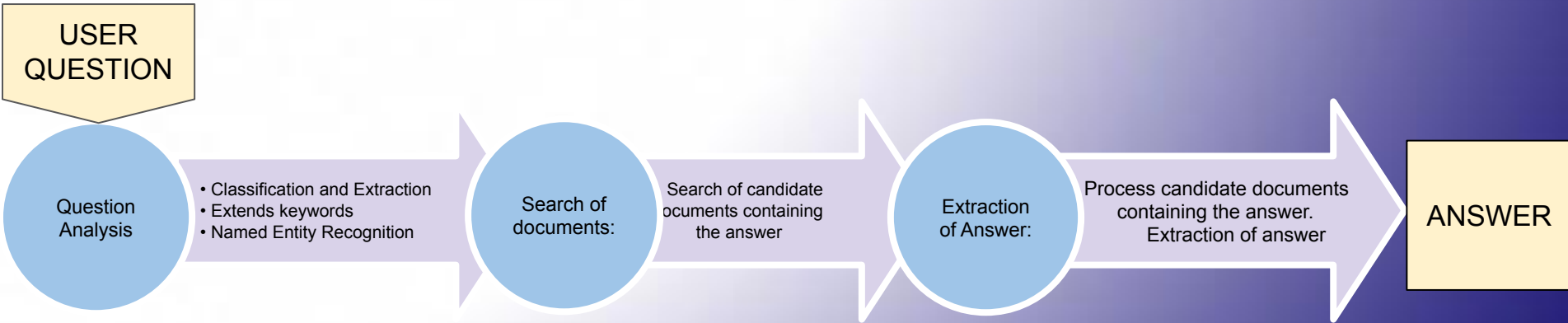


BUILDING A BERTSERINI TELEGRAM INTERFACE WITH PYTORCH LIGHTNING



Introduction

TRADITIONAL QA PIPELINE



Architecture

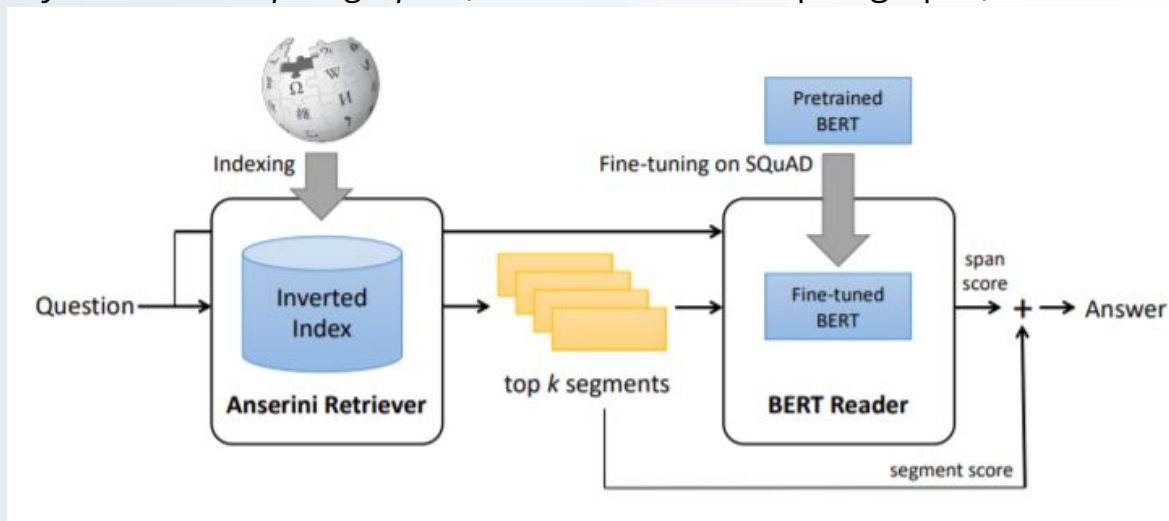
PYSERINI RETRIEVER

Given a collection of documents and the question as *bag-of-word* query, the retriever identifies the k most relevant contexts and passes them to BERT.

