



University
of Glasgow



AI4BioMed

Accelerating Biomedical Entity Linking Models

Javier Sanz-Cruzado



WORLD
CHANGING
GLASGOW

Information Access and Retrieval in the AI Age

Festival of Data Science & AI 2025

28th October 2025

A WORLD
TOP 100
UNIVERSITY



University
of Glasgow

Based on

Sanz-Cruzado, J. & Lever, J. **Accelerating Cross-Encoders in Biomedical Entity Linking**. BioNLP @ ACL 2025, pp. 136-147



Paper link





Motivation

- Around 167,000 people die from cancer in the UK per year
- Pandemics like Covid-19 have a large impact on our lives

Understanding diseases and developing effective treatments is fundamental for our healthcare!

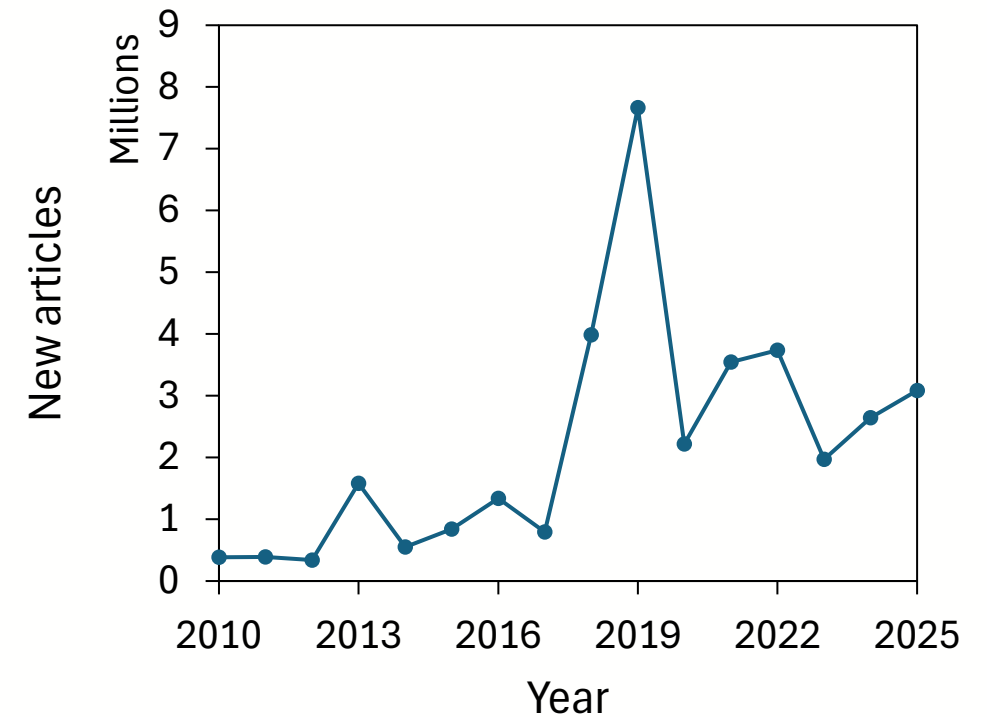


For this, it's important to know what people did before.

- What treatments were tested
- What genes / proteins are involved
- How is a disease related to others
- ...

Keeping up with prior research is becoming increasingly difficult!

Articles in PubMed Central



Biomedical databases

- Used in research for precision medicine
- Examples:
 - ClinGen: Clinical Genome Resource
 - CIViC: Clinical Interpretation of Variants in Cancer
- Summarize information about
 - Diseases
 - Treatments
 - Gene variants
 - Etc.
- Information is usually manually annotated from research papers



The screenshot displays the CIViC (Clinical Interpretation of Variants in Cancer) website interface. The top navigation bar includes a search bar, links to Home, About CIViC, Help, and Documentation, and a Sign In / Sign Up button. A sidebar on the left lists various knowledgebase categories such as Assertions, Evidence, Molecular Profiles, Features, Variants, Variant Groups, Clinical Trials, Diseases, Therapies, Phenotypes, and Sources. The main content area features a 'Welcome to CIViC' message and a 'Discover supported clinical interpretations' button. Below this, the ClinGen logo is visible, and a section titled 'Explore the clinical relevance of genes & variants' provides a brief description of ClinGen as a NIH-funded resource. A search bar is present with the placeholder text 'Enter a gene symbol or HGNC ID (Examples: ADNP, HGNC:15766)'. Below the search bar, there are filters for Curated Genes, Gene-Disease Validity, Dosage Sensitivity, Clinical Actionability, Curated Variants, Statistics, and More.

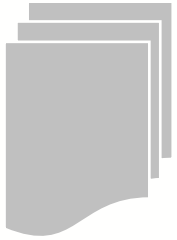
ClinGen is defining the clinical relevance of genes and variants

Founded in 2013 by the National Human Genome Research Institute, ClinGen is a growing collaborative effort, involving three grants, nine principal investigators and over 2,700 contributors from more than 72 countries. Below are a series of recent updates that ClinGen has been working on.

Natural language processing can help!



Extracting information from biomedical texts



Biomedical
documents

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.



Entities

- Varicella
- Aciclovir
- Fever



Relations

- Varicella is an infection
- Varicella causes fever
- Aciclovir treats varicella

