75 186.2

Re

rhenium

05 2024

# An open-source search engine with Large Language Models in Julia programming language

Alexandra Elbakyan

# Rhenium

**The search. Reinvented.**


**...but we have search already!**

Scopus
PubMed
Google Scholar
Semantic Scholar...

existing solutions are...

# LTD

not accessible • narrow • limited resource usage • basic functionality • closed source • frozen • not intelligent

# Time for upgrade.

modern search engine:

- intelligent
- customizable
- open

# algorithm for semantic search

```
>  transform documents to vector embeddings

   while True:

       ... wait for search query from the user

       >  transform the query to vector embedding
       o  display documents with similar vectors
```

sample
vector
embedding

«A quick brown fox jumped
over the lazy dog»

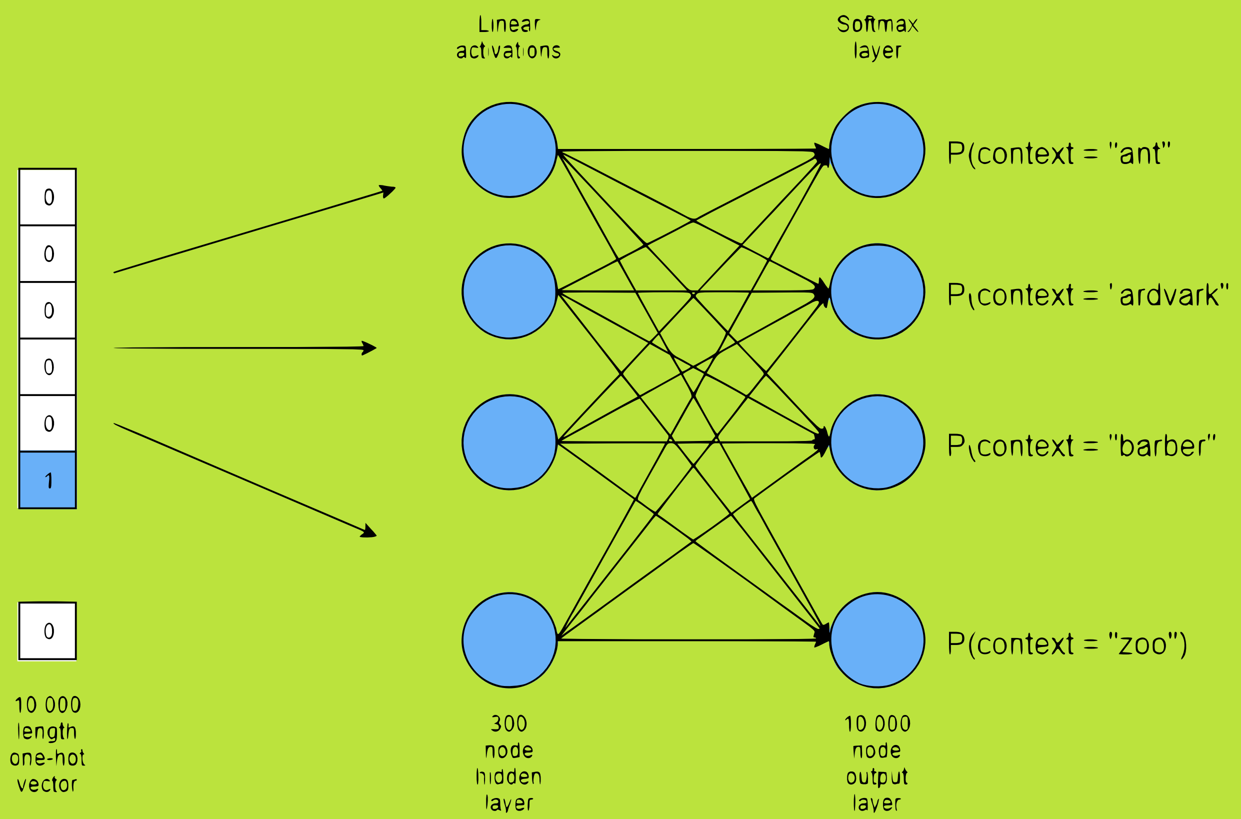
-0.025457121 0.88639474 -0.52687454 0.42610833
0.016878389 -0.1850919 0.2601459 0.5304479 -1.1237053
-0.22663605 0.9292201 -0.46498993 0.2563679 -1.003673
0.3213501 -1.2995219 -1.0151855 -0.32354406 0.6434173
-0.40887073 -0.17340171 0.34564218 0.34115812 -0.6829699
0.3323966 0.6897468 -0.41367245 0.9050691 0.27899414
0.3591251 0.06844619 -0.068081394 0.6580575 0.13972706
-1.0214432 0.21378978 -0.26289803 -0.52735114
-0.36524937 -0.29600602 0.15361847 -0.13564132
-1.0625327 -0.3357952 -1.1116723 0.46550542 -1.0217727
-0.54985404 -0.1699316 -0.033429492 -1.2249262
0.86744684 0.13178378 0.20464417 -0.07231642 -0.40423
-0.9075301 0.014228986 -0.6959451 0.17448542 -0.187875
-0.97408384 -0.30829257 -0.4895594 0.009005581 0.7310476
-0.58808756 0.32100388 -0.58030295 0.12345423
-0.32376578 0.080631174 0.42329445 -0.76645315
-0.17619115 -0.09561391 -1.2635218 0.06686572 0.24867119
0.7882624 0.15952975 0.66556424 0.3575984 -0.4402481
-0.56573486 -0.15152067 0.20452015 0.15044808 0.3790673
0.4046746 0.10349394 -0.74216646 -0.41839486 -0.1634531
-0.6980051 -0.25074166 -0.8568389 0.4902732 0.14900629
-1.0337443 -0.3640406 1.2026353 -0.604866 0.75401837
0.7514906 0.9260016 0.7186613 -0.35461208 0.17814498
-0.53443074 -0.80356896 0.4195293 -0.21487397
-0.23569633 -0.535494 0.9091784 0.4148689 0.58047
-0.71082586 1.2237266 0.5143566 0.6858553 0.16907518
-1.0889707 -1.0602703 -0.044914458 1.0562452 -0.39917028
0.53946054 0.4968267 0.18140882 -0.5325031 -0.10433626
-0.6770337 0.6820697 -0.7788533 -0.82671356 0.161646
0.13410634 0.262196 -0.5227663 0.5709462 0.17471659
0.055271886 0.42494166 0.26631585 -0.9587653
-0.030337654 0.58339536 1.0953573 0.98313093 -0.5220438
-0.10052058 0.38251966 -0.3887956 0.87375206
-0.038645677 0.4431979 0.44748423 -1.1605067 -0.6225495

