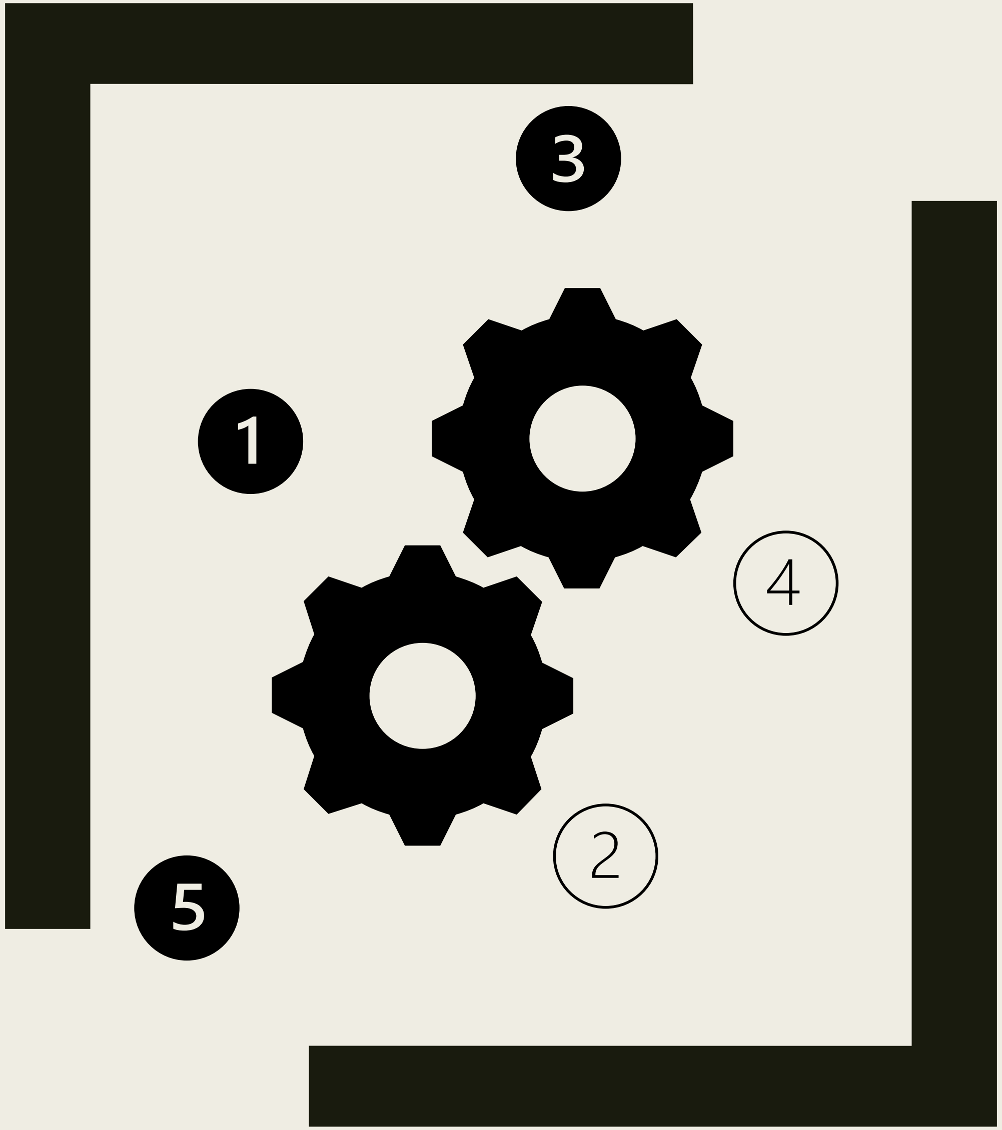


NLU & IR: NEURAL IR (I)

