The University of Queensland
Australia
CREATE CHANGE

# An Overview of Recent Advances in Neural Information Retrieval

## ielab's contributions to better neural models of search

Guido Zuccon

g.zuccon@uq.edu.au
ielab, The University of Queensland, Australia
www.ielab.io

ie
lab

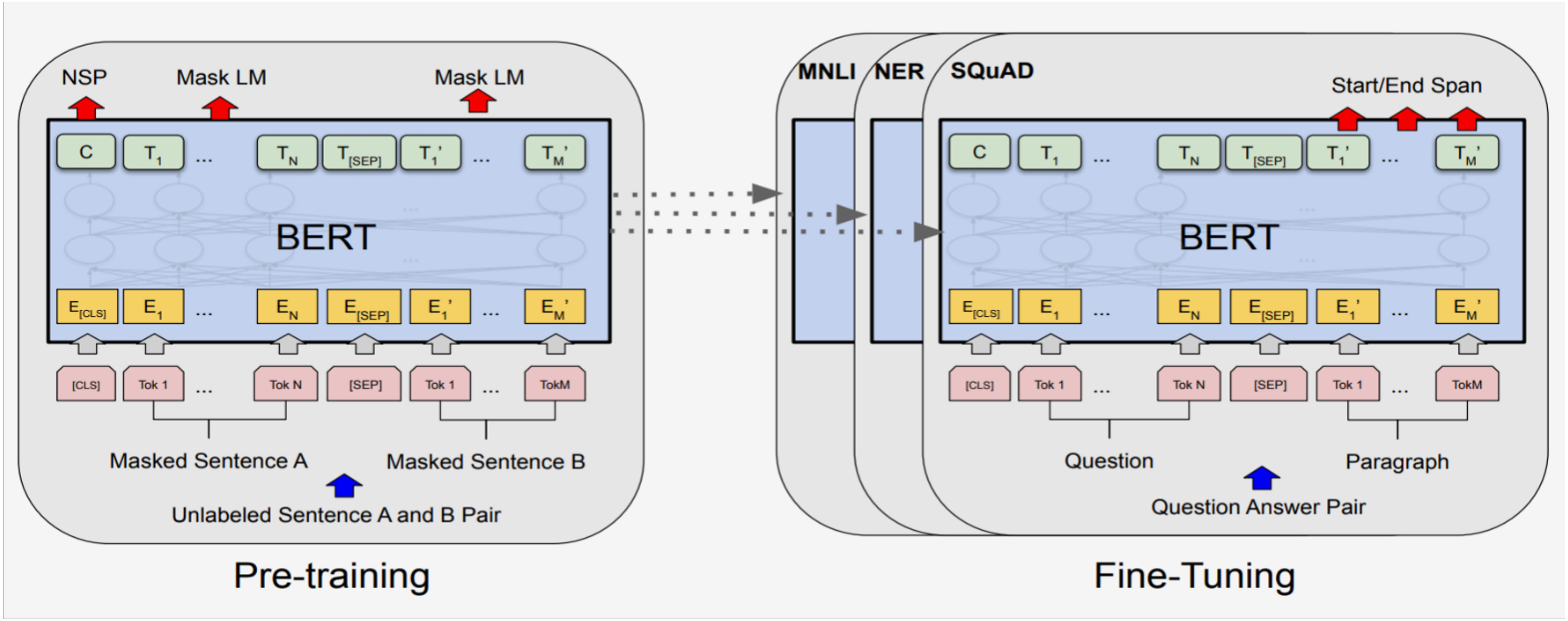
# Objective of this talk

In this talk, I will flag some of the **challenges** these methods pose and the **research my team is doing to tackle them**

ie lab

THE UNIVERSITY OF QUEENSLAND AUSTRALIA CREATE CHANGE

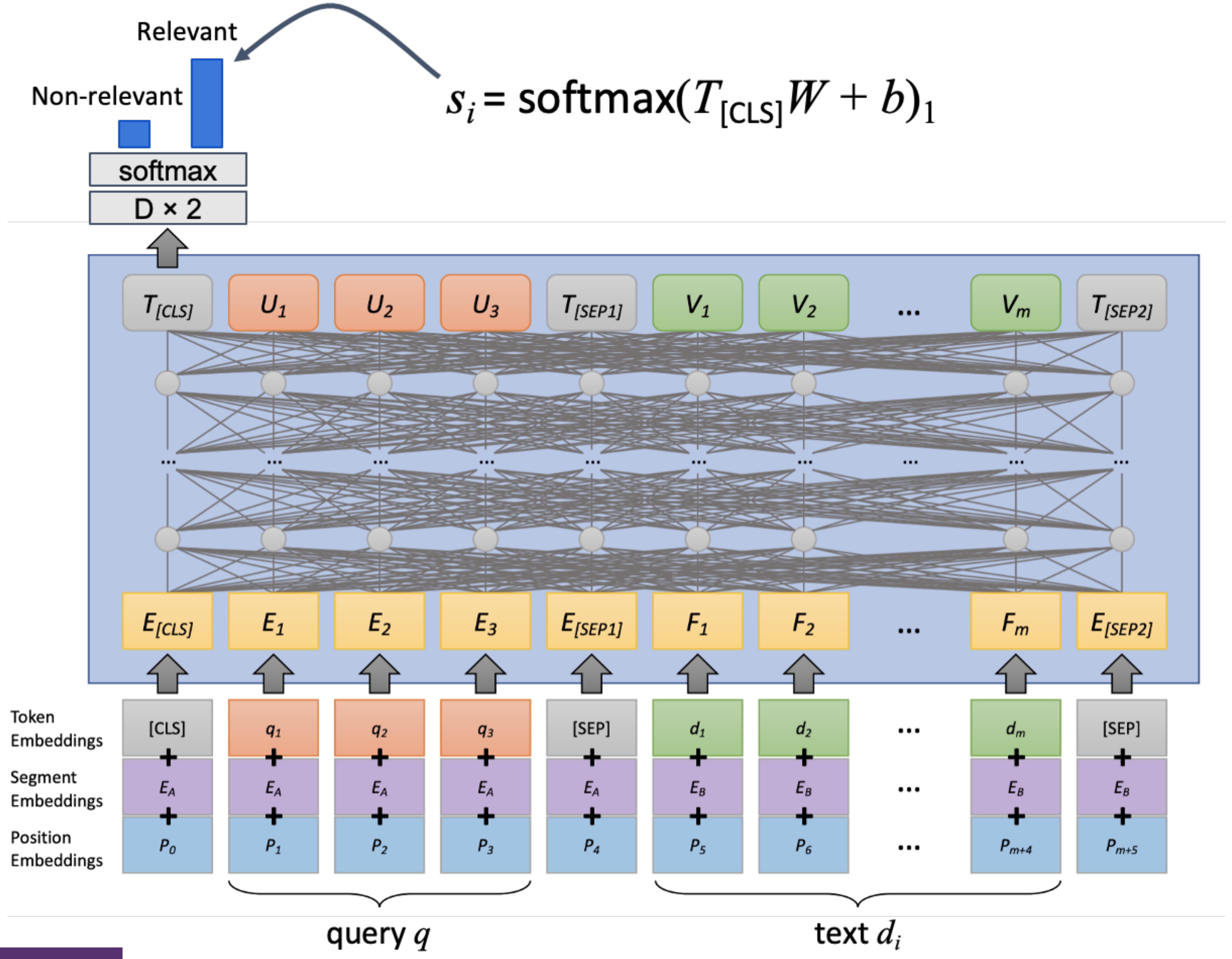
# Pre-trained Language Models (BERT)




Self-supervised: uses large amount of training data

Supervised: (few) labeled examples

# monoBERT: BERT reranker



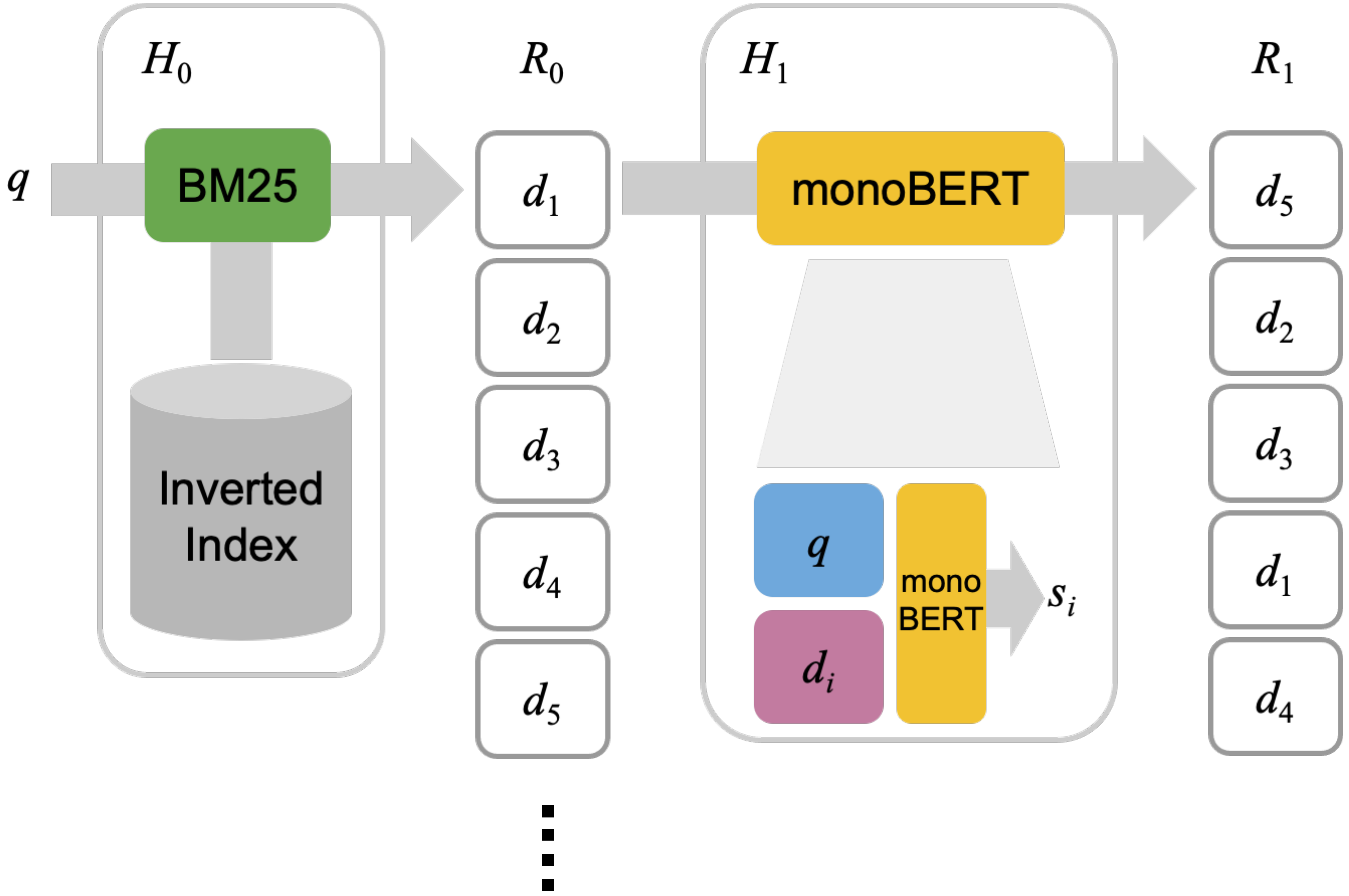

$s_i = \text{softmax}(T_{[\text{CLS}]}W + b)_1$

We want:

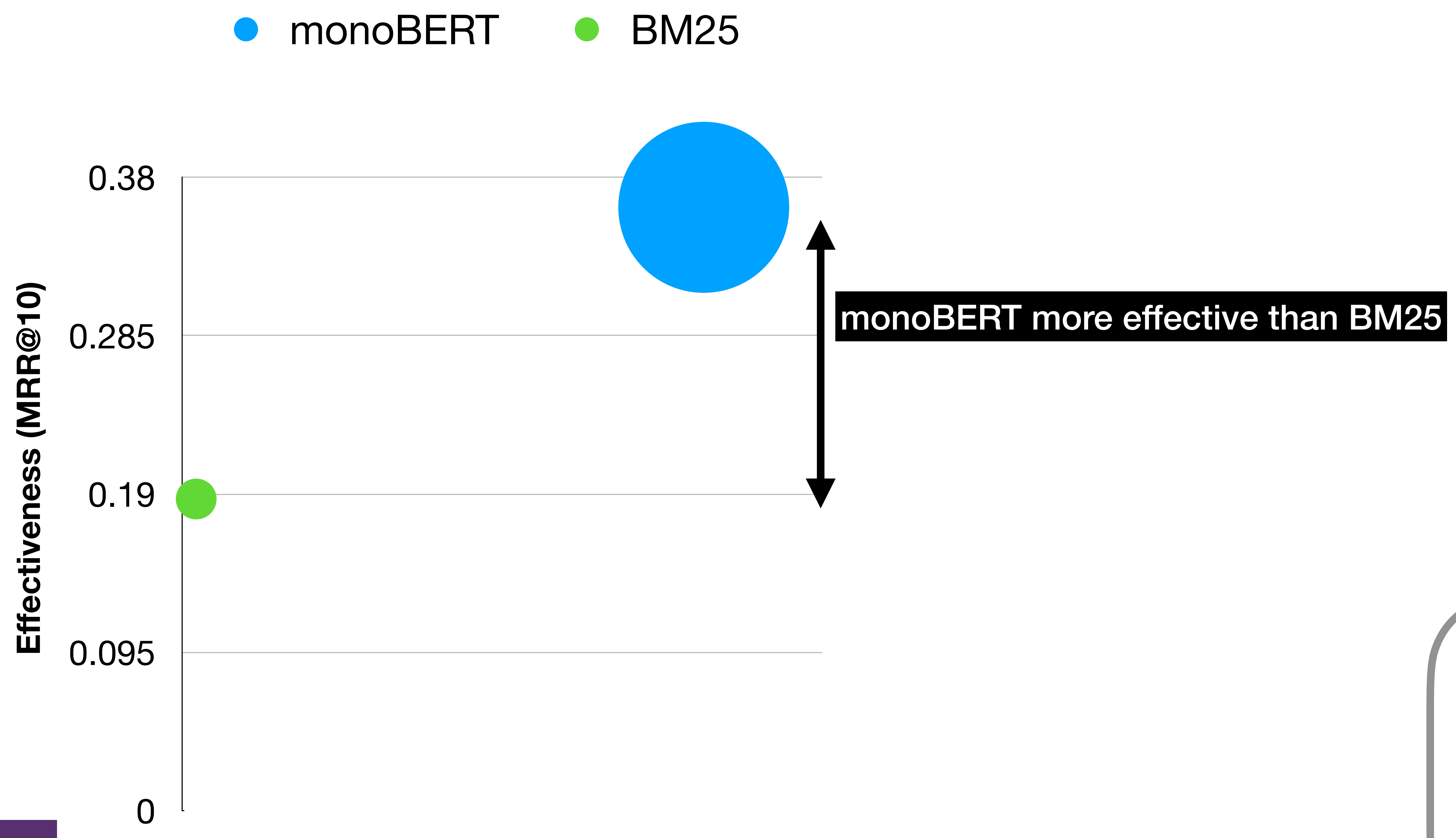
$s_i = P(\text{Relevant} = 1 | q, d_i)$

Loss: $L = -\sum_{j \in J_{\text{pos}}} \log(s_j) - \sum_{j \in J_{\text{neg}}} \log(1 - s_j)$

# Using monoBERT for ranking

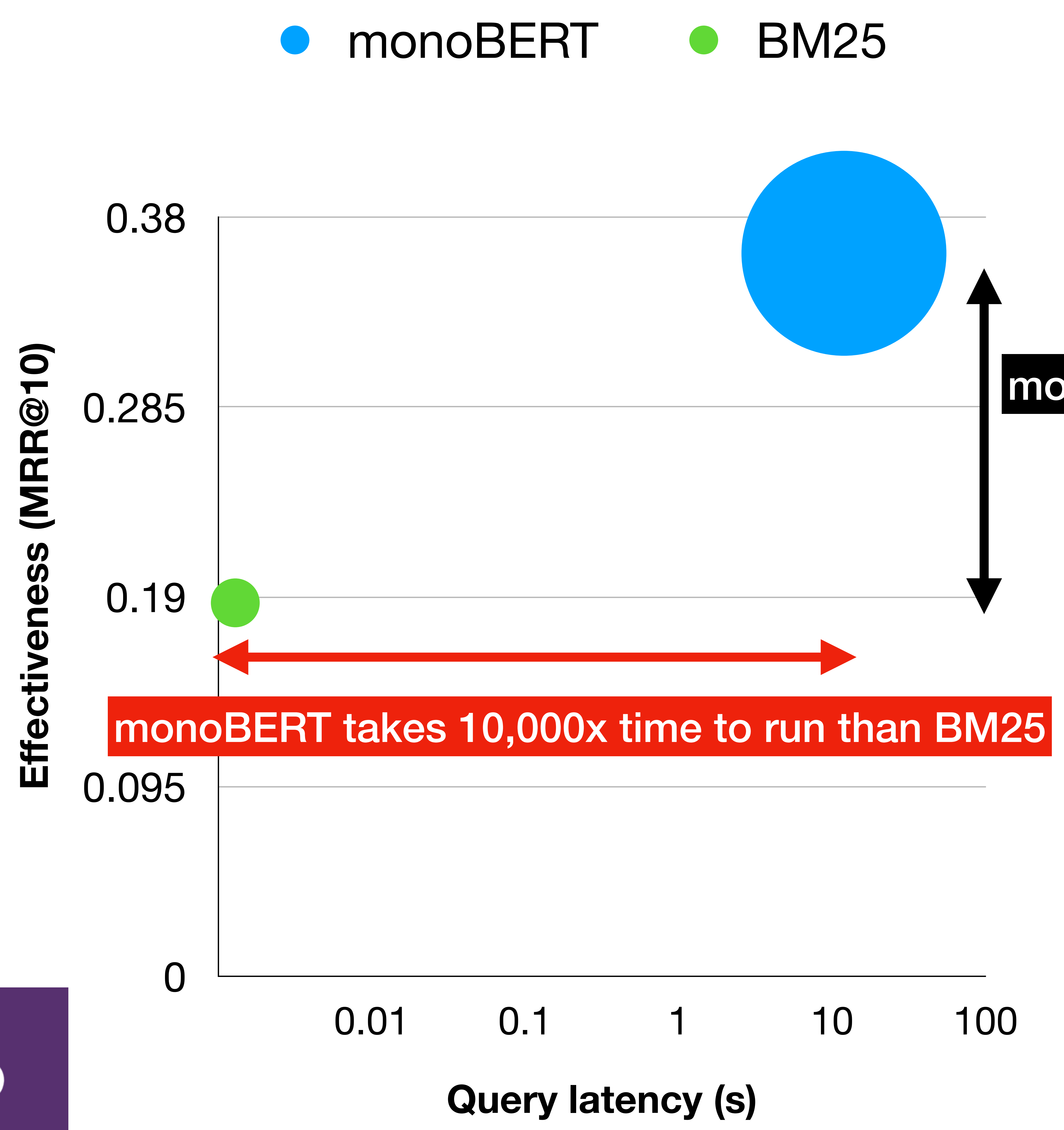


# BERT is effective!




Scells, Zhuang, Zuccon, *"Reduce, Reuse, Recycle: Green Information Retrieval Research"*, SIGIR 2022

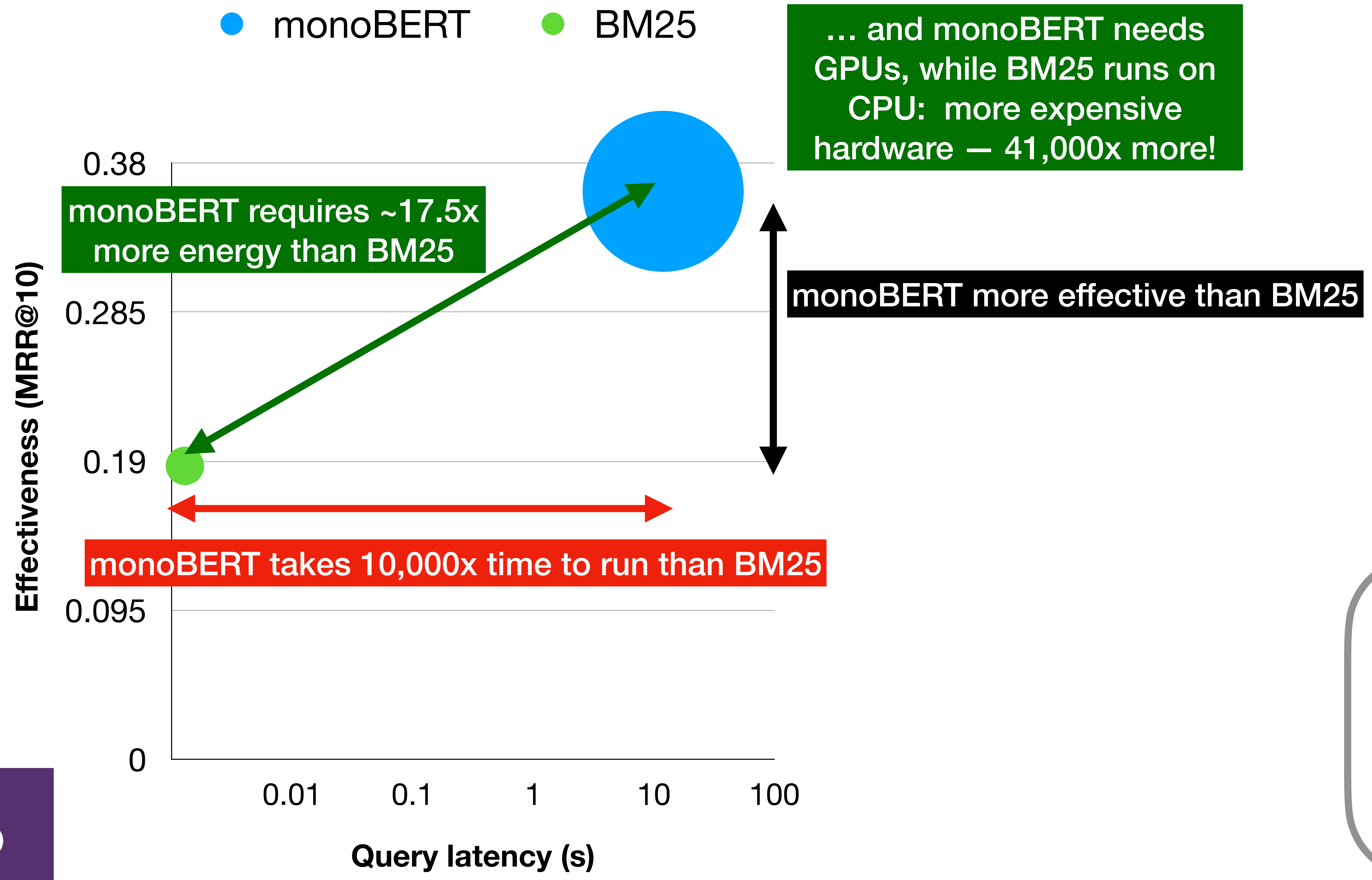
# BERT's Challenges: (1) it's slow




Scells, Zhuang, Zuccon, *"Reduce, Reuse, Recycle: Green Information Retrieval Research"*, SIGIR 2022

ie lab

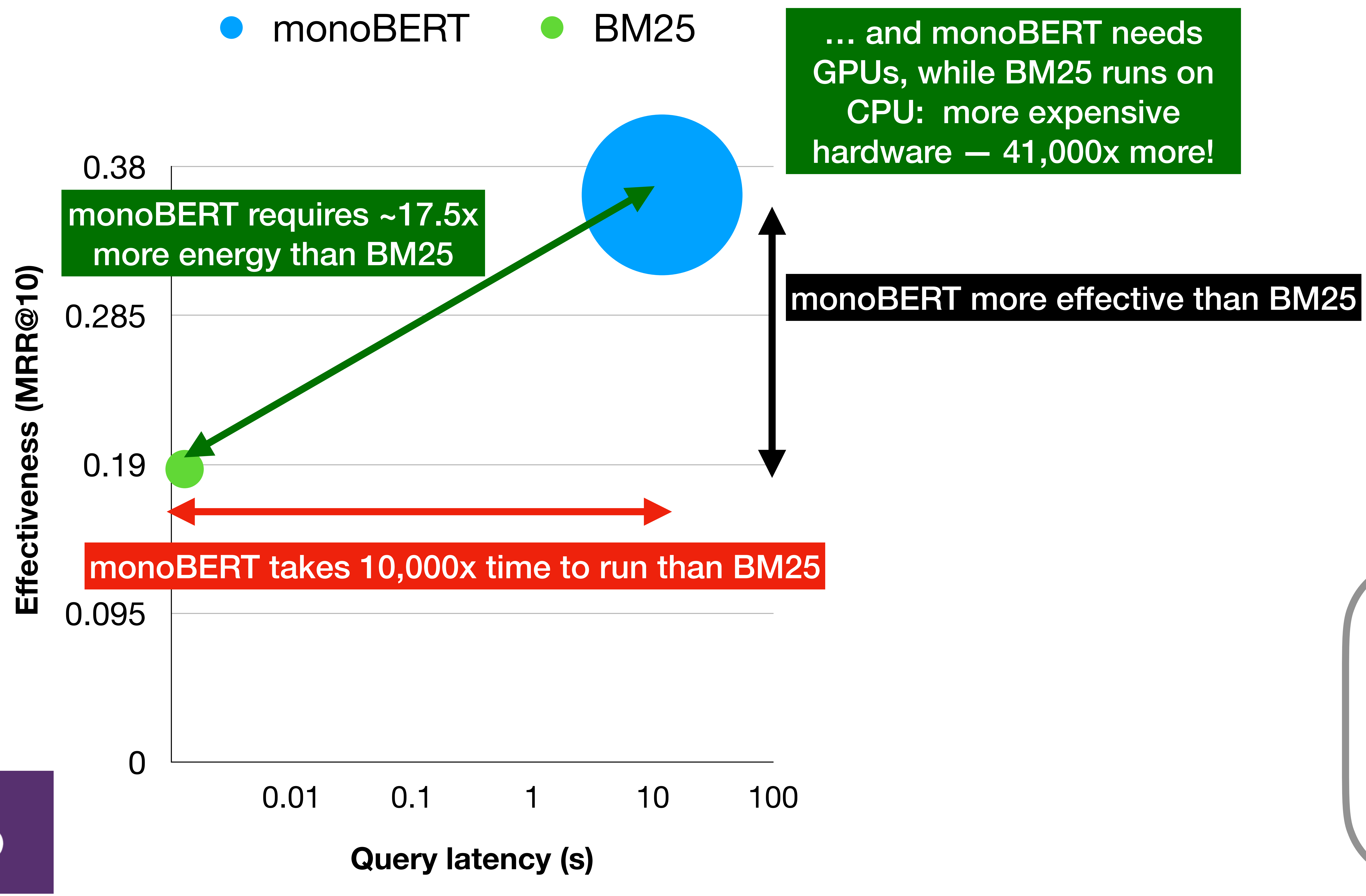
# BERT’s Challenges: (1) it’s slow, (2) it’s expensive

