University of Glasgow | School of Computing Science

AI4BioMed

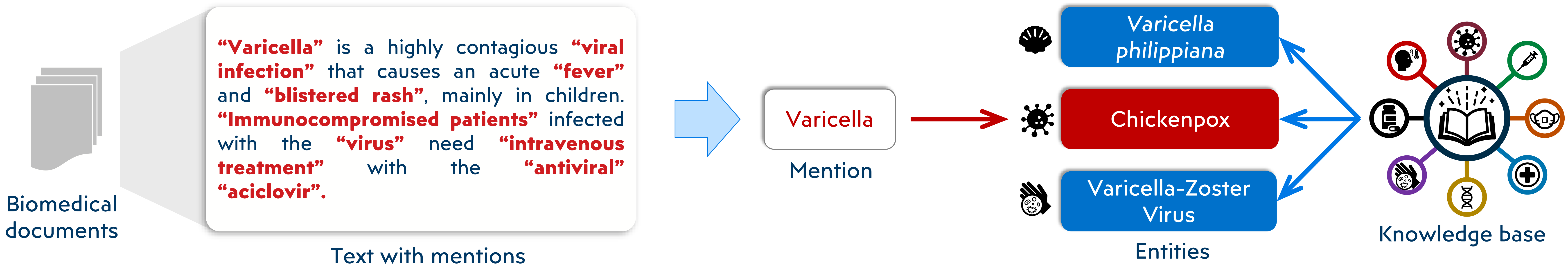
**Glasgow Computational Biology Conference 2025**
**15-16 September 2025**

# Accelerating Cross-Encoders for Biomedical Entity Linking

Javier Sanz-Cruzado and Jake Lever
{javier.sanz-cruzadopuig, jake.lever}@glasgow.ac.uk

Full text

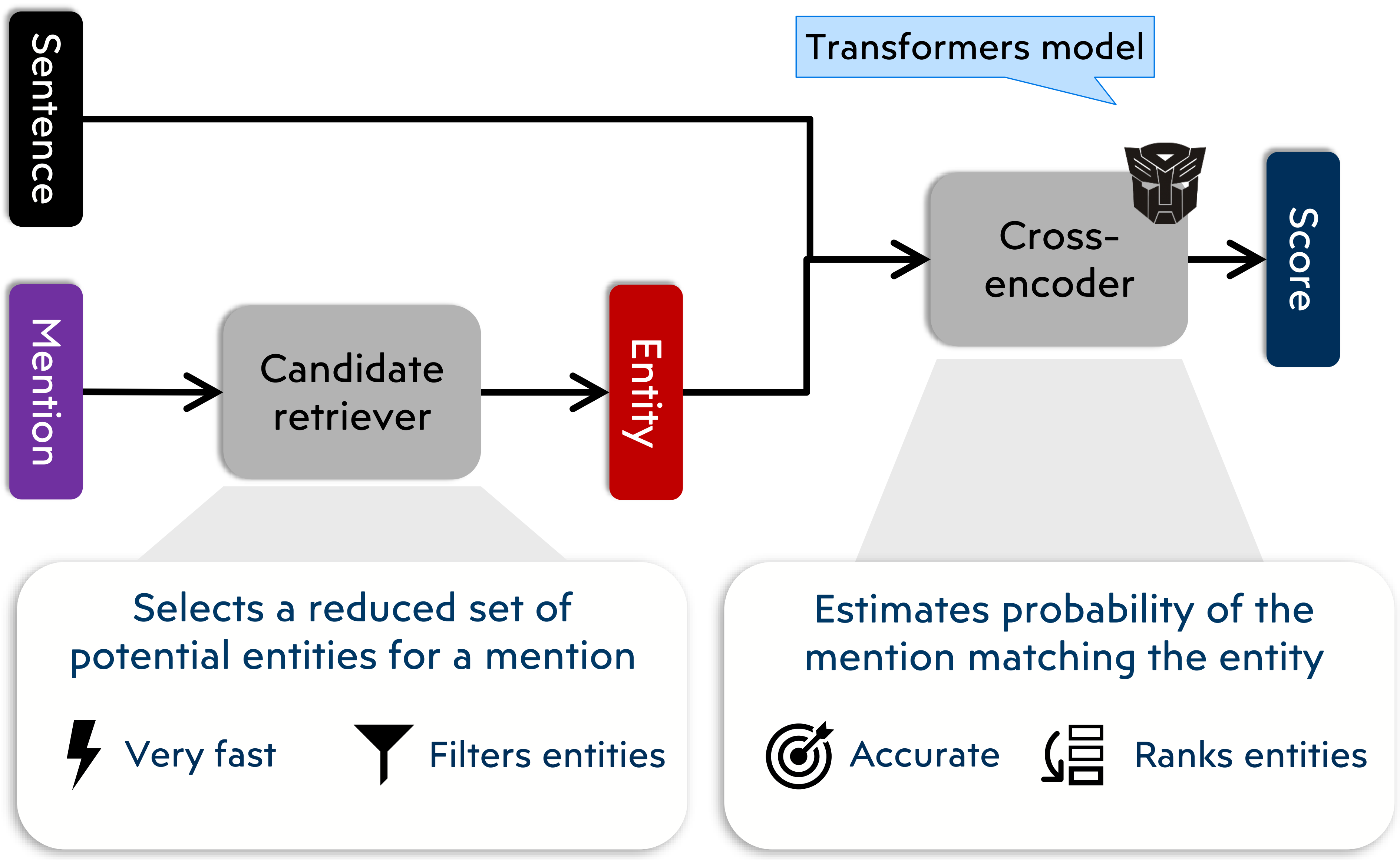
## Biomedical entity linking matches **mentions** of biomedical concepts (diseases, chemicals) in text with unique **entities** within a knowledge base



