IntraFind Software AG – at a glance

Software - Made in Germany

Leading software manufacturer for intelligent search, text analysis & knowledge management

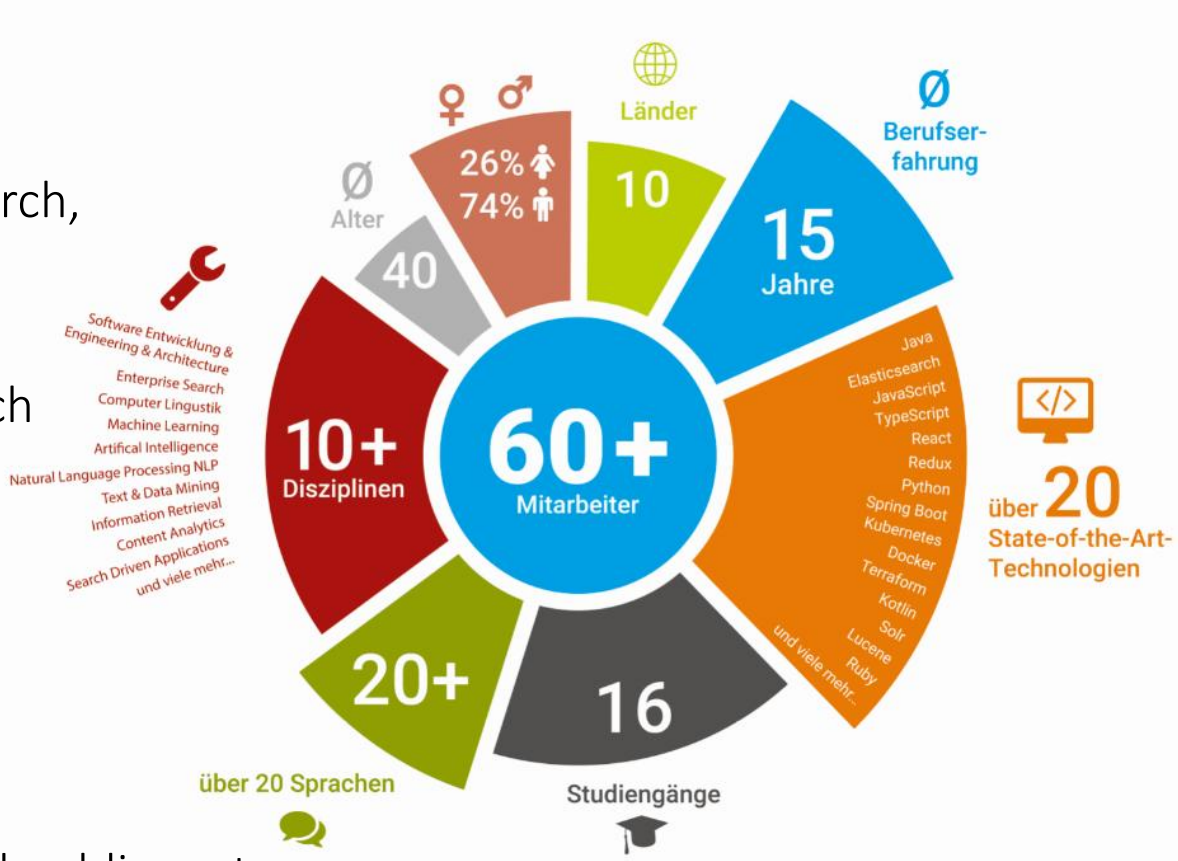
Established in 2000, headquarters: Munich, branch Berlin, Bonn, IntraFind Inc. in NY

65 employees

1,000 installations in the market, 35 partners

3.1 million daily users of iFinder in companies and public sector

WE ARE HIRING

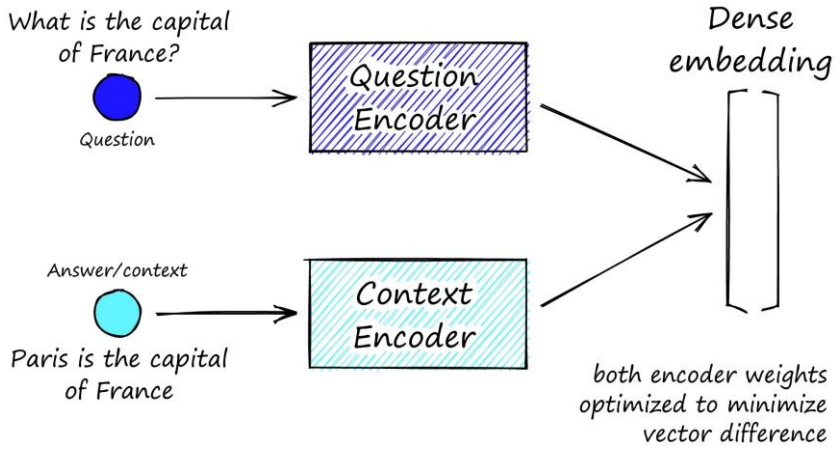


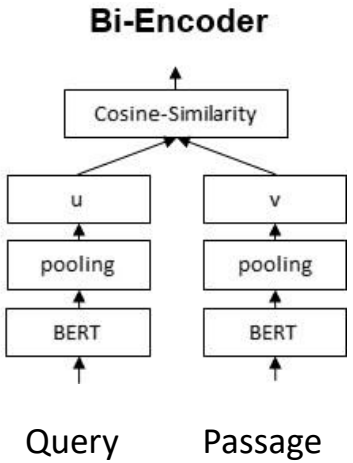
INTRAFIND

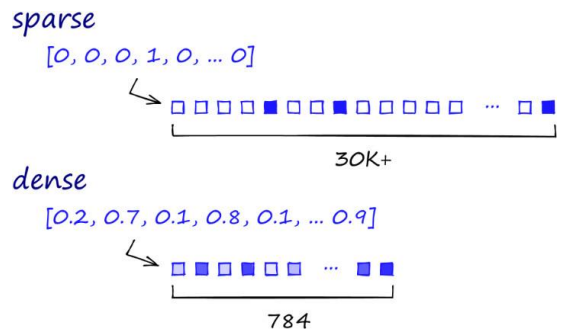
Maybe Dense Vector Search isn't as good as everybody thinks

Dr. Christoph Goller, Director Research

What is Dense Vector Search ?