# vector embeddings
## how to get 'em

### word2vec

simple feed-forward neural network

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (**2013**). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.



# vector embeddings
## how to get 'em

### BERT

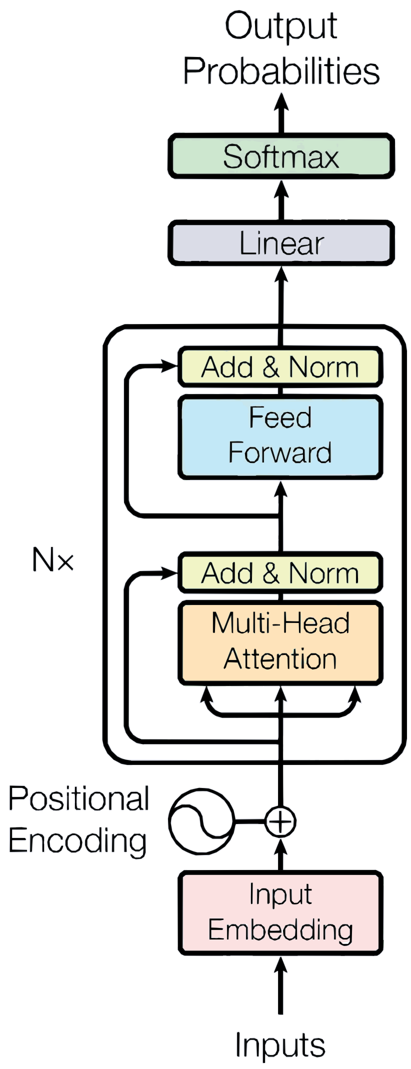
transformer-based neural network

Kenton, J. D. M. W. C., & Toutanova, L. K. (**2018**). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT (Vol. 1, p. 2)

variants:

DistilBERT
RoBERTa
DeBERTa
...