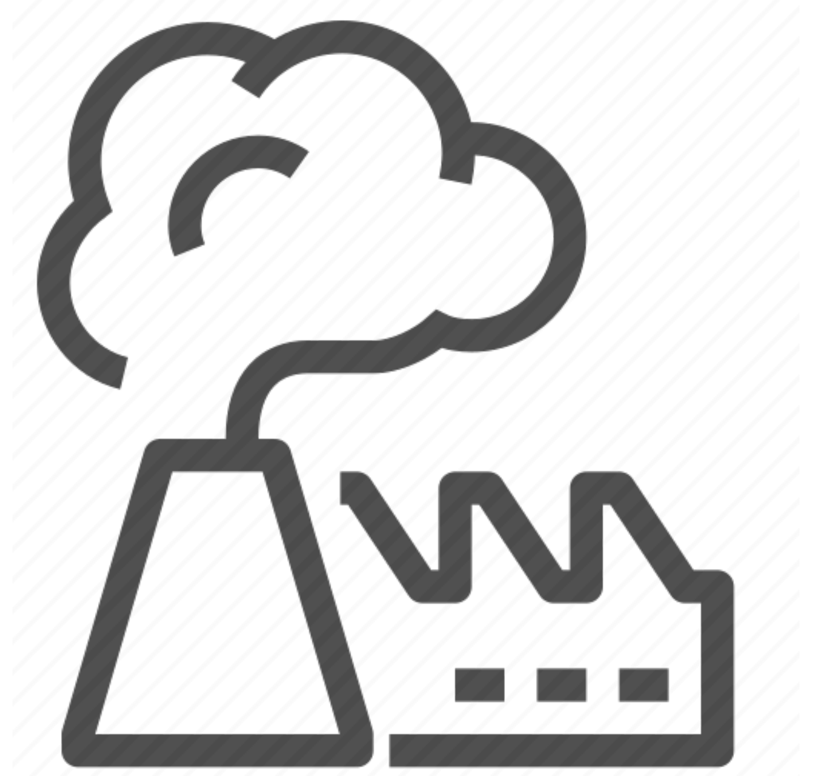



Scells, Zhuang, Zuccon, *“Reduce, Reuse, Recycle: Green Information Retrieval Research*", SIGIR 2022

ie lab

# BERT's Challenges: (1) it's slow, (2) it's expensive

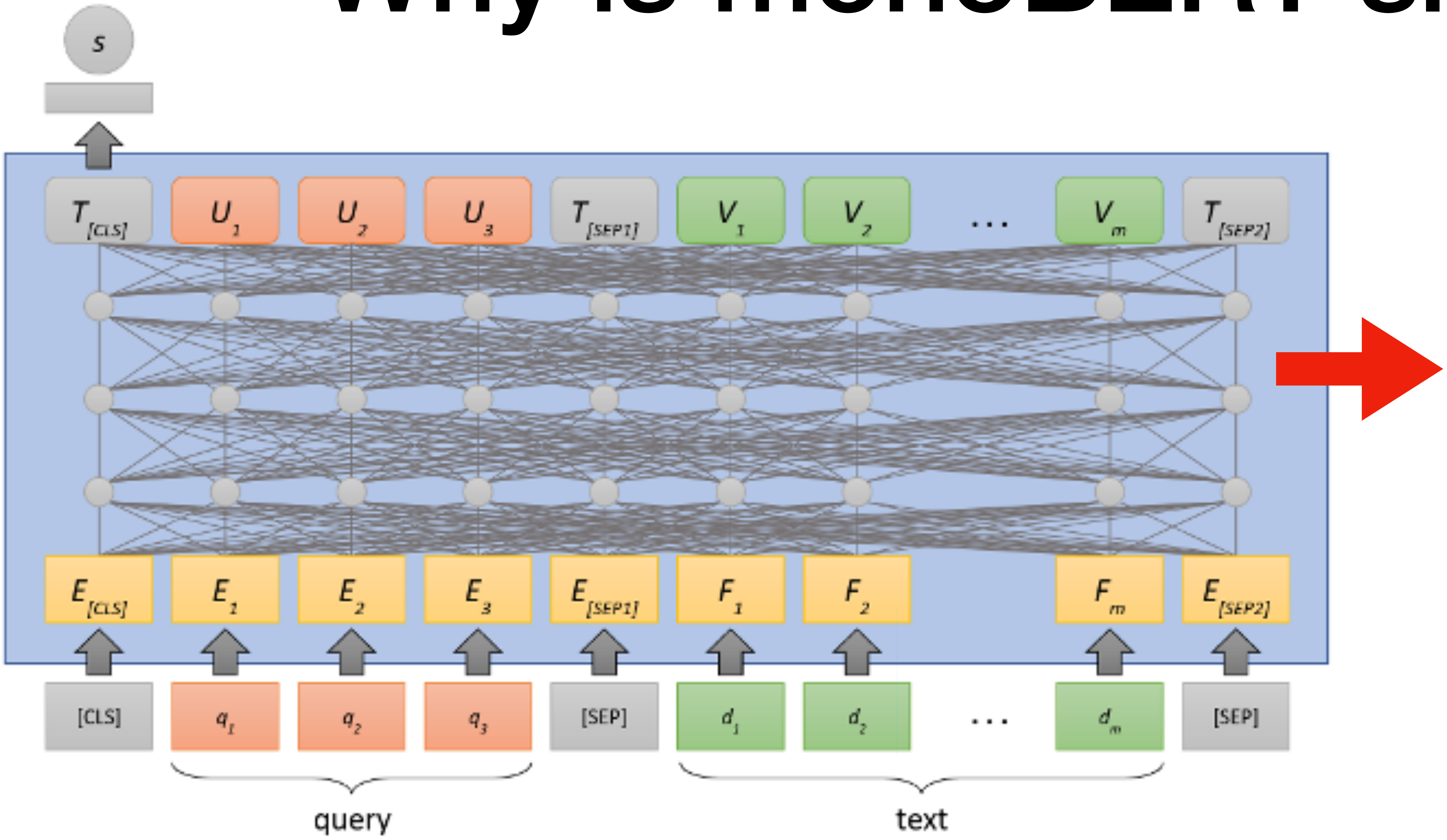


monoBERT produces 165,000x more $CO_2$ emissions than BM25

Scells, Zhuang, Zuccon, *"Reduce, Reuse, Recycle: Green Information Retrieval Research"*, SIGIR 2022

# Why is monoBERT slow?




**Computationally expensive layers**
e.g. 110+ million learned weights

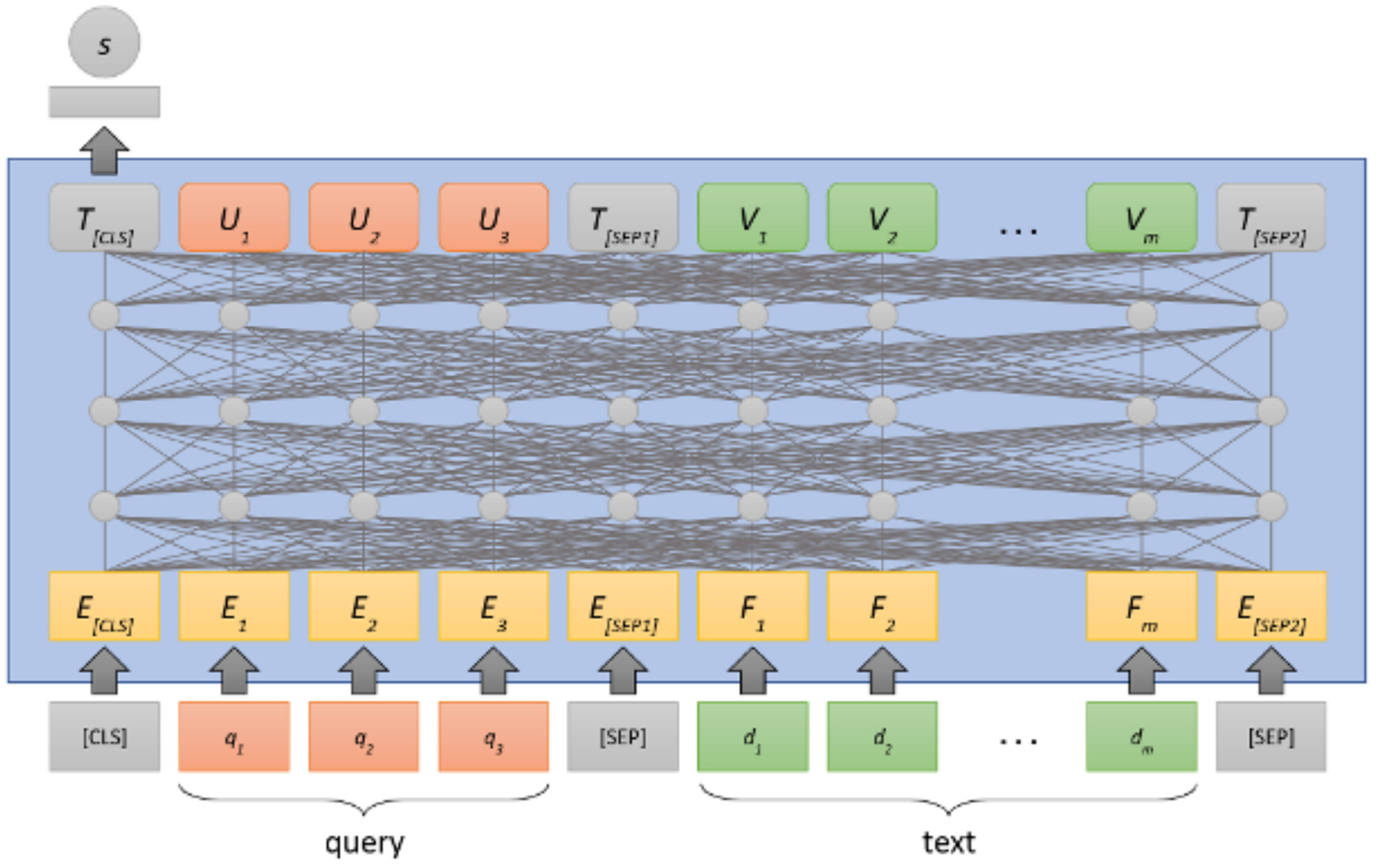
Possible Solutions:

A. Multistage ranking pipeline with limited re-ranking
B. Simplification of BERT inference to lower query latency

ie lab

The University Of Queensland Australia
CREATE CHANGE

# Cross-encoder (monoBERT) VS Bi-encoder (Dense Retriever)

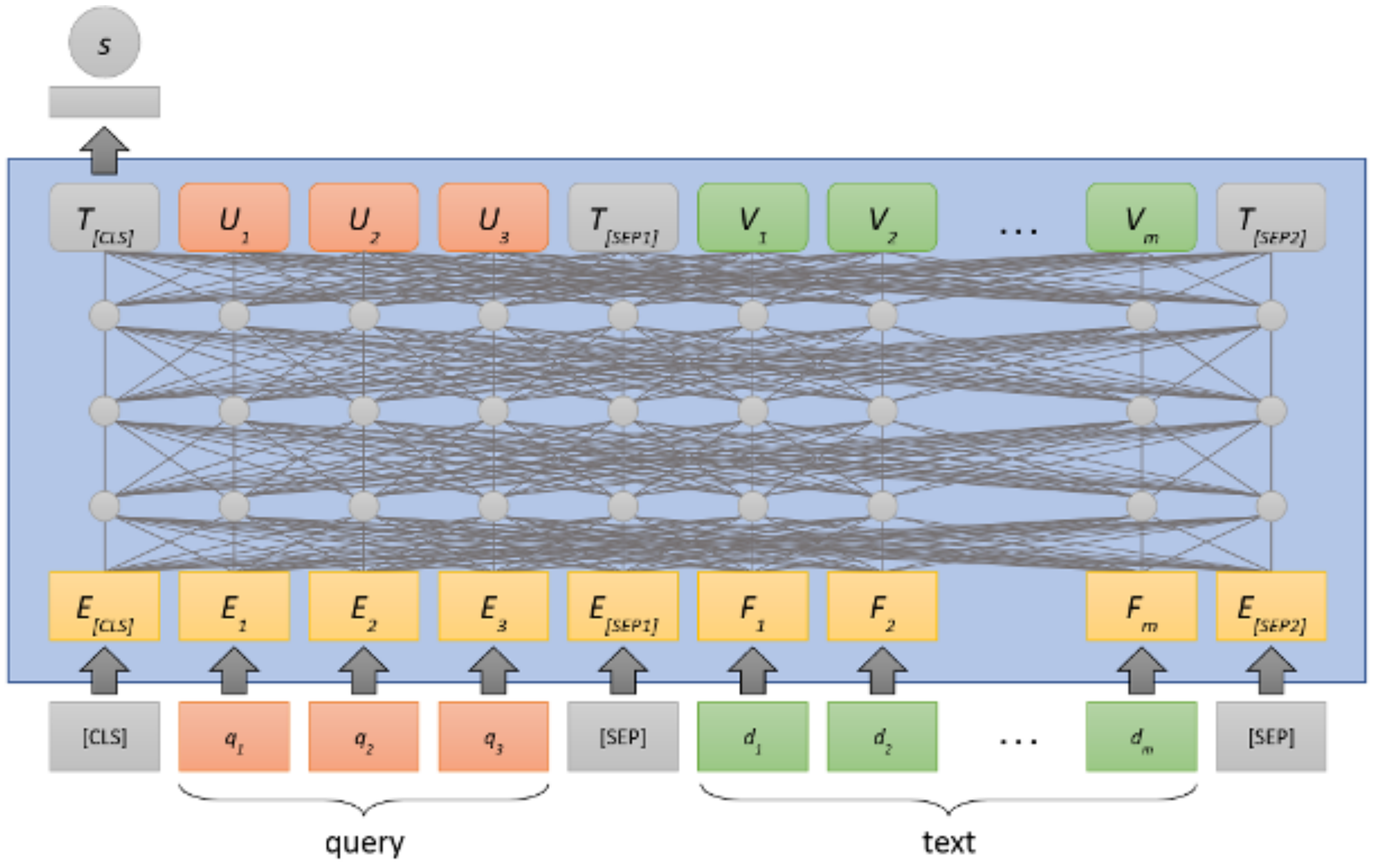


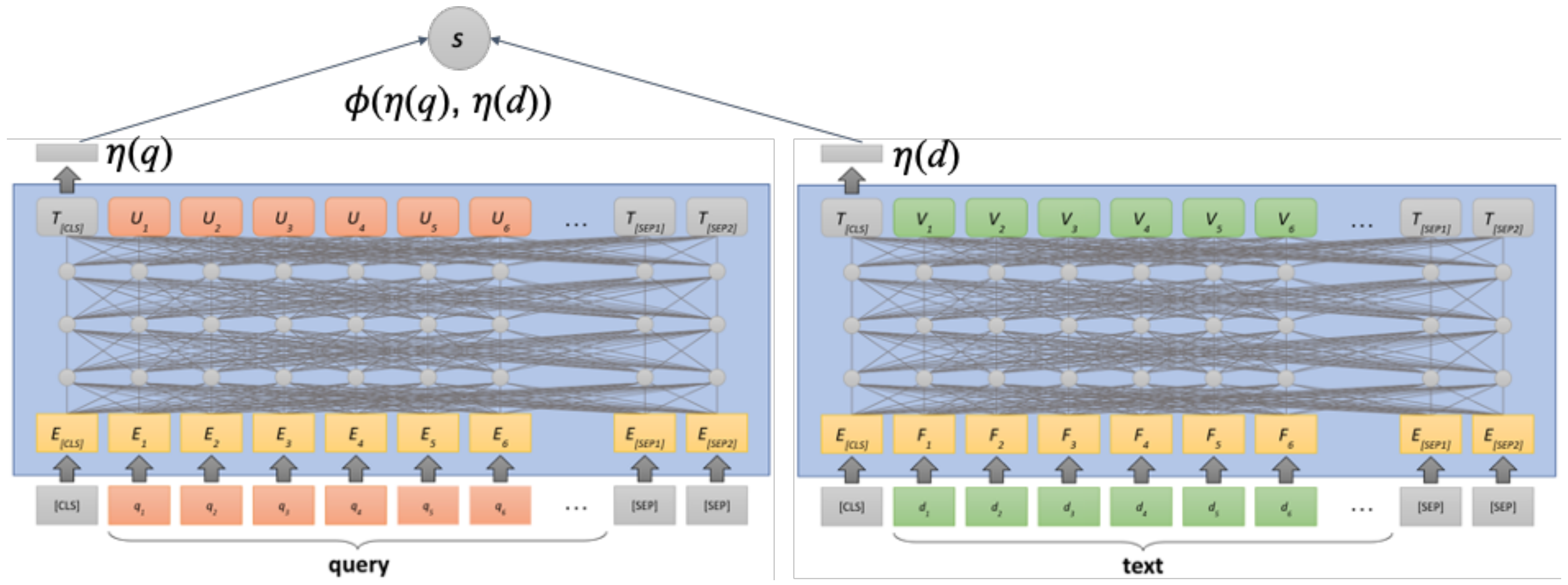

- **Cross-encoder:** at query time, need to pass the query and the document
  - Thus, the whole architecture needs to be run at run-time, for each pair of inputs
    -> k BERT inferences
  - Encoding (BERT inference) is expensive

ie lab

THE UNIVERSITY OF QUEENSLAND AUSTRALIA
CREATE CHANGE

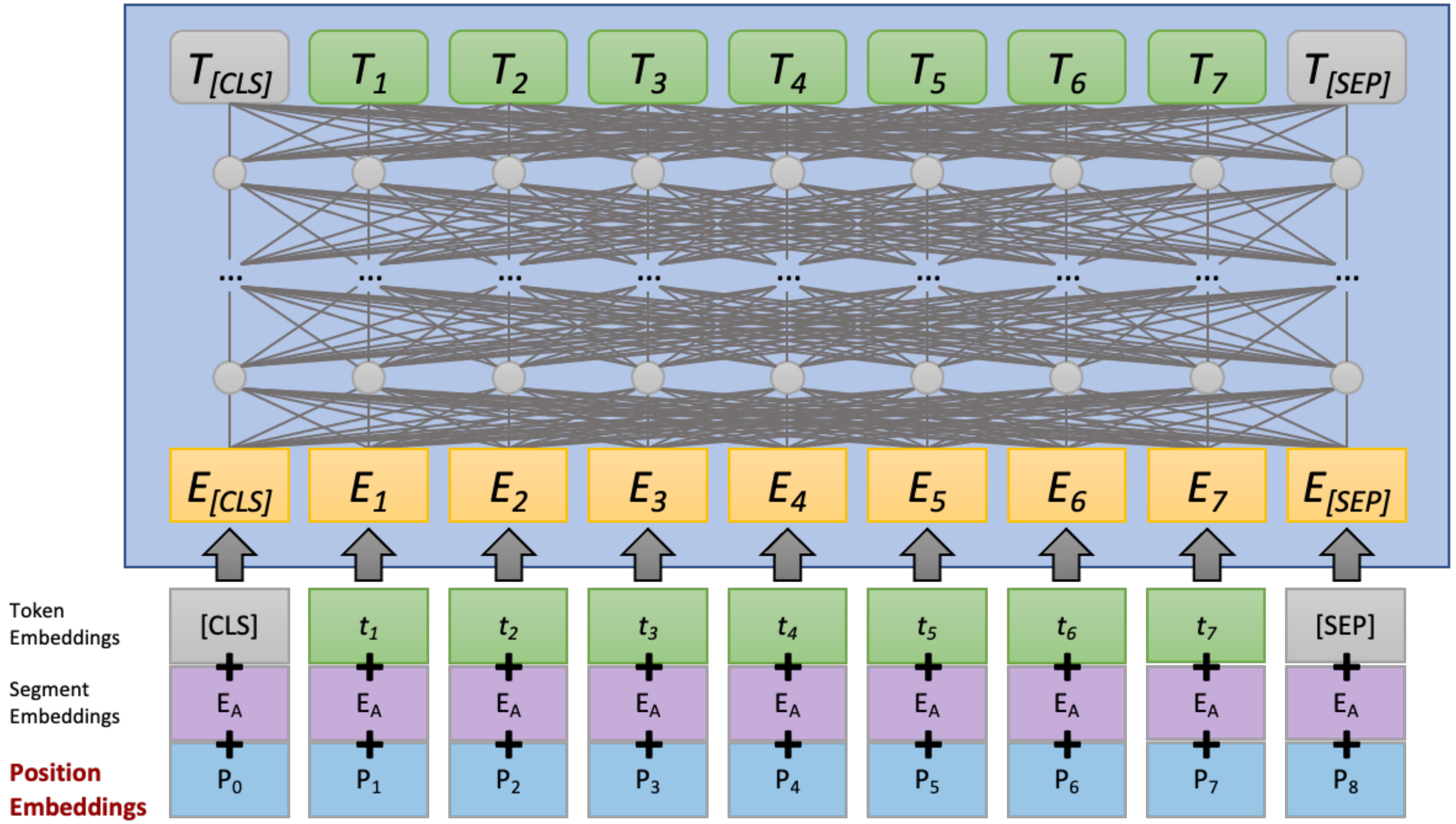
# Cross-encoder (monoBERT) VS Bi-encoder (Dense Retriever)




- **Cross-encoder:** at query time, need to pass the query and the document
  - Thus, the whole architecture needs to be run at run-time, for each pair of inputs
    -> k BERT inferences
  - Encoding (BERT inference) is expensive
  - **Bi-encoder**: at query time, document and query are passed separately
    - Thus, can encode all documents offline
    - At query time, just encode the query: one BERT inference
    - Then, just do similarity between encodings (vectors)

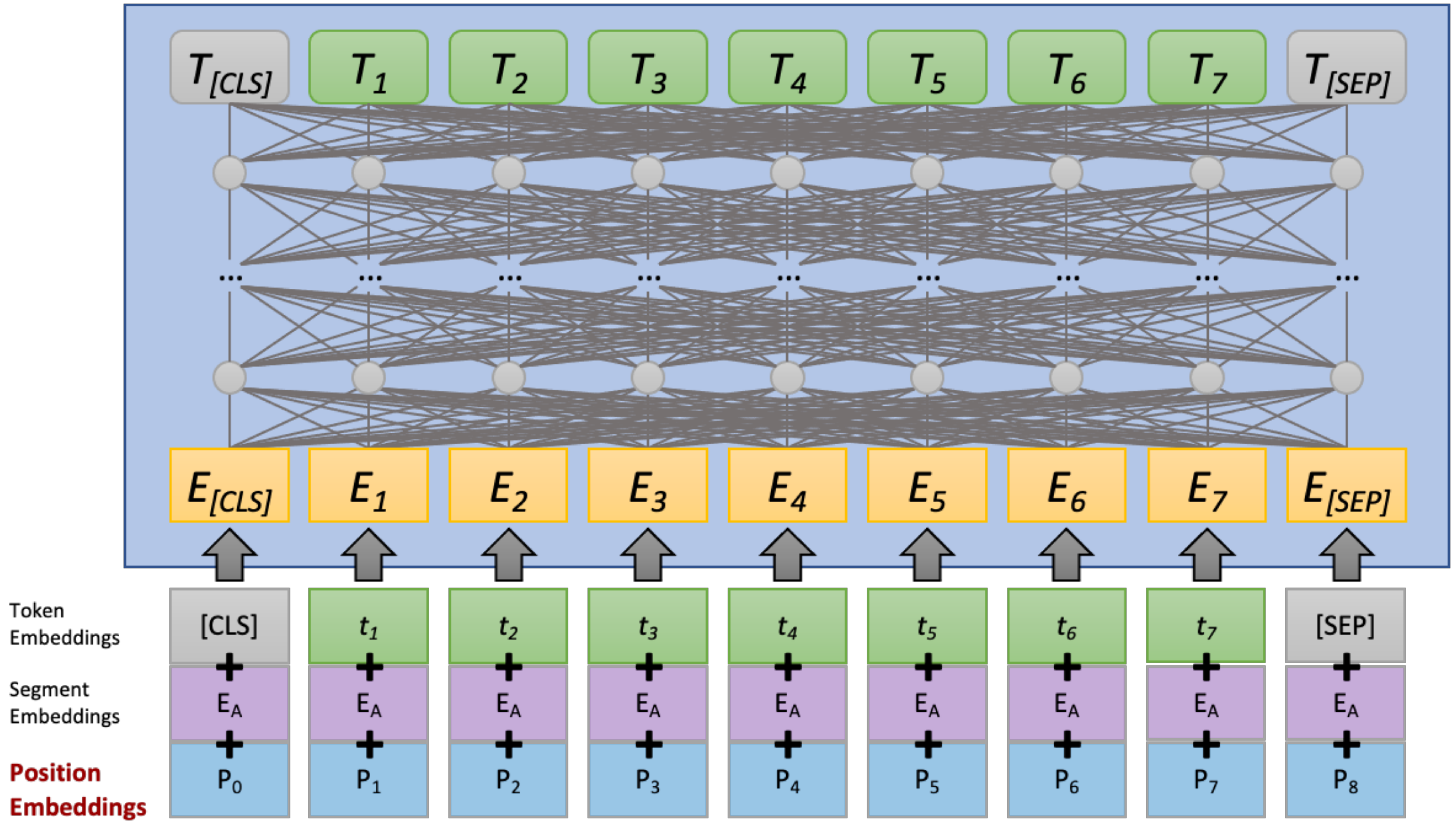
# BERT's Challenges: (3) input limit




**Need separate embedding for every possible position**

Solution: restricted to indices 0-511

# BERT's Challenges: (3) input limit




**Need separate embedding for every possible position**

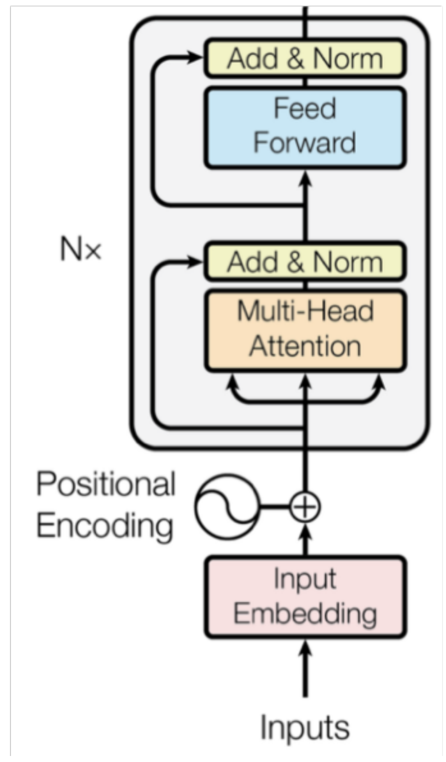
Solution: restricted to indices 0-511

What if our input size does not fit into BERT?

a) How can we do document retrieval?

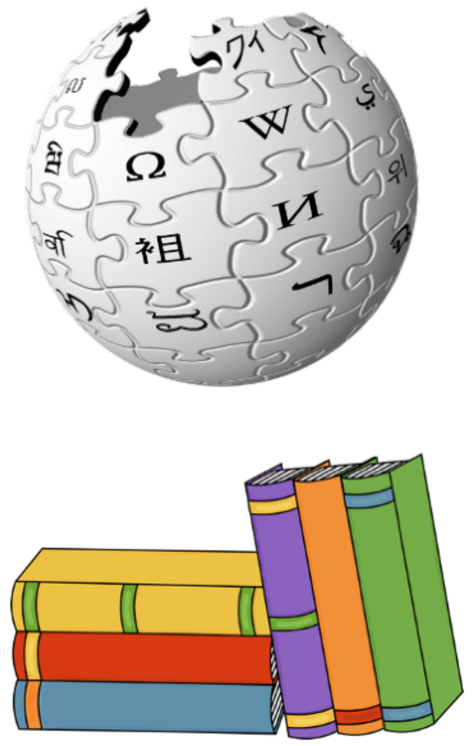
b) How can we do Pseudo-Relevance Feedback?