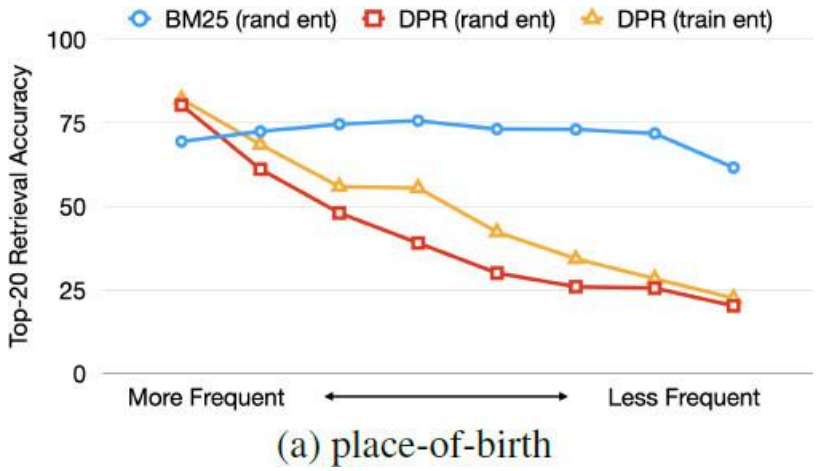
- Dense Vector Search is often used as a synonym for Semantic Search
  - What is Semantic Search?
- Vectors for query and documents (passages) are generated independently (Bi-Encoder) and document vectors can be stored in vector databases
- Similarity score for queries and documents is usually computed using the Dot-Product or Cosine-Similarity
- Embedding Models are trained on query / document pairs in a supervised way
- Multi-Modality, Multi-Lingual

Evidence that Dense Vector Search does not always work as expected

- BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models (2021) Thakur, Reimers et al.
  - Domain Shift: Embedding Models did not work in domains they had not been trained on
  - Best IR method wrt. to all benchmarks was BM25
- Simple Entity-Centric Questions Challenge Dense Retrievers (2022), Princeton
  - Dense Vector Search not very good for rare tokens → Hybrid Search
- Collapse of Dense Retrievers: Short, Early, and Literal Biases Outranking Factual Evidence, A. Modarressi, H. Schütze, ACL 2025
  - Dense retrieval models often prioritize superficial patterns, such as exact string matches, repetitive content, or information positioned early in documents, over deeper semantic understanding
- On the Theoretical Limitations of Embedding-Based Retrieval, 2025, Google Deepmind
  - Theoretical limitations, nice simple benchmark

Simple Entity-Centric Questions Challenge Dense Retrievers

	DPR (NQ)	DPR (multi)	BM25 -
Natural Questions	**80.1**	79.4	64.4
EntityQuestions (this work)	49.7	56.7	**72.0**
What is the capital of [E]?	77.3	78.9	**90.6**
Who is [E] married to?	35.6	48.1	**89.7**
Where is the headquarter of [E]?	70.0	72.0	**85.0**
Where was [E] born?	25.4	41.8	**75.3**
Where was [E] educated?	26.4	41.8	**73.1**
Who was [E] created by?	54.1	57.7	**72.6**
Who is [E]'s child?	19.2	33.8	**85.0**
(17 more types of questions)	...	...	...



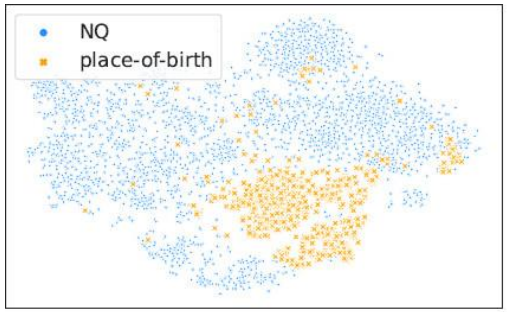
(a) place-of-birth

Top-20 retrieval accuracy, (Recall@20 ?)

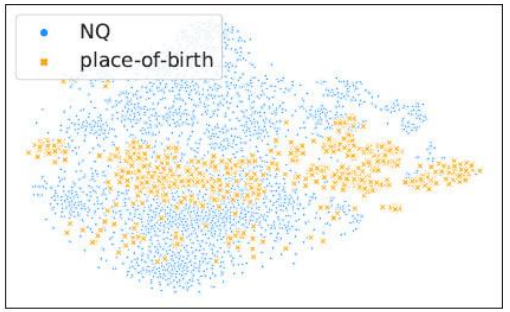
	NQ	Rel	EntityQ.
DPR-NQ	**80.1**	25.4	49.7
+ FT *p-of-birth*	62.8	**74.3**	**56.2**
+ FT NQ ∪ *p-of-birth*	70.8	52.0	47.4
DPR-NQ	**80.1**	70.0	49.7
+ FT *headquarter*	71.6	80.3	**53.3**
+ FT NQ ∪ *headquarter*	75.1	**81.3**	49.5
DPR-NQ	**80.1**	54.1	49.7
+ FT *creator*	70.8	**80.8**	**52.3**
+ FT NQ ∪ *creator*	72.6	72.3	44.1
BM25	64.4	-	**72.0**

Table 3: Top-20 retrieval accuracy on NQ and EntityQuestions. *FT*: fine-tuning. *Rel*: the performance on the relation that is used during fine-tuning.