**! cannot be used directly for retrieval**



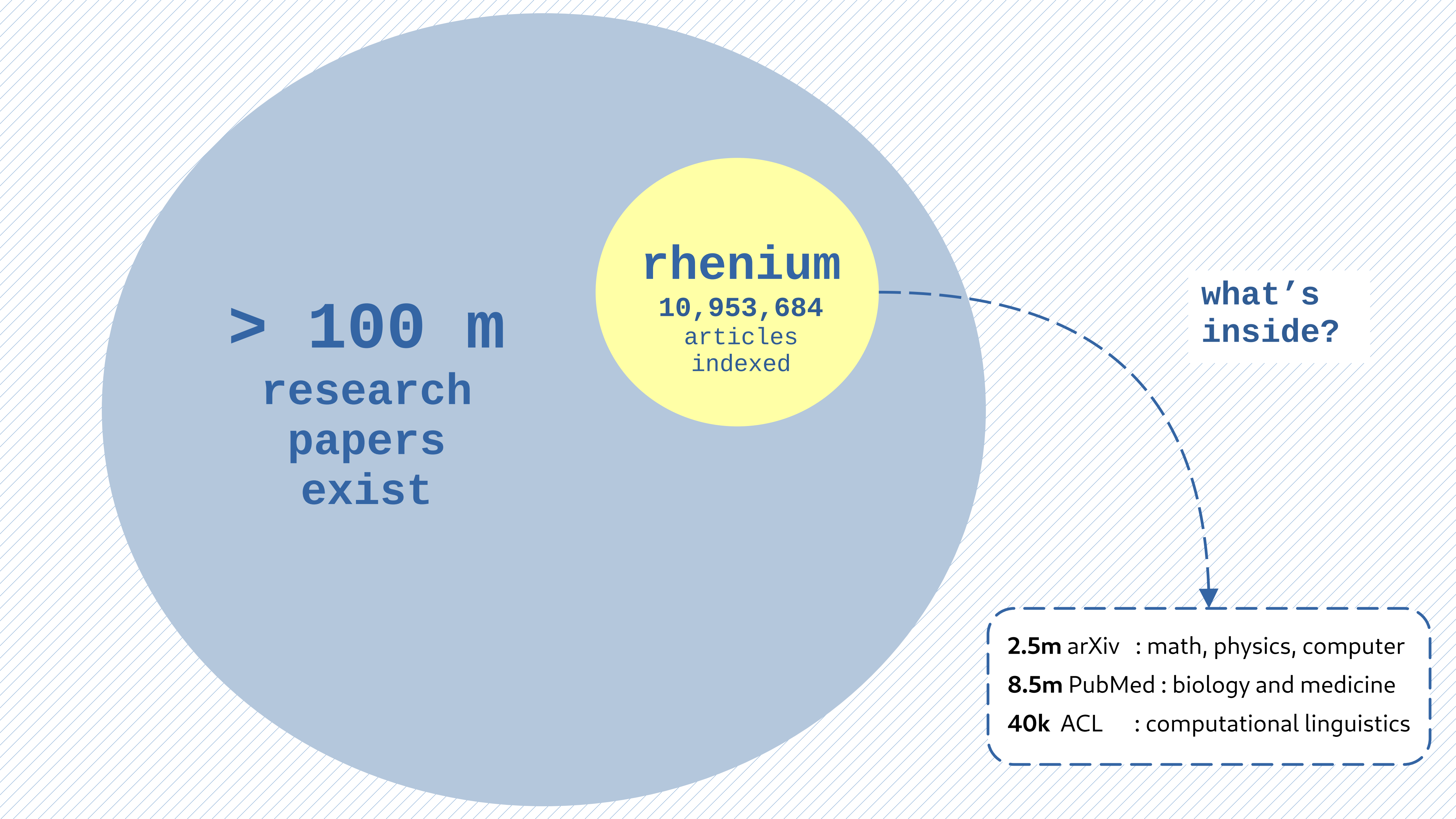
# vector embeddings
## how to get 'em

?