# BERT's Challenges: (4) training data



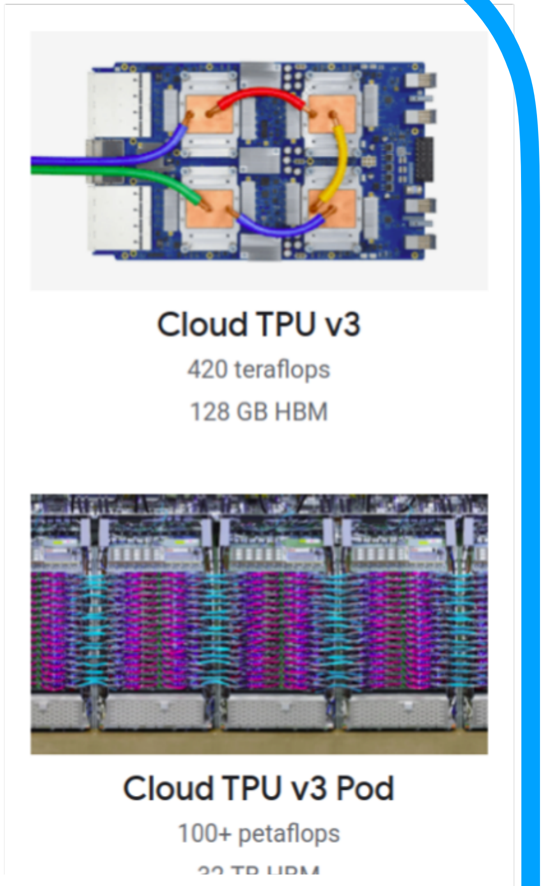

Transformer
(encoder-only)
with lots of parameters

+



Lots of texts

+




Lots of compute

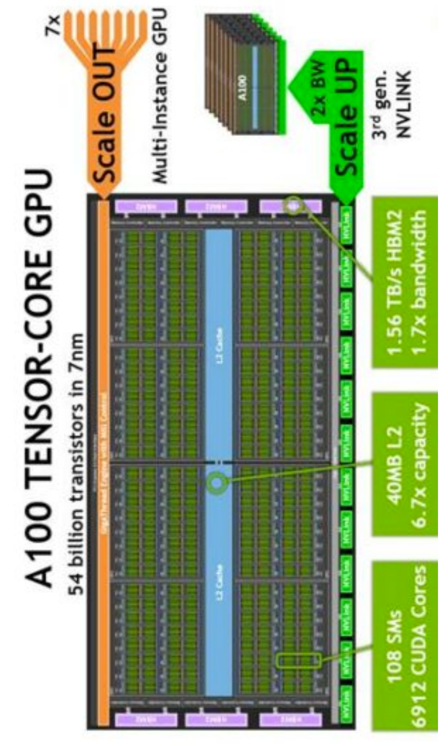
**BERT Pre-Training**

→

MS MARCO

Lots of query-passage pairs

+



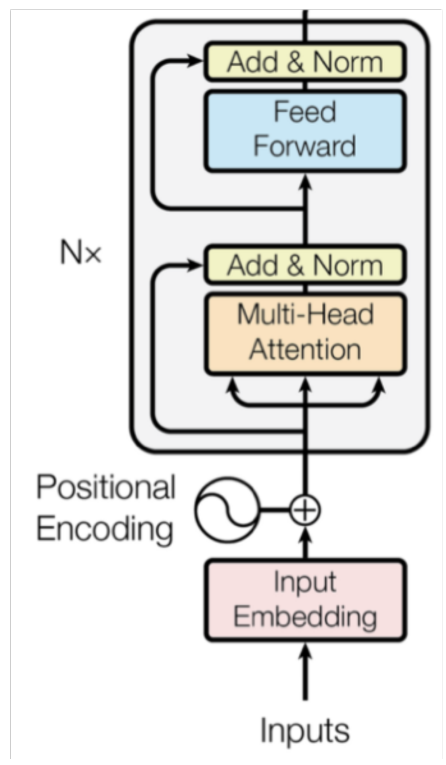

"small" GPU compute

**BERT Fine-Tuning**

**Need lots of queries with relevance assessments**

MS MARCO train: >532K query-passage pairs
(+>12K for validation)

# BERT's Challenges: (4) training data

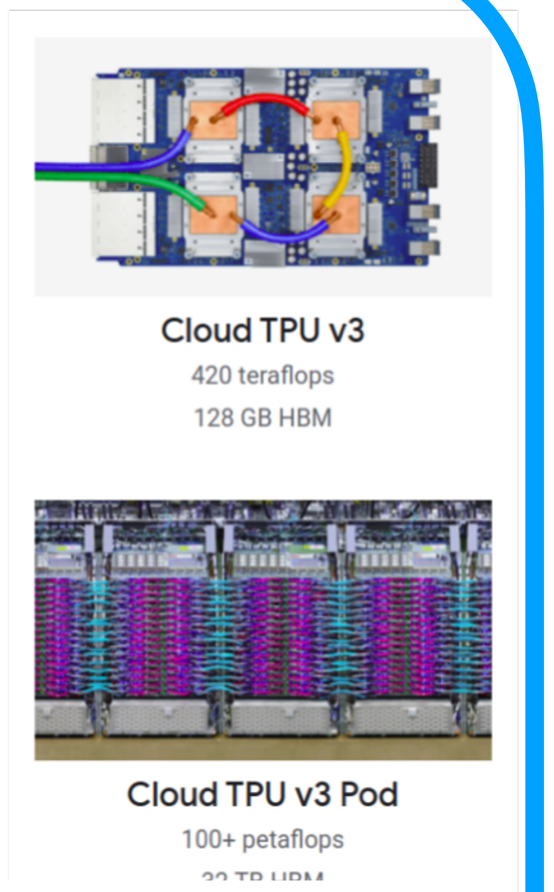



Transformer (encoder-only) with lots of parameters

+

Lots of texts

+



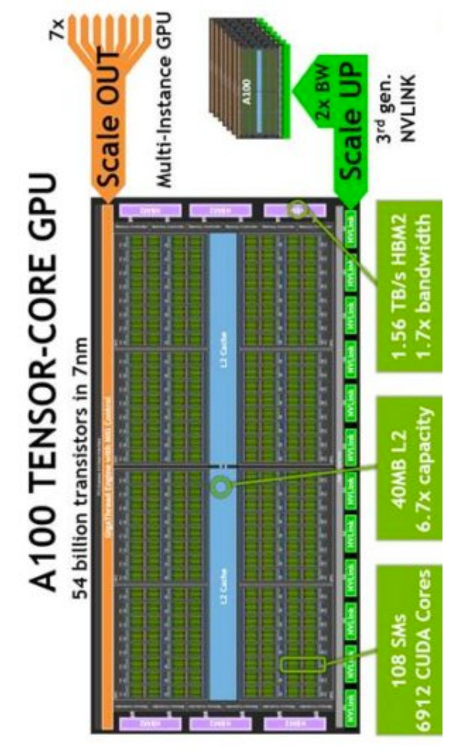

Lots of compute

**BERT Pre-Training**

MS MARCO

Lots of query-passage pairs

+




"small" GPU compute

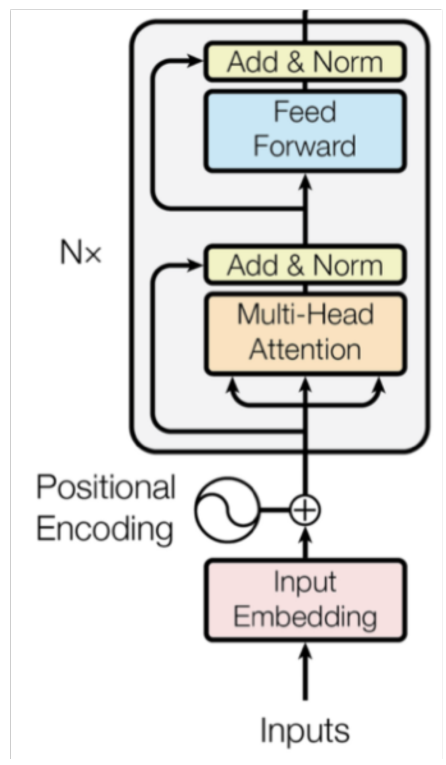
**BERT Fine-Tuning**

**Need lots of queries with relevance assessments**

MS MARCO train: >532K query-passage pairs (+>12K for validation)

How can we fine-tune these models in data-poor regimes (i.e. domain-specific IR)?

# BERT's Challenges: (5) out-of-distribution data

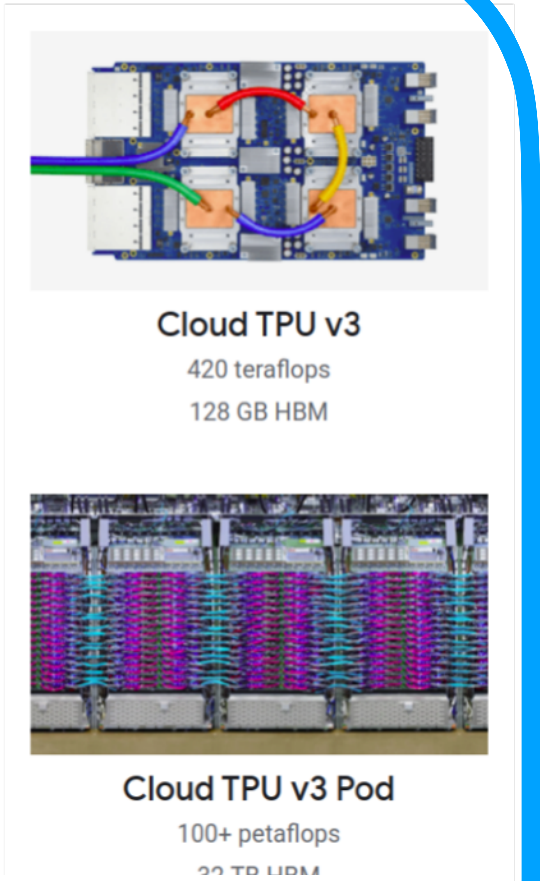



Transformer (encoder-only) with lots of parameters

+

Lots of texts

+




Lots of compute

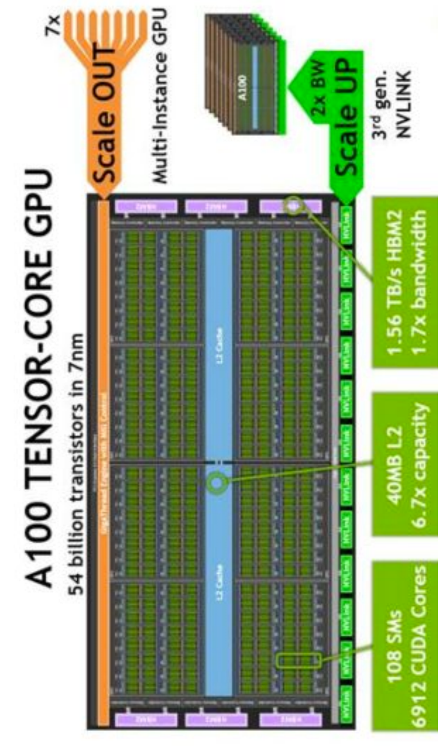
**BERT Pre-Training**

→

MS MARCO

Lots of query-passage pairs

+




"small" GPU compute

**BERT Fine-Tuning**

**Queries are quite particular in nature**

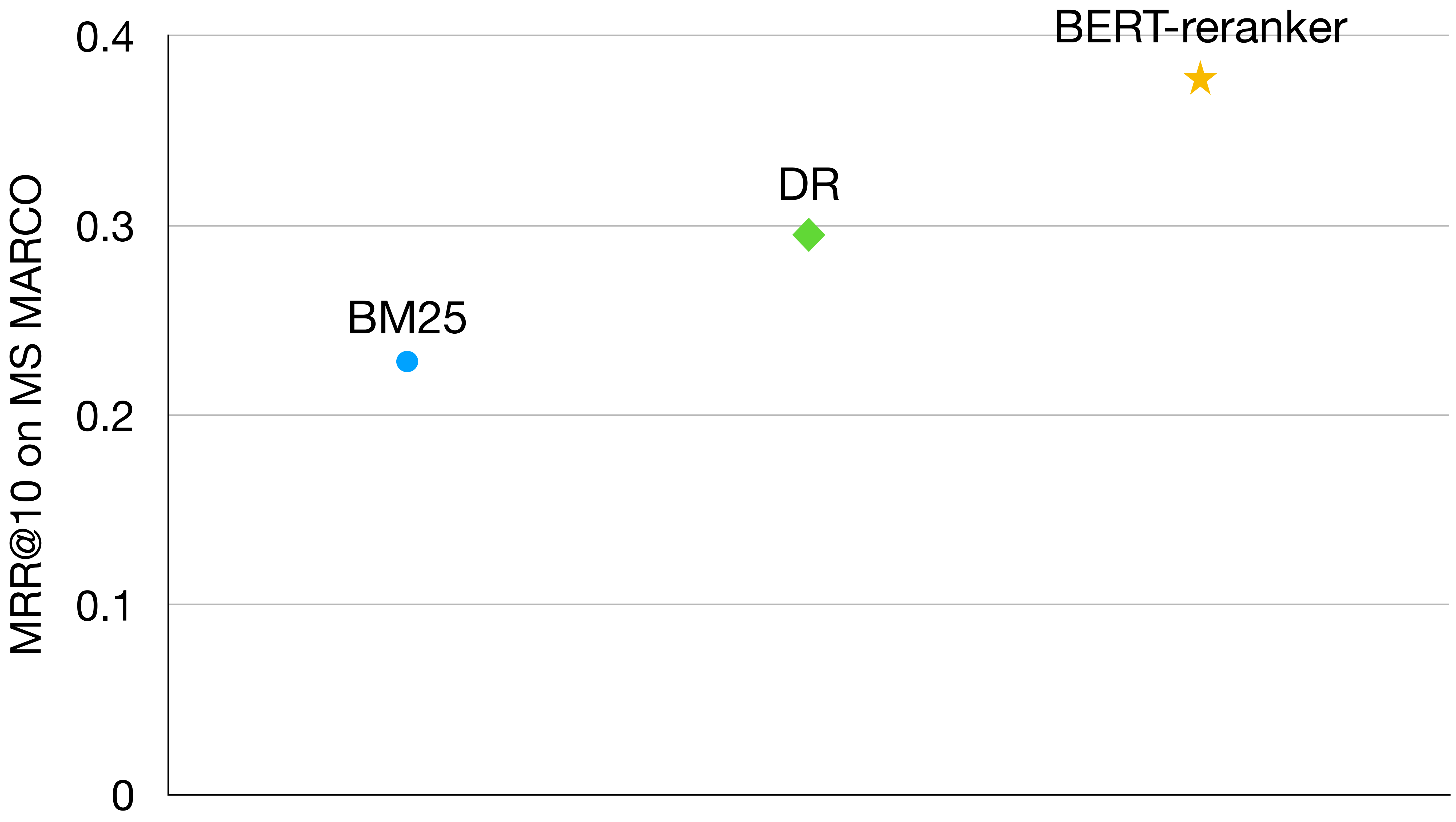
* focused on Q/A
* underwent some light curation (e.g. only few typos)

←

How do these models behave with out-of-distribution data?

The University Of Queensland
AUSTRALIA
CREATE CHANGE

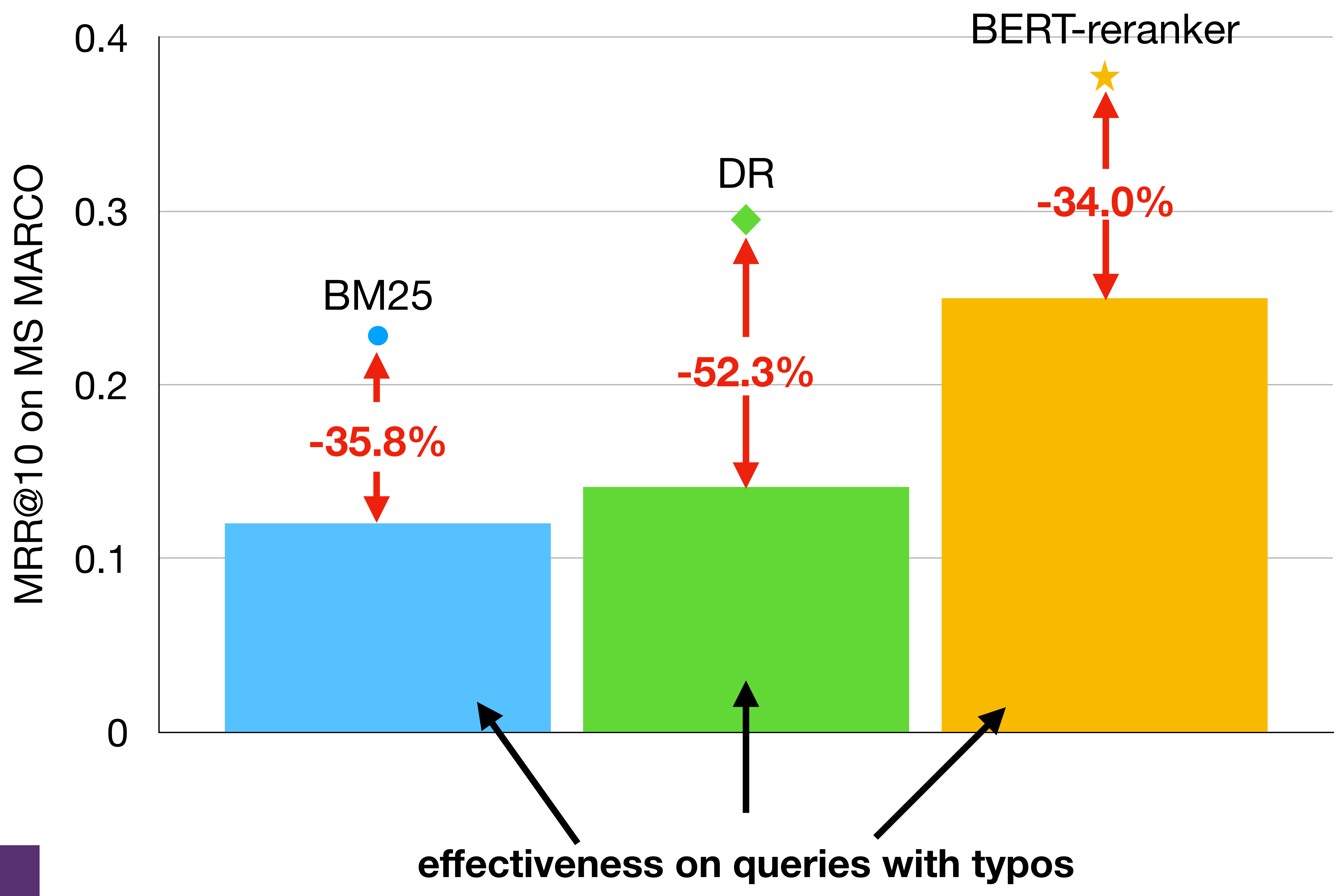
# BERT's Challenges: (5) out-of-distribution data




Effectiveness on MS MARCO queries (in distribution)

What if queries are manipulated to insert typos (out-of-distribution)?

# BERT's Challenges: (5) out-of-distribution data




**Both DR and BERT-reranker perform poorly on queries with typos**


Zhuang, Zuccon, *"Dealing with Typos for BERT-based Passage Retrieval and Ranking"*, EMNLP 2022

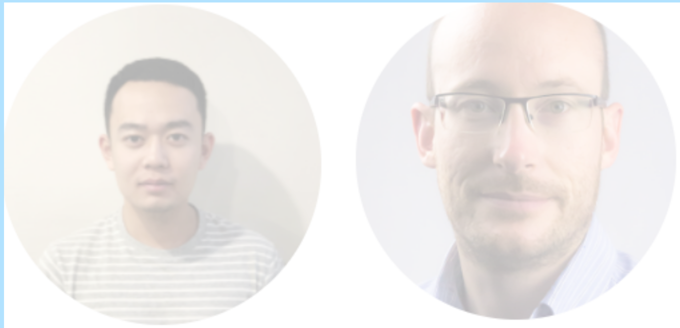
Zhuang, Zuccon, *"CharacterBERT and Self-Teaching for Improving the Robustness of Dense Retrievers on Queries with Typos"*, SIGIR 2022

ie lab

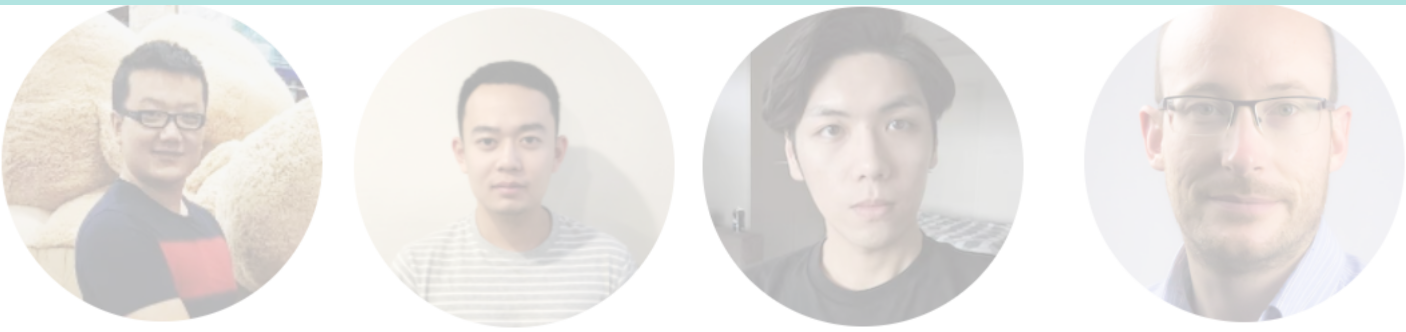
# Neural IR @ ielab

**Trade-off between effectiveness, efficiency and hardware**

**Robustness to out-of-distribution data**



**PRF integration, efficient PRF**



- **Aim: Devise methods that are effective and yet efficient**
- **Attention to methods that can be run on commodity hardware and embedded systems**

➡ Quantification of trade-offs (SIGR'22)

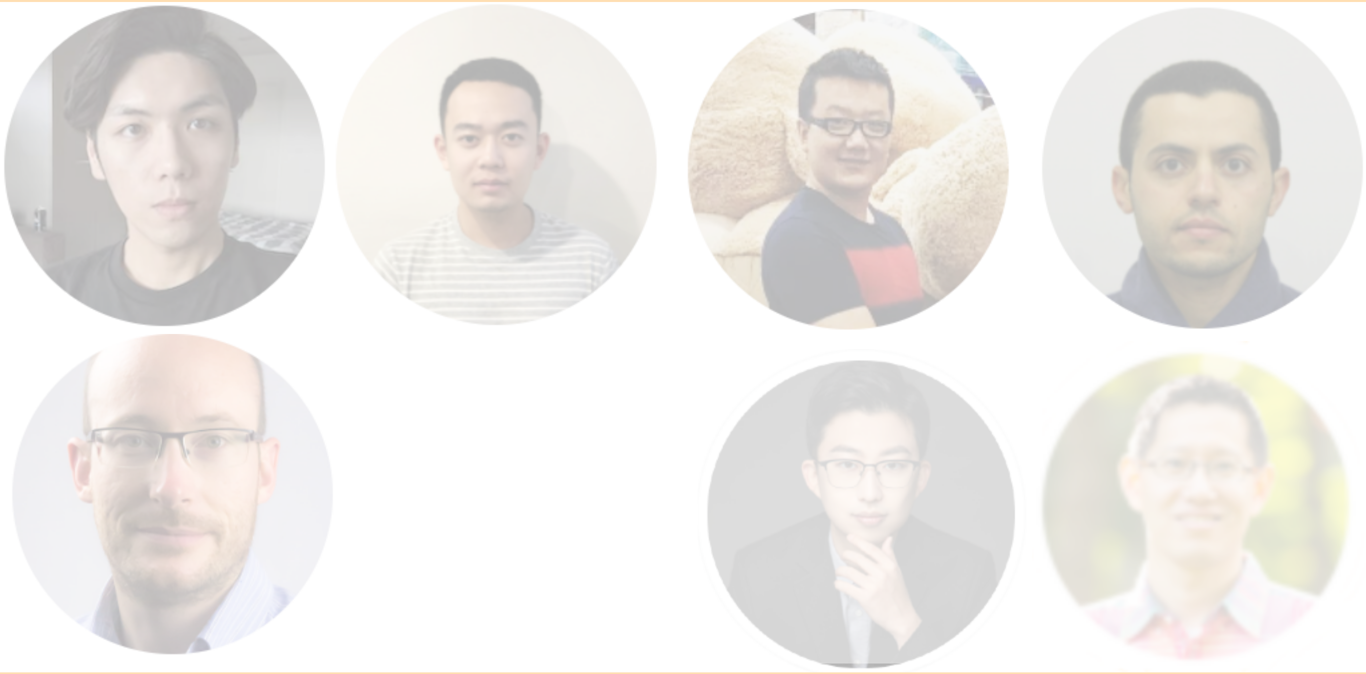
➡ Deep QLM (ECIR'21)

➡ TILDE (SIGIR'21) & TILDEv2

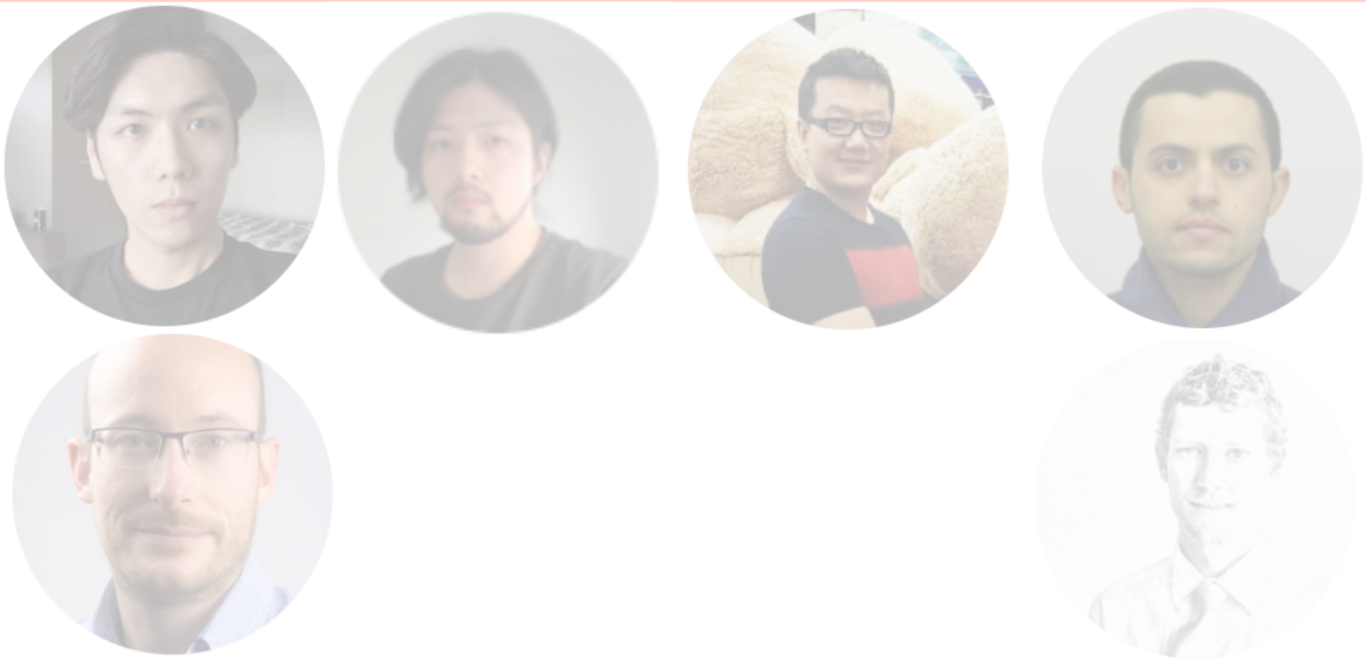
➡ Transformer PRF: An efficient learnable PRF method for embedded systems

**Data & Training efficiency**

**Hybrid sparse-dense methods**

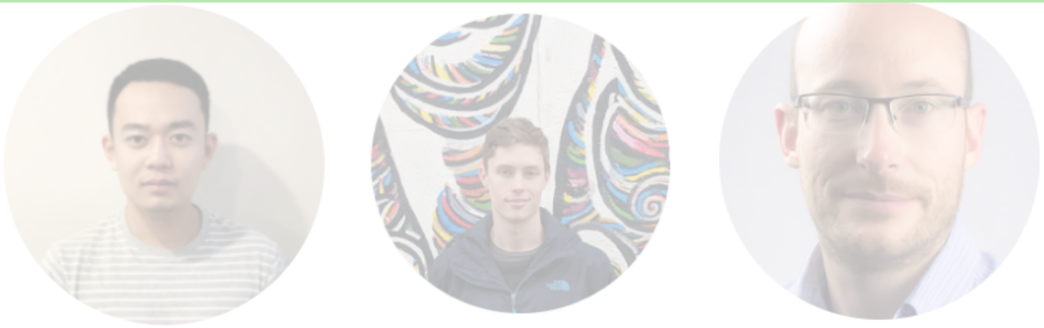


**BERT models in domain-specific tasks**

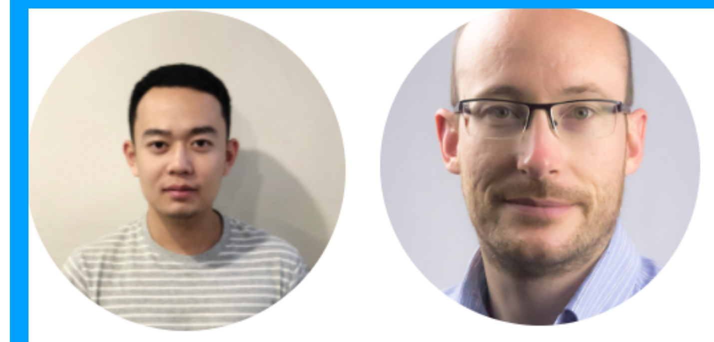


# Neural IR @ ielab

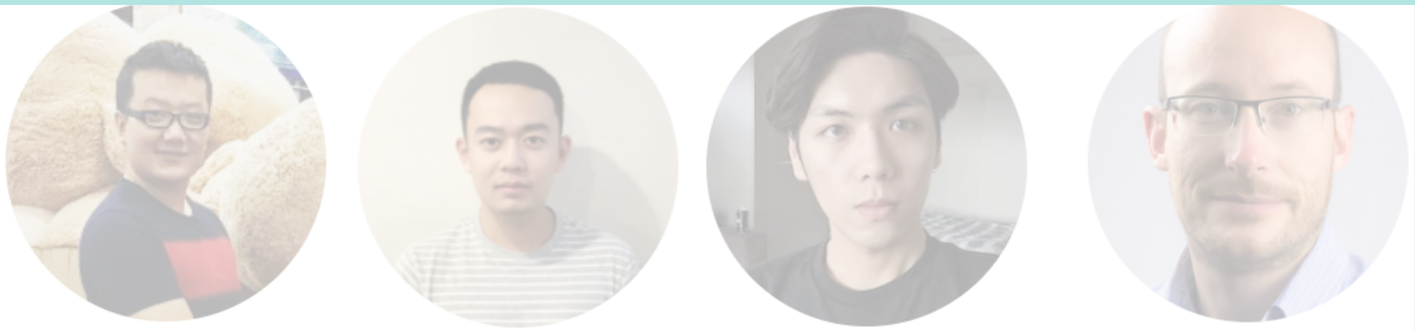
### Trade-off between effectiveness, efficiency and hardware