Fine-Tuning on Entity Queries



(a)



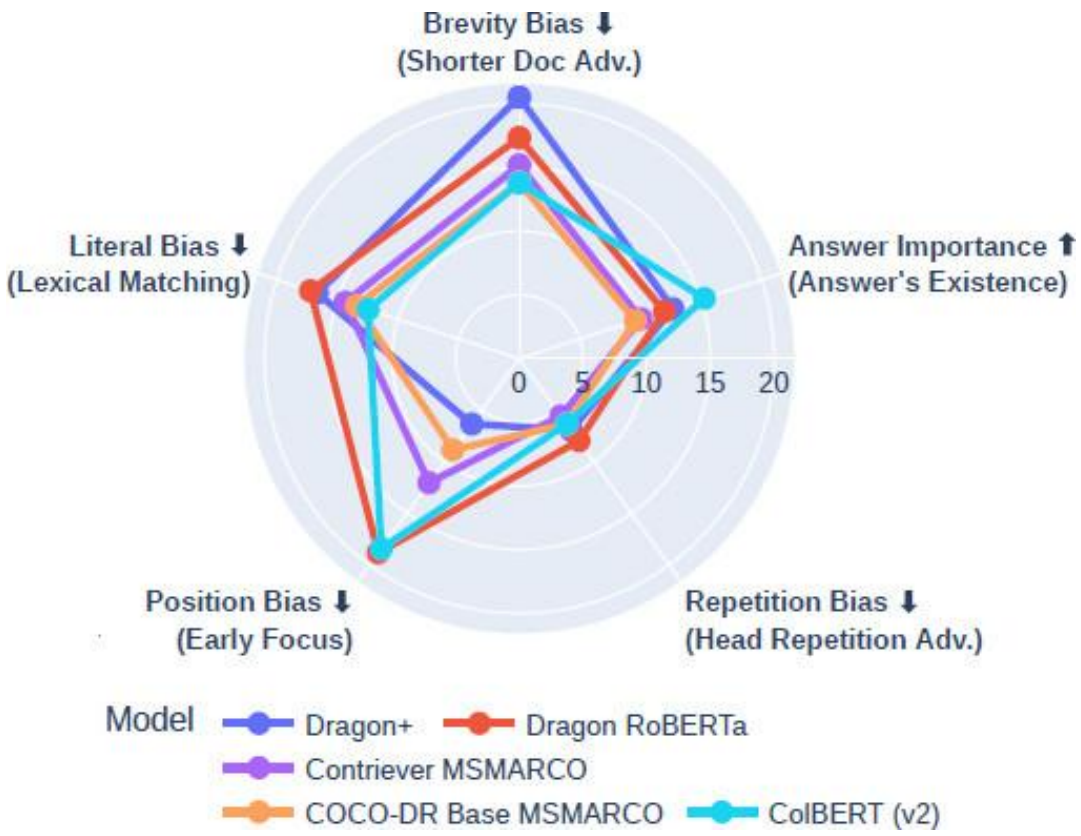
(b)

Collapse of Dense Retrievers

	Document 1 (Higher Query Document Similarity Score) - D_1	Document 2 (Lower Query Document Similarity Score) - D_2
Answer Impact	**Query:** What is the sister city of *Leonessa*? **Document:** *Leonessa* is twinned with the French town of *Gonesse* . Its population in 2008 was around 2,700 . Situated in a small plain at the foot of	**Query:** What is the sister city of *Leonessa*? **Document:** *Leonessa* is a town and comune in the far northeastern part of the Province of Rieti in the Lazio region of central Italy . Its population in 2008 was around 2,700 . Situated in a small plain at the foot of .....
Position Bias	**Query:** Which country is *Wonyong Sung* a citizen of? **Document:** *Wonyong Sung* (born 1950s) , *South Korean* professor of electronic engineering Won - yong is a Korean masculine given name ..... People with this name include : Kang Won - yong (1917 – 2006) ..... , South Korean swimmer	**Query:** Which country is *Wonyong Sung* a citizen of? **Document:** Won - yong is a Korean masculine given name ..... People with this name include : .... Jung Won - yong (born 1992) , South Korean swimmer *Wonyong Sung* (born 1950s) , *South Korean* professor of electronic engineering
Literal Bias	**Query:** When was *Seyhun* born? **Document:** *Seyhun* , (*August 22 , 1920* – May 26 , 2014) was an Iranian architect , sculptor , painter , scholar and professor . He studied fine arts at	**Query:** When was *Seyhun* born? **Document:** *Houshang Seyhoun* , (*August 22 , 1920* – May 26 , 2014) was an Iranian architect , sculptor , painter , scholar and professor . He studied fine arts at .....
Brevity Bias	**Query:** What series is *Lost Verizon* part of? **Document:** " *Lost Verizon* " is the second episode of *The Simpsons* ' twentieth season .	**Query:** What series is *Lost Verizon* part of? **Document:** " *Lost Verizon* " is the second episode of *The Simpsons* ' twentieth season . It first aired on the Fox network in the United States on October 5 , 2008 . Bart becomes jealous of his friends and their cell phones . Working at a golf course , Bart takes the cell phone of Denis Leary .....
Repetition Bias	**Query:** Where was *James Paul Maher* born? **Document:** Born in *Brooklyn , New York* , *Maher* graduated from St. Patrick 's Academy in Brooklyn . *James Paul Maher* (November 3 , 1865 – July 31 , 1946) was a U.S. Representative from New York . *Maher* was elected as a Democrat to the Sixty - second and to the four succeeding Congresses (March 4 , 1911 – March 4 , 1921) .	**Query:** Where was *James Paul Maher* born? **Document:** Born in *Brooklyn , New York* , *Maher* graduated from St. Patrick 's Academy in Brooklyn . Apprenticed to the hatter 's trade , he moved to Danbury , Connecticut in 1887 and was employed as a journeyman hatter . He became treasurer of the United Hatters of North America in 1897 .
Foil vs. Evidence	**Query:** Who is the publisher of *Assassin 's Creed Unity*? **Document:** " *Assassin 's Creed Unity* " " *Assassin 's Creed Unity* " *Assassin 's Creed Unity* received mixed reviews upon its release .	**Query:** Who is the publisher of *Assassin 's Creed Unity*? **Document:** Isa is a town and Local Government Area in the state of Sokoto in Nigeria . It shares borders with ..... *Assassin 's Creed Unity* is an action - adventure video game developed by Ubisoft Montreal and published by *Ubisoft* . Isa is a town and Local Government Area in the state of Sokoto in Nigeria . It shares borders with .....