The score associated to each retrieved context is calculated using BM-25 as ranking function:

$$\text{score}(D, q_i) = \sum_{i=1}^n \text{IDF}(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{\text{avgdl}}\right)}, \text{ where } q_i \text{ is the query term}$$

We use the pre-built index of Pyserini: *enwiki-paragraphs* (~40 million indexed paragraphs)



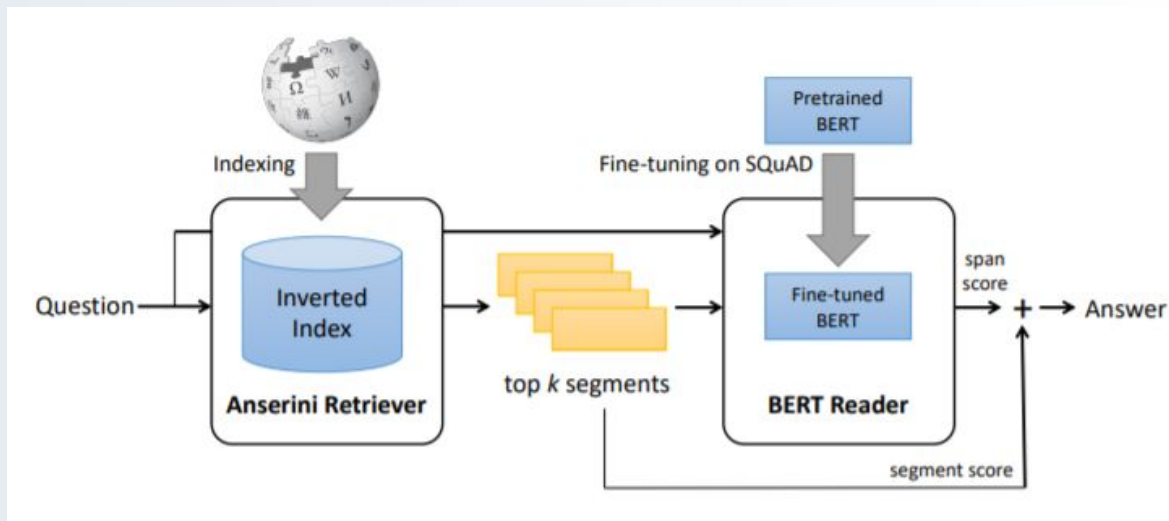
BERT-READER

BERT is a Transformer architecture that exploits a stack of encoders to learn a bidirectional contextualized representation of words and sentences.