### Robustness to out-of-distribution data



### PRF integration, efficient PRF



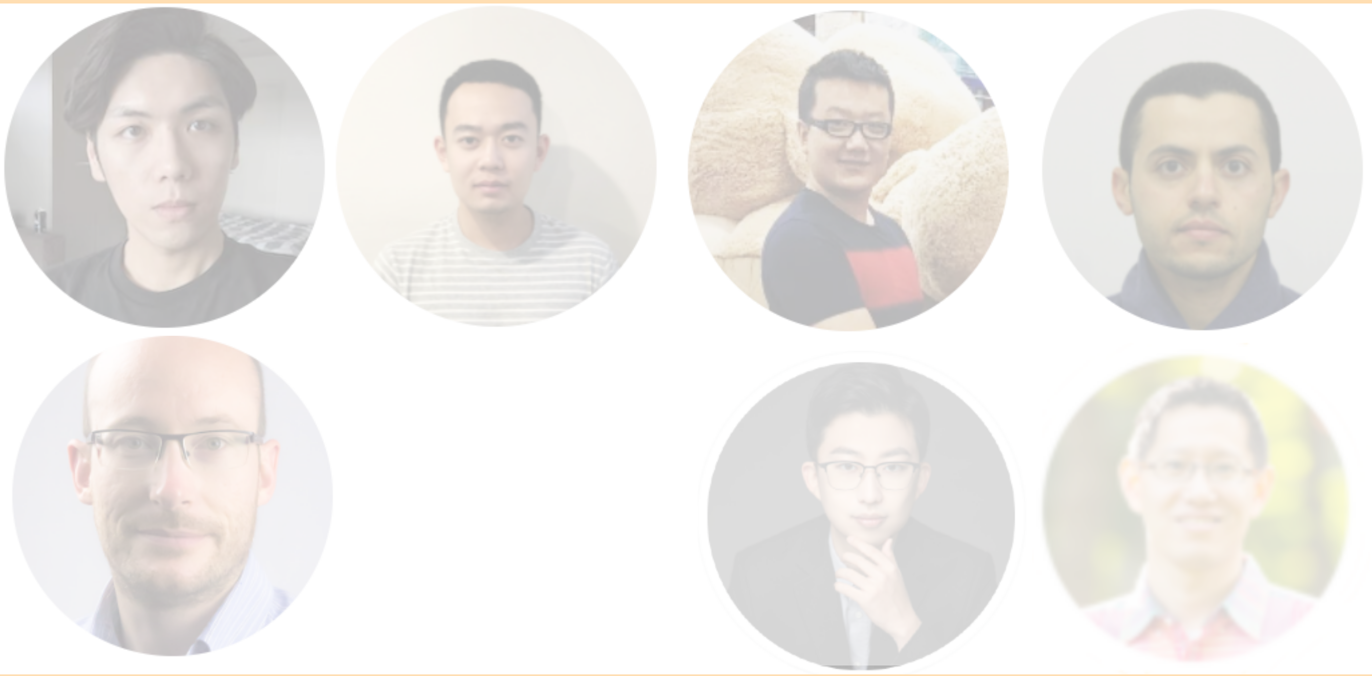
- **Aim: Training methods robust to typos in queries**

➡ Use typos to train (EMNLP’21)

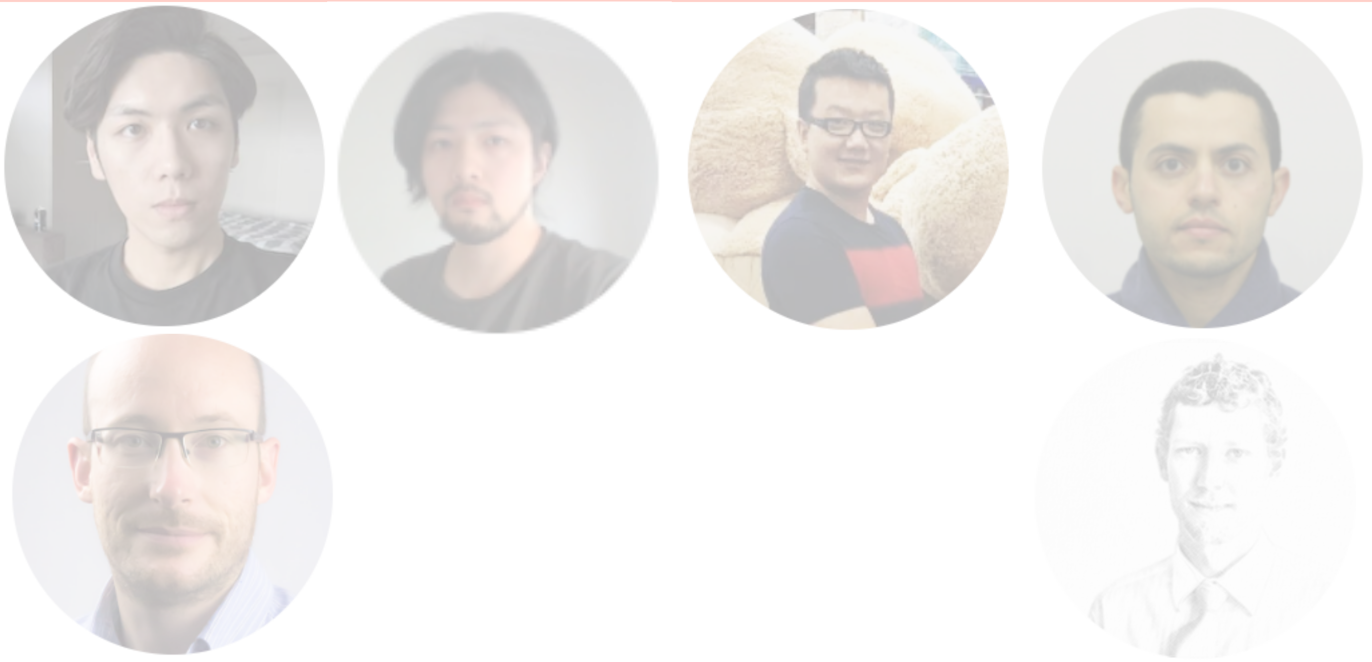
➡ Character encoder and new self-teaching (SIGIR’22)

### Data & Training efficiency

### Hybrid sparse-dense methods

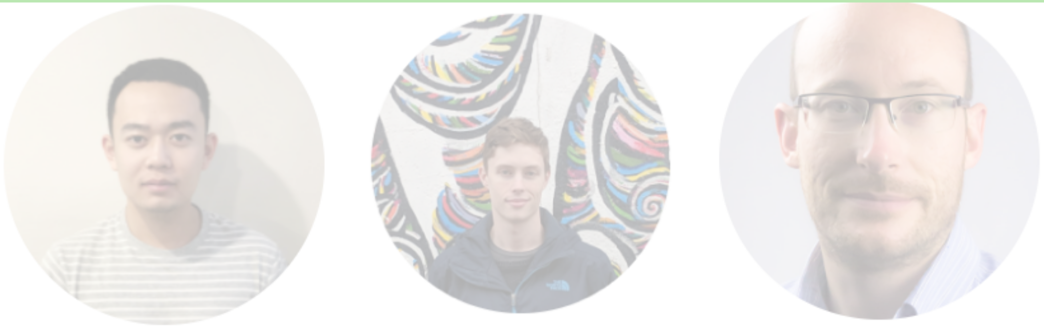


### BERT models in domain-specific tasks

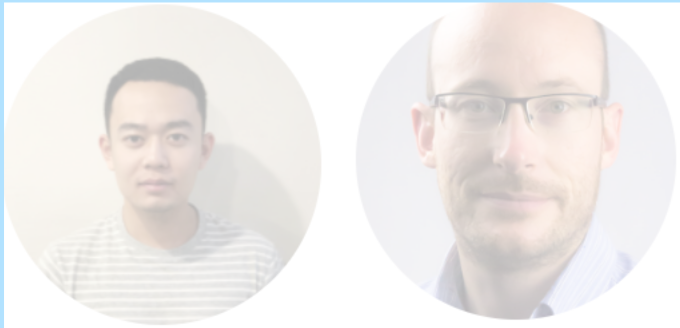


# Neural IR @ ielab

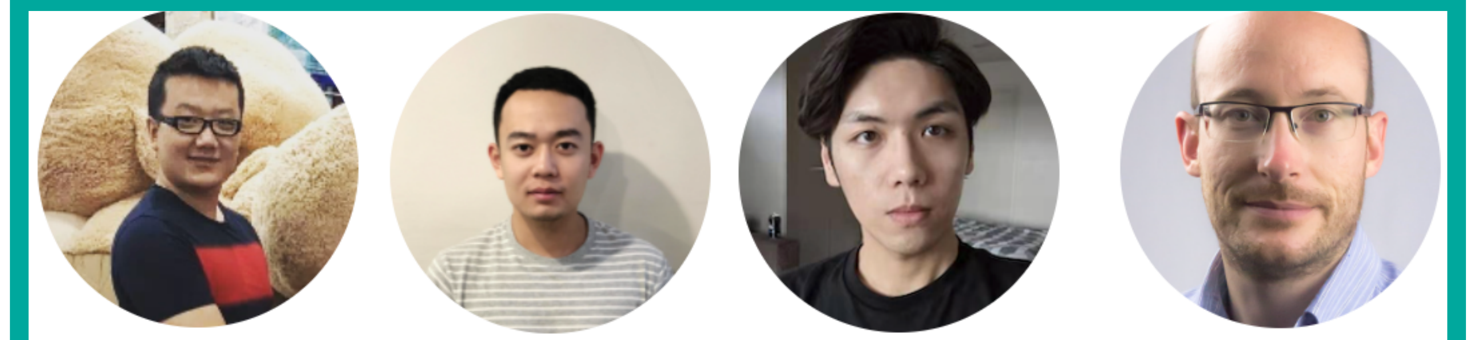
**Trade-off between effectiveness, efficiency and hardware**



**Robustness to out-of-distribution data**



**PRF integration, efficient PRF**



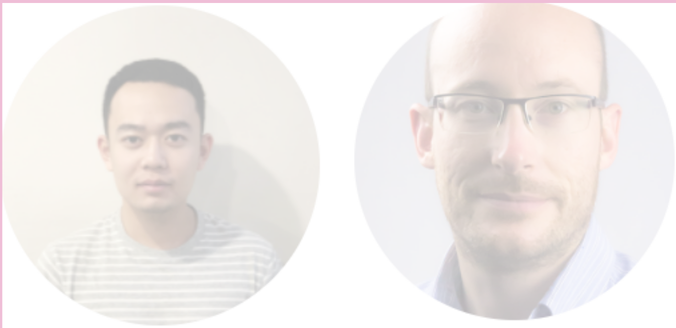
- **Aim: Integrate efficient PRF into the Neural IR pipeline**
- **Deal with inputs larger than BERT limits**

➡ Extension of a trainable PRF method to any dense retriever (ECIR'21)

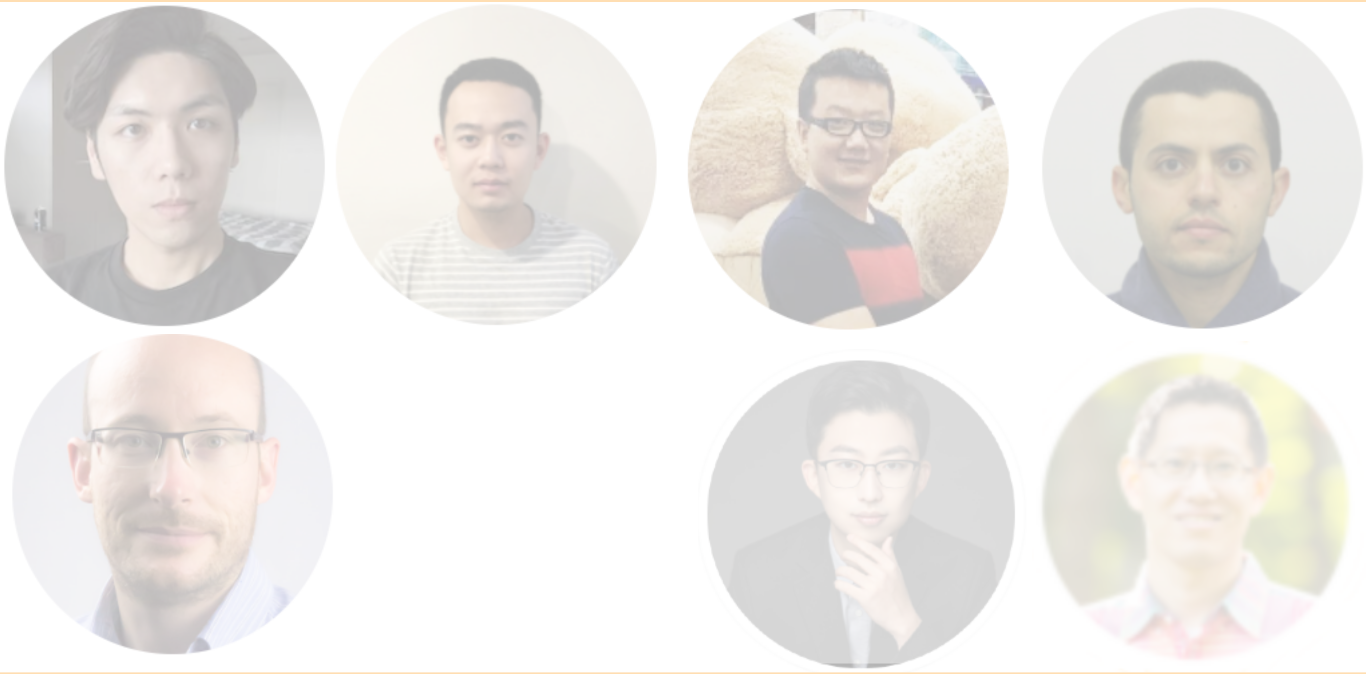
➡ Scalable, efficient and effective vector-based PRF method (TOIS'22)

➡ Transformer PRF: An efficient learnable PRF method for embedded systems

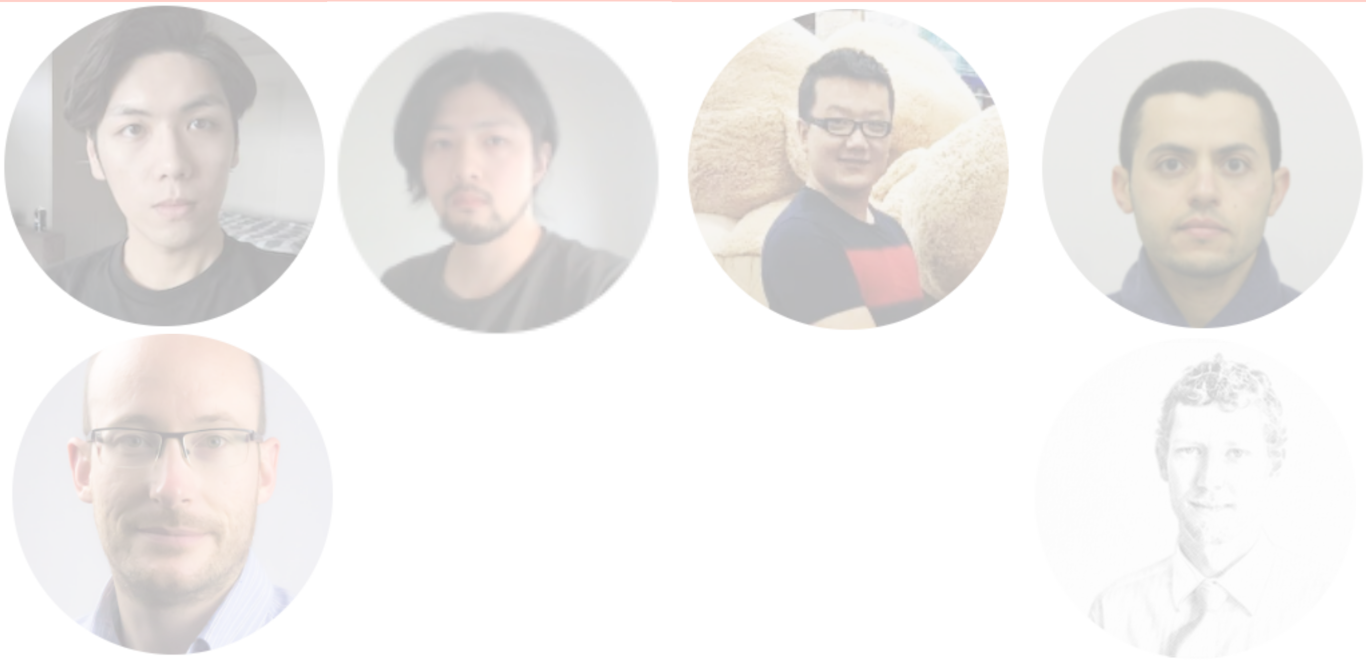
**Data & Training efficiency**



**Hybrid sparse-dense methods**

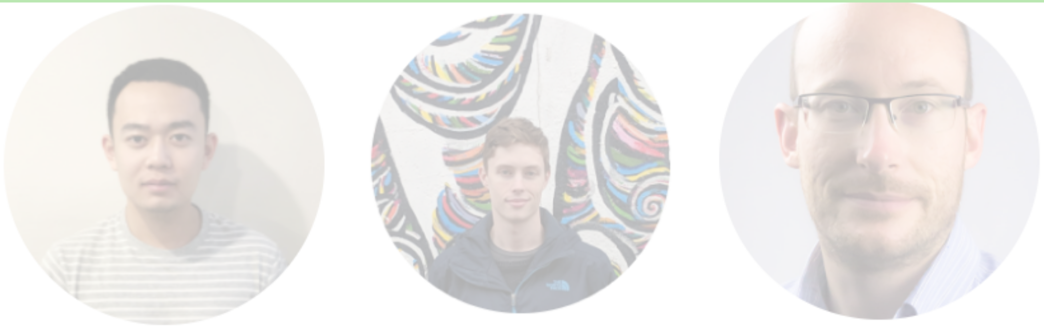


**BERT models in domain-specific tasks**

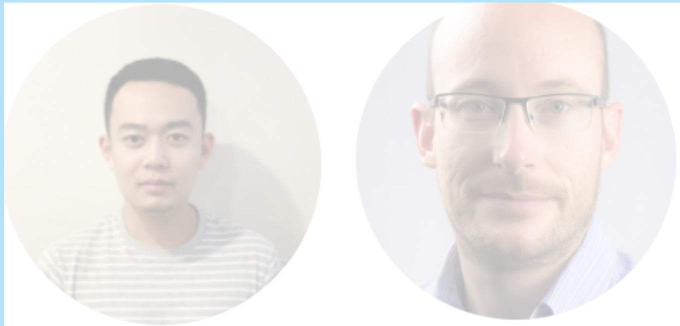


# Neural IR @ ielab

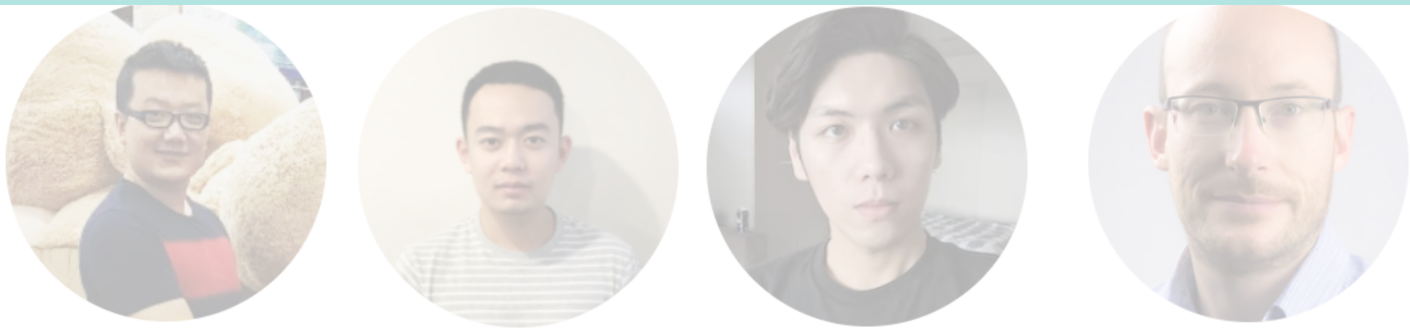
Trade-off between effectiveness, efficiency and hardware



Robustness to out-of-distribution data



PRF integration, efficient PRF



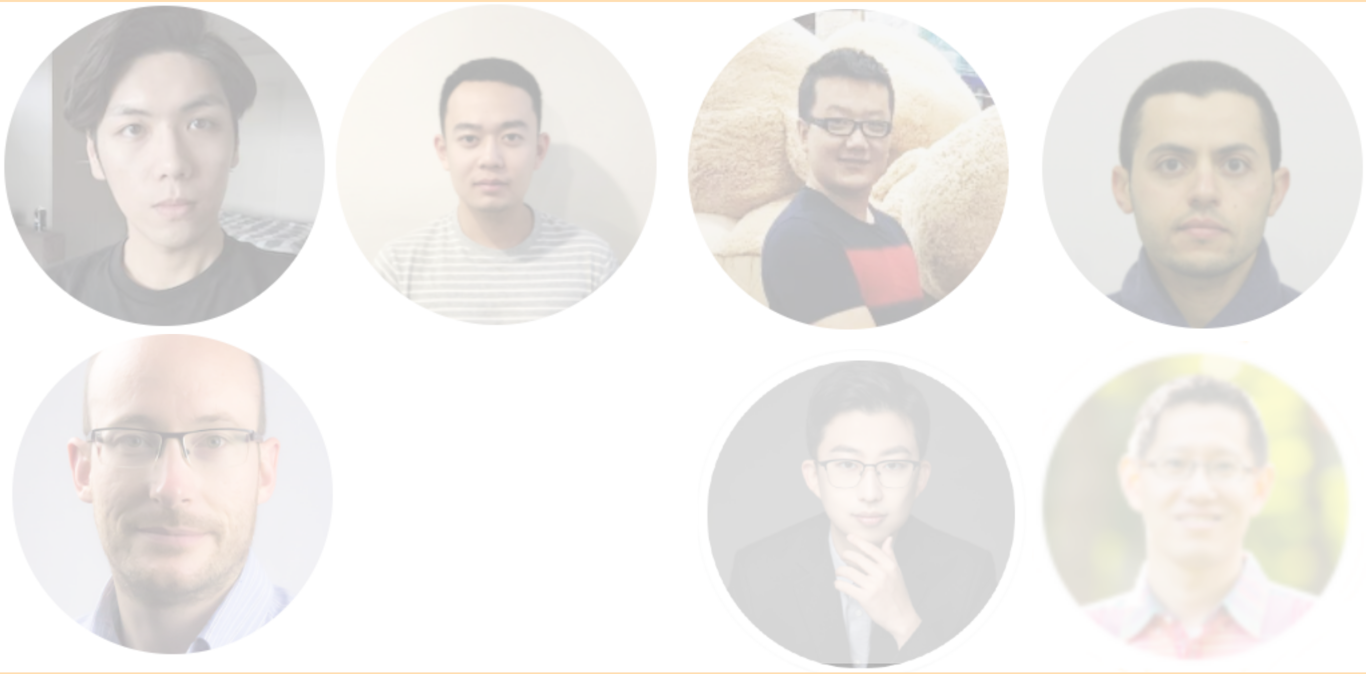
- **Aim: Investigate data-poor regimes**

➡ Exploit signals from implicit user interactions to train Dense Retrievers (SIGIR’22)

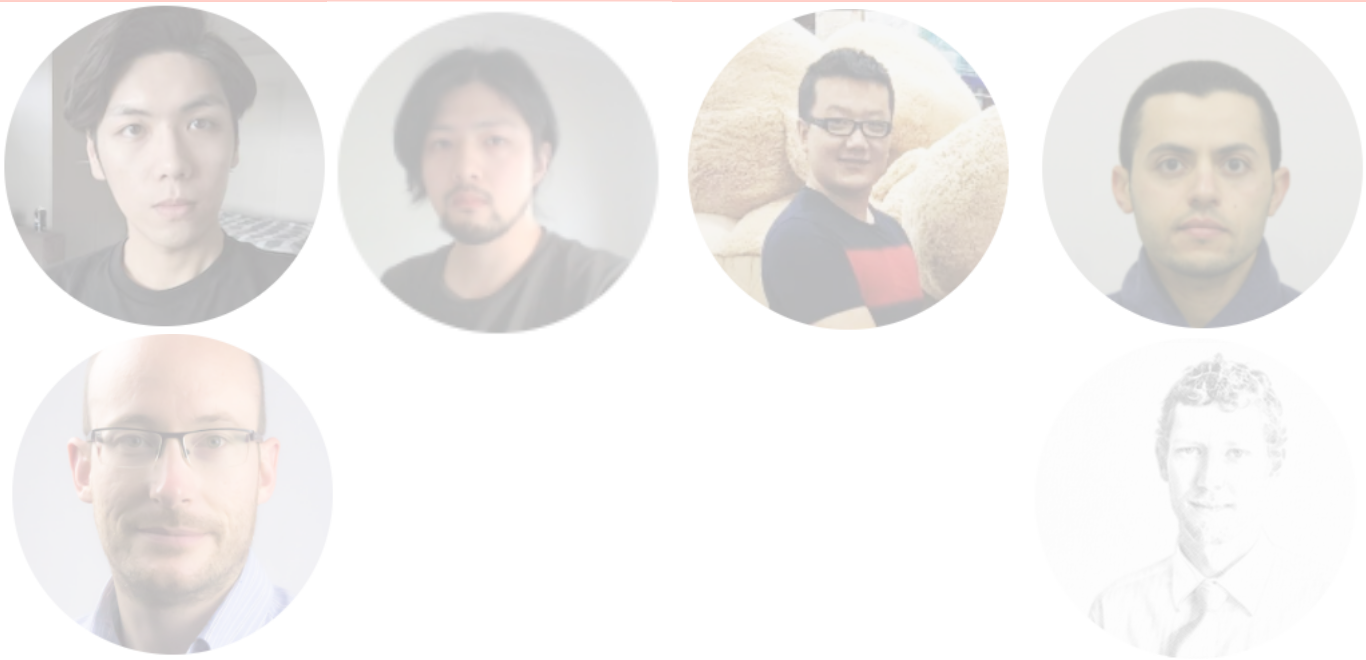
➡ Reduce validation time, compute & data when training Dense Retrievers (SIGIR’22)

Data & Training efficiency

Hybrid sparse-dense methods

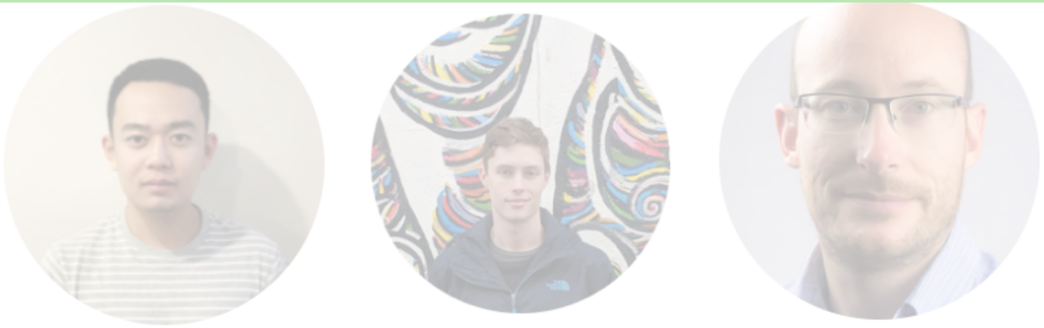


BERT models in domain-specific tasks

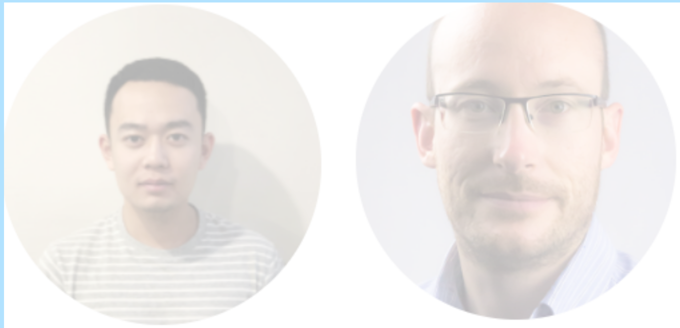


# Neural IR @ ielab

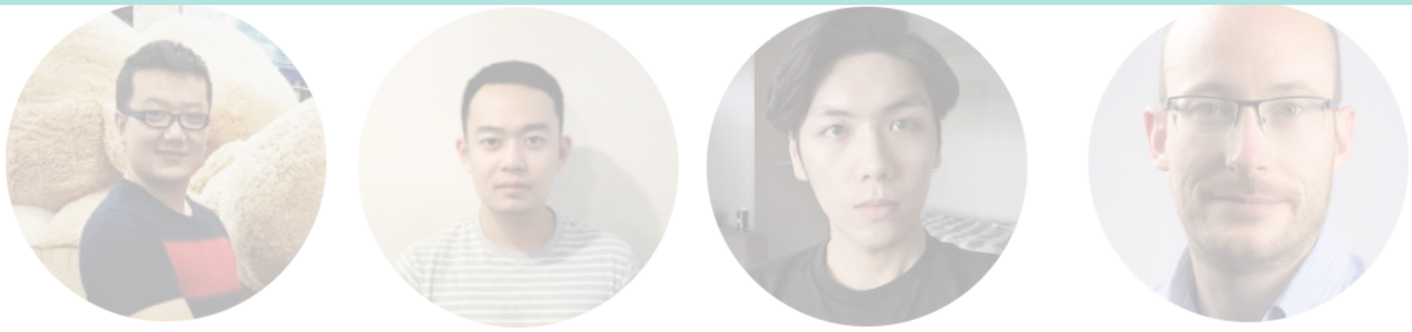
**Trade-off between effectiveness, efficiency and hardware**



**Robustness to out-of-distribution data**



**PRF integration, efficient PRF**

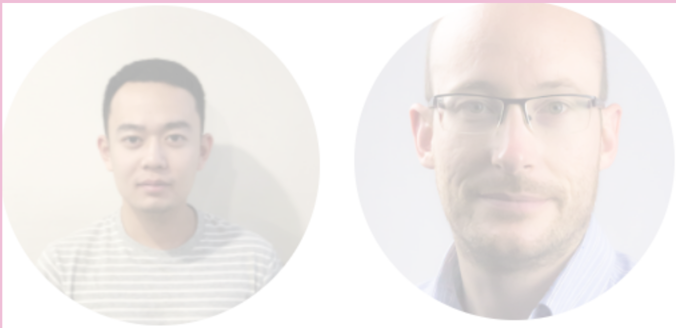


- **Aim: Investigate combination of dense and sparse architectures: best-of-breed**