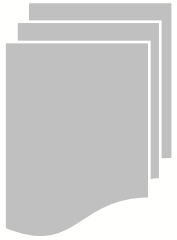


More complex information

- Aciclovir treats varicella via intravenous treatment

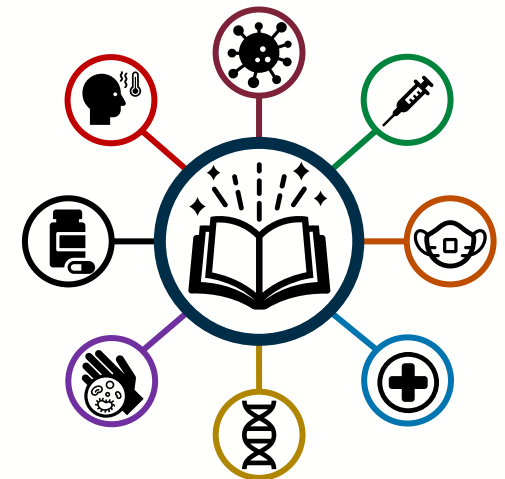
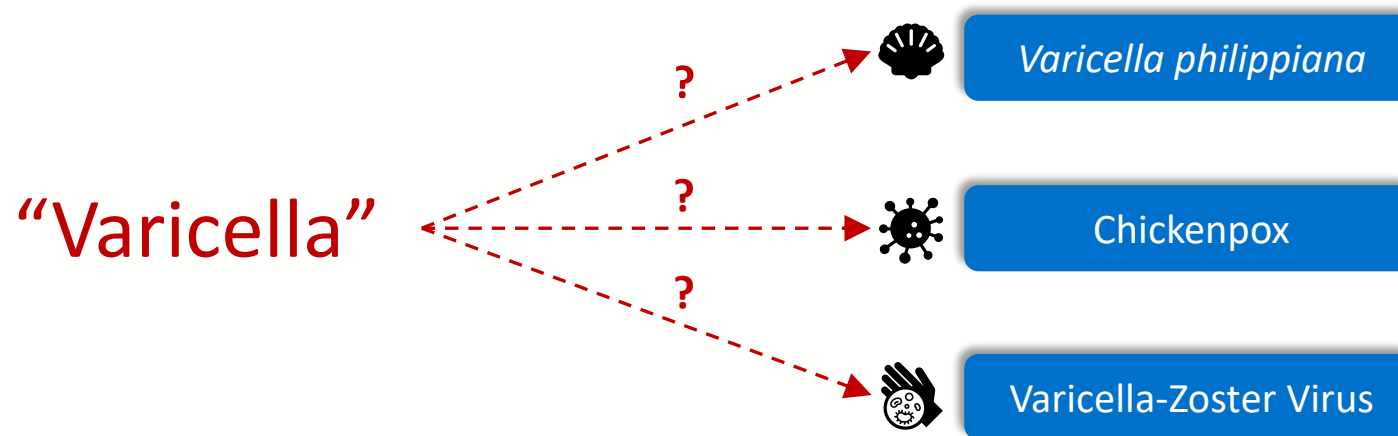


What is biomedical entity linking?



Biomedical
documents

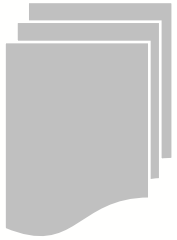
“**Varicella**” is a highly contagious “**viral infection**” that causes an acute “**fever**” and “**blistered rash**”, mainly in children.
“**Immunocompromised patients**” infected with the “**virus**” need “**intravenous treatment**” with the “**antiviral**” “**aciclovir**”.



Knowledge base



What is biomedical entity linking?



Biomedical
documents

“**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**”.

“**Varicella**”



Varicella philippiana



Chickenpox



Varicella-Zoster Virus

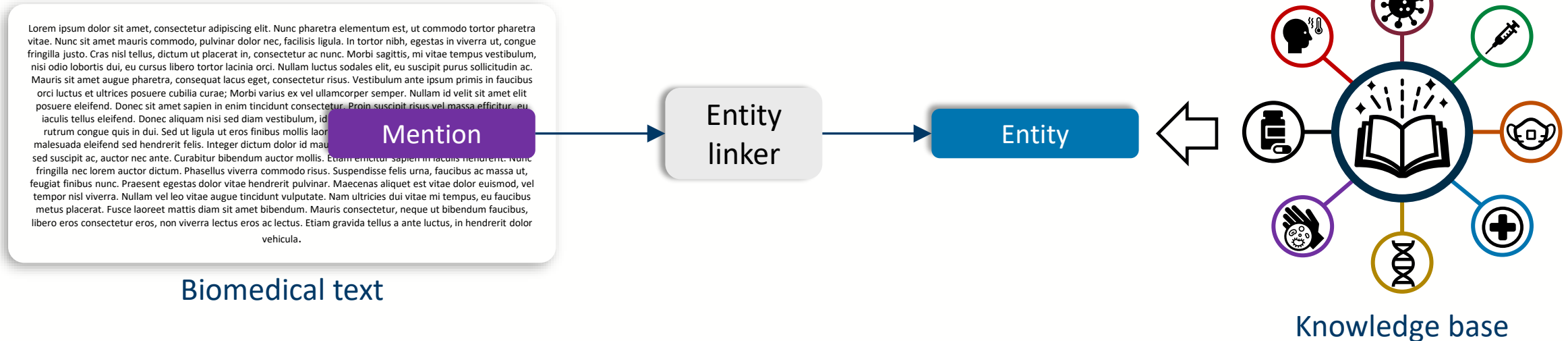


Knowledge base



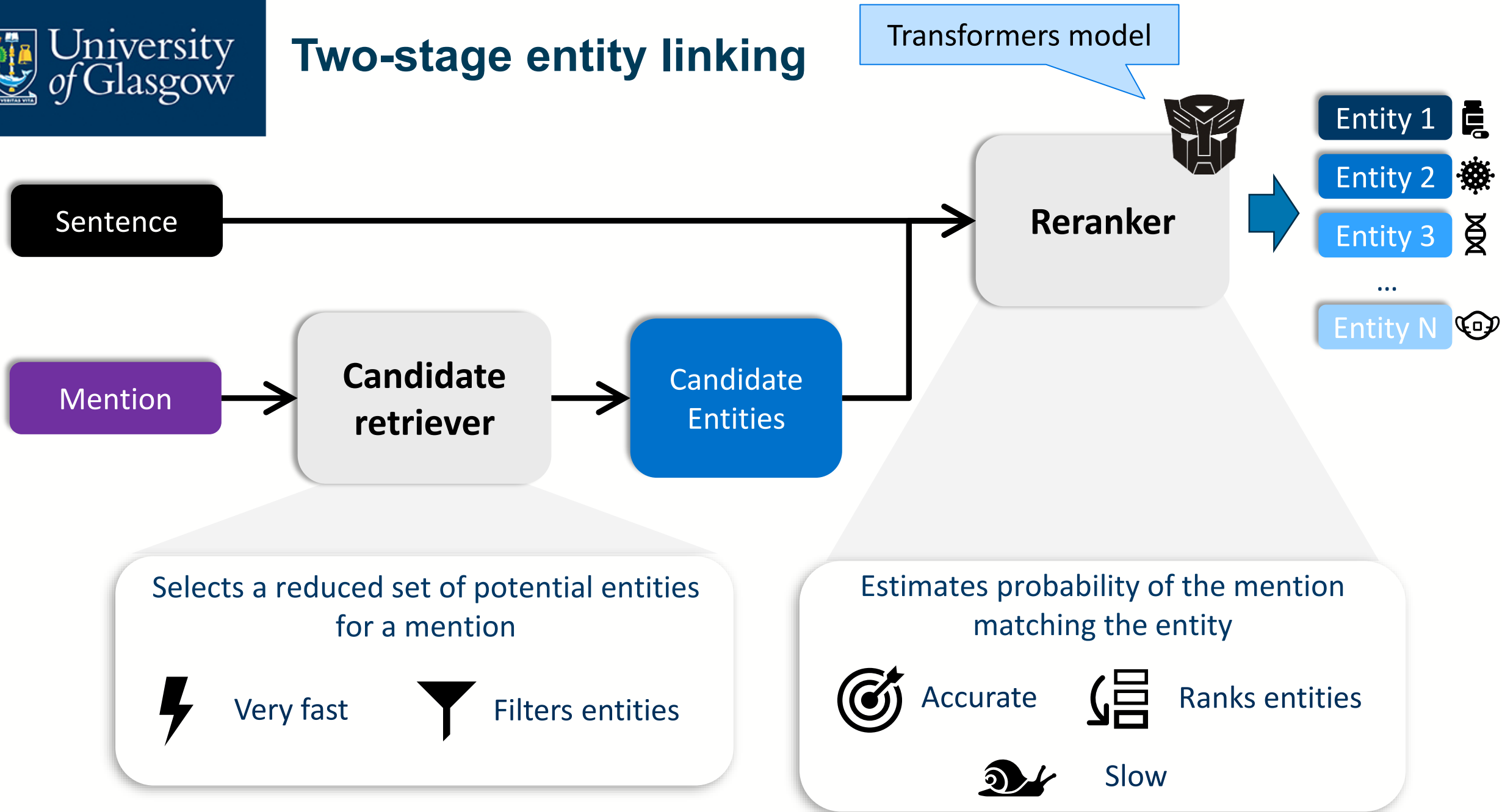
What is biomedical entity linking?