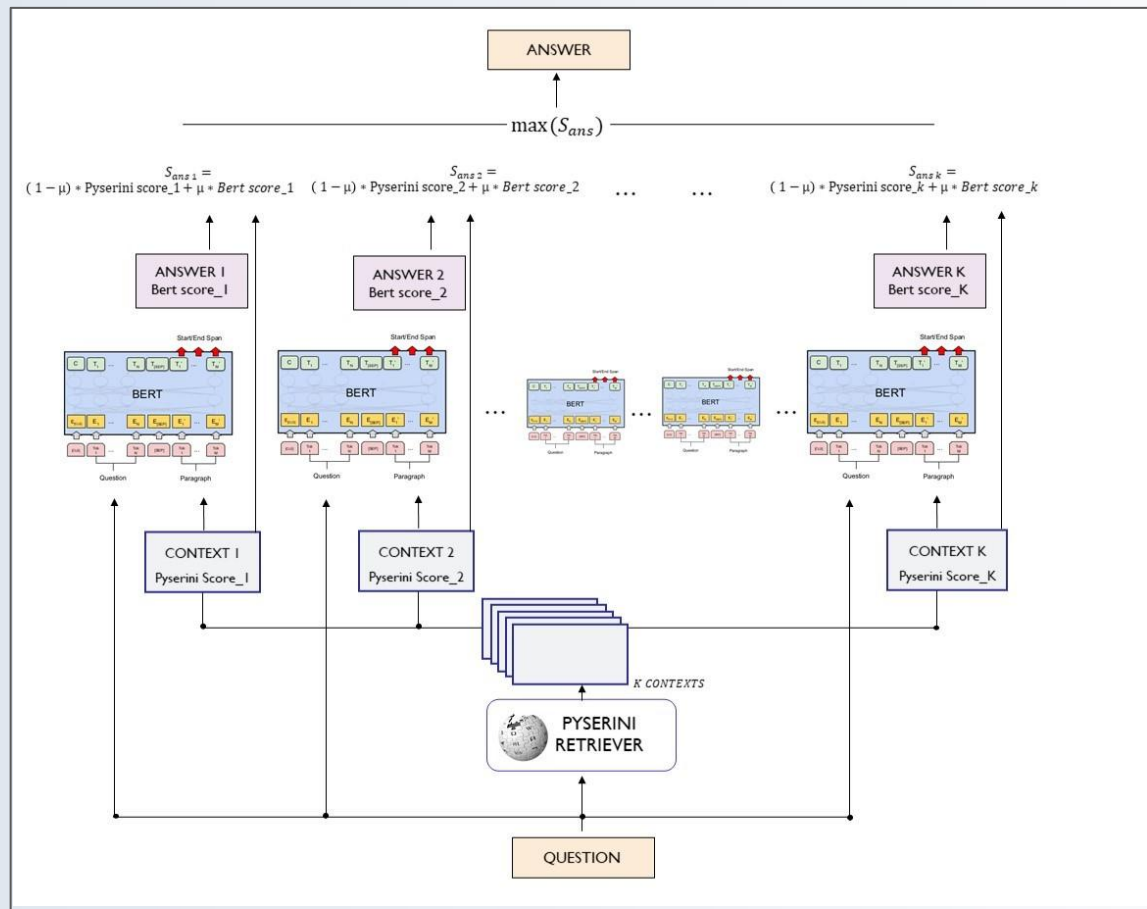
For Question-Answering systems, BERT is fine tuned to output 2 vectors (so called *start vector* and *end vector*), marking the probability of each token of being the start-word and end-word of the output span

We use BERT fine-tuned on SQuAD:

- 12 layers
- 12 attention heads
- 110M parameters
- Hidden representation: 768



INTERPOLATION FUNCTION: $S = (1 - \mu) * S_{pyserini} + \mu * SBERT$



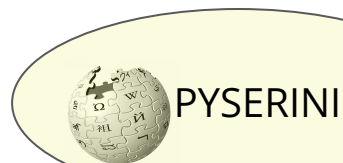
Experiments

METRICS

- TOP-1 EXACT MATCH
- F1
- TOP-K EXACT MATCH
- RECALL

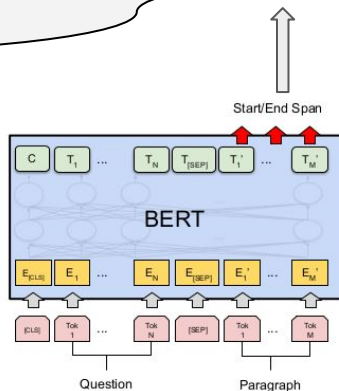
Gerhard Domagk"focus to tuberculosis and chemotherapy against cancer. He continued to live and work in Wuppertal. Domagk died from a heart attack in Burgberg near Königsfeld, Schwarzwald. Gerhard Domagk Gerhard Johannes Paul Domagk (30 October 1895 – 24 April 1964) was a German pathologist and bacteriologist. He is credited with the discovery of Sulfonamidochrysoidine (KI-730), the first commercially available antibiotic and marketed under the brand name Prontosil, for which he received the 1939 Nobel Prize in Physiology or Medicine. Domagk was born in Lagow, Brandenburg, the son of a school headmaster. Until he was 14, he attended school in Sommerfeld university's first library which was hosted on the third floor in the east

OUTPUT: "1939 Nobel Prize in Physiology or Medicine"



QUESTION: "What prize did Domagk get for his work?"