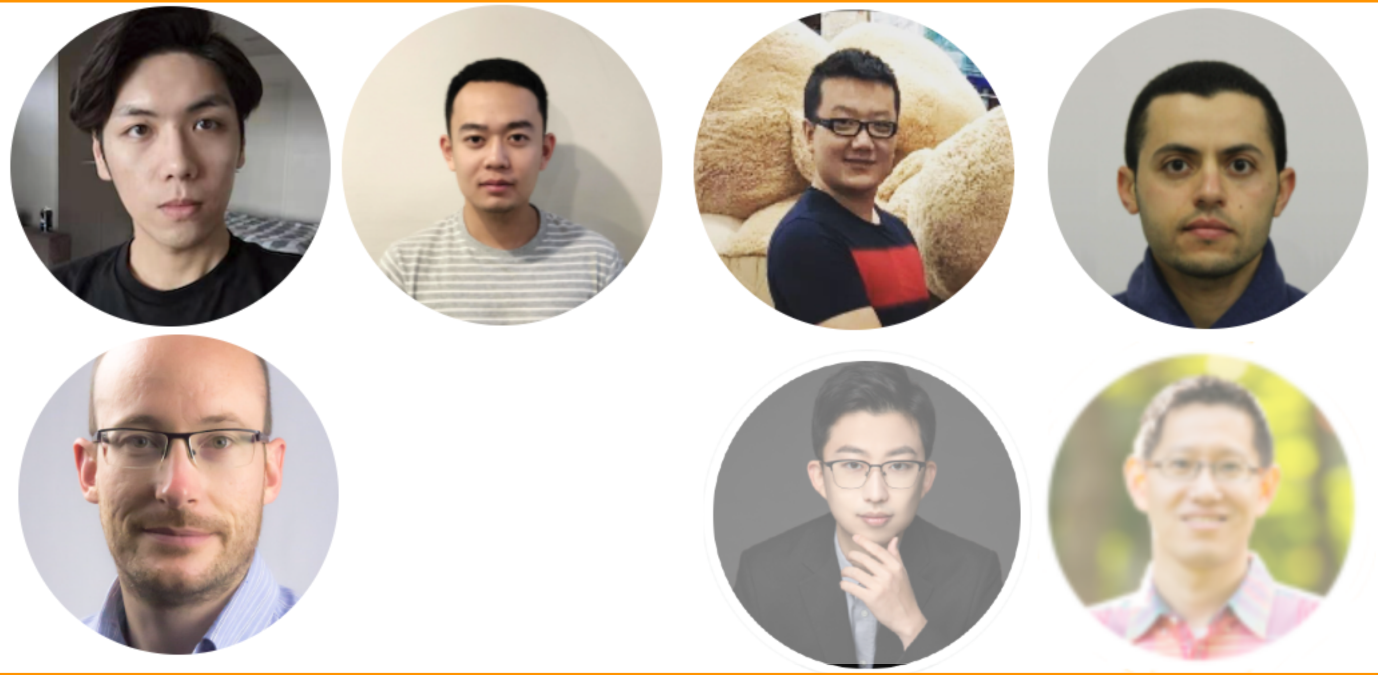
➡ Dense Retrievers benefit from hybrid architecture (ICTIR'21)

➡ PRF methods benefit from hybrid dense-sparse architectures (SIGIR'22)

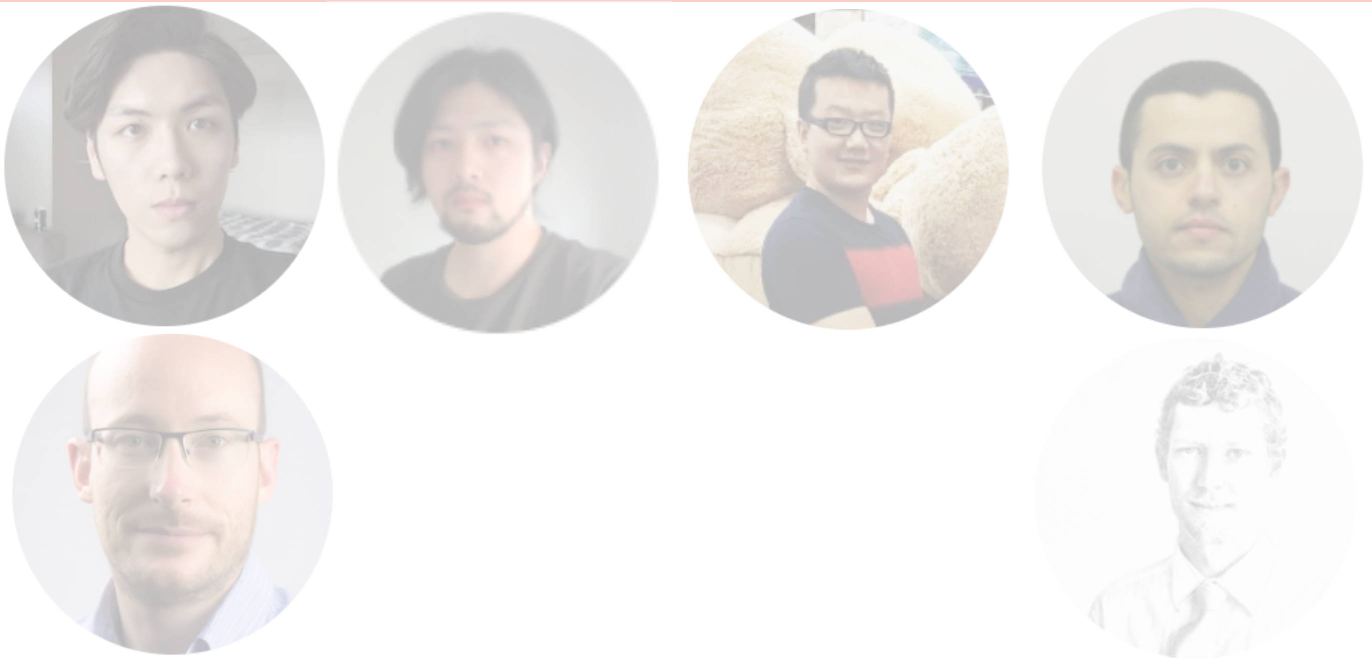
**Data & Training efficiency**



**Hybrid sparse-dense methods**

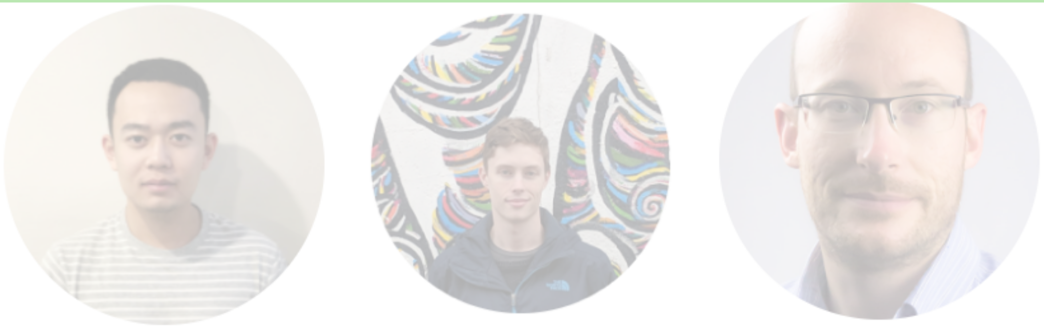


**BERT models in domain-specific tasks**

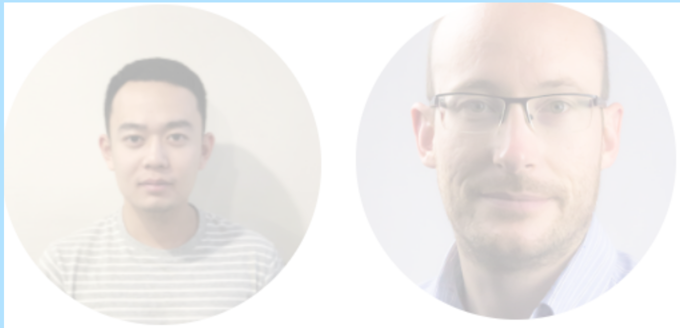


# Neural IR @ ielab

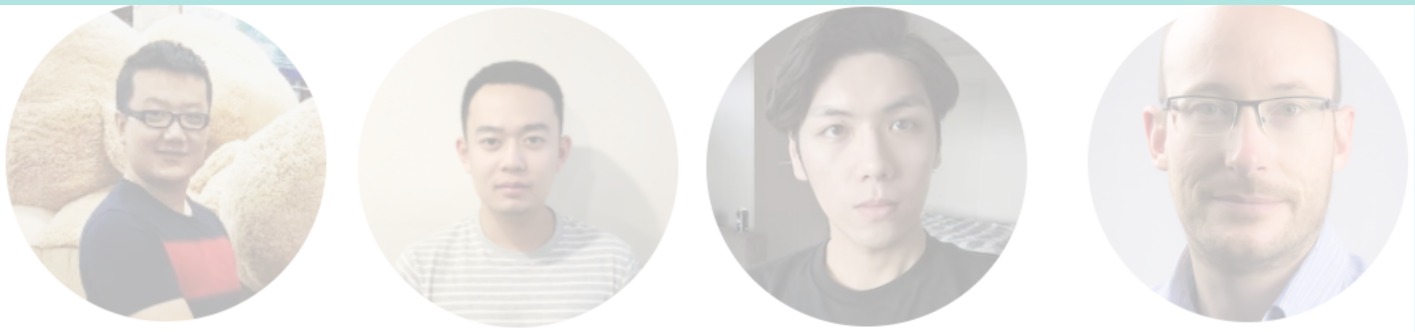
### Trade-off between effectiveness, efficiency and hardware



### Robustness to out-of-distribution data



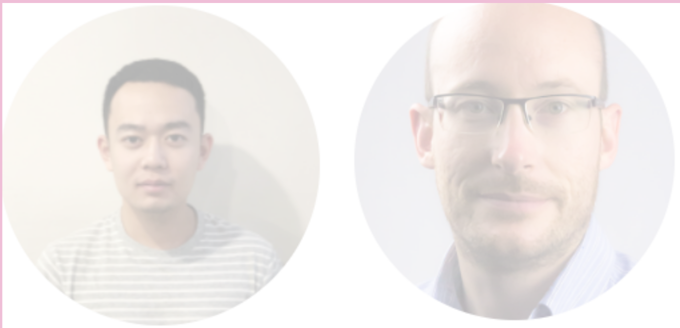
### PRF integration, efficient PRF



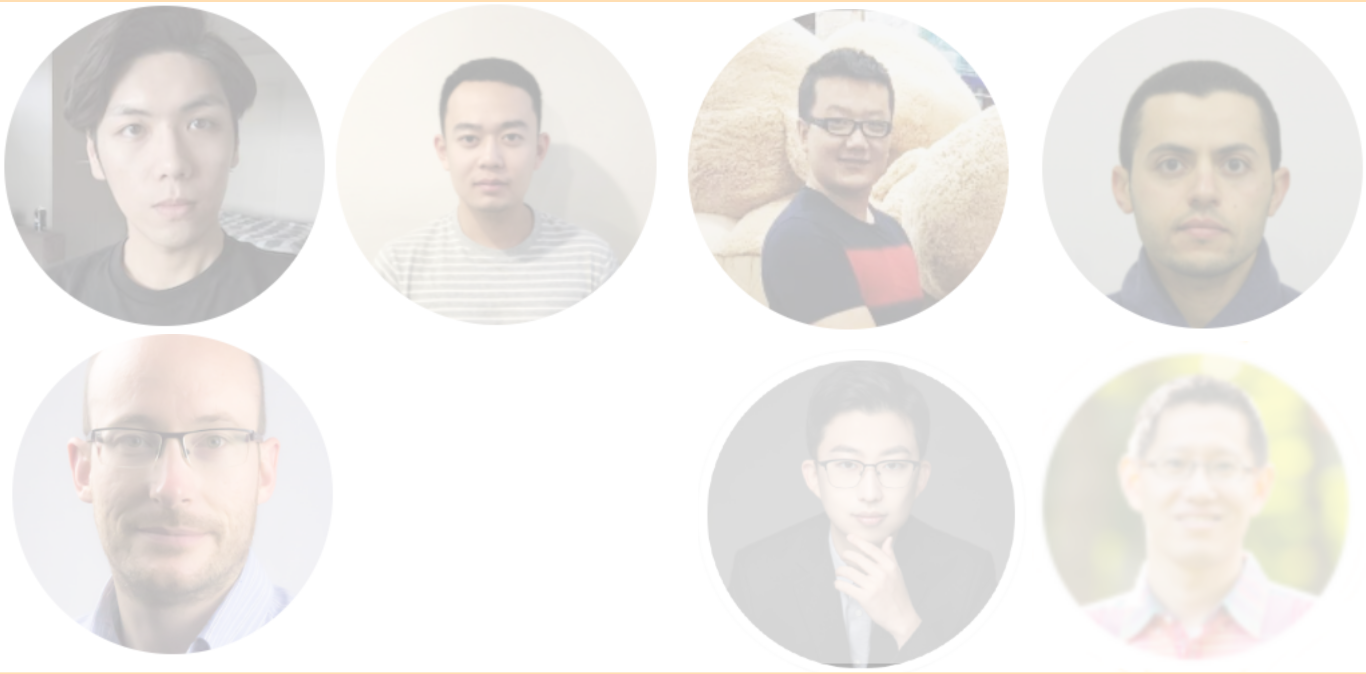
- **Aim: Best practice when using neural rankers in domain-specific search**
- Transfer of models / Transfer Learning
- Finetuning regimes
- Integration of domain-knowledge

➡ AgAsk conversational agent service

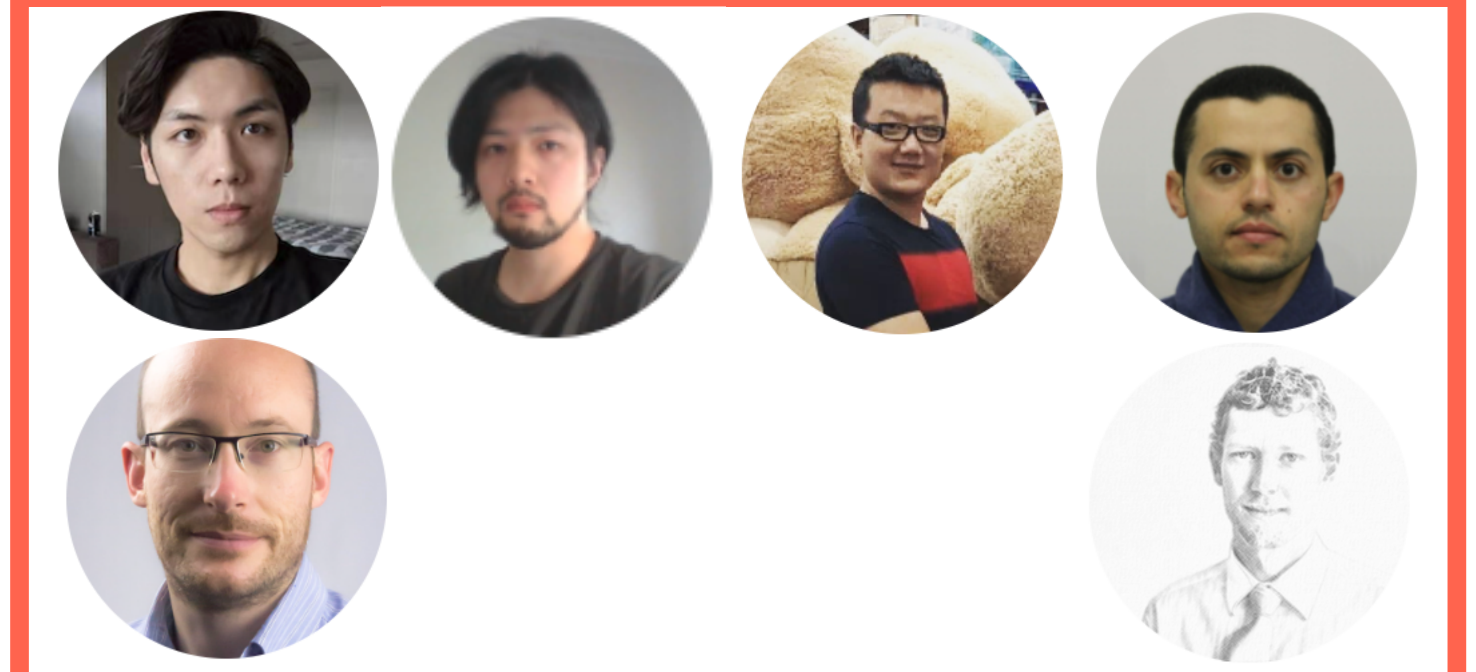
### Data & Training efficiency



### Hybrid sparse-dense methods



### BERT models in domain-specific tasks

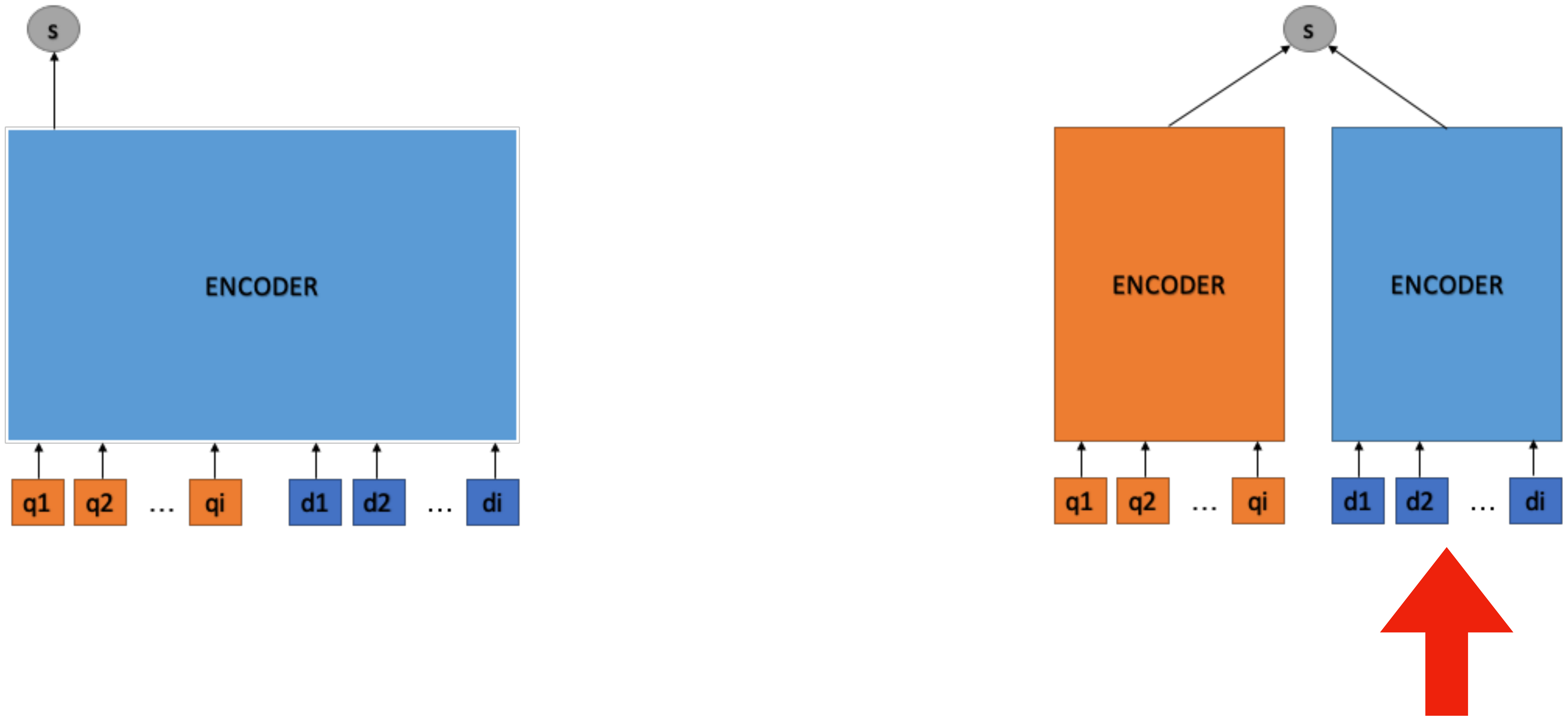


# Trade-off between effectiveness, efficiency and hardware



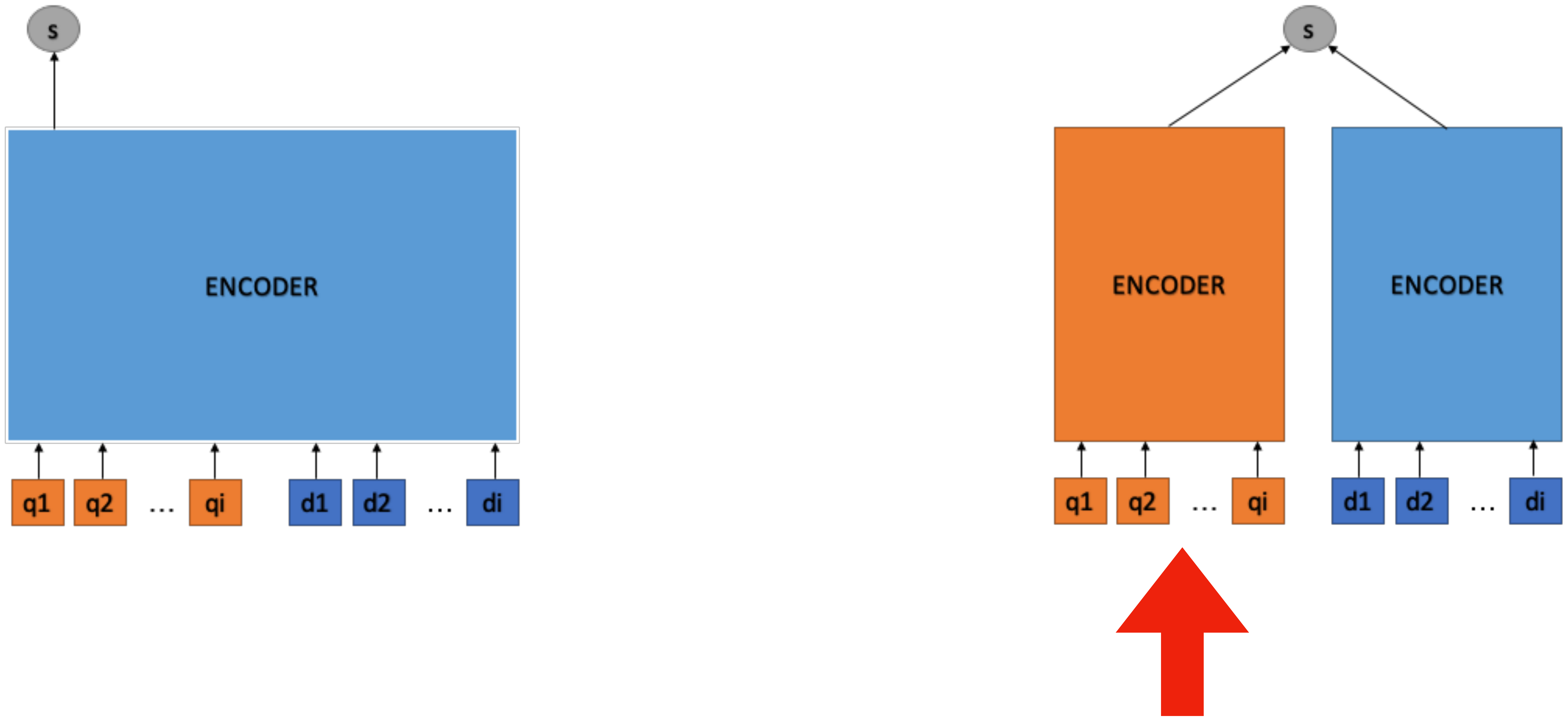
# Cross-Encoder, Bi-encoder, … can we simply further?

# Cross-Encoder, Bi-encoder, … can we simply further?



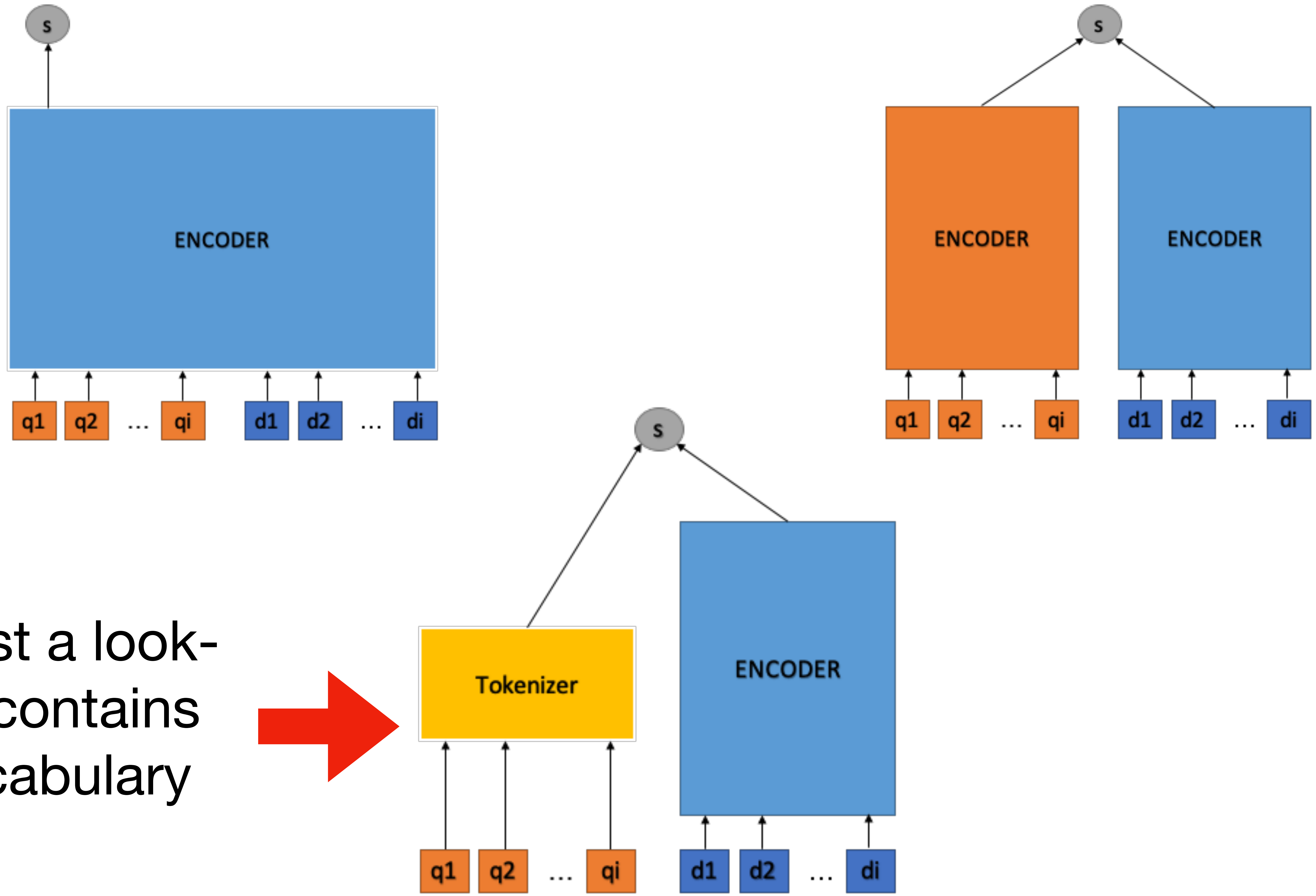

If wanting to reduce query latency, then no need to modify this

# Cross-Encoder, Bi-encoder, … can we simply further?




This influences query latency — can we make this faster?