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





Two-stage entity linking





How slow are entity linking rerankers?

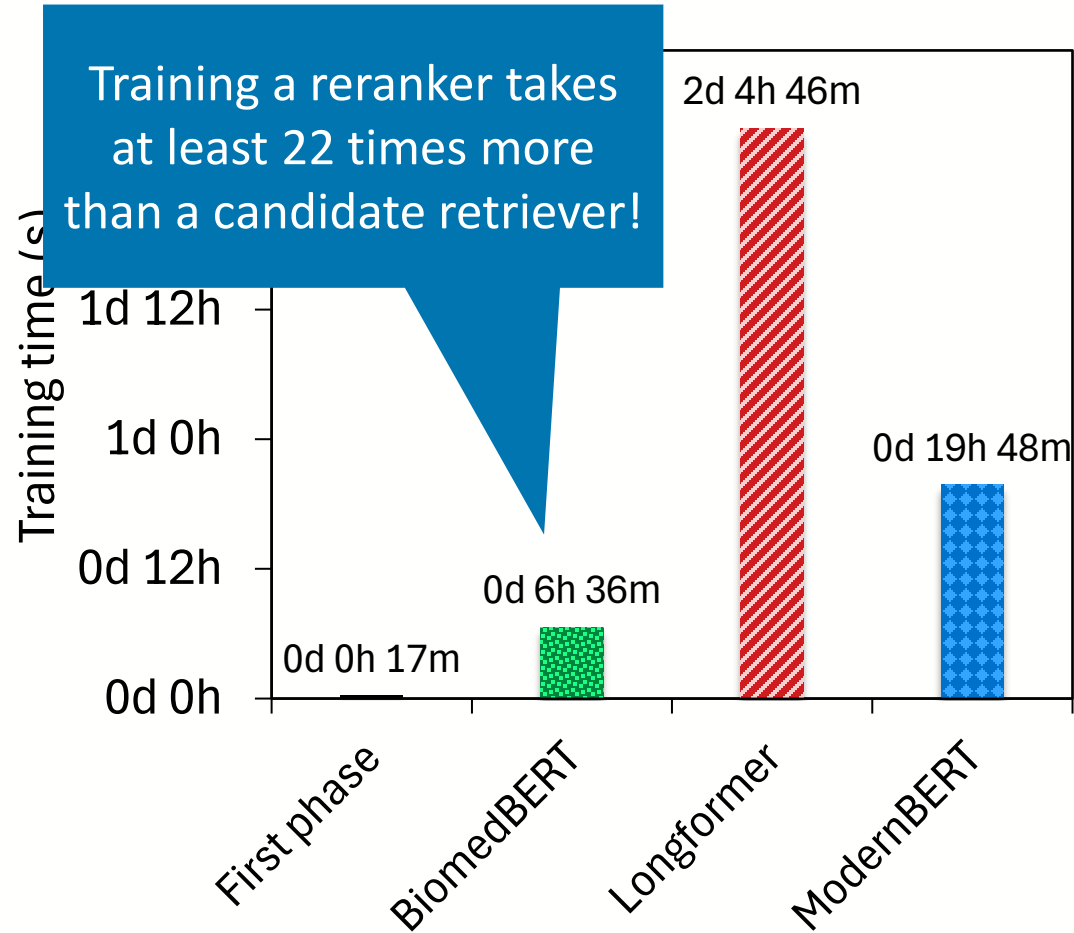
- **We compare:**
 - A first-stage candidate retrieval model
 - With three cross-encoders (state-of-the-art rerankers)
- **We use the same hardware**
- **Dataset:** MedMentions
 - Training data:
 - 2,635 biomedical paper abstracts
 - 211, 029 mentions
 - Test data:
 - 879 biomedical paper abstracts
 - 70,405 mentions

Far from what we might find
on the complete PubMed!