GROUND TRUTH: "1939 Nobel Prize for Medicine"



TOP1-EXACT MATCH: percentage of predictions that **exactly match** one ground truth
In this example: EM = 0

F1: takes into account **precision** and **recall**
In this example: F1 = 0.66

Tragedy: In English, the most famous and most successful tragedies are those of William Shakespeare and his Elizabethan contemporaries. Shakespeare's tragedies include: A contemporary of Shakespeare, Christopher Marlowe, also wrote examples of tragedy in English, notably: (1580?–1634), also wrote famous plays of the genre: Shakespeare was taken

score = 15.59

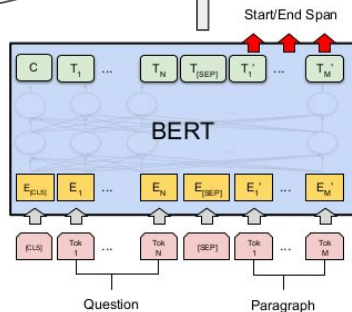
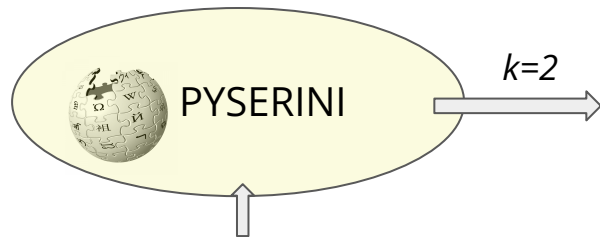
CANDIDATE OUTPUT:
"Christopher Marlowe"

TOP
SCORING

Richard Burbage"star of William Shakespeare's theatre company, the Lord Chamberlain's Men which became the King's Men on the ascension of James I in 1603. He played the title role in the first performances of many of Shakespeare's plays, including "Hamlet", "Othello", "Richard III", and "King Lear". But he was in great demand and also appeared in the plays of many of the great contemporary writers, such as Ben Jonson (the title role in "Volpone", and Subtle in "The Alchemist"), John Marston ("The Malcontent")

score = 13.77

CANDIDATE OUTPUT:
"Ben Jonson"



QUESTION: "What contemporary of Williams Shakespeare wrote a play entitled The Alchemist?"

GROUND TRUTH: "Ben Jonson"

TOPK-EXACT MATCH: the percentage of answers correctly identified in **any retrieved segment**.

In this example: TOP1-EM = 0
but TOPK-EM = 1

Abbey Mills Mosque"Abbey Mills Mosque The Abbey Mills Mosque, also known as the London Markaz or Masjid-e-Ilyas, is a temporary mosque located in **Stratford**, east London, accommodating around 2,500 people. Plans were made to expand the capacity of the mosque to what would have been the largest religious building in Britain three times the size of St Paul's Cathedral. The mosque is one of the largest in Western Europe. Further plans for the building is offered.

score = 13.09

TOP
SCORING

East London Mosque"East London Mosque The East London Mosque (ELM) is situated in the **London Borough of Tower Hamlets** between Whitechapel and Aldgate. It serves Great Britain's largest Muslim community. Combined with the adjoining London Muslim Centre and Maryam Centre, it is one of the largest mosques in Europe, accommodating more than 7,000 worshippers for congregational prayers. The mosque was one of the first in the UK to be allowed to use loudspeakers to broadcast

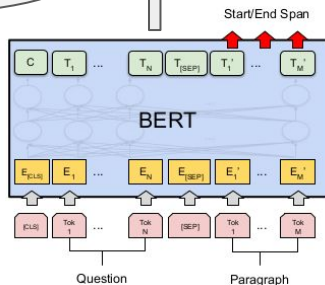
score = 12.44

The ground truth appears at least in one of the retrieved contexts, arbitrarily in the text!