# Cross-Encoder, Bi-encoder, … can we simply further?




Now, this is just a look-up table that contains the BERT vocabulary

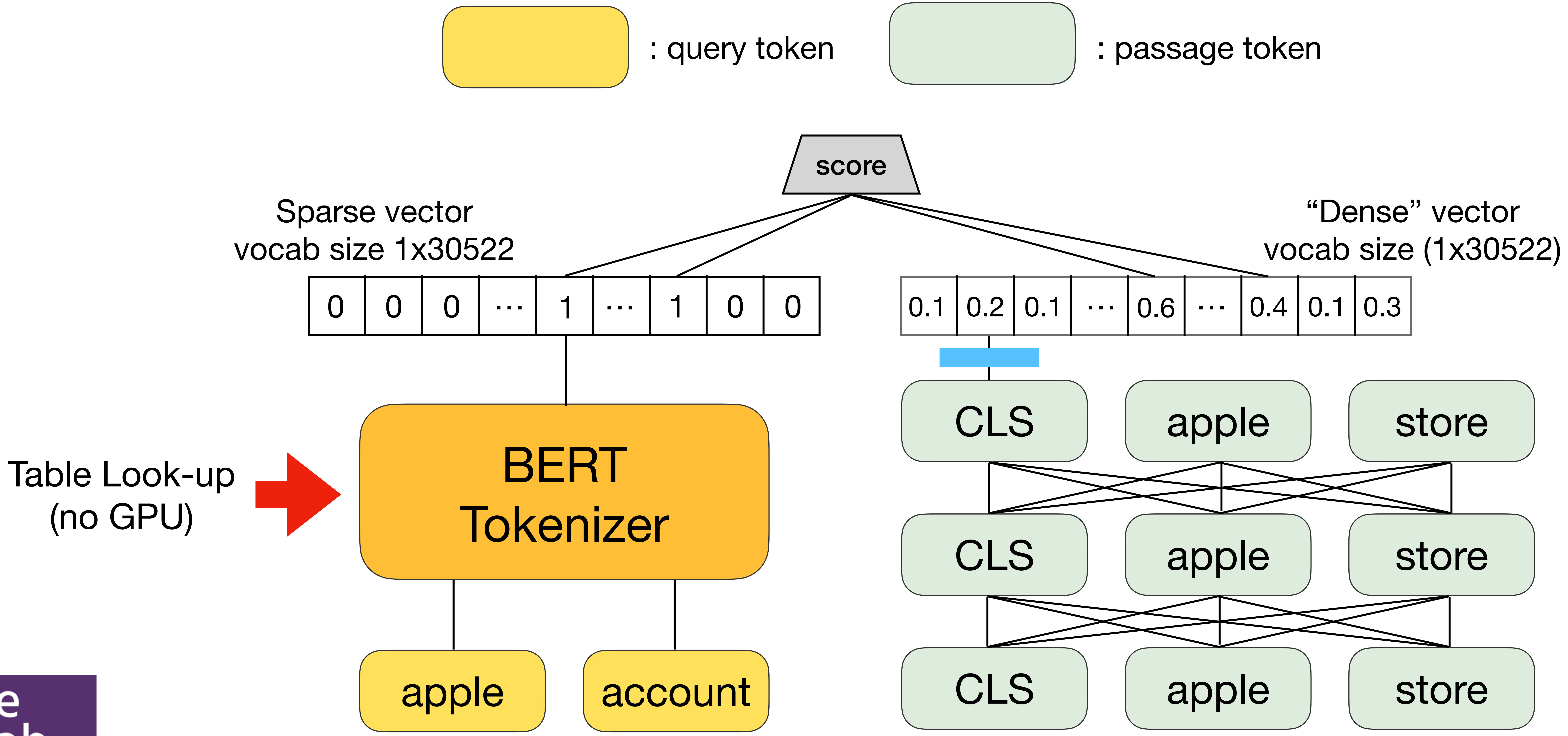
# TILDE: Term Independent Likelihood moDEl

- Use BERT tokeniser to obtain sparse query encoding
- Use CLS token to encode document
- Project CLS token embedding to |V| vector
- Inner product between sparse query vector and document vector, in the |V| space


Zhuang, Zuccon, *“TILDE: Term Independent Likelihood moDEl for passage re-ranking*",
SIGIR 2021

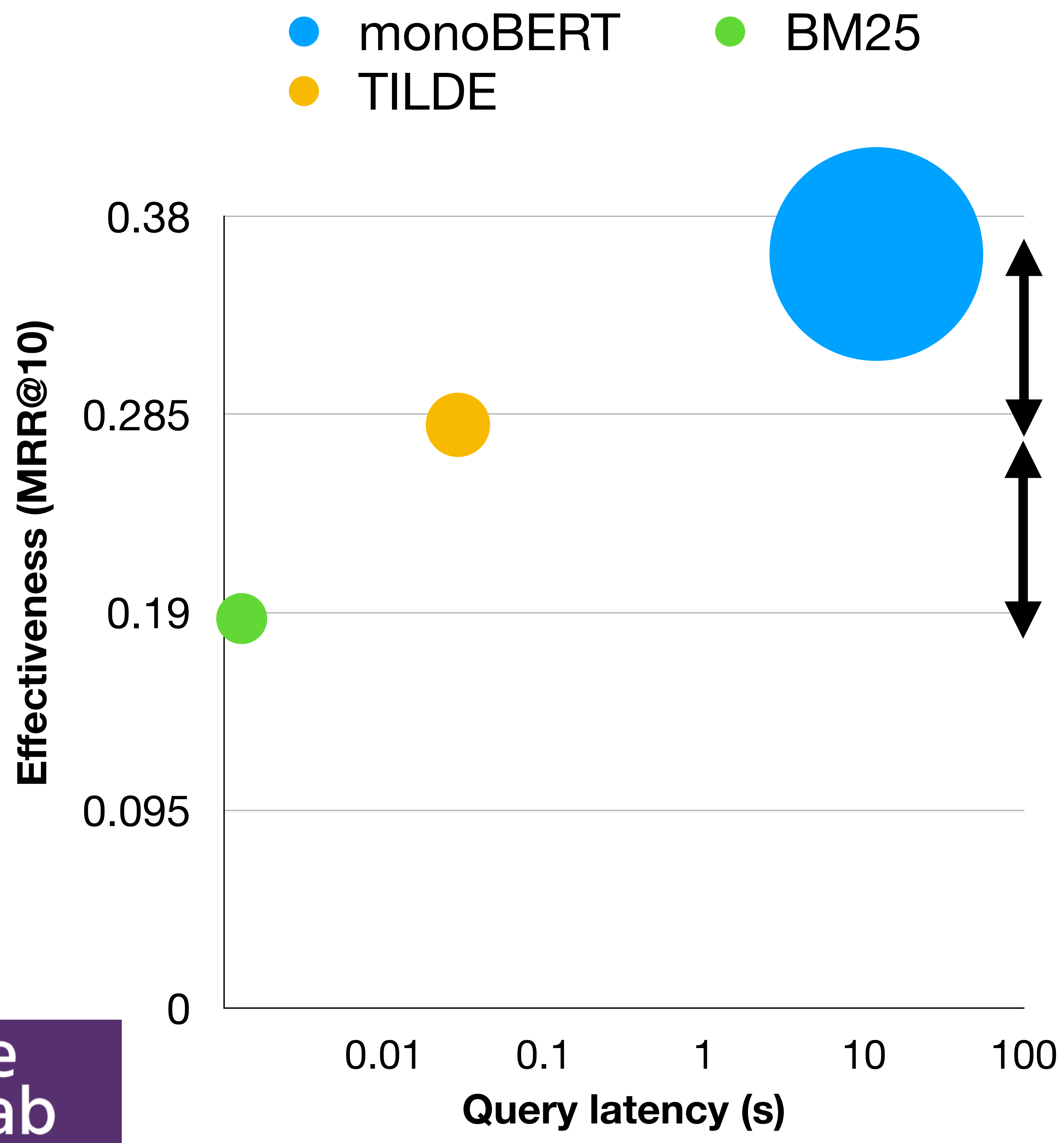
ie
lab

# TILDE



: query token
: passage token
score
Sparse vector
vocab size 1x30522
0 0 0 ··· 1 ··· 1 0 0
"Dense" vector
vocab size (1x30522)
0.1 0.2 0.1 ··· 0.6 ··· 0.4 0.1 0.3
Table Look-up
(no GPU)
BERT
Tokenizer
apple
account
CLS
apple
store
CLS
apple
store
CLS
apple
store

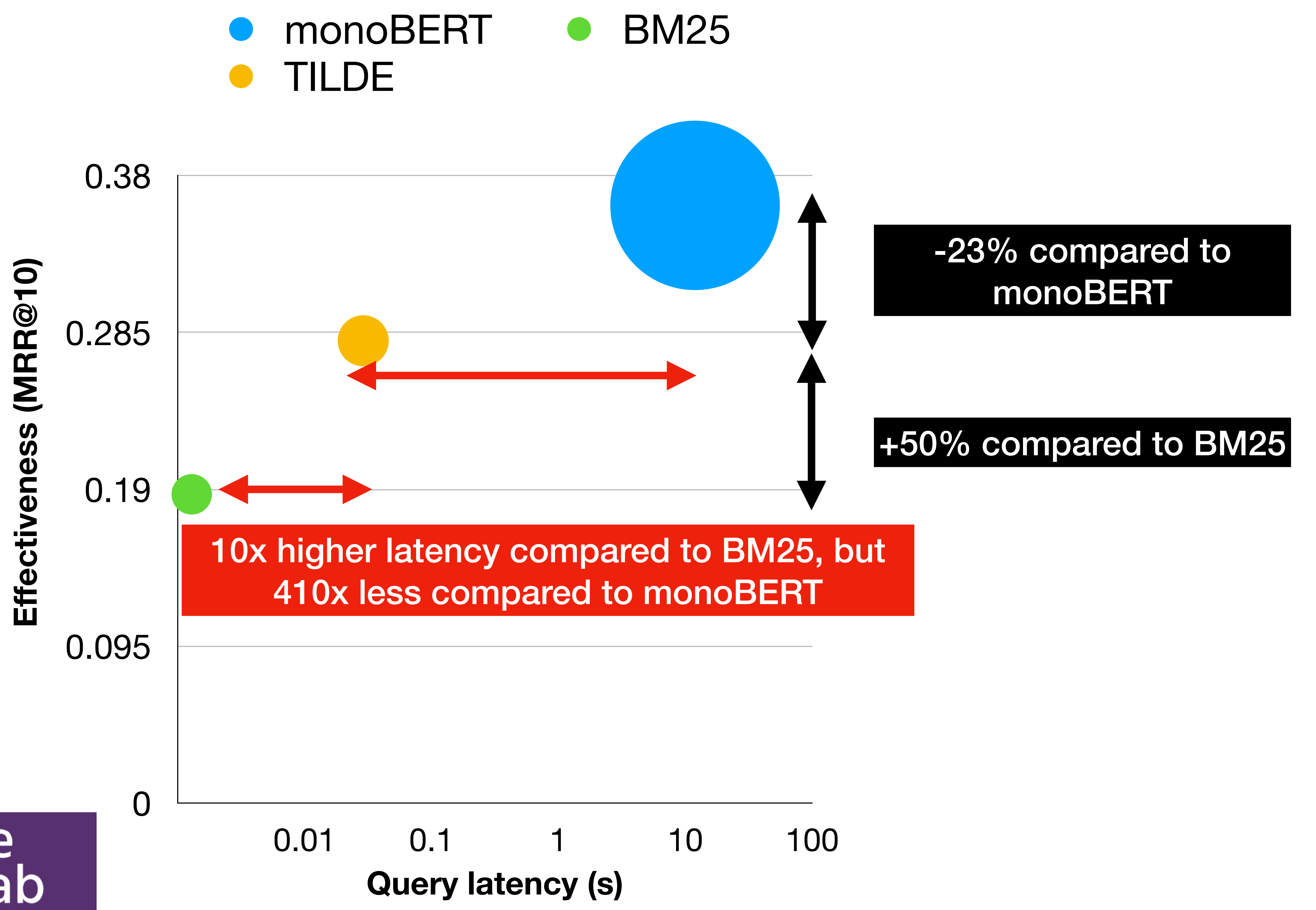
# Effectiveness/Efficiency/Hardware Trade-offs




-23% compared to monoBERT

+50% compared to BM25

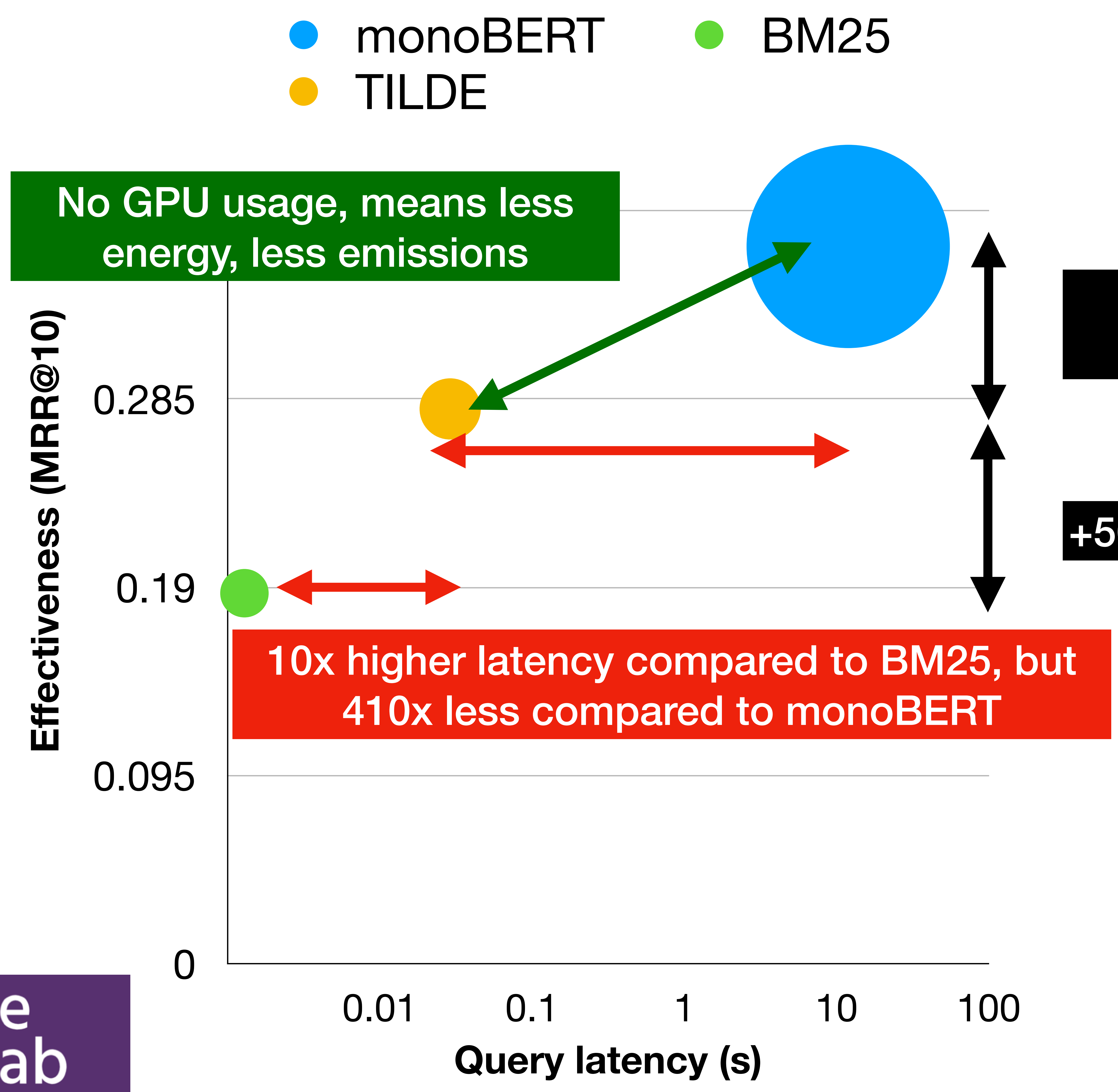
# Effectiveness/Efficiency/Hardware Trade-offs




ie lab

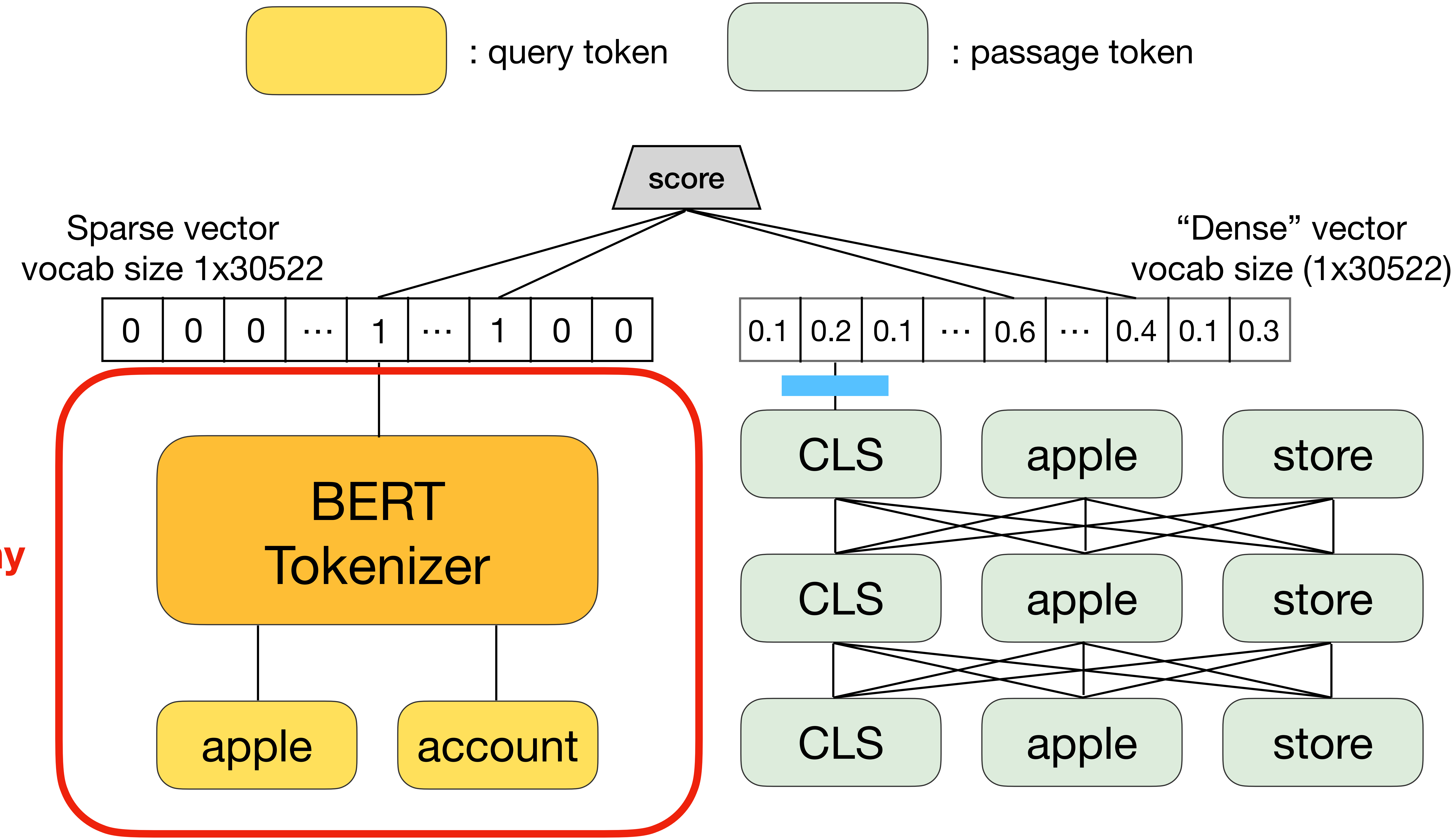
The University Of Queensland
AUSTRALIA
CREATE CHANGE

# Effectiveness/Efficiency/Hardware Trade-offs



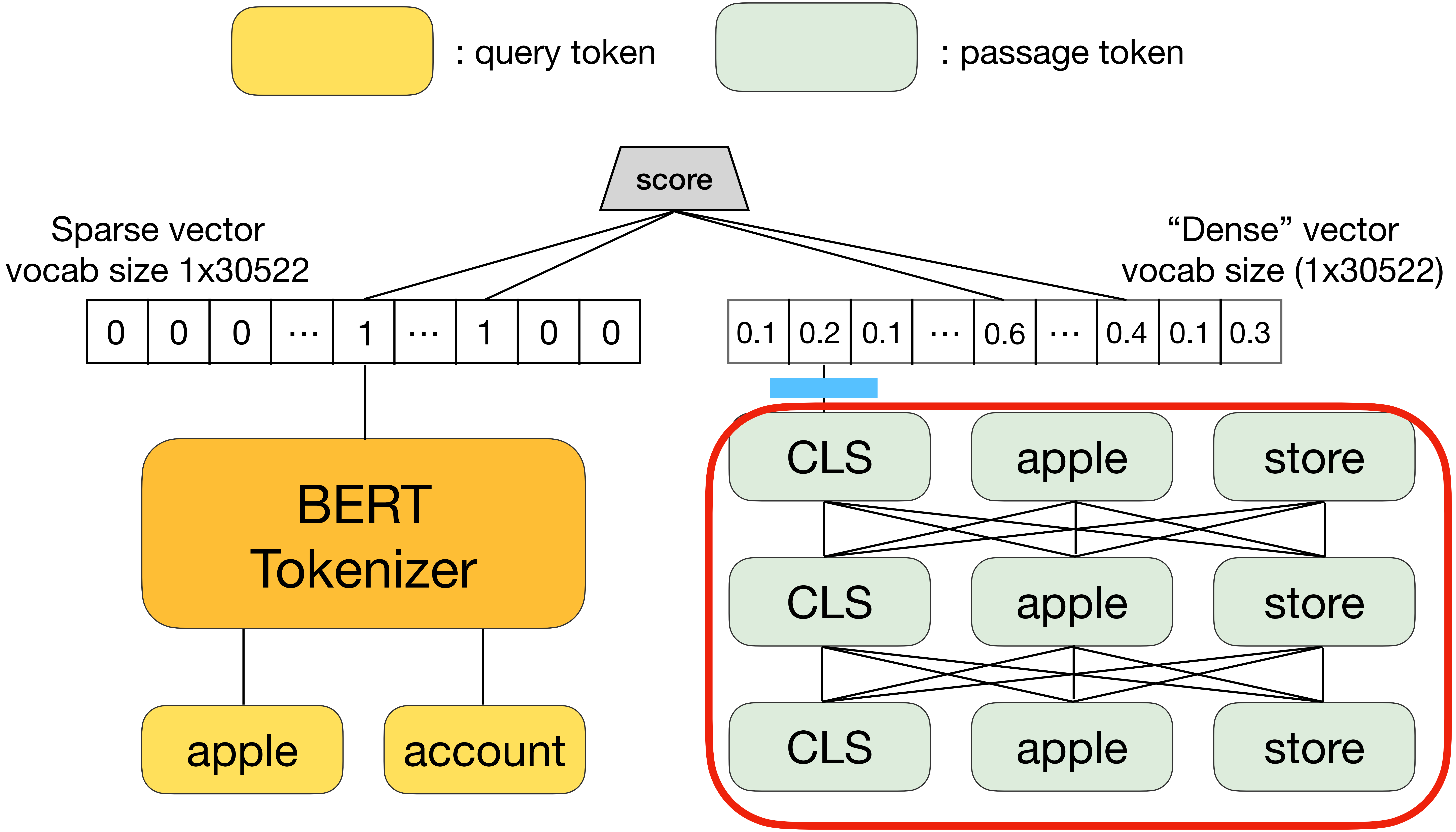

- TILDE provides a trade-off between effectiveness and efficiency
- Does not require GPU at query time
- Yet, losses in effectiveness w.r.t. monoBERT are significant
- Less effective than DRs
- 10x less latency than DRs ran on GPU, 100x less latency than DRs ran on CPU

# Can we do any better w.r.t. efficiency & effectiveness?



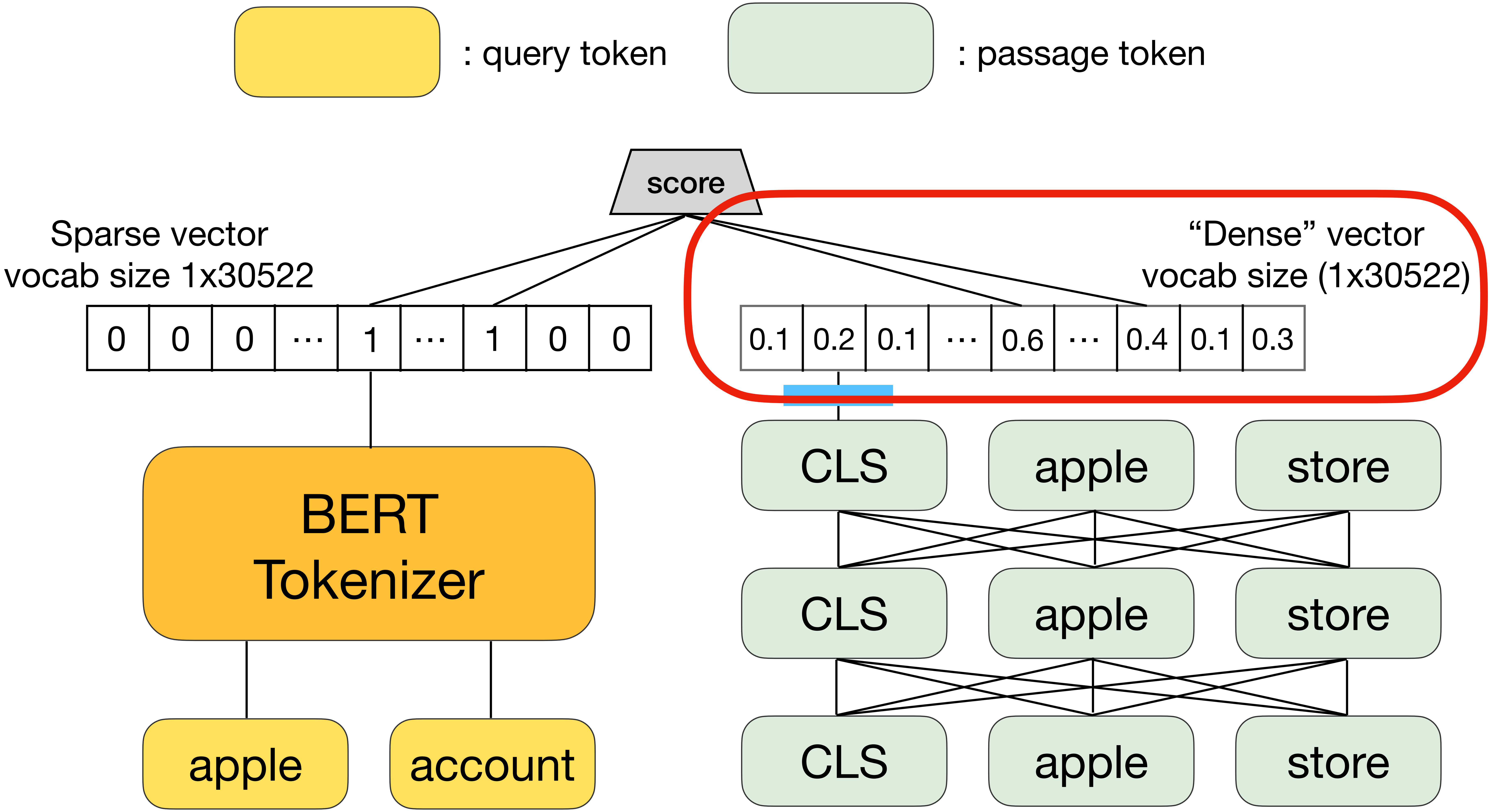

Already reduced to bare minimum; any increase of complexity means increased latency

# Can we do any better w.r.t efficiency & effectiveness?




Offline computation: increase of complexity does not affect latency. REFINE THIS!

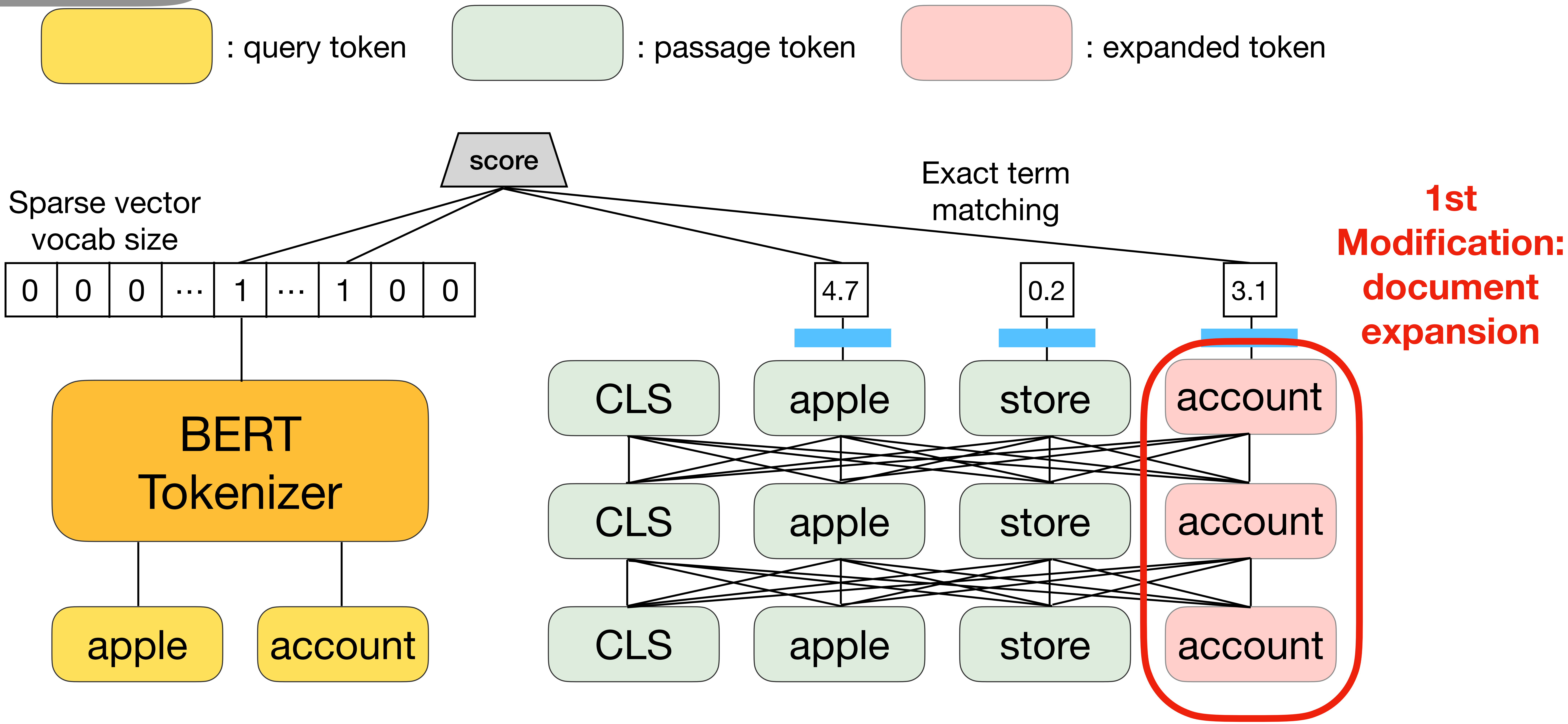
# Can we do any better w.r.t efficiency & effectiveness?