**"Varicella"** is a highly contagious **"viral infection"** that causes an acute **"fever"** and **"blistered rash"**, mainly in children. **"Immunocompromised patients"** infected with the **"virus"** need **"intravenous treatment"** with the **"antiviral"** **"aciclovir"**.

Text with mentions

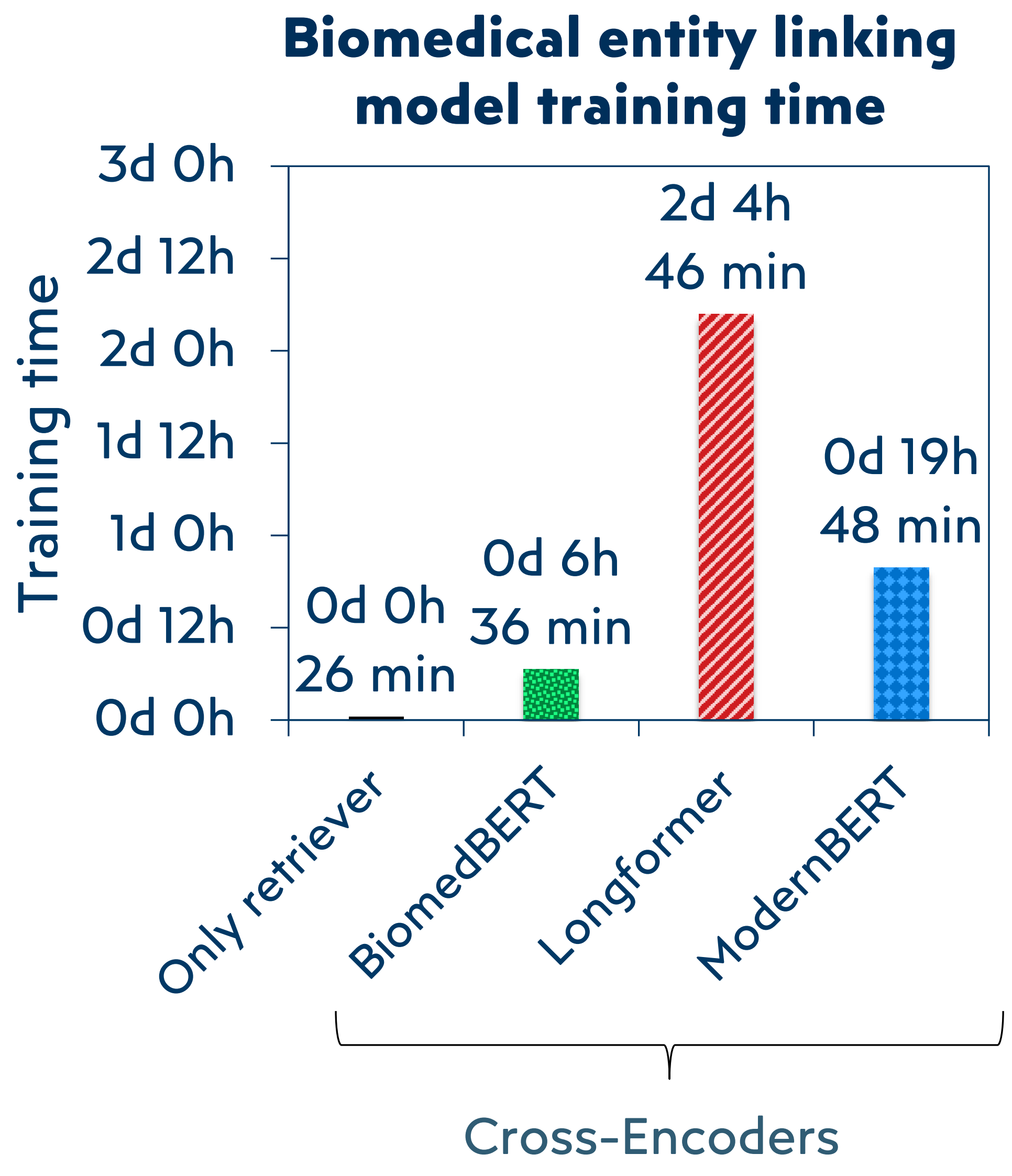
## Cross-encoders are effective solutions for biomedical entity linking



Candidate retriever: Selects a reduced set of potential entities for a mention
- Very fast
- Filters entities

Cross-encoder: Estimates probability of the mention matching the entity
- Accurate
- Ranks entities