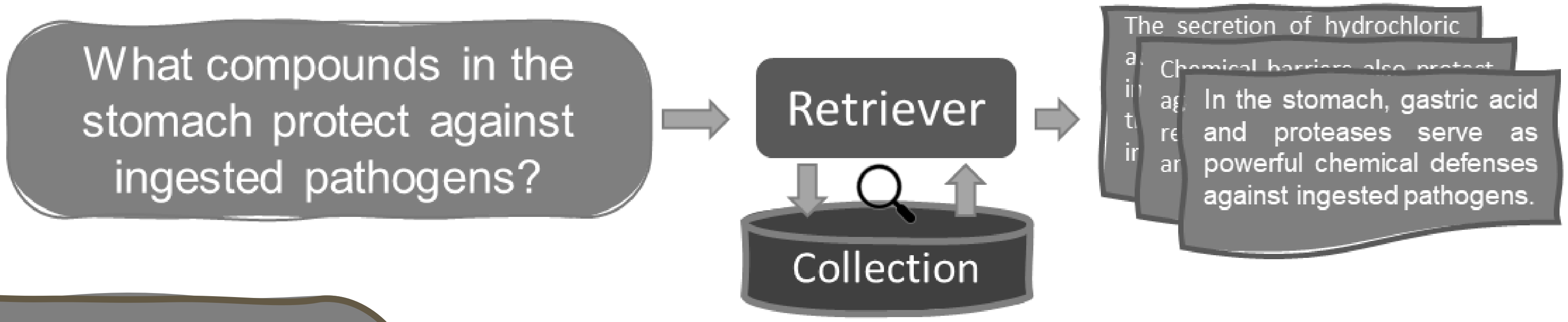
Omar Khattab

CS224U: Natural Language Understanding

Spring 2021

Ranked Retrieval

- Scope: A large corpus of text documents (e.g., Wikipedia)
- Input: A textual query (e.g., a natural-language question)
- Output: **Top-K Ranking** of **relevant** documents (e.g., top-100)



What compounds in the stomach protect against ingested pathogens?

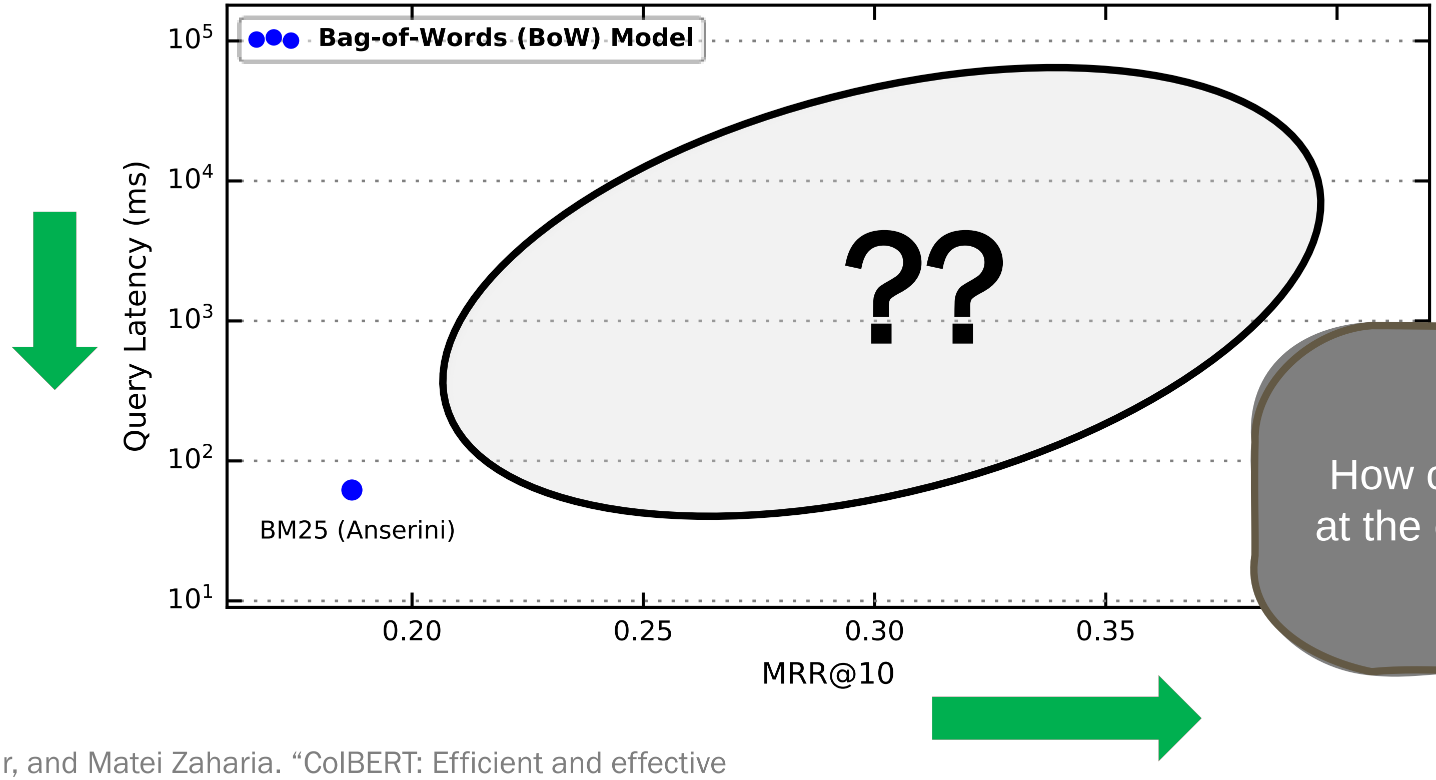
In the stomach, gastric acid and proteases serve as powerful chemical defenses against ingested pathogens.