PYSERINI

$k=2$



QUESTION: "Where is the East London Mosque located?"

GROUND TRUTH: "Tower Hamlets"

RECALL: the percentage of questions for which the answer appears **in any retrieved context**.

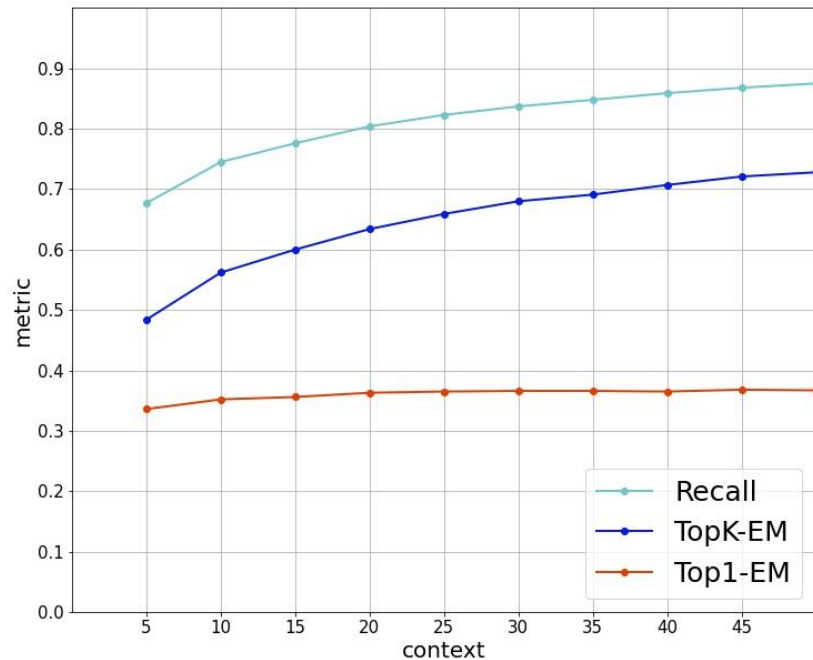
In this example: TOP1-EM = 0

TOPK-EM = 0

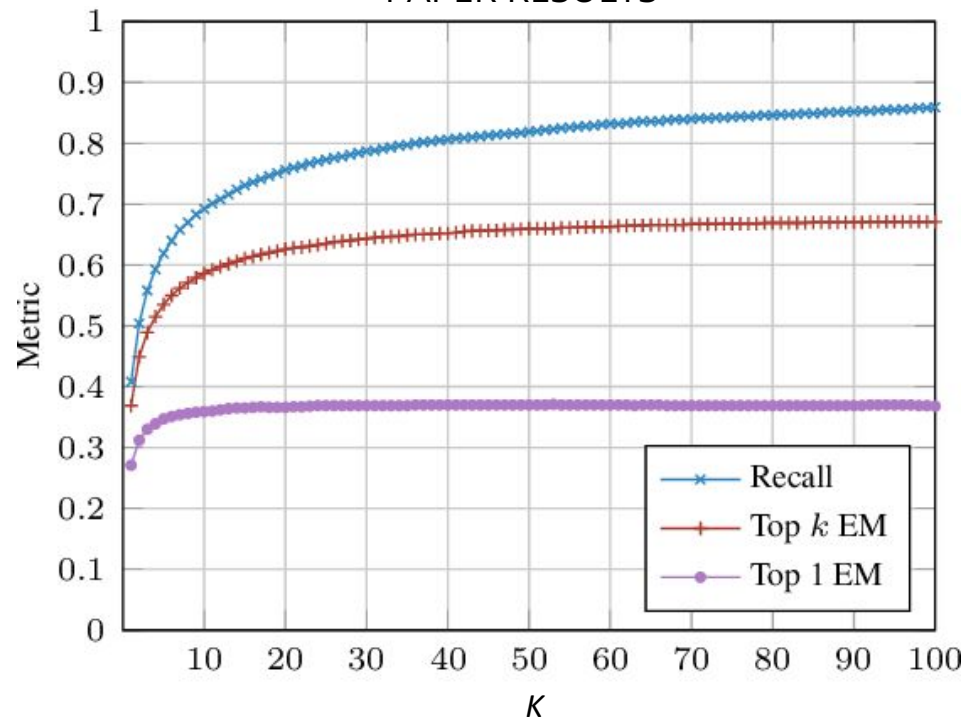
but RECALL = 1

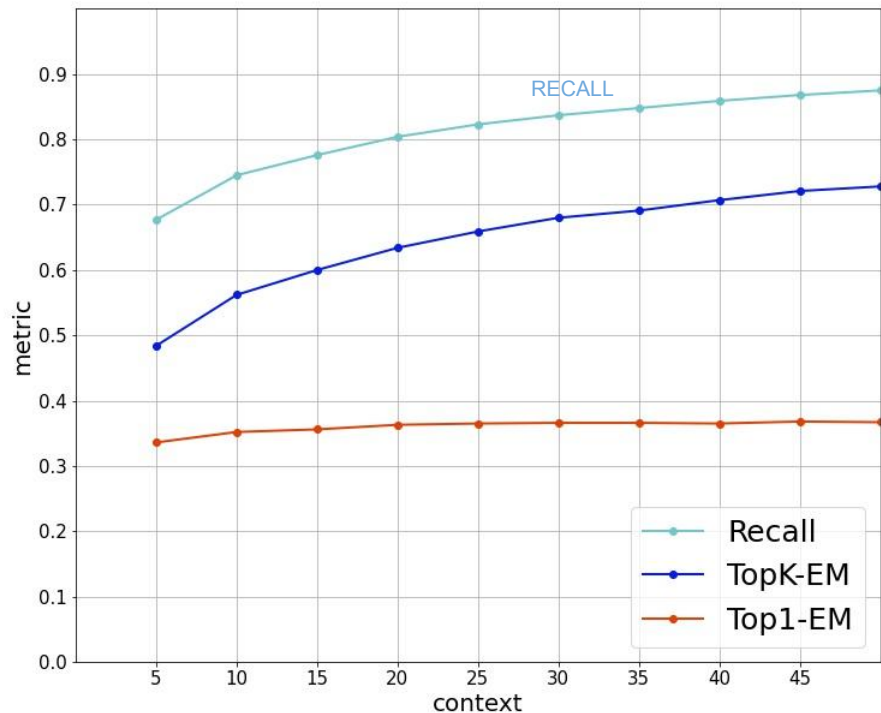