## But very slow!



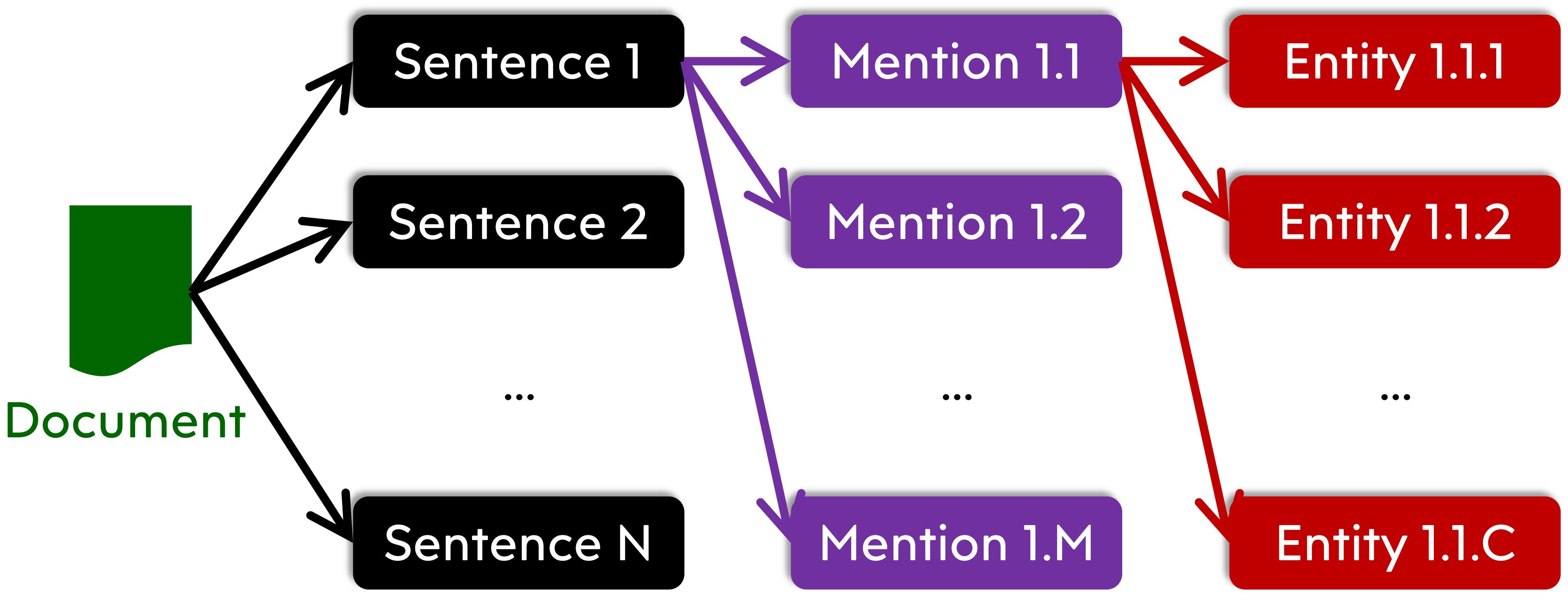
Cross-encoders take from 6 hours to 2 days to train

That's between 15-120 times more than the retriever!