Should we just use BM25?

Why not?

Efficiency–Effectiveness Tradeoff

- MS MARCO: Bing Queries, 9M Passages from the Web
  - Effectiveness in **MRR@10** and Efficiency in Latency (**milliseconds**; in log-scale!)

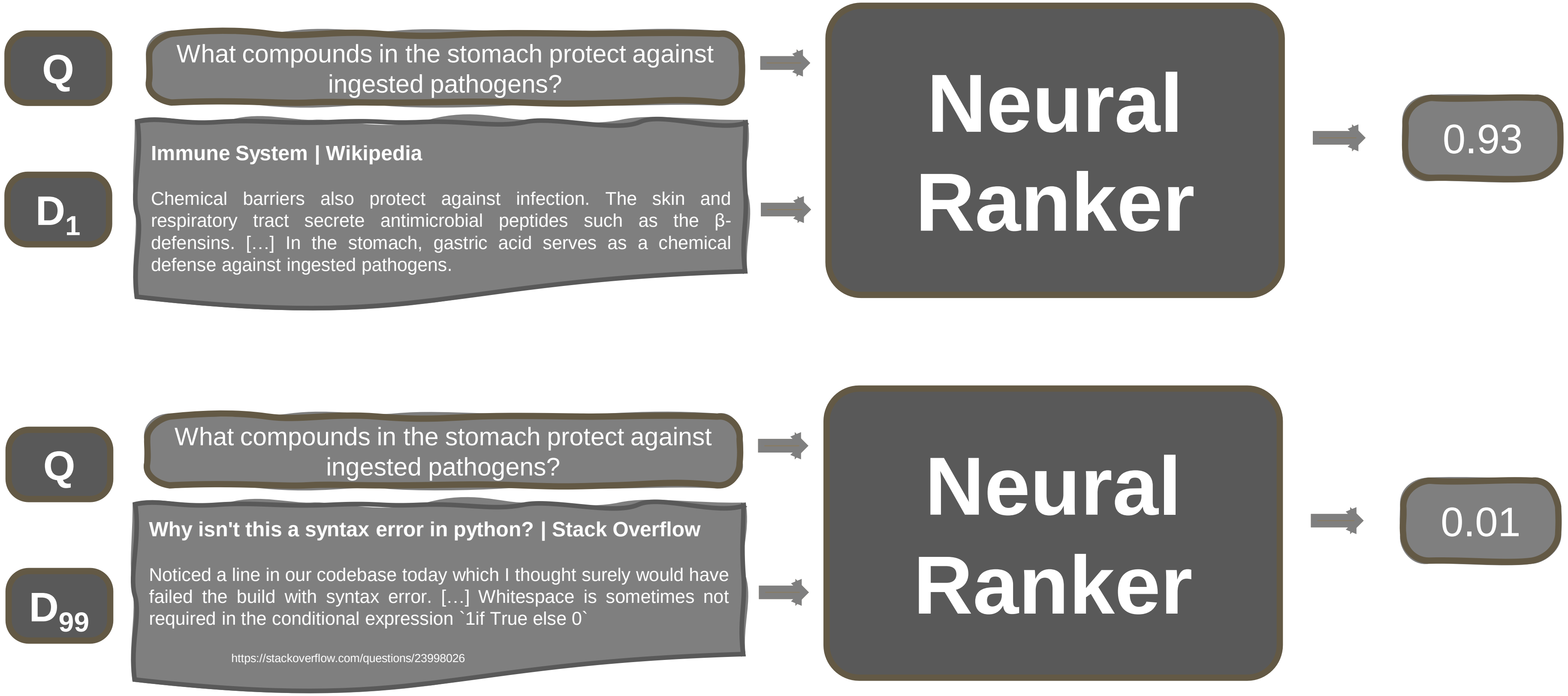


How can we increase MRR@10, possibly at the expense of some increased latency?

Khattab, Omar, and Matei Zaharia. “ColBERT: Efficient and effective passage search via contextualized late interaction over BERT." SIGIR’20.

Neural Ranking: Functional View

- All we need is a score for every query–document pair
  - We’ll sort the results by decreasing score



What compounds in the stomach protect against ingested pathogens?