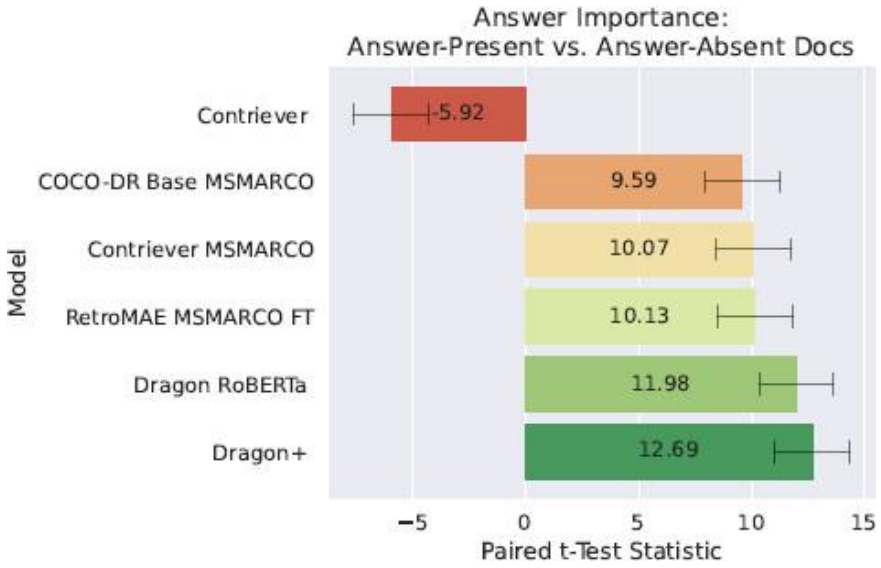
Explored Biases

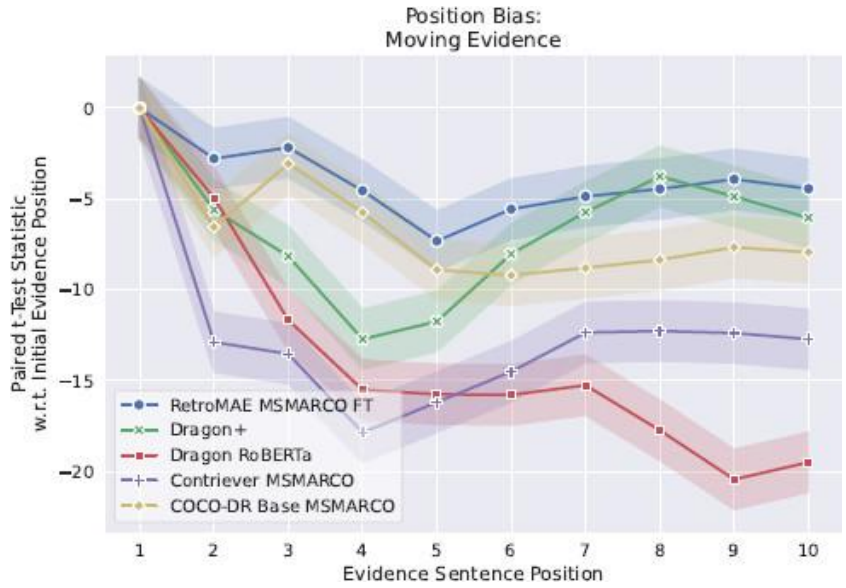
Collapse of Dense Retrievers



Paired t-test statistics comparing retriever scores between document pairs (D1 vs. D2) M(Query,D1) – M(Query,D2) where M is the retrieval score of the model)

Collapse of Dense Retrievers





Paired t-test statistics comparing retrieval scores between exact name matches (Q1-D1) and variant name pairs (Q2-D2)

			Model	Contriever MSMARCO	Dragon+
Q1	*D1*	*Q2*	*D2*		
long	**long**	**long**	**short**	+21.05	+21.04
		short	**long**	+22.04	+13.40
short	**short**	**long**	**short**	+4.62	+9.04
		short	**long**	+14.37	+16.62