Similar patterns occur during inference

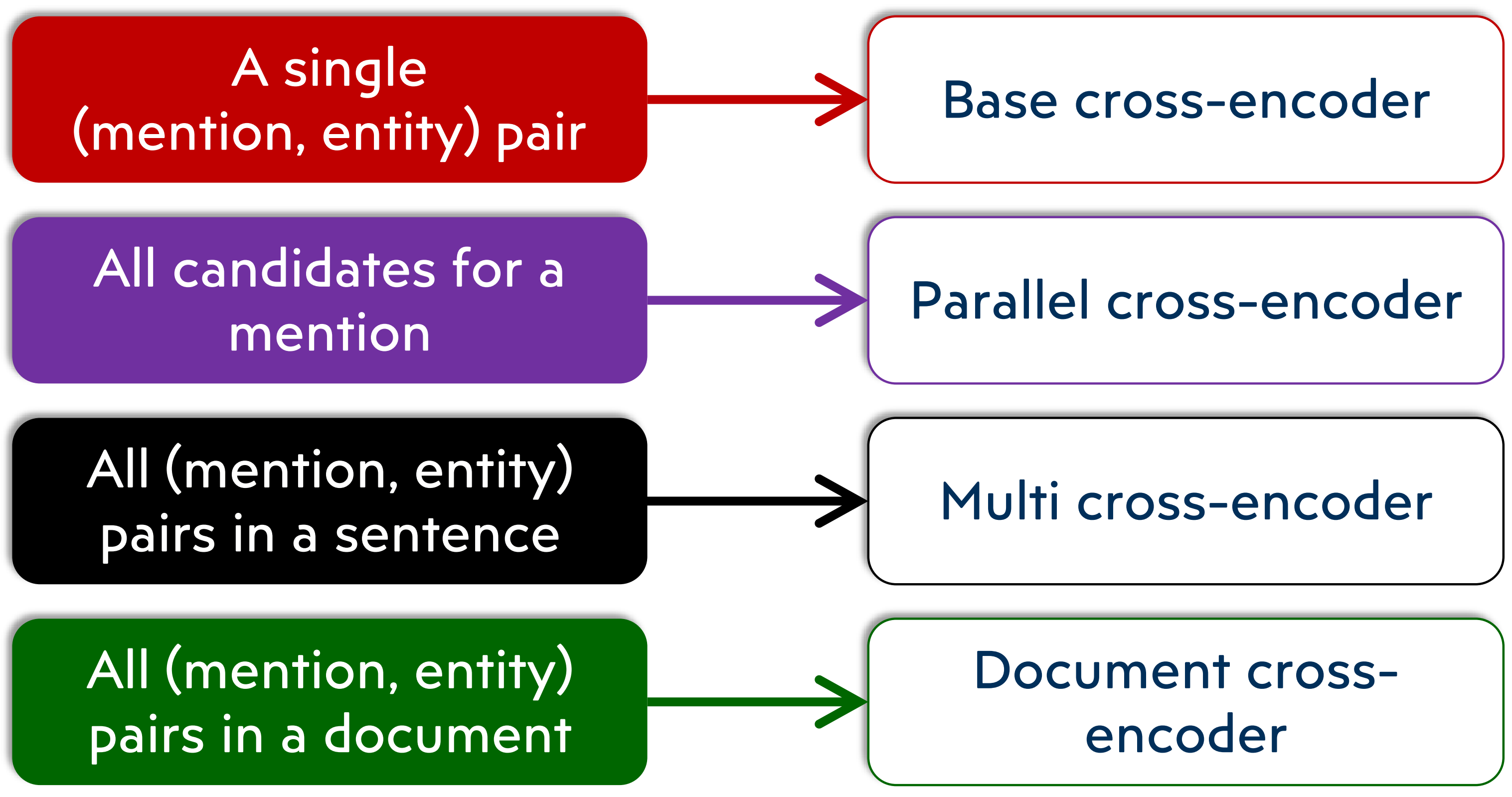
## Can we accelerate cross-encoders without losing accuracy?