### Sentence-BERT

siamese neural network BERT plus BERT

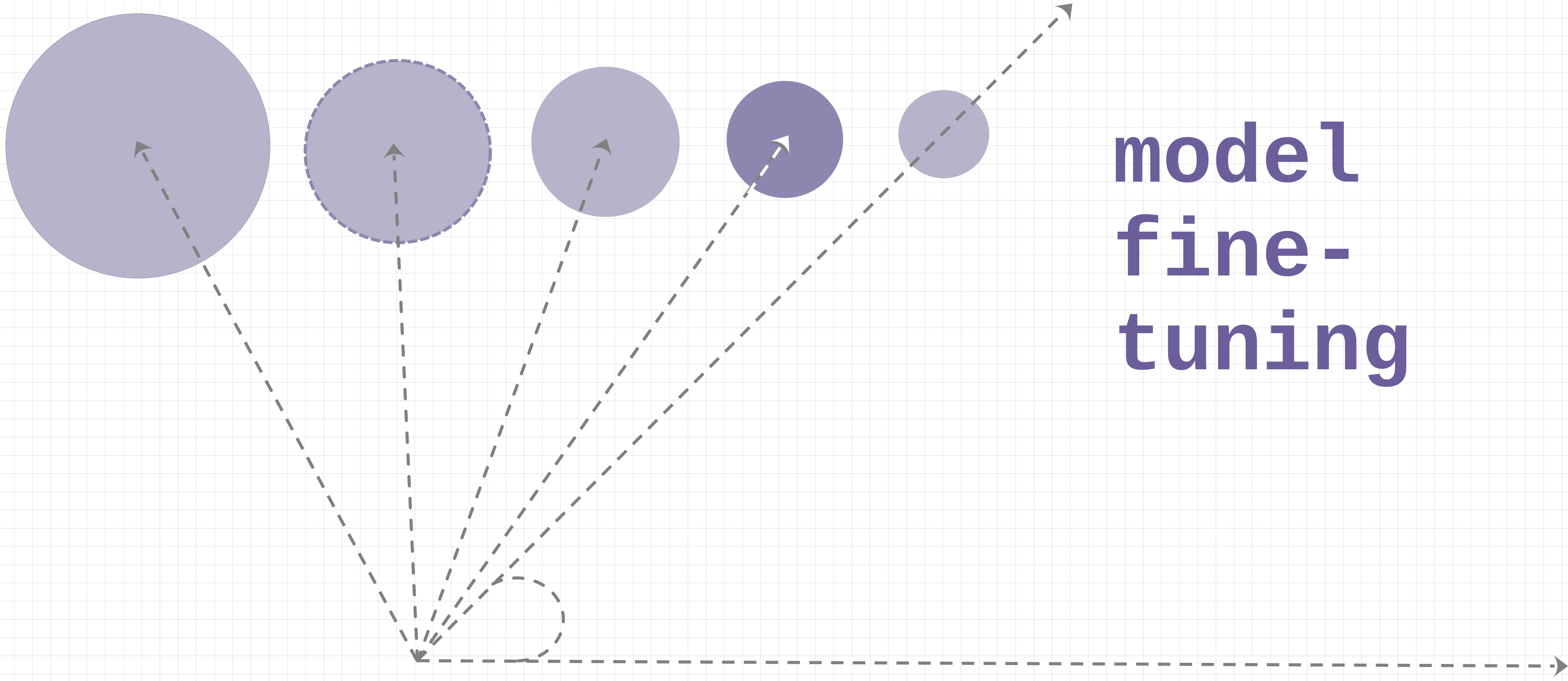
Reimers, N., & Gurevych, I. (**2019**). Sentence-BERT: Sentence embeddings using siamese BERT-networks. arXiv preprint arXiv:1908.10084.



> 100 m
research
papers
exist
rhenium
10,953,684
articles
indexed
what's
inside?
**2.5m** arXiv : math, physics, computer
**8.5m** PubMed : biology and medicine
**40k** ACL : computational linguistics