Collapse of Dense Retrievers

$$D_1 := 2 \times h + S_{-t}^{+h}$$

$$D_2 := 4 \times \tilde{S}_{-t}^{-h} + S_{ev} + 4 \times \tilde{S}_{-t}^{-h}$$

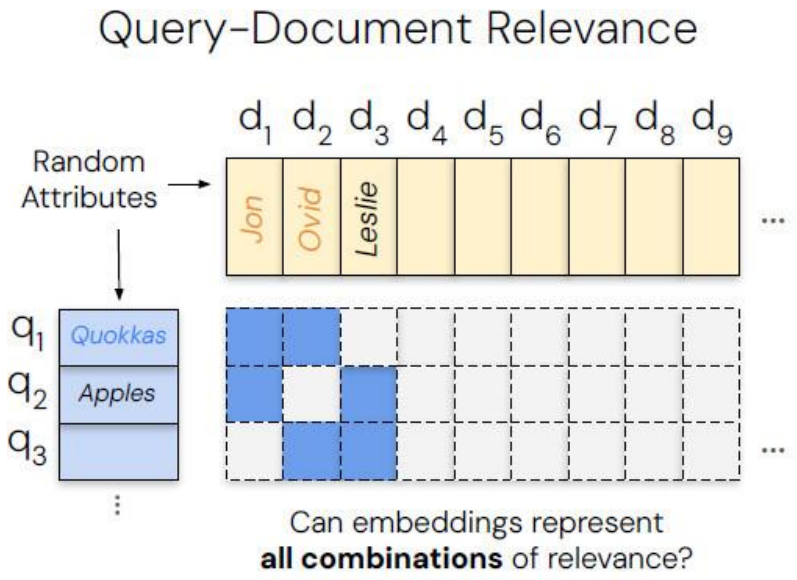
Interplay Between Bias Types

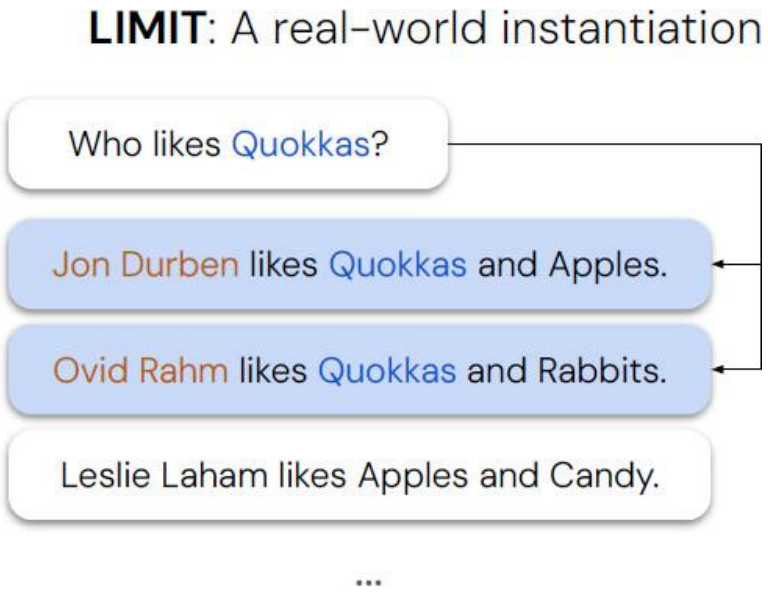
The accuracy and paired t-test comparing a foil document D1 to a second document D2 with evidence embedded in unrelated sentences. Accuracy is the proportion of 250 example pairs where M(Q,D2) > M(Q,D1)

All retrieval models perform extremely poorly (<10% accuracy)

Model	Accuracy	Paired t-Test Statistic	p-value
Contriever	0.4%	-34.58	< 0.01
RetroMAE MSMARCO FT	0.4%	-41.49	< 0.01
Contriever MSMARCO	0.8%	-42.25	< 0.01
Dragon RoBERTa	0.8%	-36.53	< 0.01
Dragon+	1.2%	-40.94	< 0.01
COCO-DR Base MSMARCO	2.4%	-32.92	< 0.01
ColBERT (v2)	7.6%	-20.96	< 0.01
ReasonIR-8B	8.0%	-36.92	< 0.01

On the Theoretical Limitations of Embedding-Based Retrieval





- The query-document relevance matrix represents the gold-standard (**binary ground-truth relevance matrix**)
- A **score matrix** for a particular embedding model consists of the scores (dot-product of query and document vectors)
- There is a **minimum rank for a score matrix** that represents the relative order of entries of documents for each query, and **this rank is the minimum dimension for the embedding model!**
- For any embedding dimension we can construct a query-document relevance matrix for which there is no proper score matrix based on these embeddings. BUT: No analytical results