**Large vector to store. Additional projection, no direct relation to vocabulary. REFINE THIS!**

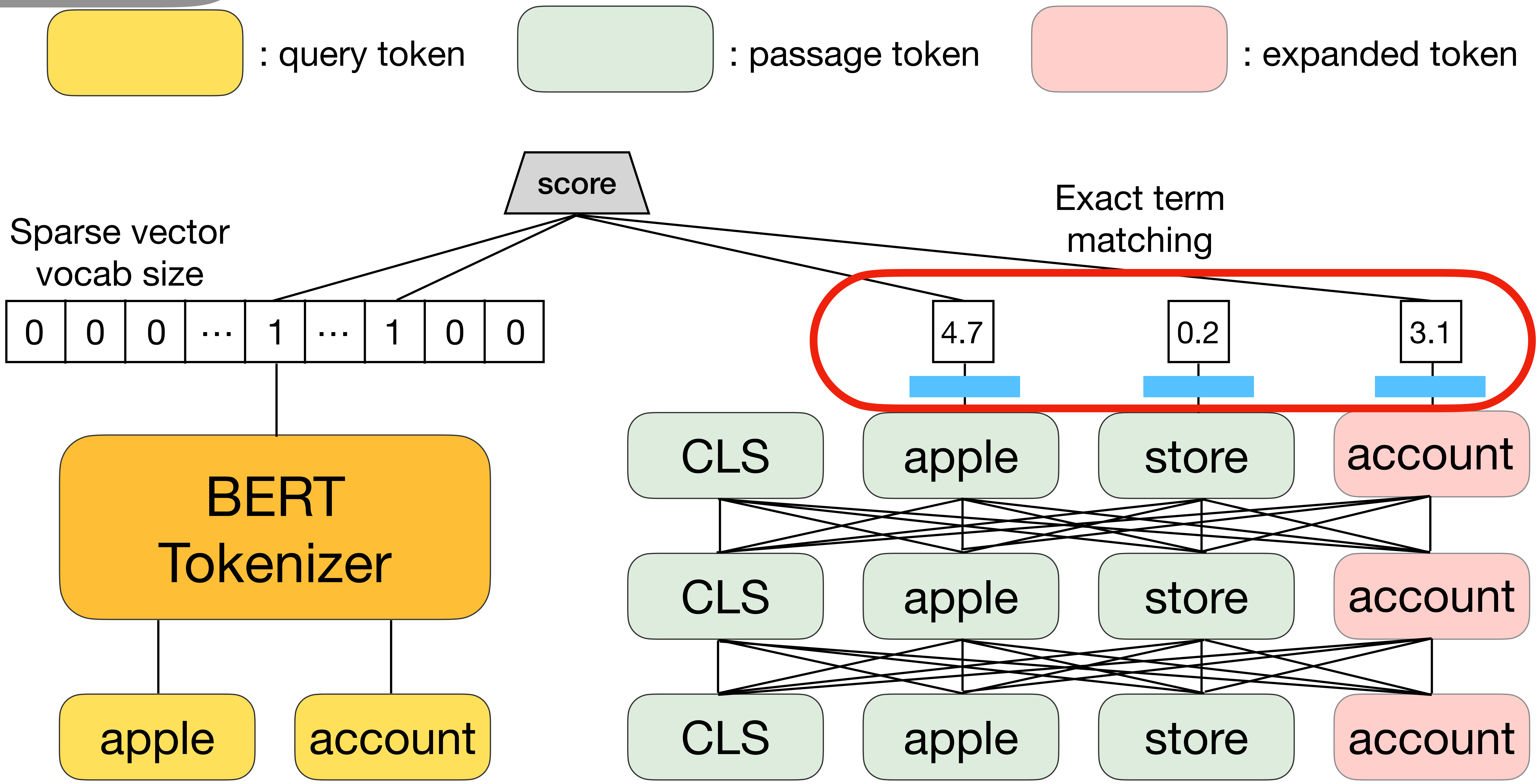
Zhuang, Zuccon, “*Fast passage re-ranking with contextualized exact term matching and efficient passage expansion*"

# TILDEv2



# TILDEv2

Zhuang, Zuccon, *"Fast passage re-ranking with contextualized exact term matching and efficient passage expansion"*

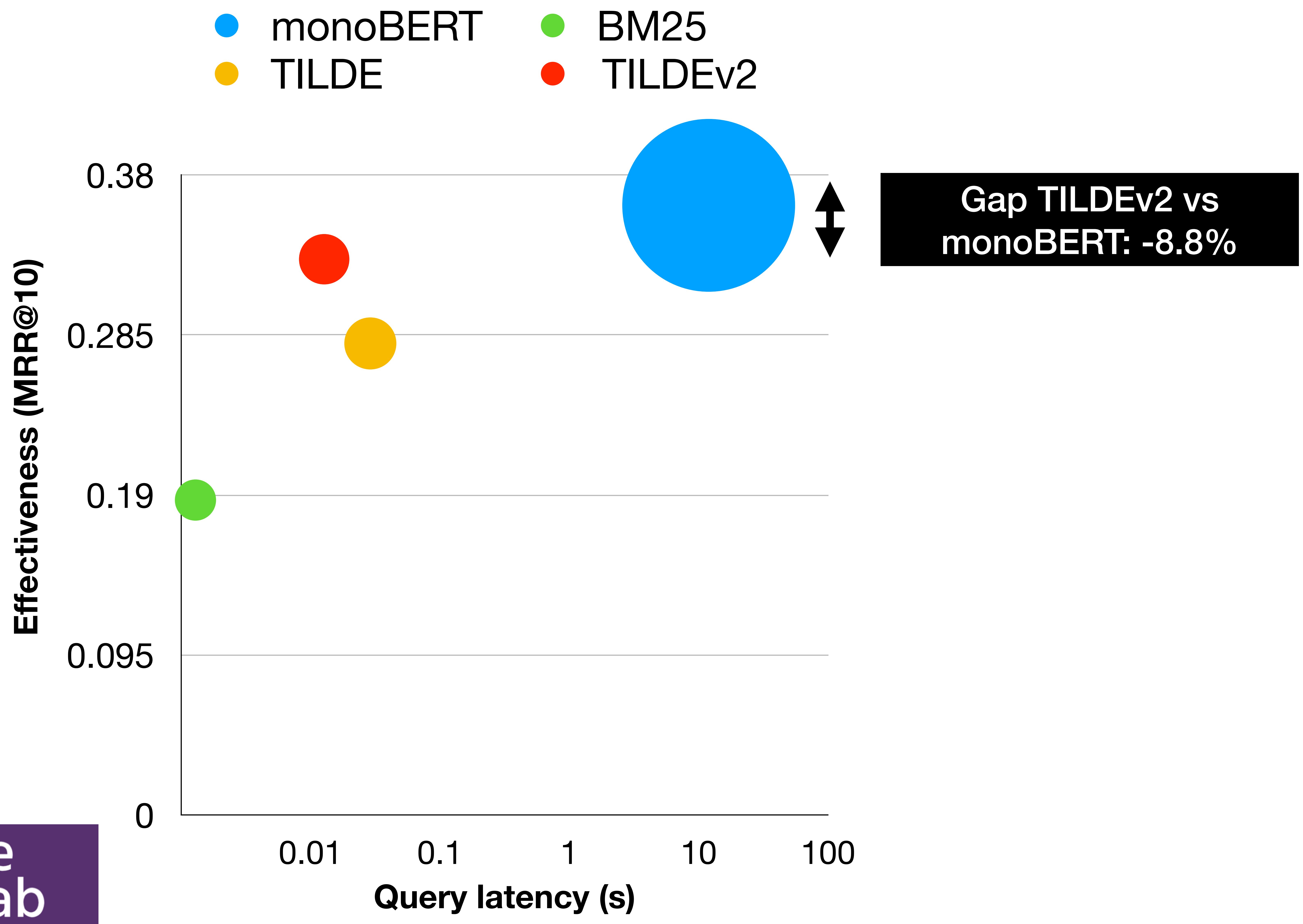



**2nd Modification: projection token embedding into single score & exact matching with query**

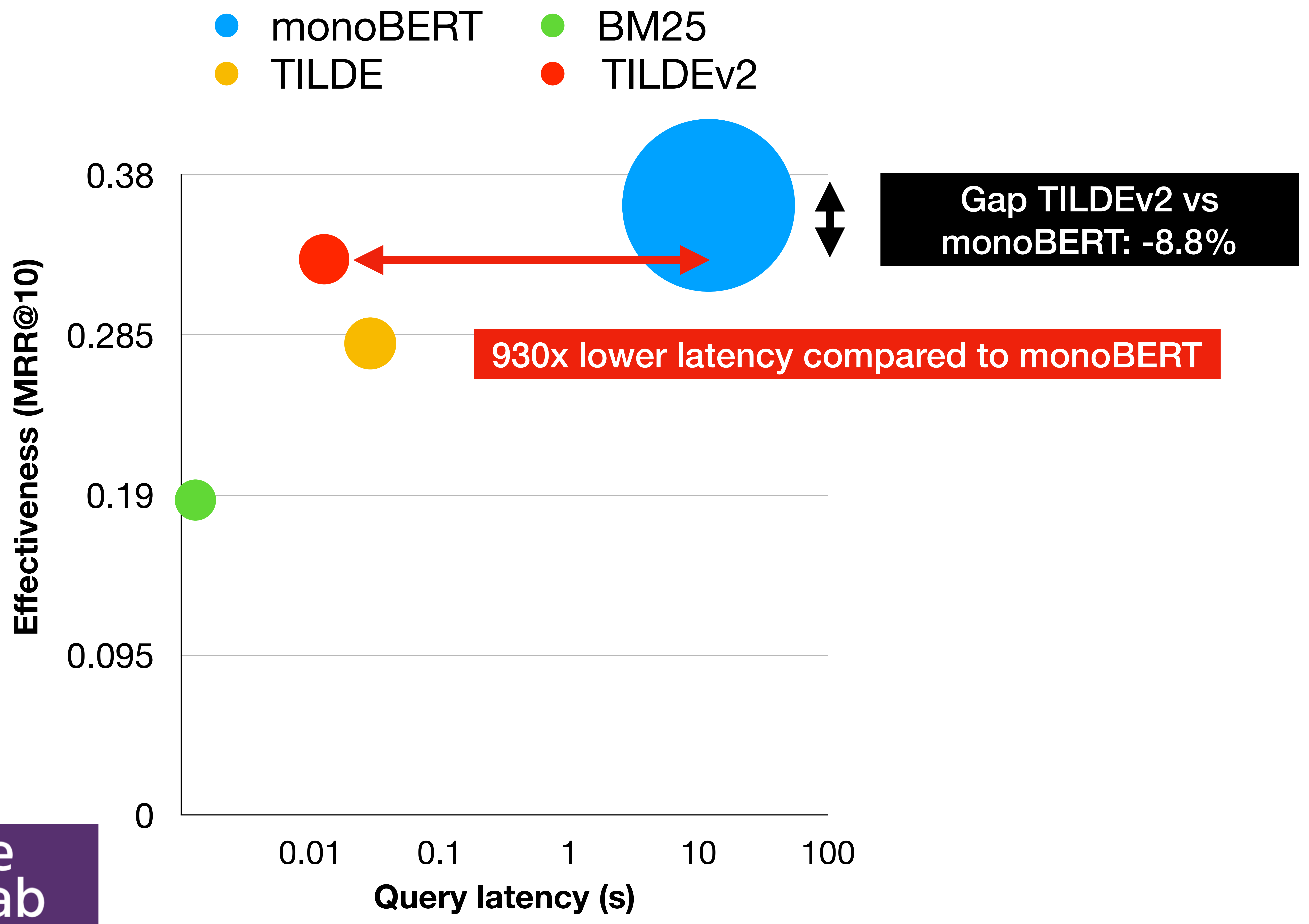
ie lab

THE UNIVERSITY OF QUEENSLAND AUSTRALIA
CREATE CHANGE

# Effectiveness/Efficiency/Hardware Trade-offs



# Effectiveness/Efficiency/Hardware Trade-offs




- TILDEv2 achieved a great trade-off
- Can be run in production, on commodity hardware: it does not even require a GPU
- More effective & efficient than DRs
- Effectiveness at par to other sparse models (uniCOIL, SPLADE)

ie lab

# Importance of Expansion in TILDEv2

| | No Exp |
|---|---|
| MRR@10 | 0.299 |
| Avg added tokens | - |
| Index size | 4.3 GB |

# Importance of Expansion in TILDEv2

| | No Exp | Doc2query |
|---|---|---|
| MRR@10 | 0.299 | 0.333 |
| Avg added tokens | - | 19.0 |
| Index size | 4.3 GB | 5.2 GB |

ie lab

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

CREATE CHANGE

# Importance of Expansion in TILDEv2

| | No Exp | Doc2query | TILDE m=128 | TILDE m=150 | TILDE m=200 |
|---|---|---|---|---|---|
| MRR@10 | 0.299 | 0.333 | 0.326 | 0.327 | 0.330 |
| Avg added tokens | - | 19.0 | 13.0 | 25.2 | 61.6 |
| Index size | 4.3 GB | 5.2 GB | 5.2 GB | 5.6 GB | 6.9GB |

# However, expansion comes at a cost

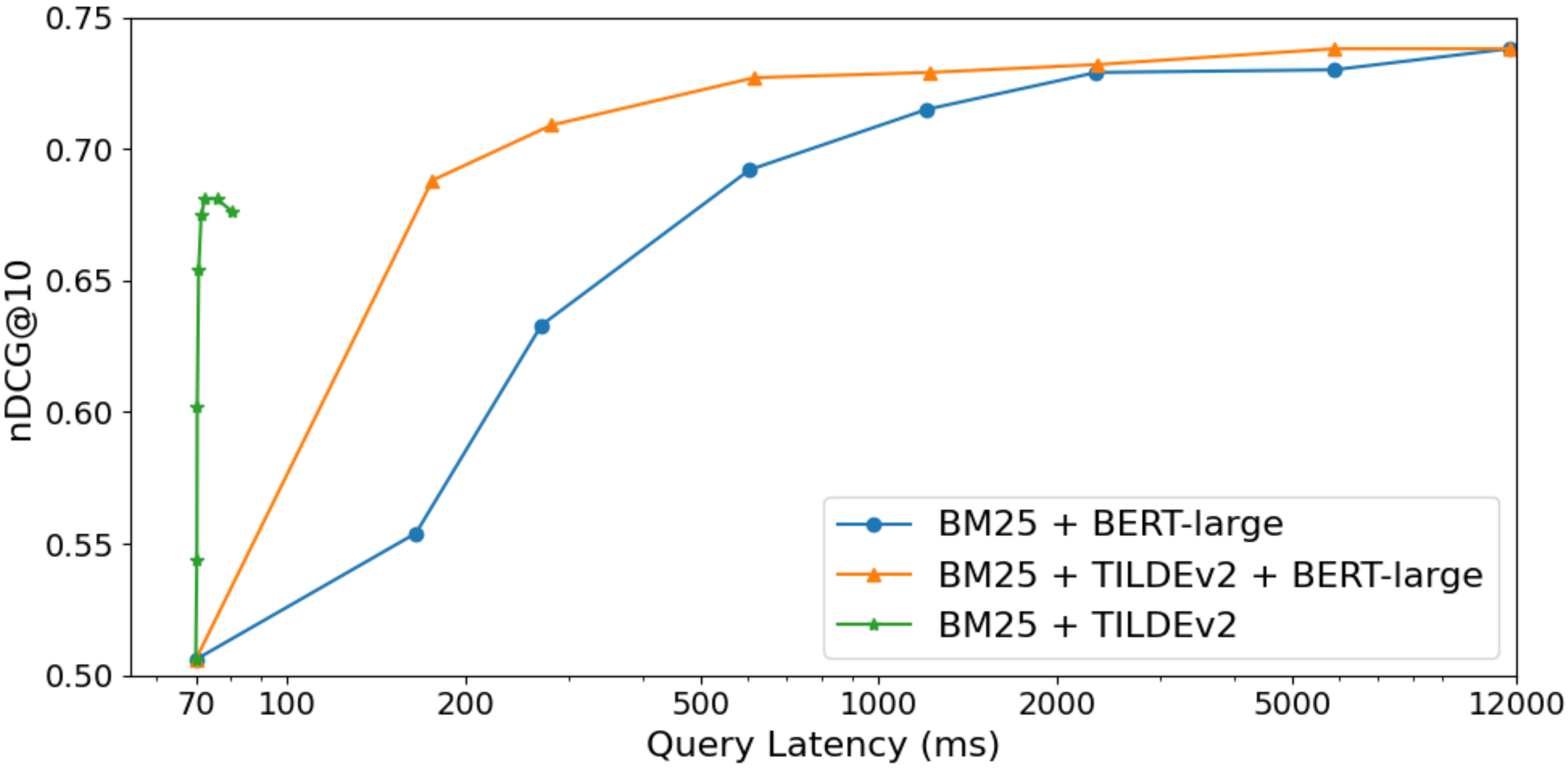
| | No Exp | Doc2query | TILDE m=128 | TILDE m=150 | TILDE m=200 |
|---|---|---|---|---|---|
| **MRR@10** | 0.299 | 0.333 | 0.326 | 0.327 | 0.330 |
| **Avg added tokens** | - | 19.0 | 13.0 | 25.2 | 61.6 |
| **Index size** | 4.3 GB | 5.2 GB | 5.2 GB | 5.6 GB | 6.9GB |
| **Expansion cost** | - | 320 hours, $768 | 7.22 hours, $5.34 | 7.25 hours, $5.37 | 7.33 hours, $5.42 |

ie lab

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

CREATE CHANGE

# TILDEv2 Cost-Quality Trade-off



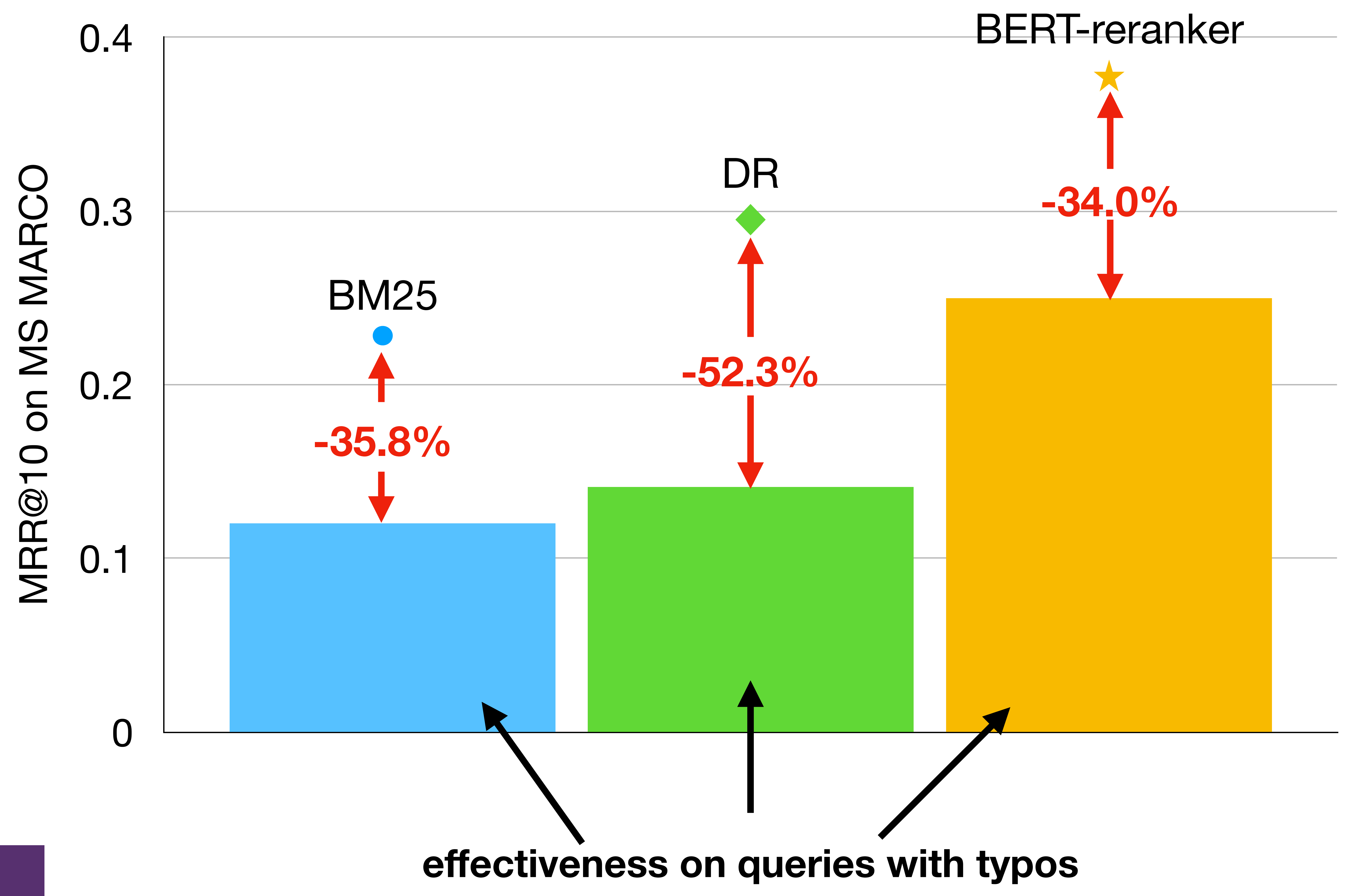

ie lab

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

CREATE CHANGE

# Robustness to out-of-distribution data

# BERT's Challenges: (5) out-of-distribution data



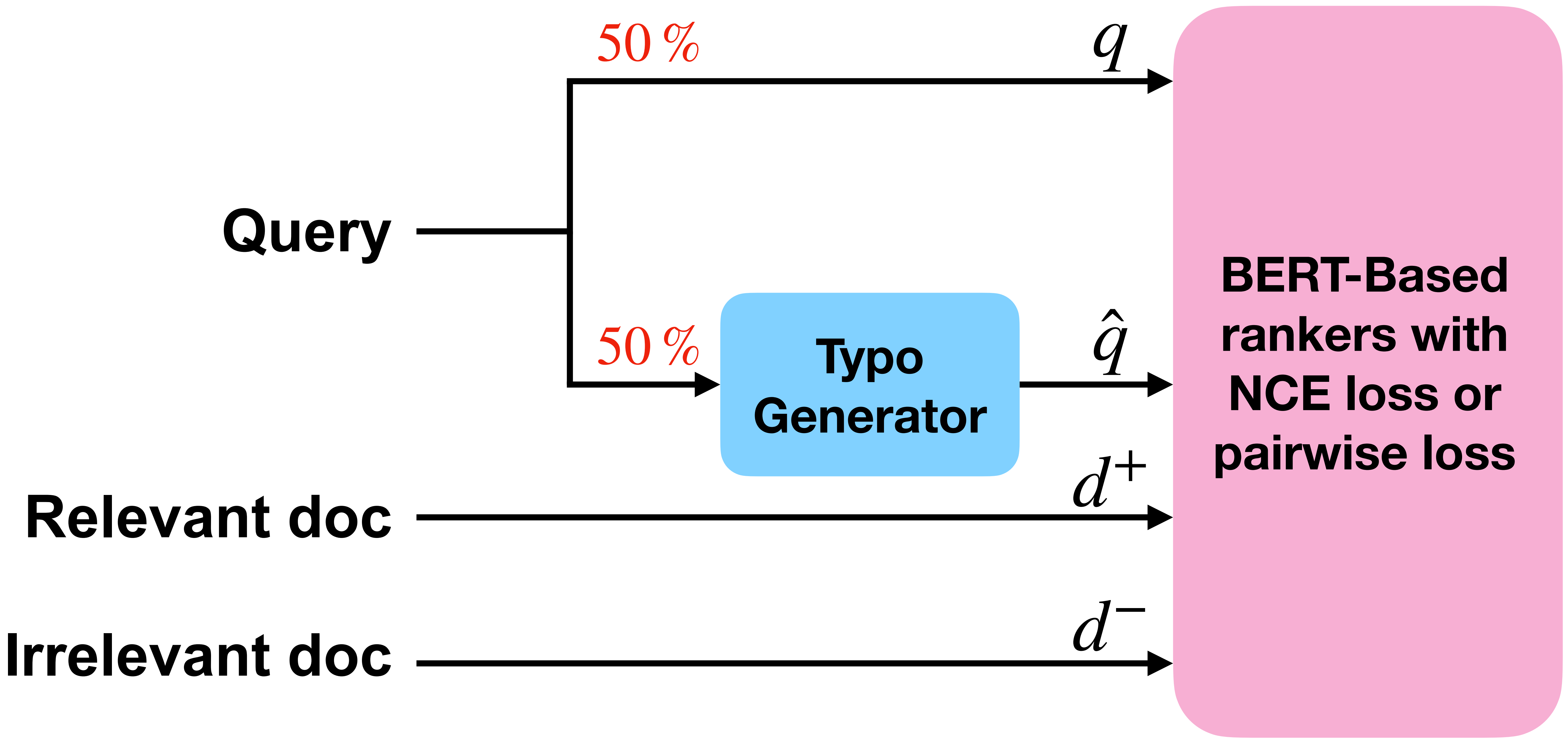

**Both DR and BERT-reranker perform poorly on queries with typos**

Zhuang, Zuccon, *"Dealing with Typos for BERT-based Passage Retrieval and Ranking*", EMNLP 2022