Immune System | Wikipedia

Chemical barriers also protect against infection. The skin and respiratory tract secrete antimicrobial peptides such as the β-defensins. […] In the stomach, gastric acid serves as a chemical defense against ingested pathogens.

What compounds in the stomach protect against ingested pathogens?

Why isn't this a syntax error in python? | Stack Overflow

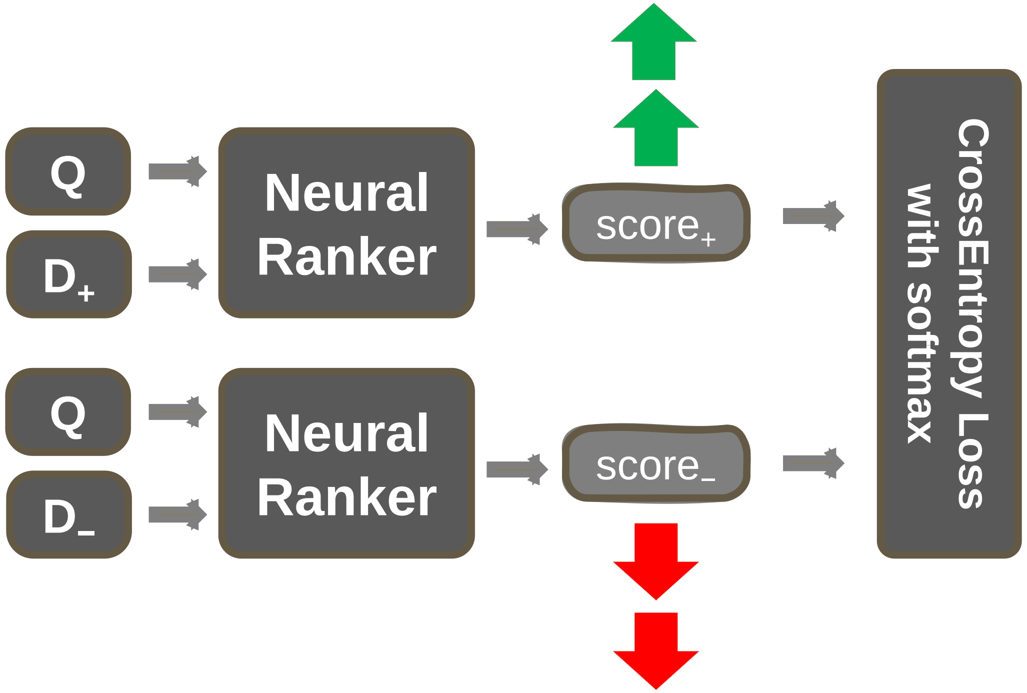
Noticed a line in our codebase today which I thought surely would have failed the build with syntax error. […] Whitespace is sometimes not required in the conditional expression `1if True else 0`