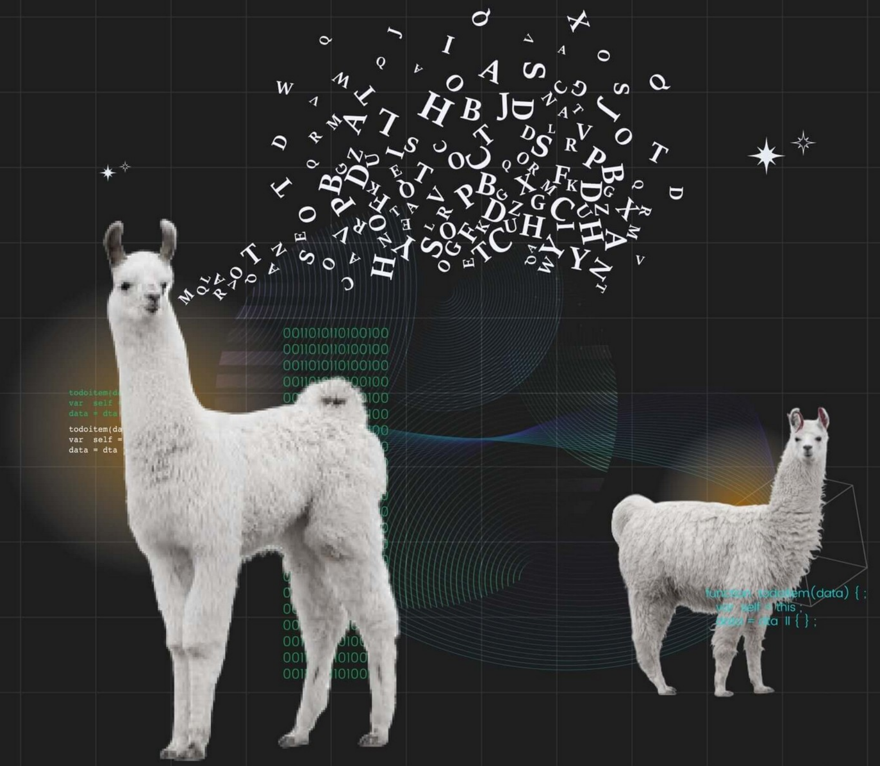
# ranking of language models

| Rank | Model | Model Size (Million Parameters) | Memory Usage (GB, fp32) | Average | ArguAna |
|---|---|---|---|---|---|
| 2 | gte-large-en-v1.5 | 434 | 1.62 | 57.91 | 72.11 |
| 5 | voyage-lite-02-instruct | 1220 | 4.54 | 56.6 | 70.28 |
| 3 | GritLM-7B | 7242 | 26.98 | 57.41 | 63.24 |
| 7 | LLM2Vec-Mistral-supervised | 7111 | 26.49 | 55.99 | 57.48 |
| 12 | text-embedding-3-large | | | 55.44 | 58.05 |
| 1 | SFR-Embedding-Mistral | 7111 | 26.49 | 59 | 67.17 |
| 18 | LLM2Vec-Llama-supervised | 6607 | 24.61 | 54.6 | 56.53 |
| 13 | GritLM-8x7B | 46703 | 173.98 | 55.09 | 59.49 |
| 21 | gte-base-en-v1.5 | 137 | 0.51 | 54.09 | 63.49 |
| 4 | e5-mistral-7b-instruct | 7111 | 26.49 | 56.89 | 61.88 |
| 42 | gte-base | 109 | 0.41 | 51.14 | 57.12 |


model
fine-
tuning

# why fine-tuning?

- special terms
- results tailored to the field of research
- expert knowledge



# biology
# medicine

use BioBERT!

