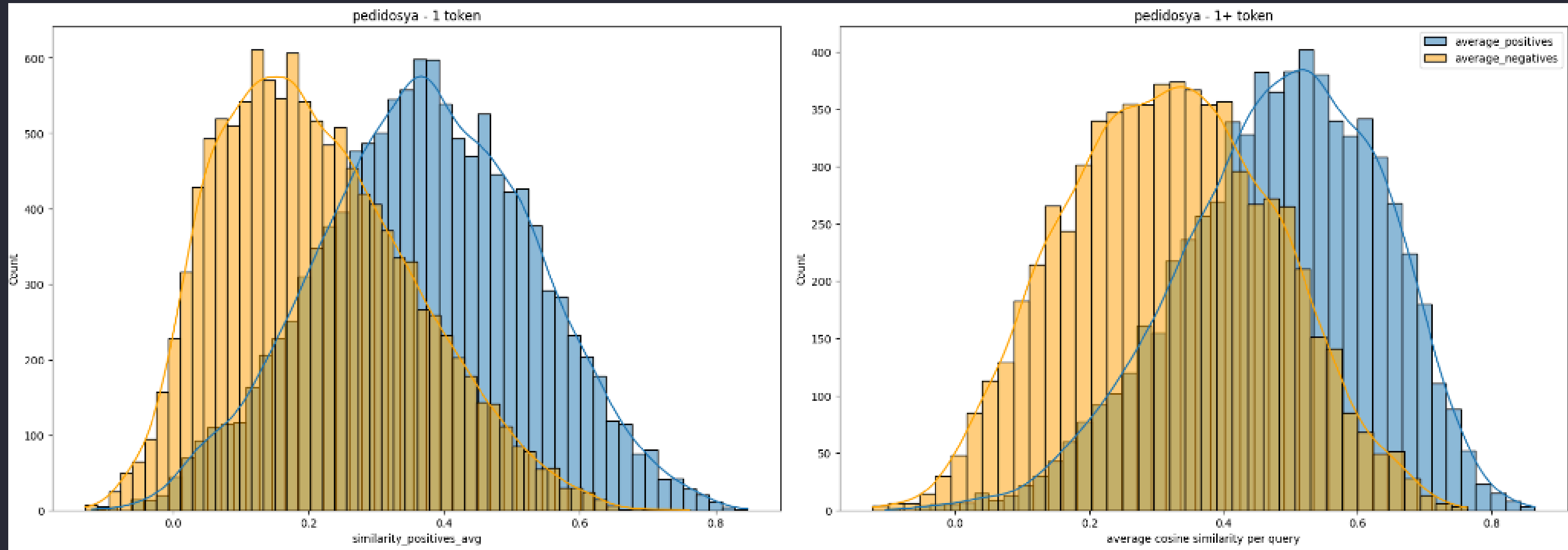
FAIL: too much noise

Threshold depends on the model!



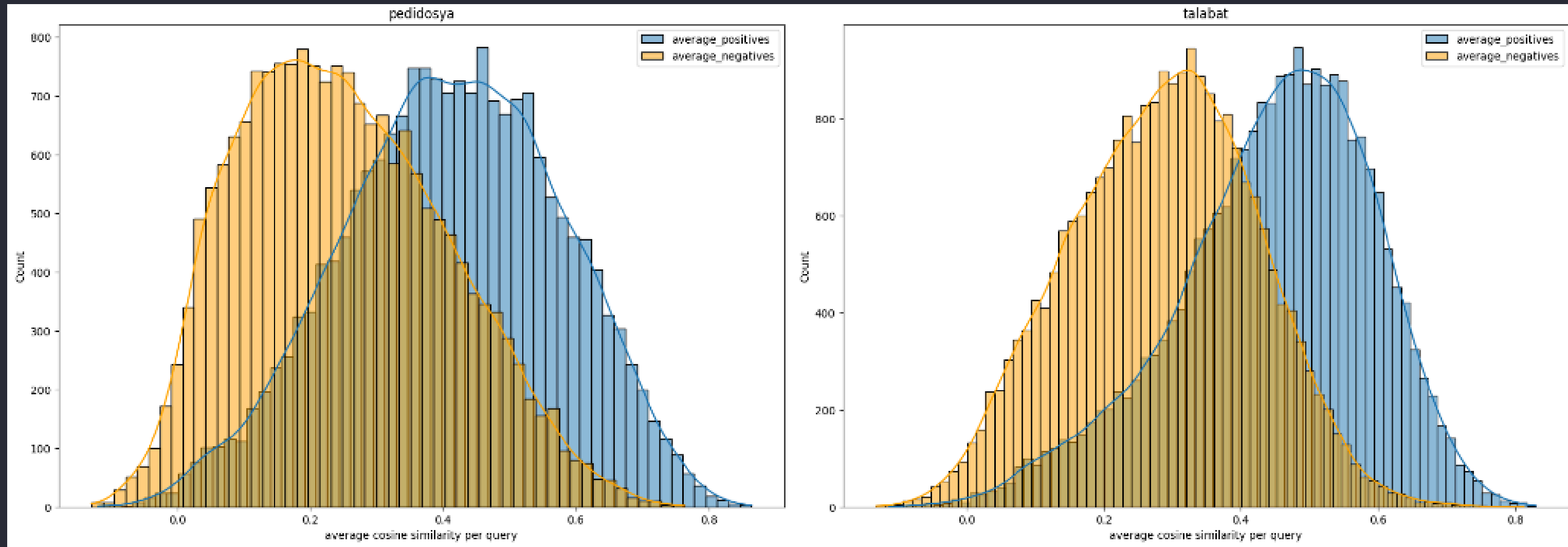
InfoNCE training temperature: model confidence level

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/](https://www.linkedin.com/in/romangrebennikov/)
- **MTEB Leaderboard:** huggingface.co/spaces/mteb/leaderboard
- **Sentence-transformers v3:** [sbert.net/](https://www.sbert.net/)

questions?