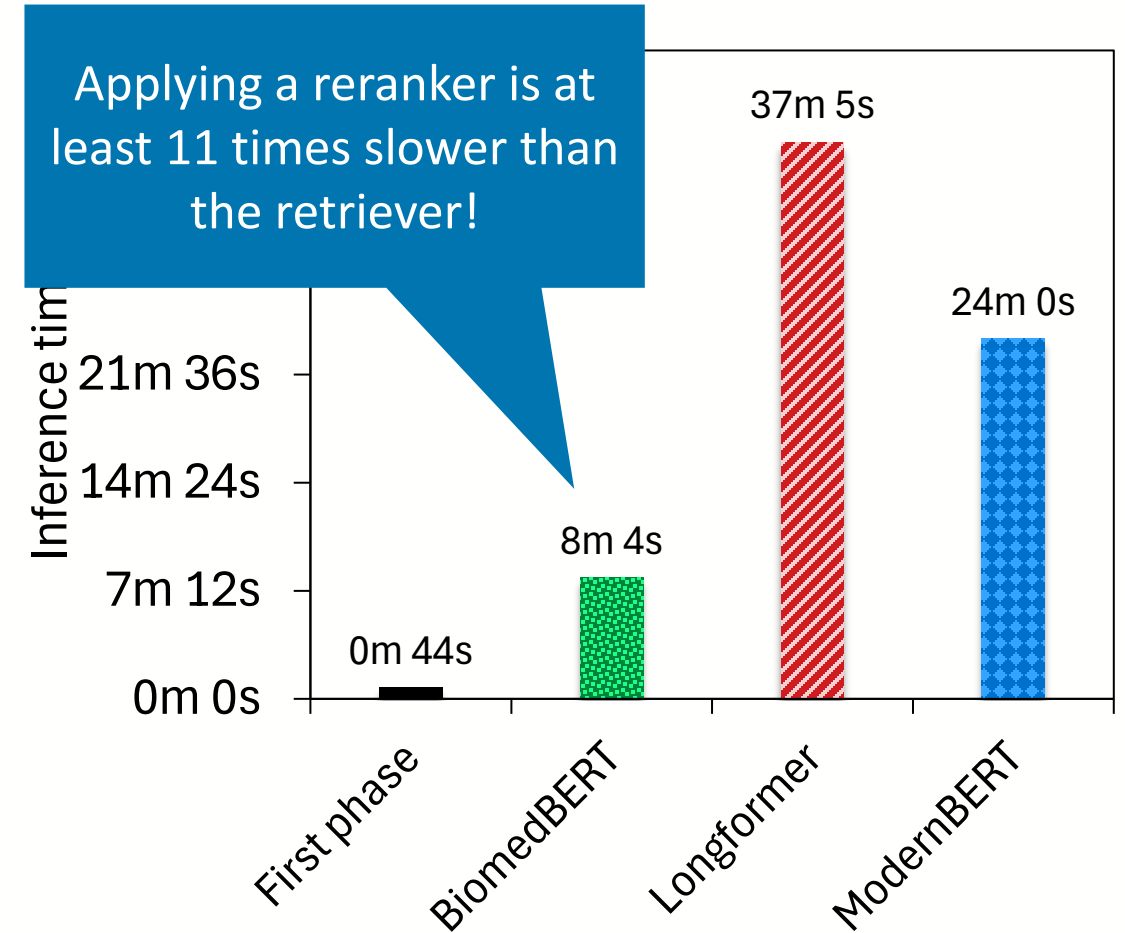


How slow are entity linking rerankers?

Training time



Inference time





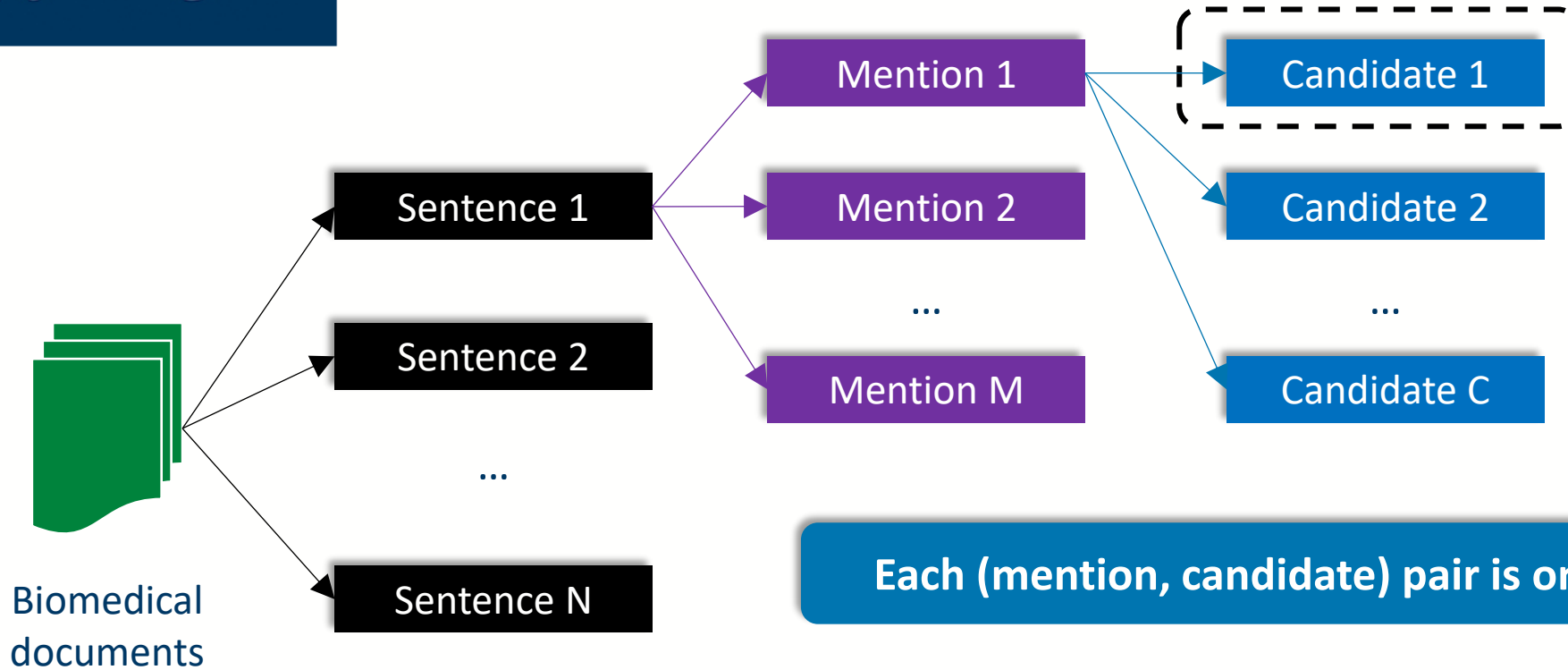
Research question

Cross-encoder rerankers are very slow...

Can we make them faster without harming performance?