On the Theoretical Limitations of Embedding-Based Retrieval

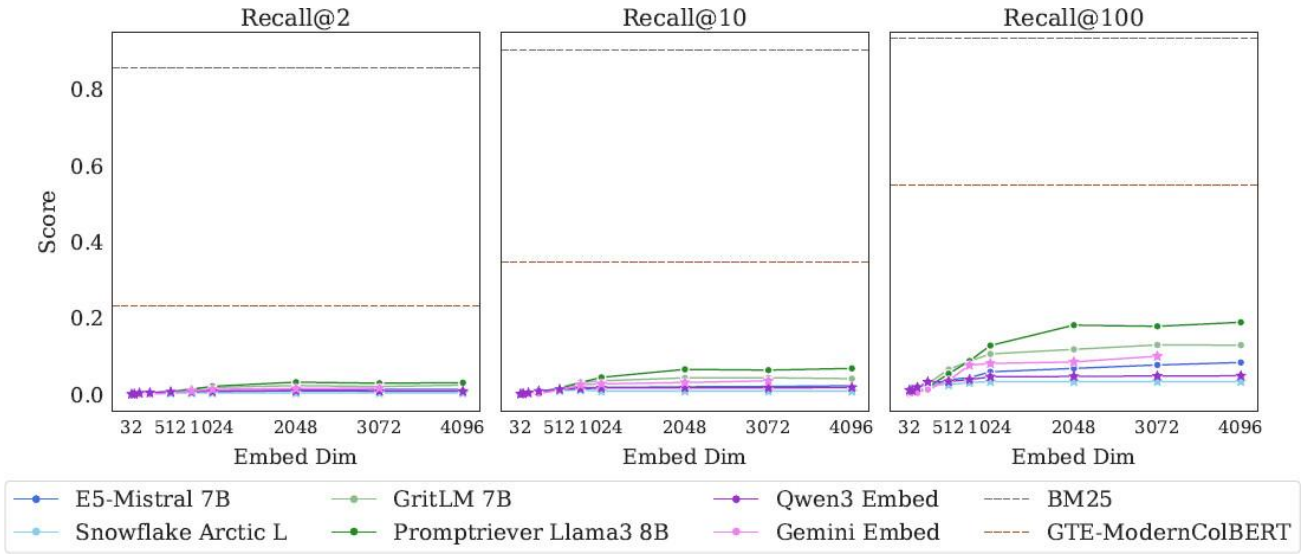
The Limit Dataset

- List of things a person could like: no duplicates / synonyms / hypernyms (size 1850)
- Documents describe what users like: keeping the documents short (less than 50 per user)
- 50k documents / each represents one person: random first and last name, randomly selecting 50 things they like
- Each query should only ask for one item to keep the task simple (i.e. “who likes X”)
- $\binom{46}{2}$ = 1035 queries each having 2 relevant documents: only 46 documents are relevant for these queries
- Query-Document matrix is very dense

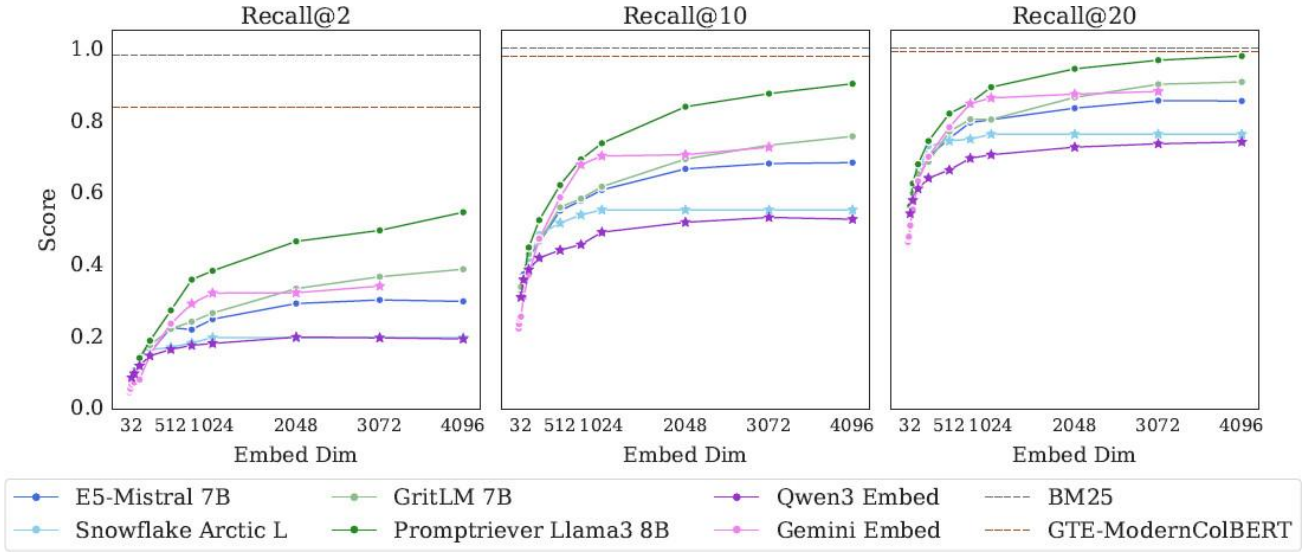
Example Document:

{"_id": "Olinda Posso", "title": "", "text": "Olinda Posso likes Bagels, Hot Chocolate, Pumpkin Seeds, The Industrial Revolution, Cola Soda, Quinoa, Alfajores, Rats, Eggplants, The Gilded Age, Pavements Ants, Cribbage, Florists, Butchers, Eggnog, Armadillos, Scuba Diving, Bammy, the Texas Rangers, Grey Parrots, Urban Exploration, Wallets, Rainbows, Juggling, Green Peppercorns, Dryers, Pulled Pork, Holland Lops, Blueberries, The Sound of Wind in the Trees, Apple Juice, Markhors, Philosophy, Orchids, Risk, Alligators, Peonies, Birch Trees, Stand-up Comedy, Cod, Paneer, Environmental Engineering, Caramel Candies, Lotteries and Levels."}

On the Theoretical Limitations of Embedding-Based Retrieval



Scores on the LIMIT task



Scores on the LIMIT small task (N=46)