VALIDATION

OUR RESULTS



PAPER RESULTS

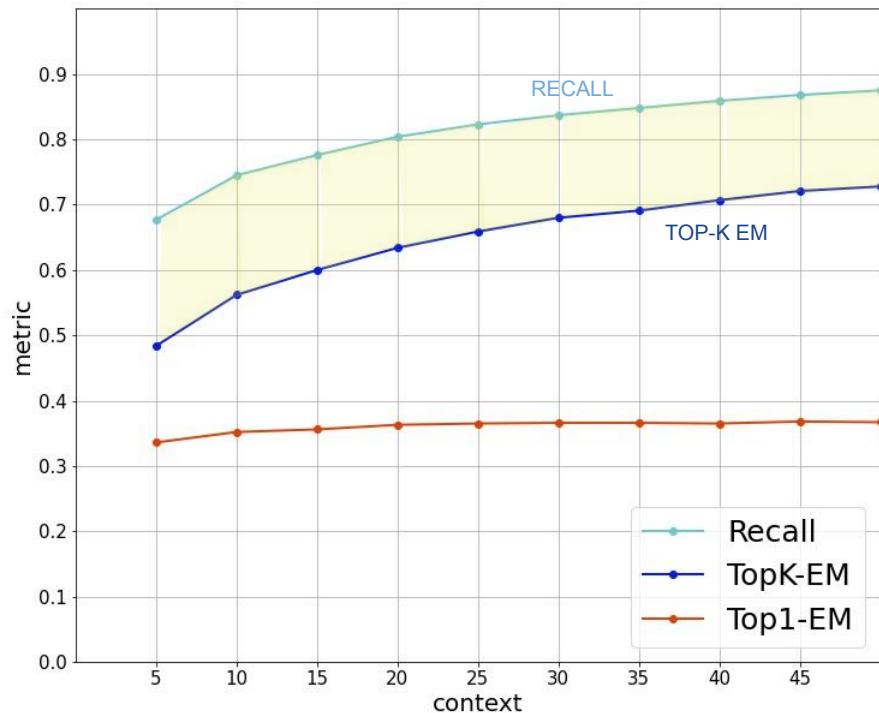




WHAT DOES THE RECALL CURVE TELL US?

When retrieving $k = 20$ contexts onward, Pyserini successfully returns 80% of the time **at least one paragraph containing the correct answer**

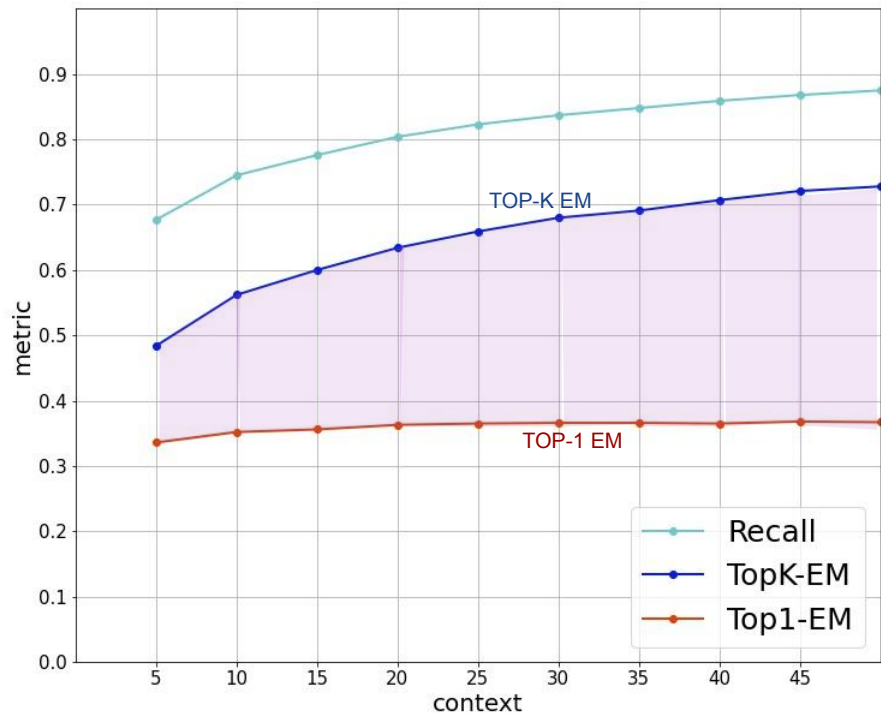
PYSERINI RETRIEVAL IS NOT THE BOTTLENECK



FIRST POINT OF FAILURE

20% of the time the reference answer is present somewhere in the context but **BERT is not able to find it.**

**THIS GAP IS THE ROOM
FOR IMPROVEMENT
FOR BERT**



SECOND POINT OF FAILURE

BERT is able to identify a set of candidate spans containing the reference answer, but then we **fail in selecting the top-scoring one**