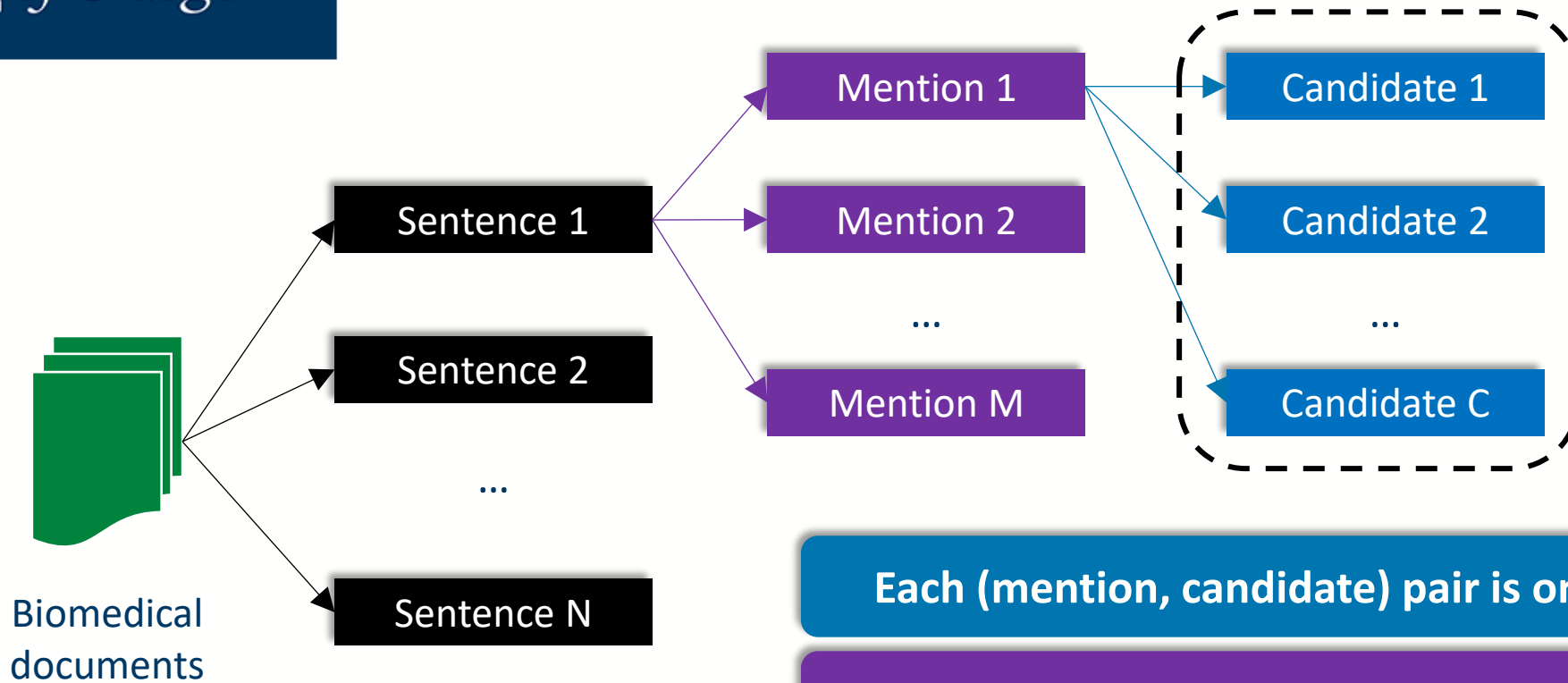


Let's review the cross-encoder working





Let's review the cross-encoder working

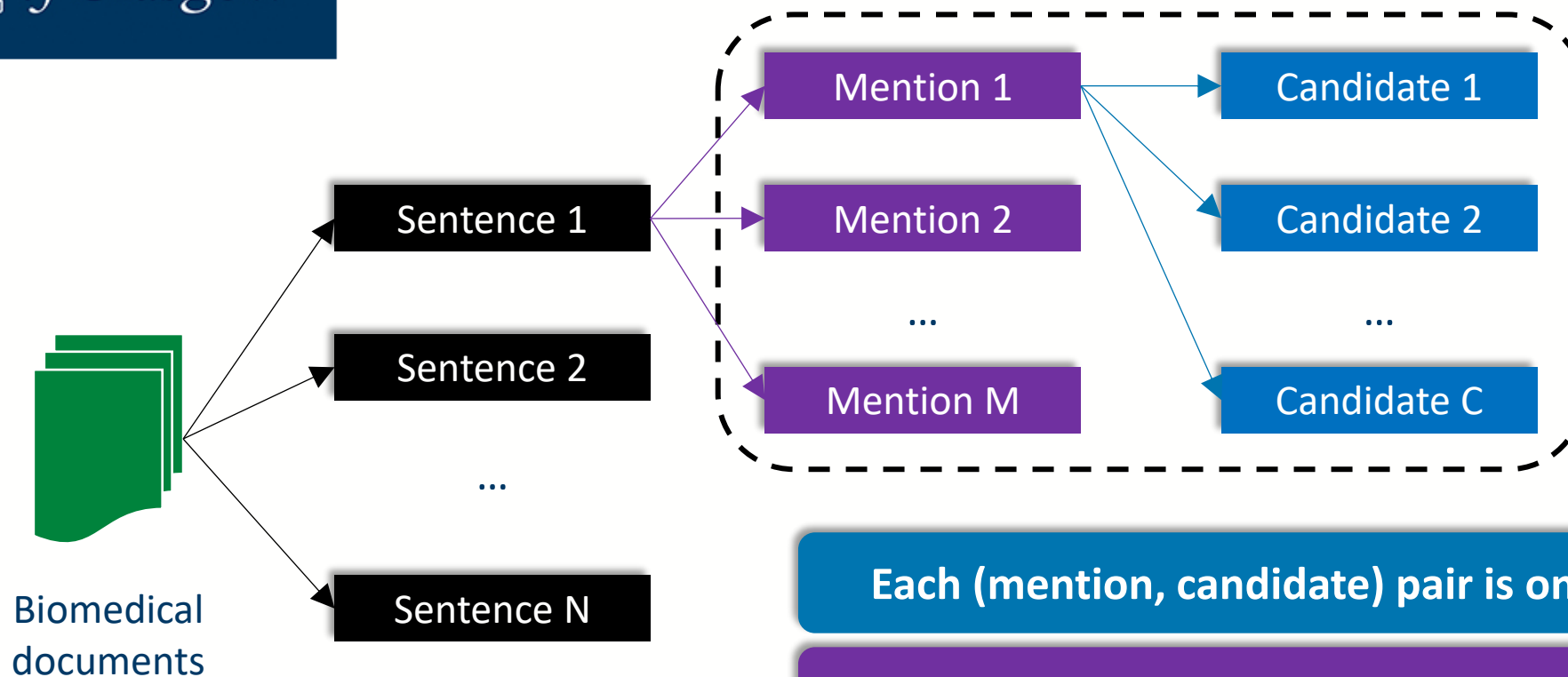


Each (mention, candidate) pair is only processed once.

The same mention is being processed C times!