### Every time we use the cross-encoder, we only provide as input a single (mention, entity) pair



We are calling a cross-encoder multiple times for (a) the same mention, (b) the same sentence and (c) the same document

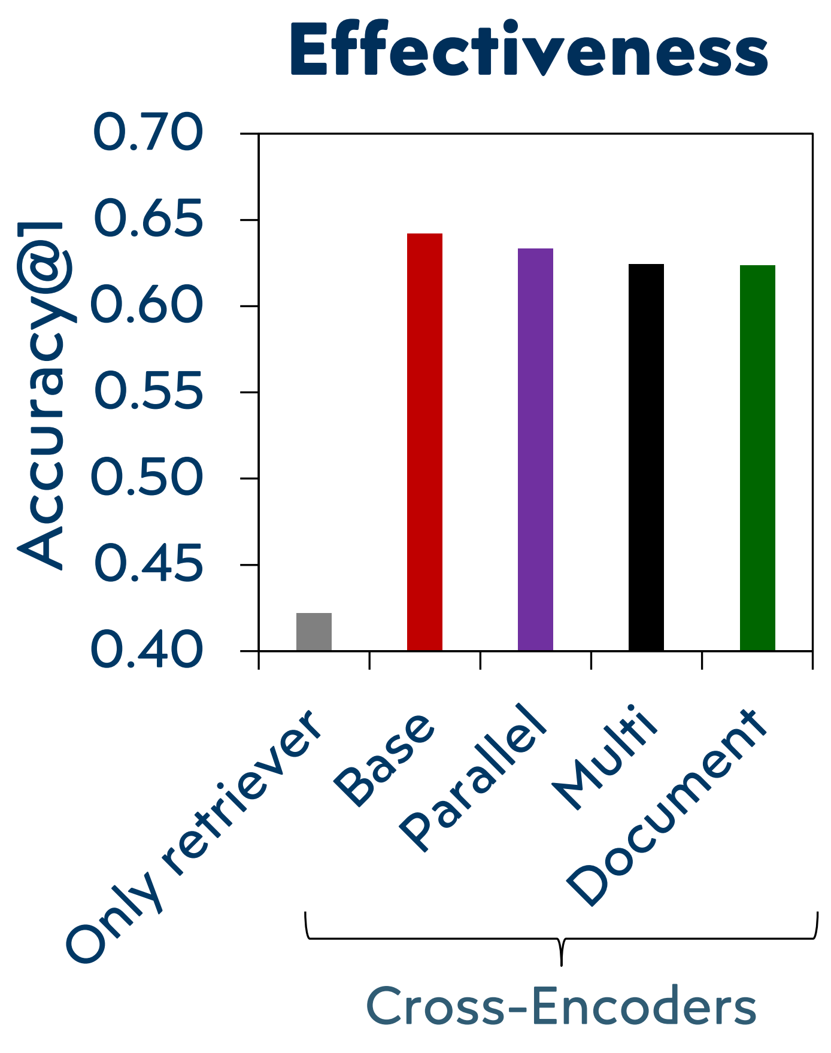
### What if we show multiple (mention, entity) pairs simultaneously to the cross-encoder?