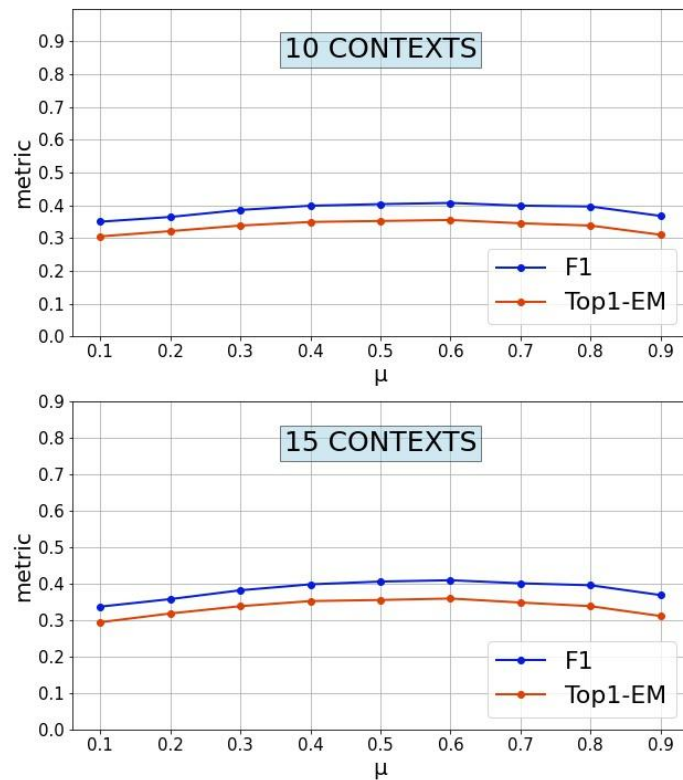
THE BOTTLENECK IS IN THE INTERPOLATION FUNCTION

EFFECT OF VARYING THE INTERPOLATION FACTOR μ

$$S = (1 - \mu) * S_{pyserini} + \mu * S_{BERT}$$

The maximum, both in terms of TOP1-EM and F1, is reached with:

$$\mu=0.6$$



Extensions and Contributions

Our contributions

- PyTorchLightning Framework
- Telegram Interface
- Language Detection + Automatic Translation

PytorchLightning Framework

It's the new Pytorch research framework, that increase the structure of the code, allowing for faster prototyping and scaling, without the boilerplate.

The main components:

- **LightningDataModule:** deals with the loading, cleaning, transformation of the dataset
- **LightningModule:** contains the model's training logic, and many useful hooks
- **LightningCLI:** bridges the modules above and handles all the model's actions
- **.yaml config files:** offer a neat and organized way of storing parameters



Telegram Interface



To improve the user experience, we wrapped our BERT Q&A project into Telegram APIs. This allowed us to have an easy and familiar way to interact with the system.

- The “BOT” is hosted on the same system where an instance of BERT is spawned
- The user can interact with the bot by simply starting the message with the character “!”
- The BOT catches the message, and feeds it to the BERT Q&A pipeline
- BERT’s answer is then delivered back as a response message sent by the BOT itself

Language Detection & Machine Translation

To enable a sort of multi-lingual support for our framework, we created an “automated translation” pipeline in order to translate the user questions from their native languages (if supported) to english, and translate back BERT’s answer.

- FastText detects the language of the user’s question (up to 176 supported languages)
- mBART translates the question from the detected language (if supported) to English
- BERT then predicts an answer from the PySerini indexed documents
- mBART translates the predicted answer back to the user’s language

fastText





B

BertBot

bot



/start 11:52 ✓

Hi! I'm a telegram chatbot powered by BERT!
If you ask me a question I will try to answer you based on a couple of matches found on wikipedia
To interact with me it's easy, just start any message with the character '!' and I'll try to answer your question!

11:52

! Which major war ended in 1945? 11:52 ✓

World War II 11:53

! Quale guerra mondiale è terminata nel 1945? 11:53 ✓

Seconda Guerra Mondiale 11:53

! quelle grande guerre s'est terminée en 1945 ? 11:53 ✓

Deuxième guerre mondiale 11:53

! أي حرب كبرى انتهت عام 1945 ؟ 11:53 ✓

الحرب العالمية الثانية 11:53

! Cila luftë përfundoi në 1945? 11:54 ✓

Sorry, Albanian is not supported by Mbart50 :(

11:54