Let's review the cross-encoder working



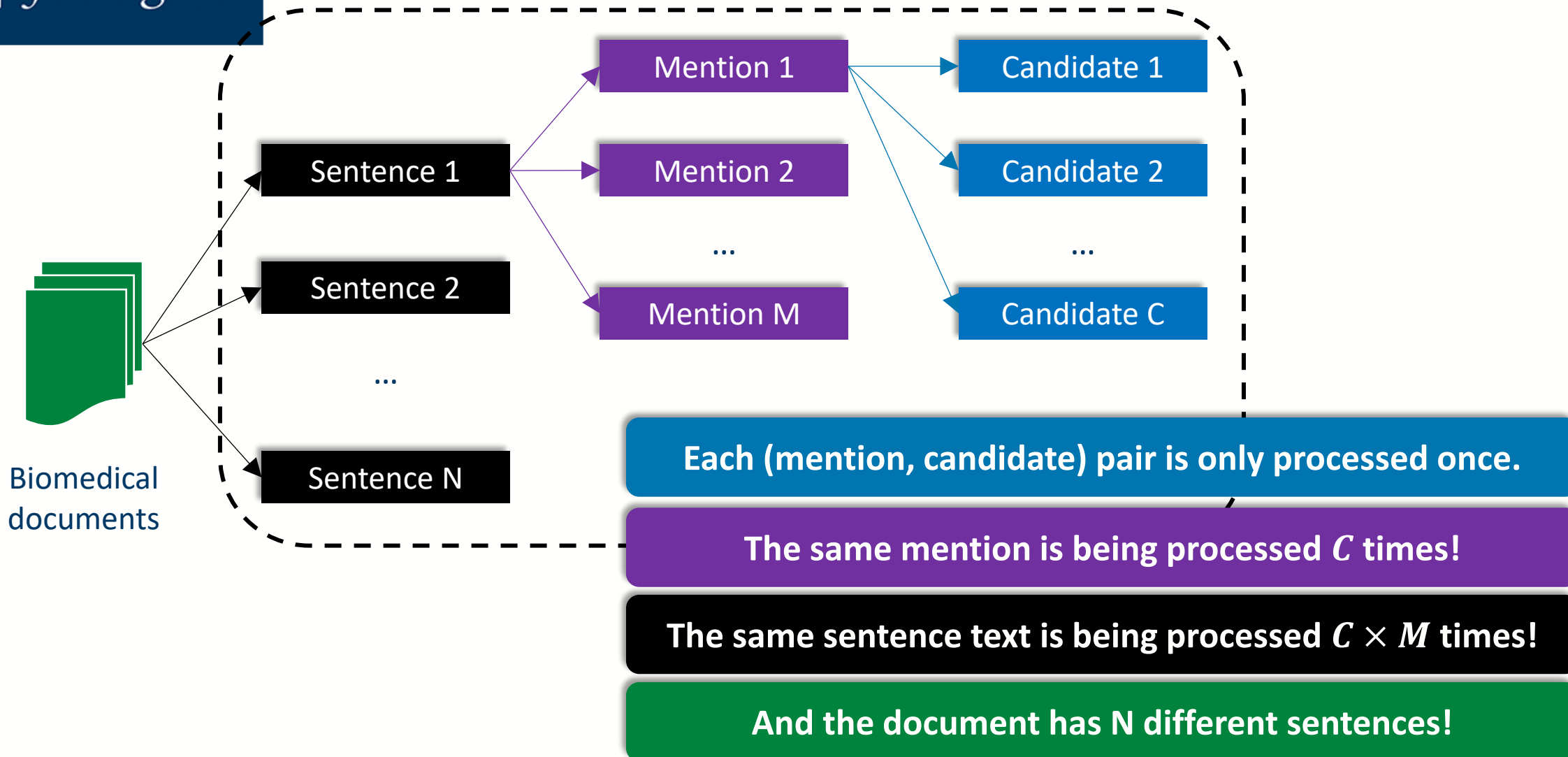
Each (mention, candidate) pair is only processed once.

The same mention is being processed C times!

The same sentence text is being processed $C \times M$ times!



Let's review the cross-encoder working





Accelerating cross-encoders

Each (mention, candidate) pair is only processed once.

The same mention is being processed C times!

The same sentence text is being processed $C \times M$ times!

And the document has N different sentences!

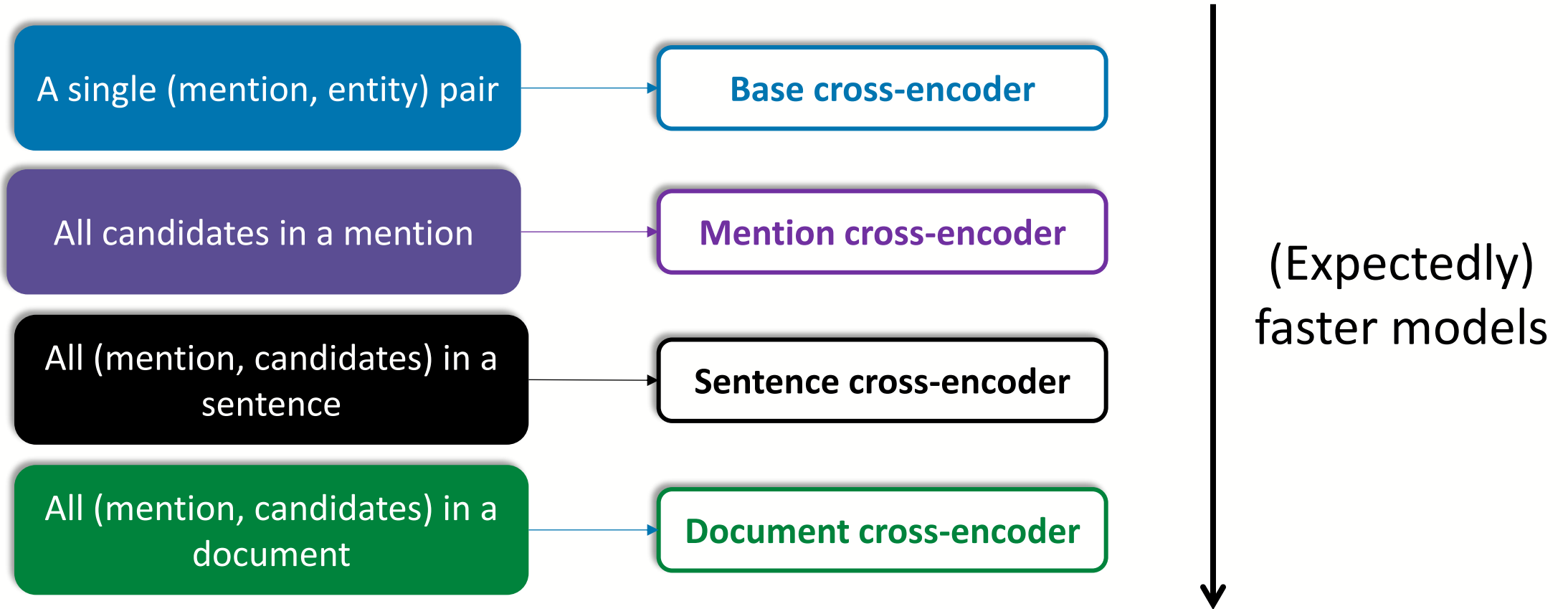
Both in training
and inference

Idea: What if, instead of showing one (mention, candidate) pair on each call, we show several simultaneously?