On the Theoretical Limitations of Embedding-Based Retrieval

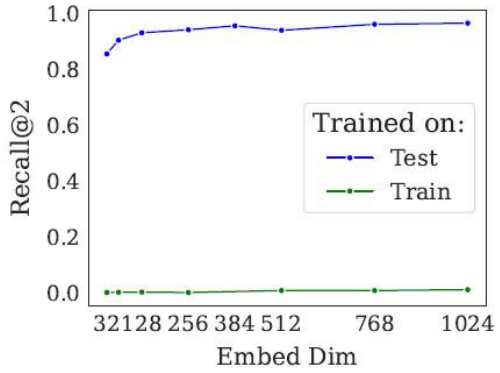
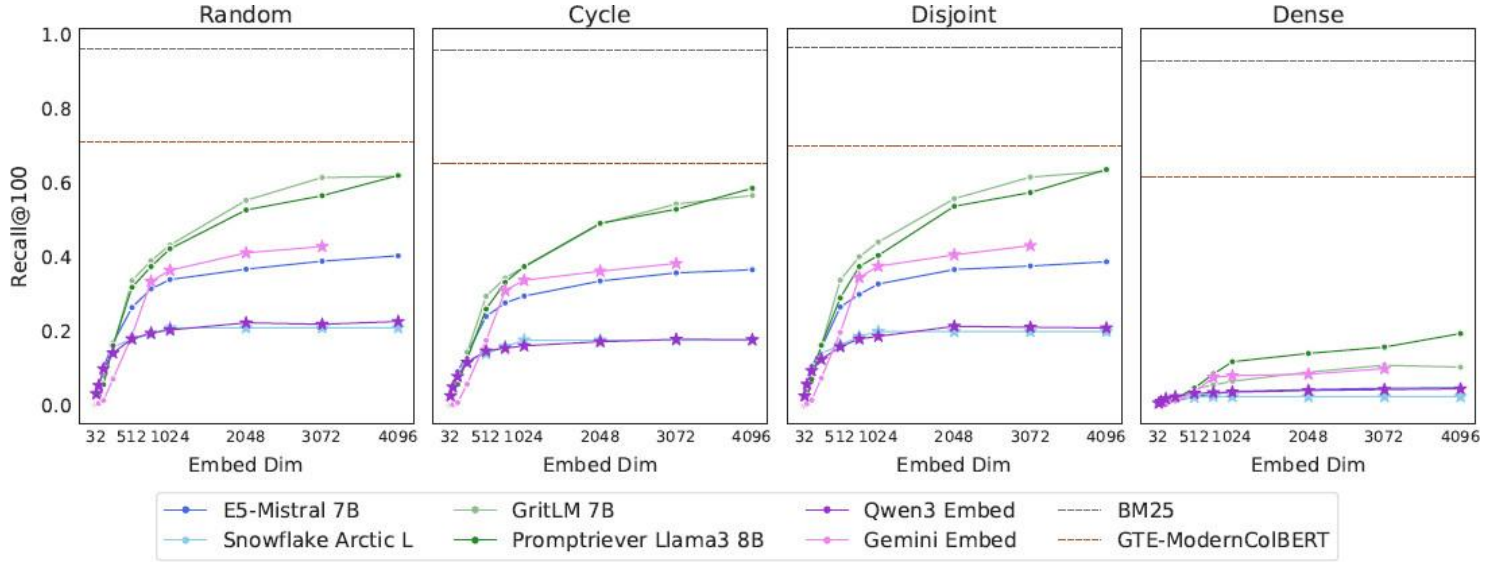
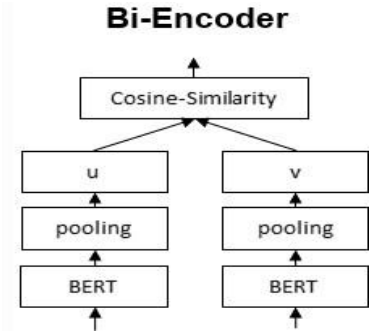


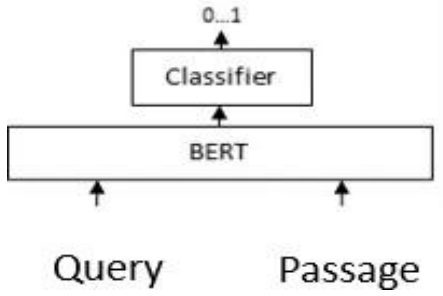
Figure 5 | Training on LIMIT train does not significantly help, indicating the issue is not domain shift. But models can solve it if they overfit to the test set.

It's not domain shift!



It's the "denseness" of the query-document relevance matrix



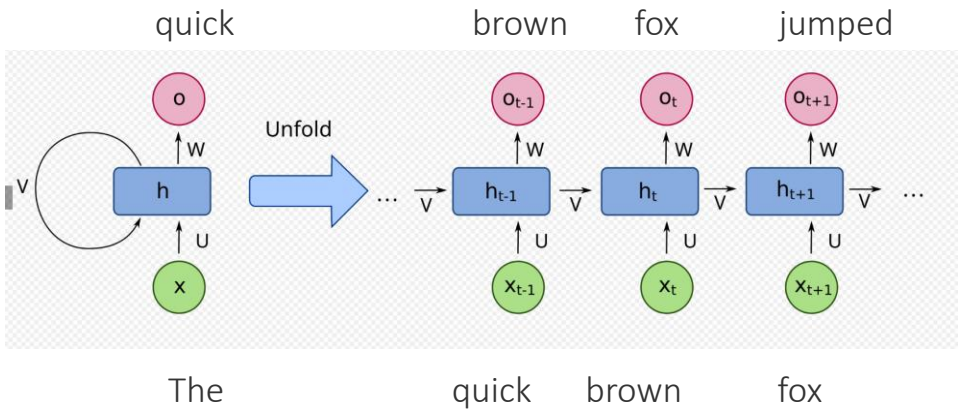


LIMIT benchmark is no problem for a cross-encoder!

What do these Results mean?

- Dense Vector Search may be not as great as we currently think!
- Why haven't we noticed?
  - How meaningful are results on existing IR benchmarks?
    - There seems to be a strong focus on the most popular (common knowledge) results!
    - Some relevant answers may have been overlooked
    - Average number of answers per test query:
      - MSMARCO test (95) dev (1.0), nq (1.2), Dbpedia (35)
    - Different concepts for "good passages":
      - MSMARCO: Focus on results that actually provide an answer and not only repeat the question
      - Dbpedia: Any entity that somehow relates to the question, passage not necessary contains answer?
  - We never really used classical search for passage retrieval
    - BM25 is optimized for IR on longer documents

LLMs are so great, so why isn't Dense Vector Search?