https://stackoverflow.com/questions/23998026

Neural Ranking: Training

- Many possible choices, but **2-way classification** is often effective!
  - Each training instance is a **triple**

    < query, positive document, negative document >



Q

D_+

Neural Ranker

$score_+$

Q

D_-

Neural Ranker

$score_-$

CrossEntropy Loss with softmax

Recall that we can get positives for each query from our relevance assessments.

Every non-positive can often be treated as an implicit negative.

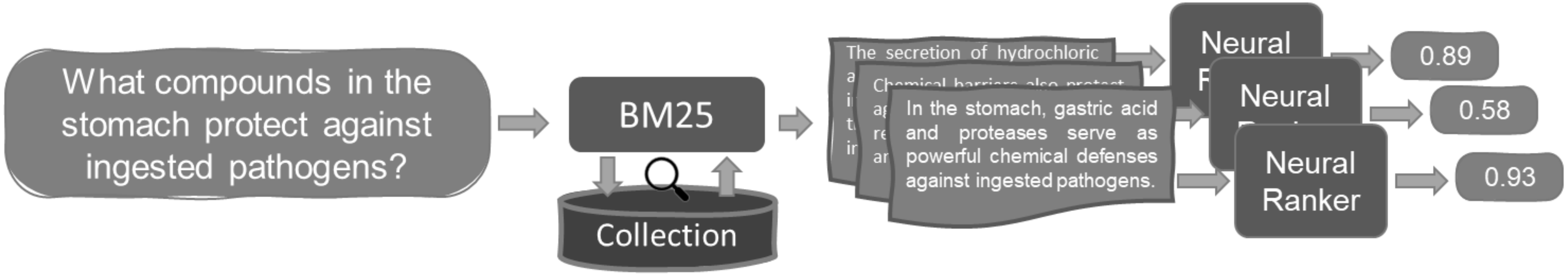
Neural Ranking: Inference

- Given a query Q, pick each document d and pass <Q, d> through the network. Sort all by score, returning the top-k results!

- But collections often have many millions of documents
  - MS MARCO has 9M passages
  - Even if you model runs in 1 <u>micro</u>second per passage, that’s 9 seconds per query!

Neural Re-Ranking: Pipelines

- BM25 top-1000 -> Neural IR reranker



What compounds in the stomach protect against ingested pathogens?
BM25
Collection
The secretion of hydrochloric
Chemical barriers also protect
In the stomach, gastric acid and proteases serve as powerful chemical defenses against ingested pathogens.
Neural Ranker
Neural Ranker
Neural Ranker
0.89
0.58
0.93

- Cuts the work on 10M documents by factor of 10k!
  - But introduces an artificial recall ceiling.

Can we do better?

Yes! Later, we'll discuss **end-to-end retrieval.**

References

Khattab, Omar, and Matei Zaharia. “ColBERT: Efficient and effective passage search via contextualized late interaction over BERT." Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2020.