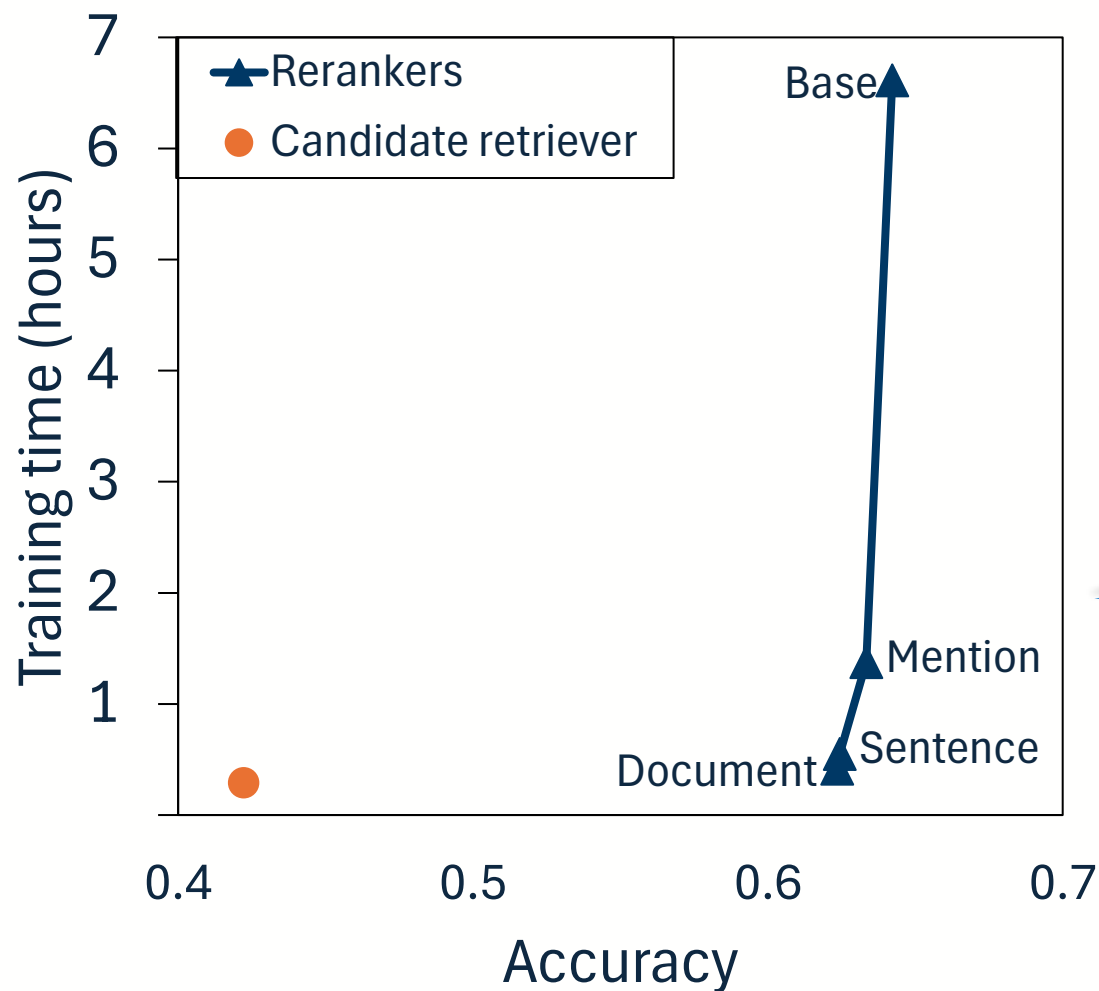
Accelerating cross-encoders

We build re-rankers where we provide different amounts of (mention, candidate) pairs to process





How do these models work?



All rerankers are more effective than the candidate retriever, as expected

Variation in effectiveness is small

But, as we add more (mention, entity) pairs, training time notably diminishes

Similar results were obtained on other datasets and models, and also on inference speed



Conclusions

Processing more (mention, entity) pairs simultaneously has the following effects on entity linking cross-encoder rerankers

Small variations in 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

**We can have a major
training / inference
speed boost at a small
accuracy cost!**

**Our solutions are suitable for environments where speed is crucial
(or data is huge)**



Paper link

Questions?



Dr. Javier Sanz-Cruzado

AI4BioMed Group, University of Glasgow



javier.sanz-cruzadopuig@glasgow.ac.uk



[JavierSanzCruza](#)



[Javiersanzcruza.bsky.social](#)