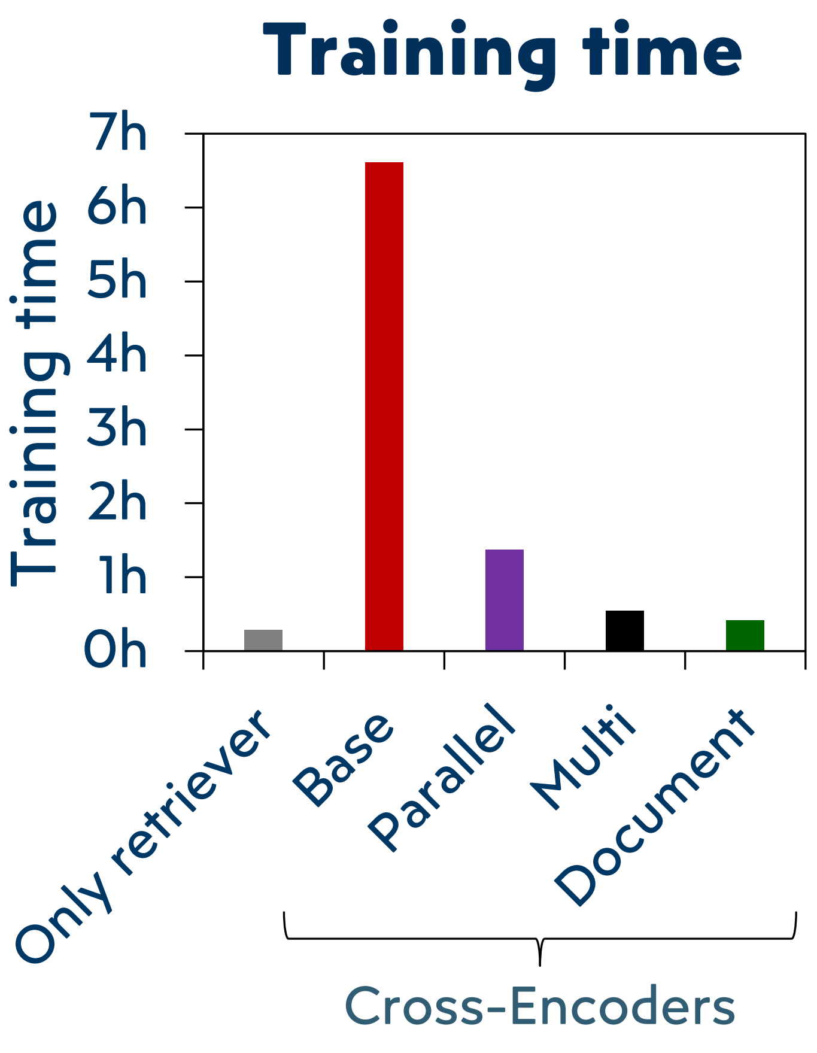


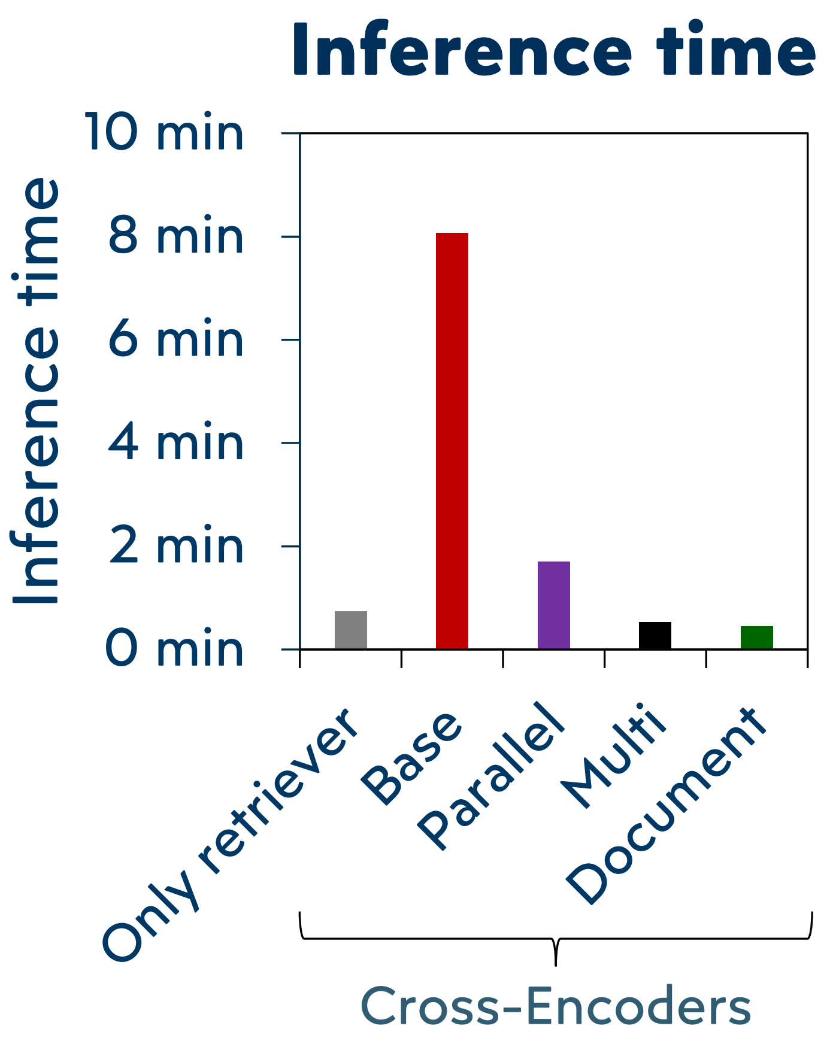
| | |
|---|---|
| A single (mention, entity) pair | Base cross-encoder |
| All candidates for a mention | Parallel cross-encoder |
| All (mention, entity) pairs in a sentence | Multi cross-encoder |
| All (mention, entity) pairs in a document | Document cross-encoder |

## Experiment

We evaluate the different BiomedBERT cross-encoders on the Medmentions dataset (more datasets and Transformers models in the paper)







## Conclusions

Processing more (mention, entity) pairs simultaneously has the following effects

**Small variations on accuracy**
-3.42 to 2.76% differences with base model

**Major improvements in training speed**
2.68x – 36.97x faster training than base model

**Major improvements in inference speed**
3.8x – 26.47x faster inference than base model

**Our solution is suitable on environments where speed is crucial**