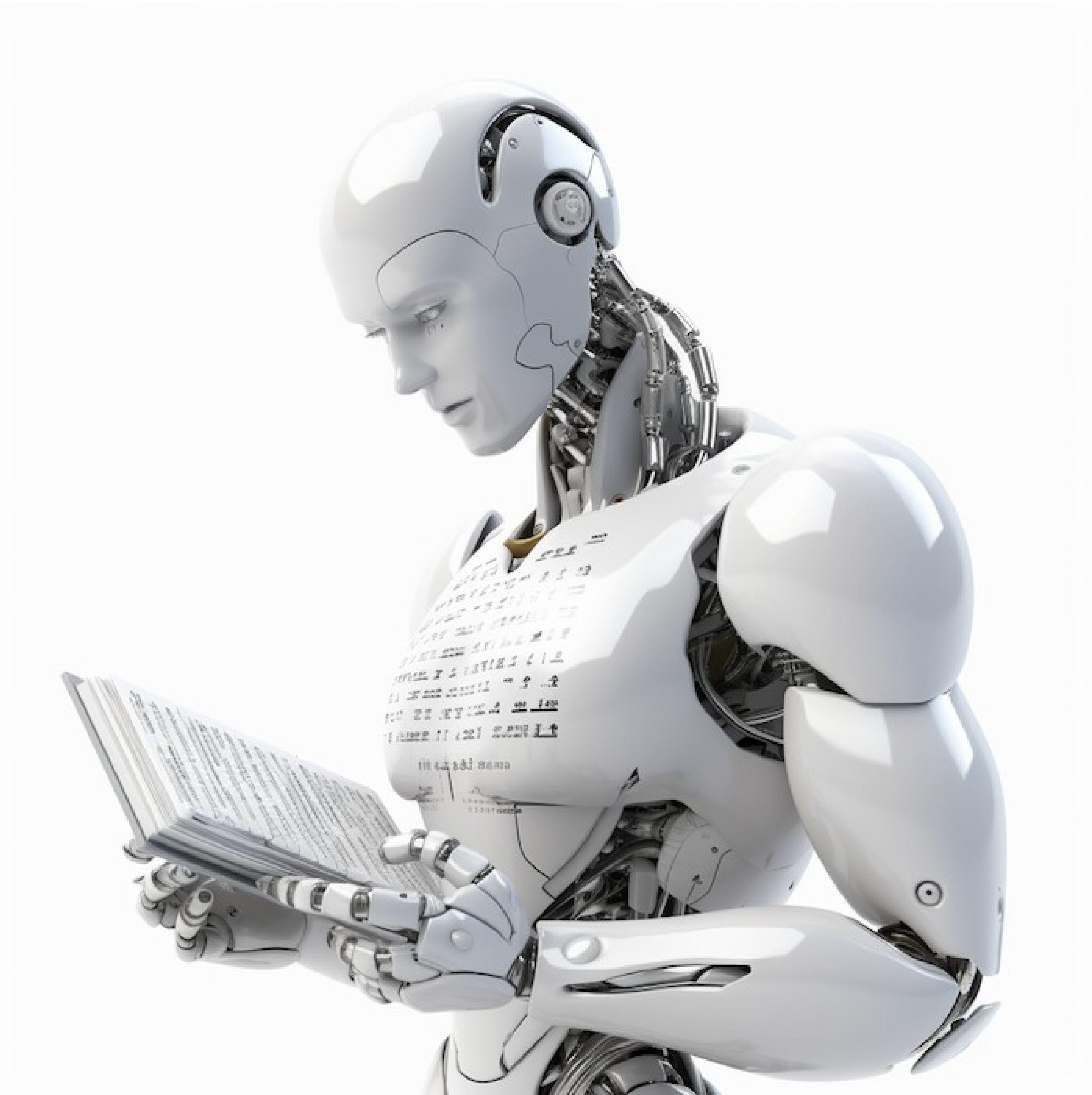
available on

pritamdeka/BioBERT-mnli-snli-scinli-scitail-mednli-stsb

# fine-tuning for the AI domain

**dataset**

1 million articles relevant to «artificial intelligence»



## AI-RoBERTa

base model: **DistilRoBERTa**

training stages:

1. unsupervised MLM on a corpus of AI texts
2. supervised training on NLI datasets

training time: 8-9 hours

* MLM – Masked Language Modeling
* NLI – Natural Language Inference
* GPL – Generative Pseudo Labeling

## AI-BERT-GPL

base model: **DistilBERT**

training stages:

1. generating queries from texts in the corpus
2. retrieve negative passages
3. calculate similarity scores
4. supervised training on resulting dataset

training time: > 40 hours

performance

evaluation

| | gte-large | BioBERT | AI-RoBERTa | AI-BERT-GPL |
|---|---|---|---|---|
| **MRR@10** | 0.99283333 | 0.91068095 | 0.90570238 | 0.93495714 |
| **NDCG@10** | 0.95424540 | 0.77984771 | 0.78139278 | 0.80093784 |
| **MAP** | 0.92281592 | 0.71697288 | 0.71533471 | 0.74431640 |
| **cosine distance accuracy** | 0.94180943 | 0.85071644 | 0.86098910 | 0.79142114 |
| **processes** | **gte-large** | **BioBERT** | **AI-RoBERTa** | **AI-BERT-GPL** |
| 1 | 4857 | 12485 | 16456 | 17123 |
| 2 | 8266 | 17947 | 27651 | 28721 |
| 4 | **9015** | 22072 | 35919 | 37423 |
| 6 | 8513 | **22601** | 39599 | 40988 |
| 7 | 8339 | 22183 | **40656** | 41260 |
| 8 | 8261 | 21986 | 40462 | **41776** |
| **max. query per second** | 150 | 377 | 677 | 696 |

# future work

: index 100m research articles

: support for many research fields

: improve performance of existing models

...

thank you!