<https://www.linkedin.com/in/javier-sanz-cruzado-puig/>





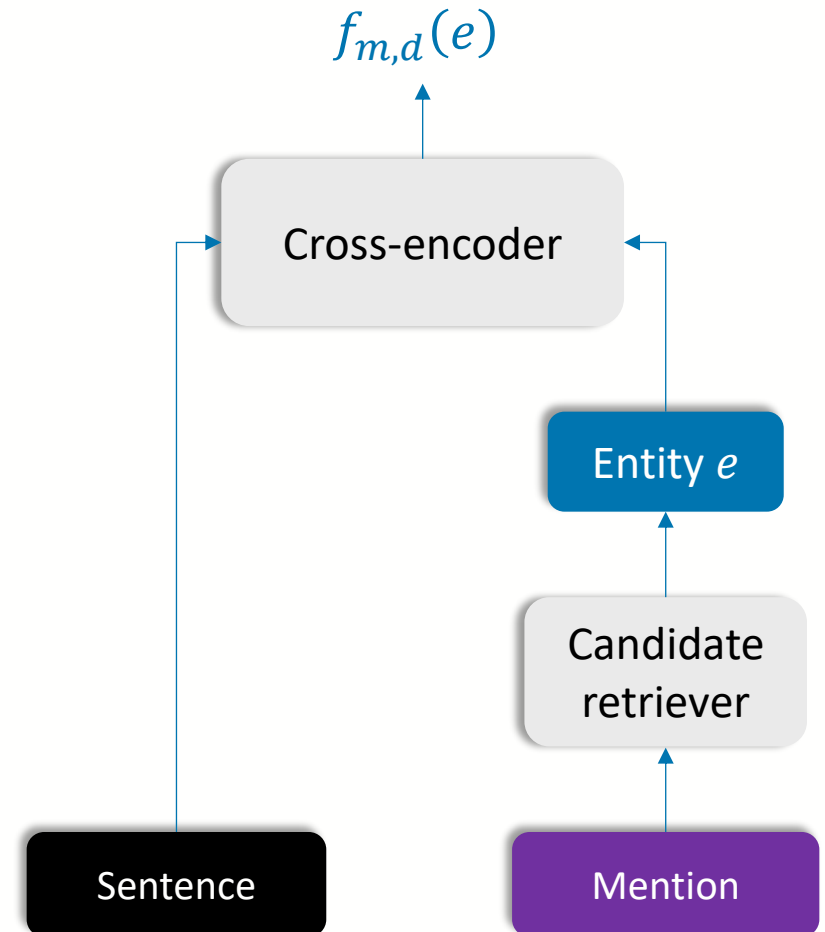
Solution 1: Mention cross-encoder

Each (mention, candidate) pair
is only processed once.

The same mention is being
processed C times!

The same sentence text is
being processed $C \times M$ times!

And the document has N
different sentences!





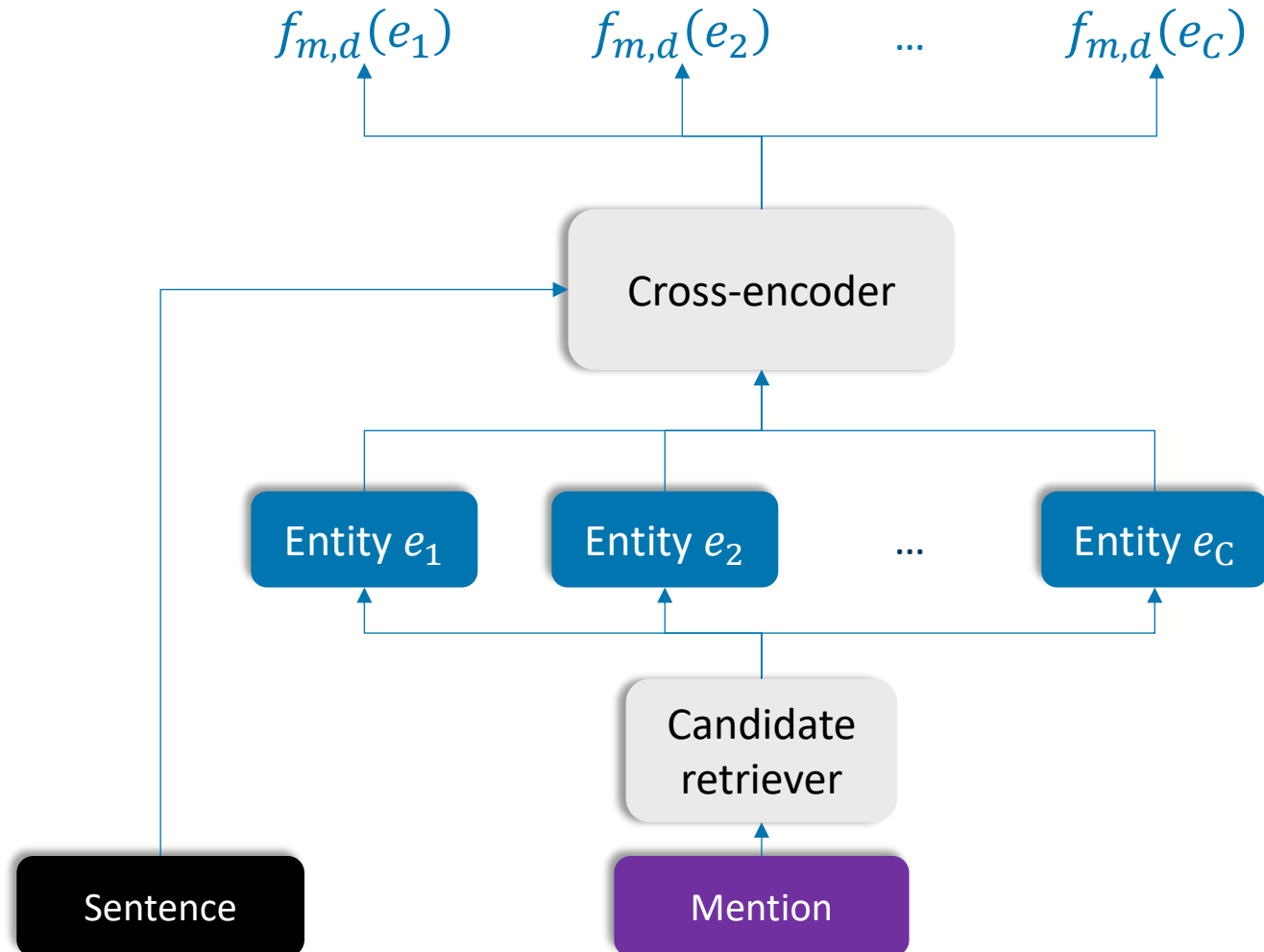
Solution 1: Mention cross-encoder

Each (mention, candidate) pair
is only processed once.

The same mention is being
processed C times!

The same sentence text is
being processed $C \times M$ times!

And the document has N
different sentences!





Solution 1: Mention cross-encoder

- While the cross encoder uses template for each candidate:

Each (mention, candidate) pair
is only processed once.

The same mention is being
processed C times!

The same sentence text is
being processed $C \times M$ times!

And the document has N
different sentences!

Text [SEP] Mention [MASK] Entity e name



Solution 1: Mention cross-encoder

Each (mention, candidate) pair
is only processed once.

The same mention is being
processed C times!

The same sentence text is
being processed $C \times M$ times!

And the document has N
different sentences!

- The mention cross-encoder receives a sentence using the following a template for each mention:

Text [SEP] Mention [MASK] Entity e_1 name
[SEP] Mention [MASK] Entity e_2 name
...
[SEP] Mention [MASK] Entity e_C name

- Therefore, the score of the entity e_i is the probability of its [MASK] token taking value 1



Solution 1: Mention cross-encoder

Each (mention, candidate) pair
is only processed once.

The same mention is being
processed once.

The same sentence text is
being processed $C \times M$ times!

And the document has N
different sentences!

- The mention cross-encoder receives a sentence using the following a template for each mention:

Text [SEP] Mention [MASK] Entity e_1 name
[SEP] Mention [MASK] Entity e_2 name
...
[SEP] Mention [MASK] Entity e_C name

- Therefore, the score of the entity e_i is the probability of its [MASK] token taking value 1

And so on for the other models