Deep Learning for NLP until 2017 / 2018:

- Elman Network 1990 -> LSTM (1997) / xLSTM
- Sequence processing, BPTT
- State which represents context read so far is ONE vector

- Dense Vector Search works like Deep Learning until 2017 bevor the Attention Mechanism
  - Text is represented as one dense vector
- Uses Simple Distance Measure

Deep Learning and the Attention Mechnism

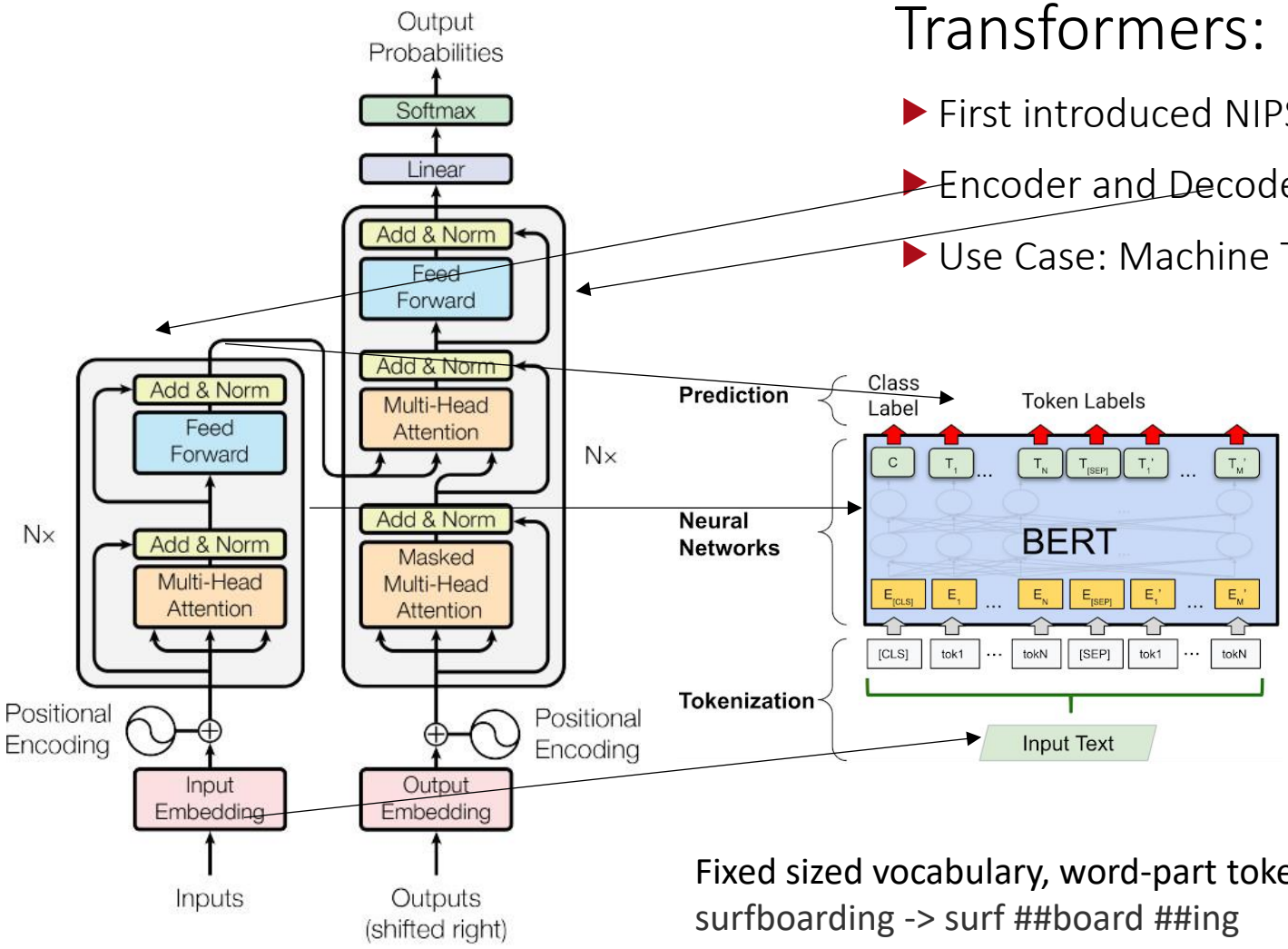
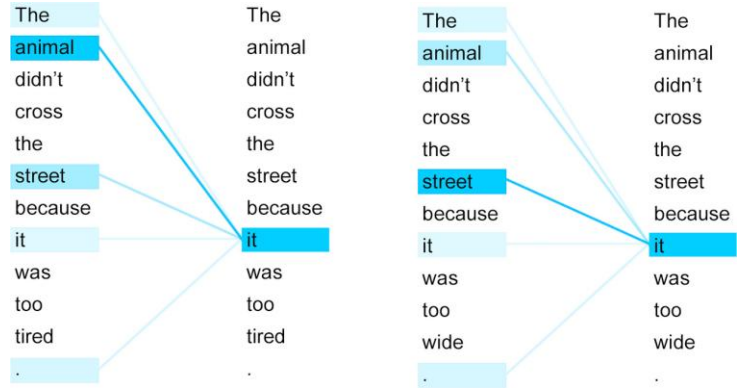


Figure 1: The Transformer - model architecture.

Transformers:

- First introduced NIPS 2017: “Attention Is All You Need” (Google)
- Encoder and Decoder consist of 6 layers
- Use Case: Machine Translation



Fixed sized vocabulary, word-part tokenization:
surfboarding -> surf ##board ##ing

The attention mechanism

Future of Semantic Search?

- Better Training for Encoding Models?
- Multi-Vector Models?
- Cross-Encoder: not suitable for first stage retrieval at scale?
- Sparse Semantic Search:
  - Reconsider “Semantic Search” based on Traditional Index:
    - Synonyms, Query Reformulations by LLMs
  - Sparse Semantic Search like Splade / SparTerm

#IntraFindSummit

Thanks for listening

Dr. Christoph Goller, Director Research

Tel: +49 89 3090446-0
E-Mail: christoph.goller@intrafind.com
Web: www.intrafind.com

IntraFind Software AG
Landsberger Straße 368
80687 München

Collapse of Dense Retrievers