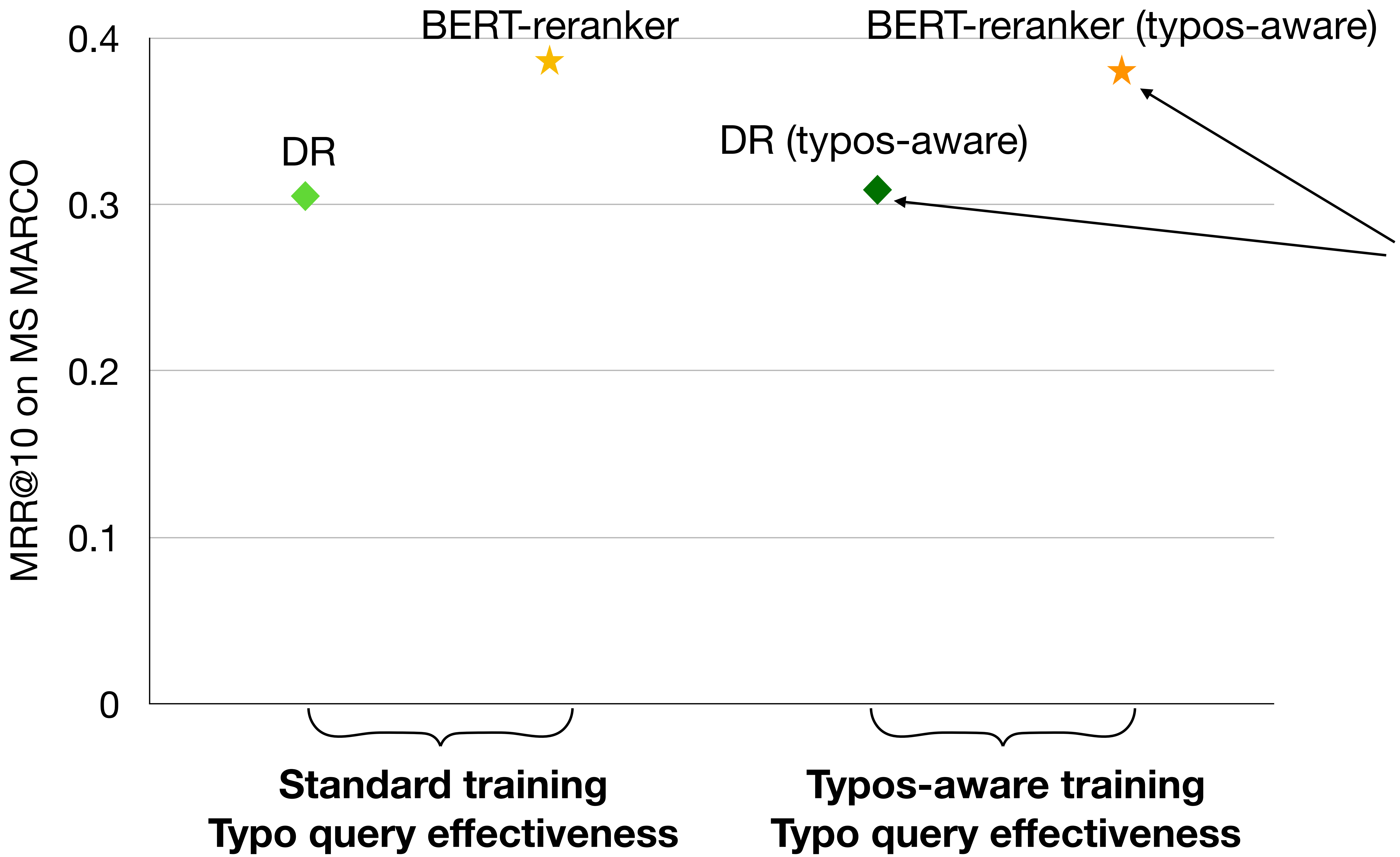
Zhuang, Zuccon, *"CharacterBERT and Self-Teaching for Improving the Robustness of Dense Retrievers on Queries with Typos*", SIGIR 2022

ie lab

# Typos-aware training



# Effectiveness on original queries

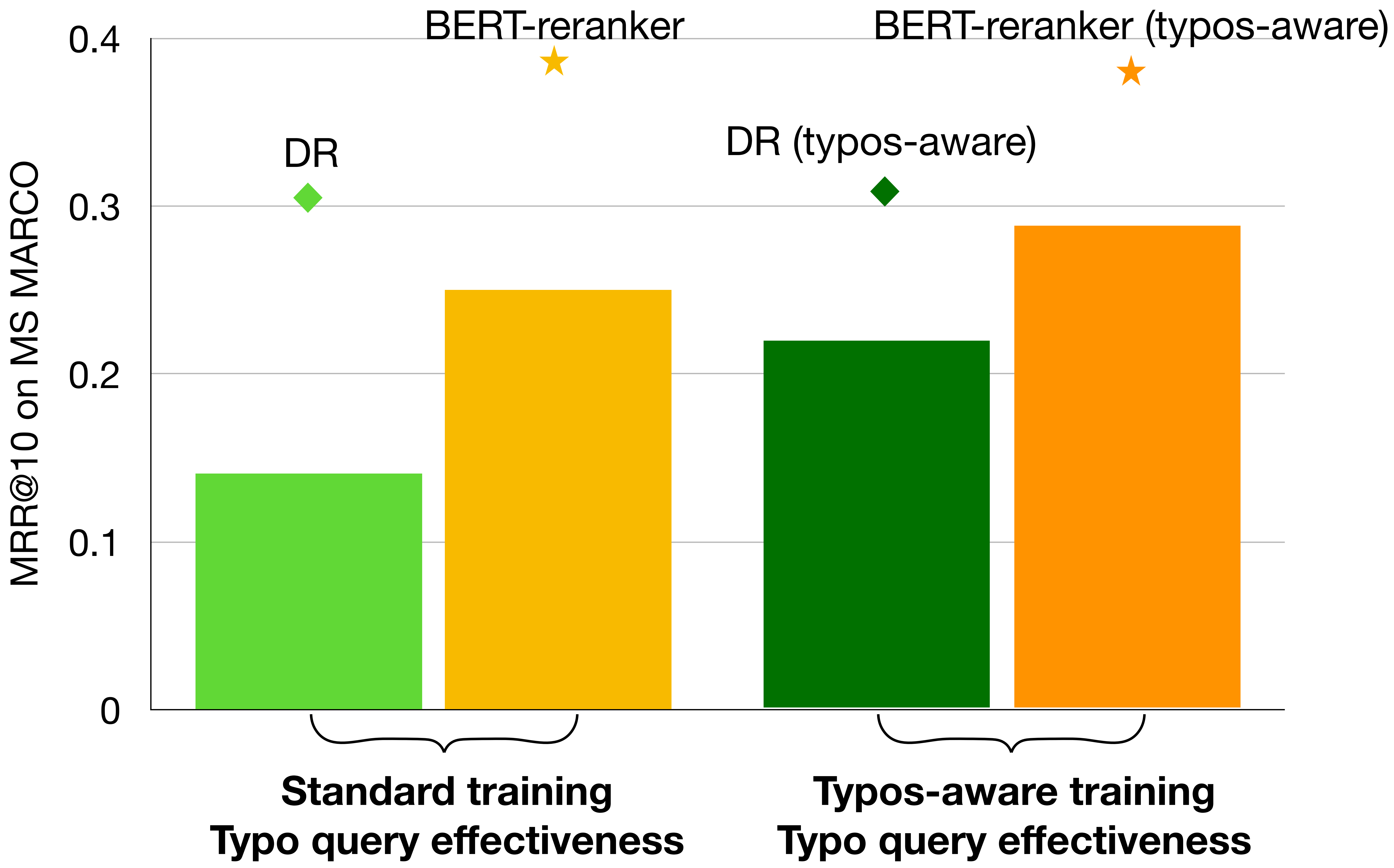



**Typos-aware training does not produce significant difference on original queries**

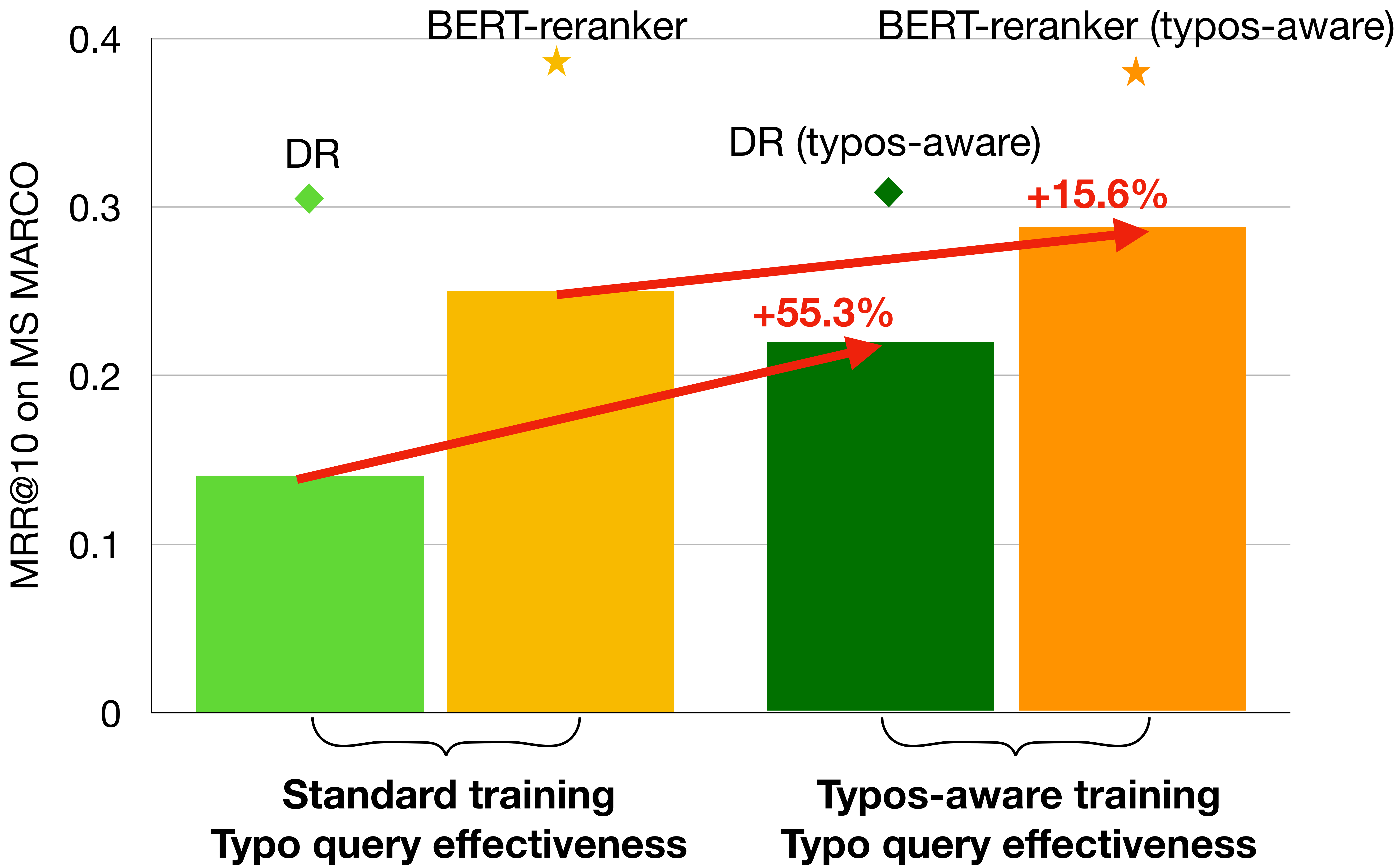
ielab

The University Of Queensland Australia
Create Change

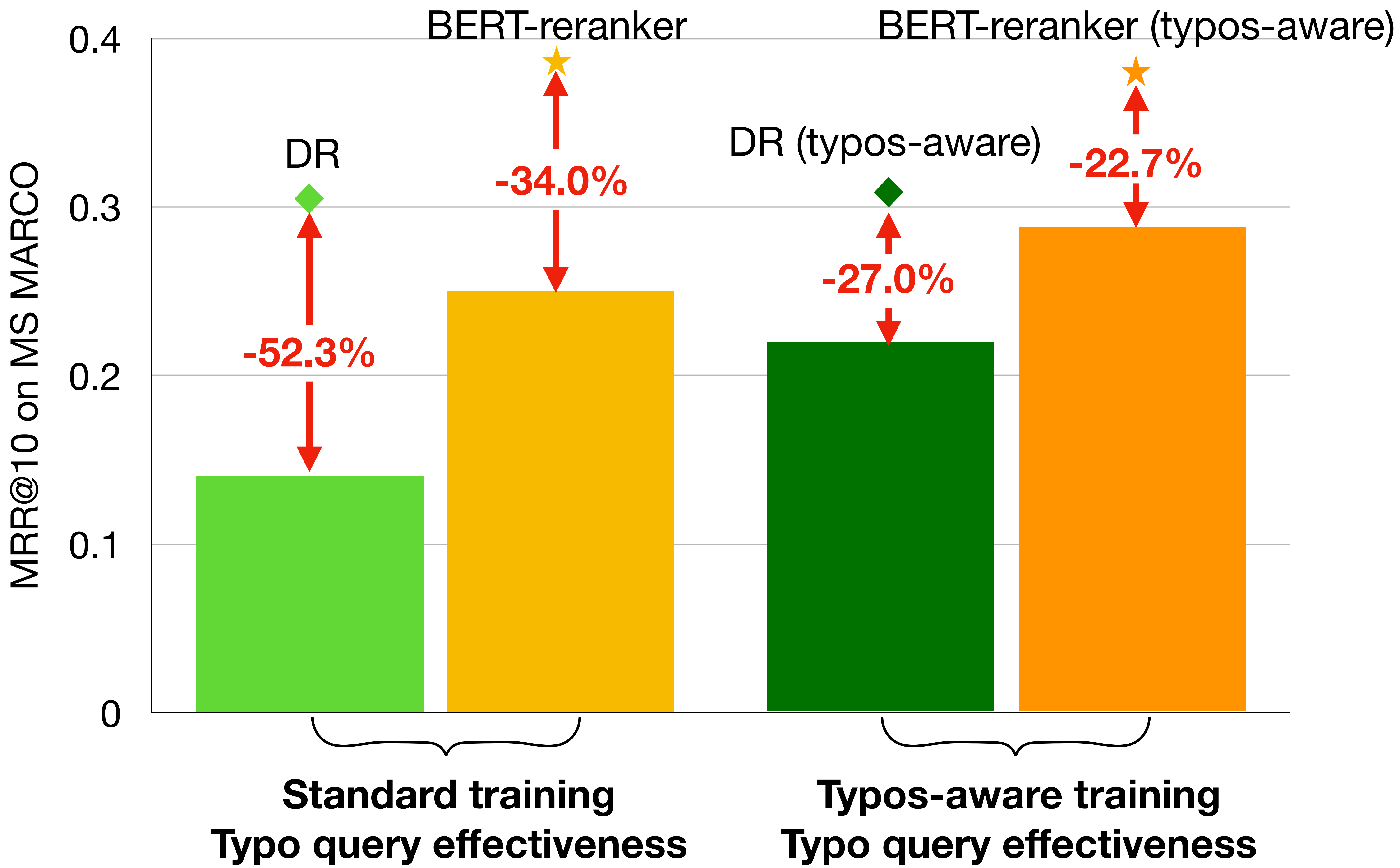
# Effectiveness on typo queries



# Effectiveness on typo queries



# Effectiveness on typo queries




ie lab

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

CREATE CHANGE

# Why BERT-based DRs cannot deal with queries with typos?

- BERT is pre-trained on curated text, and MS MARCO dataset has no or very few queries with typos.
- WordPiece Tokenization based on small vocabulary which contains common terms + common subtokens
- What is the difference in output from the WordPiece Tokenization in presence of a typo?

**‘information retrieval’** $\xrightarrow{\text{tokenize}}$ **[‘information’, ‘retrieval’]** $\xrightarrow{\text{input\_ids}}$ **[2592, 26384]**

**‘infromation retrieval’** $\xrightarrow{\text{tokenize}}$ **[‘in’, ‘fr’, ‘oma’, ‘tion’, ‘retrieval’]** $\xrightarrow{\text{input\_ids}}$ **[1999, 19699, 9626, 3508, 26384]**

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

CREATE CHANGE

# Why BERT-based DRs cannot deal with queries with typos?

- BERT is pre-trained on curated text, and MS MARCO dataset has no or very few queries with typos.

- WordPiece Tokenization based on small vocabulary which contains common terms + common subtokens

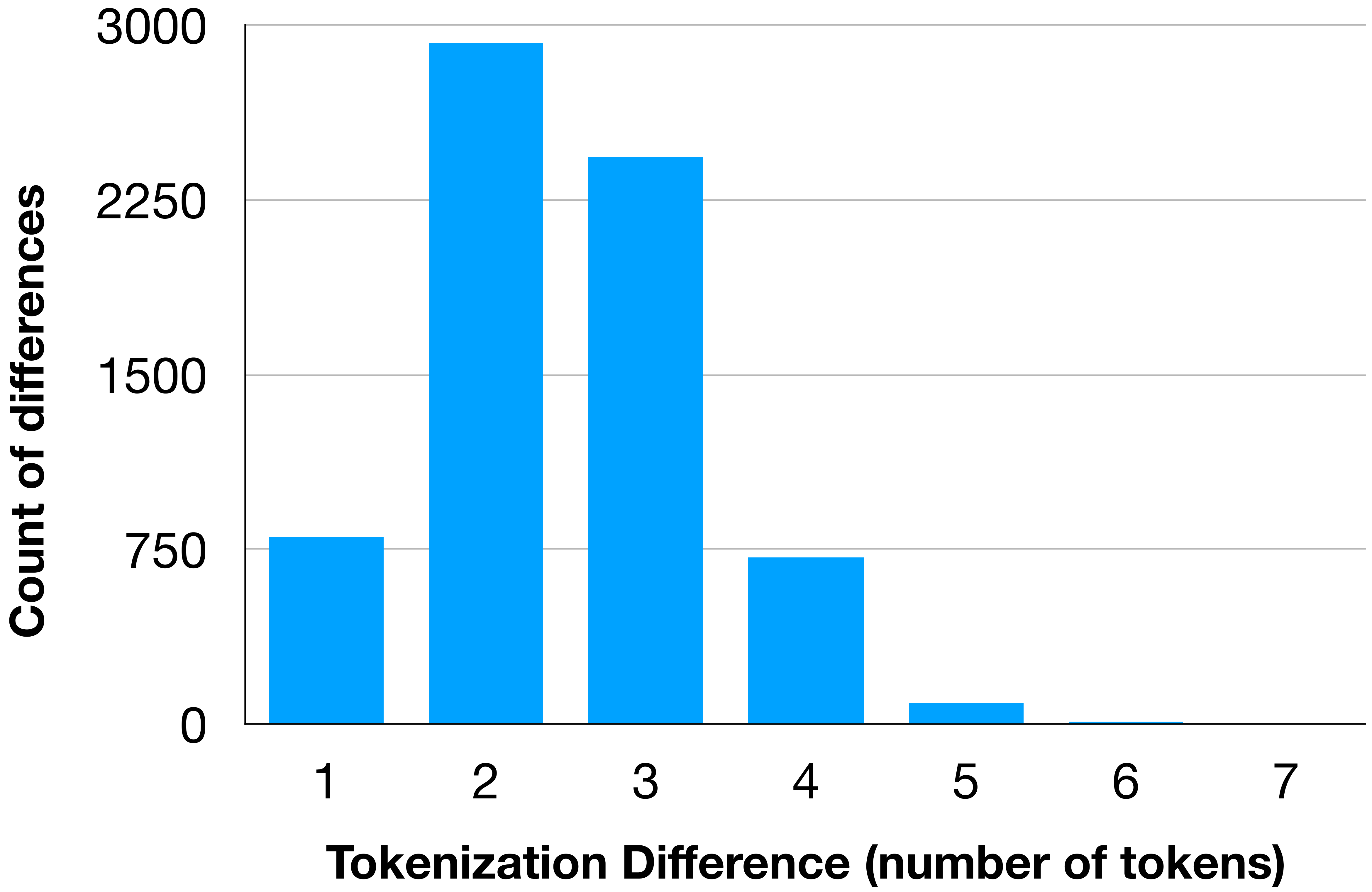
- What is the difference in output from the WordPiece Tokenization in presence of a typo?

‘information retrieval’ $\xrightarrow{\text{tokenize}}$ [‘information’, ‘retrieval’] $\xrightarrow{\text{input\_ids}}$ [2592, 26384]

‘infromation retrieval’ $\xrightarrow{\text{tokenize}}$ [‘in’, ‘fr’, ‘oma’, ‘tion’, ‘retrieval’] $\xrightarrow{\text{input\_ids}}$ [1999, 19699, 9626, 3508, 26384]

**A typo resulted in the query being represented by 4 additional tokens (and lost 1)**

# How large are the tokenization differences caused by typos in queries?

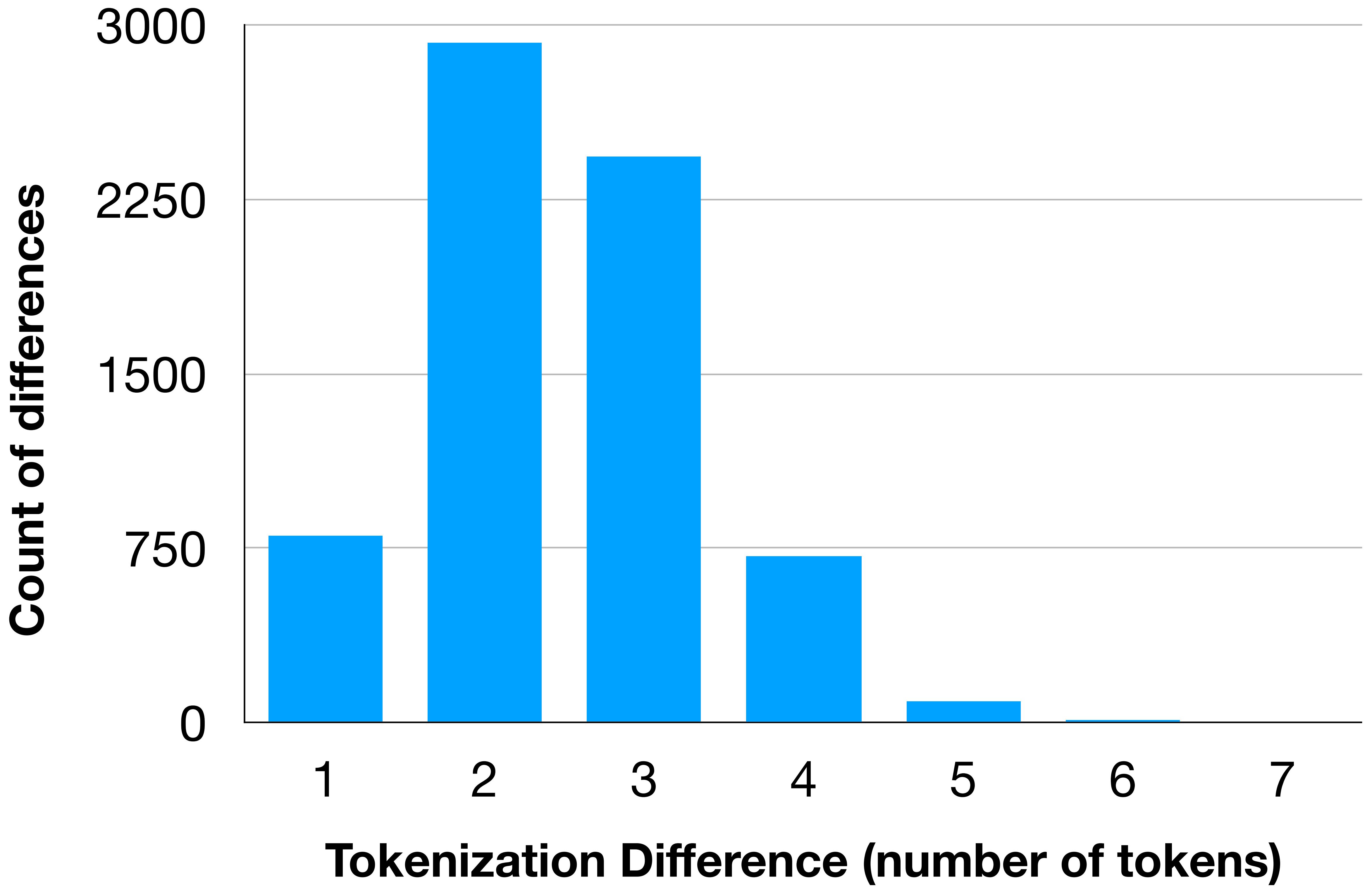



ie lab

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

CREATE CHANGE

# How large are the tokenization differences caused by typos in queries?

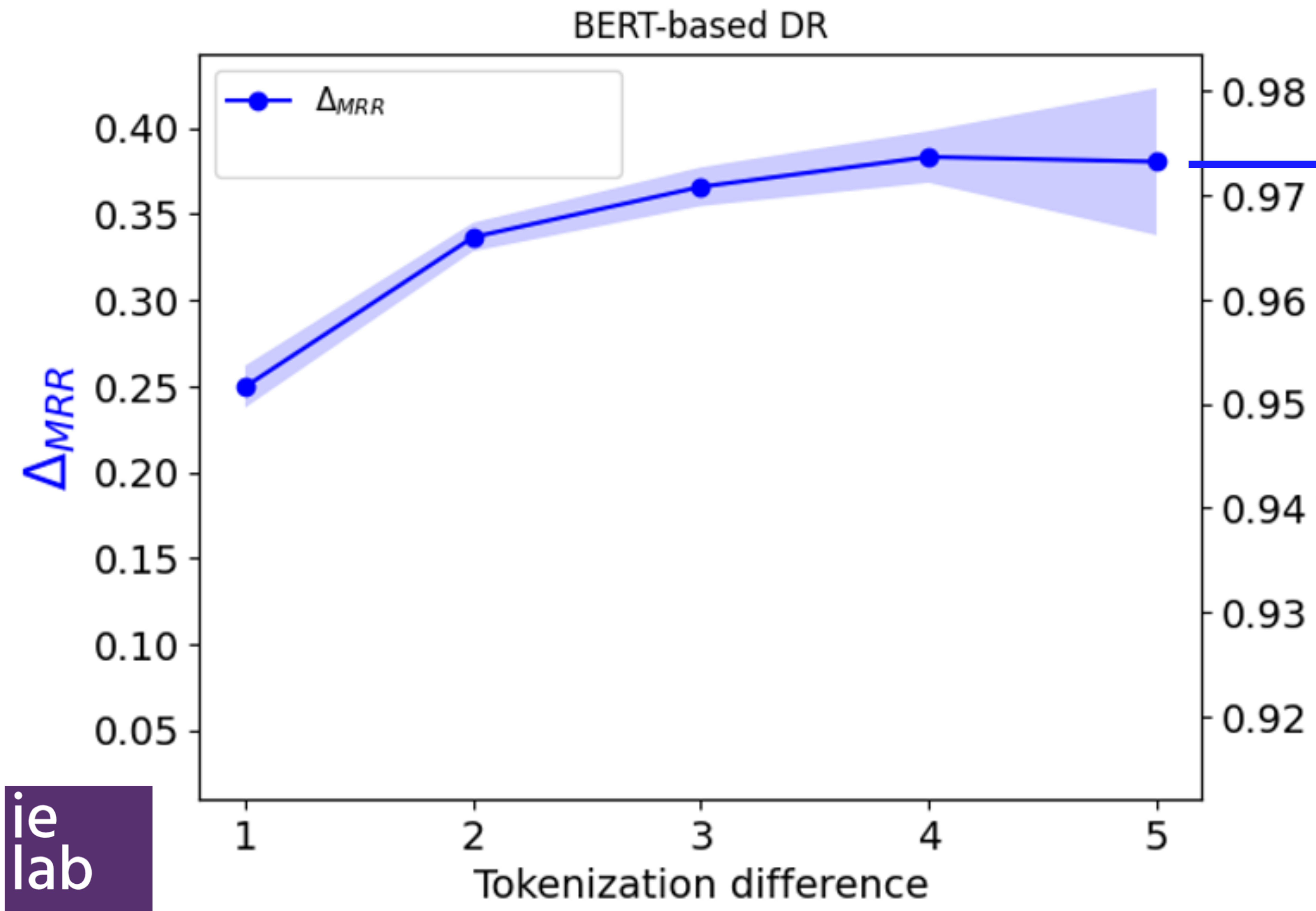



**What are the effects of such tokenization difference?**

ie lab

THE UNIVERSITY OF QUEENSLAND AUSTRALIA
CREATE CHANGE

# Tokenization Difference produces effectiveness losses




**Effectiveness drops**

**=**

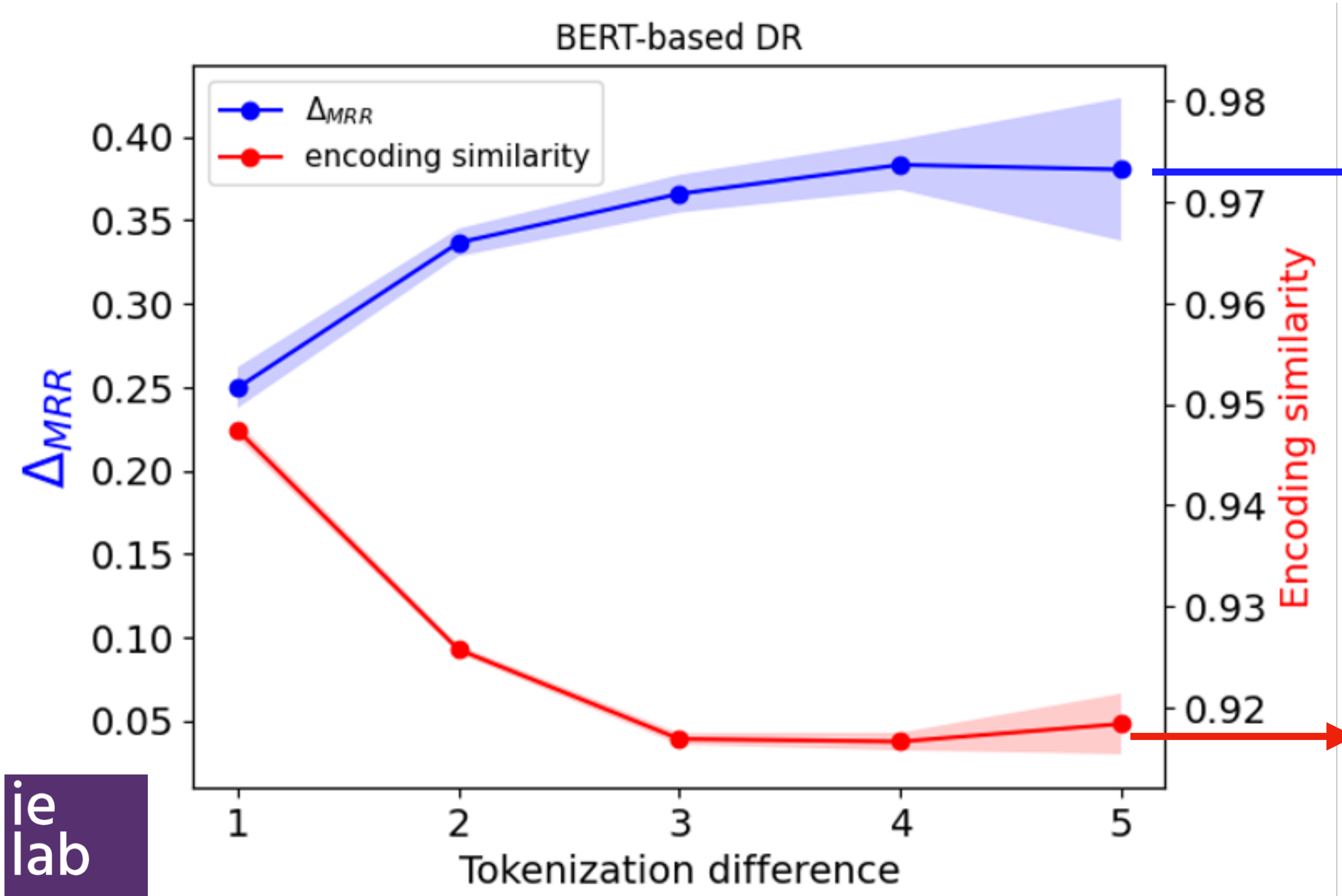
**The higher the loss, the worse the DR does on typos**

ie lab

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

CREATE CHANGE

# Tokenization Difference produces effectiveness losses & encodings different from query without typo




**Effectiveness drops**

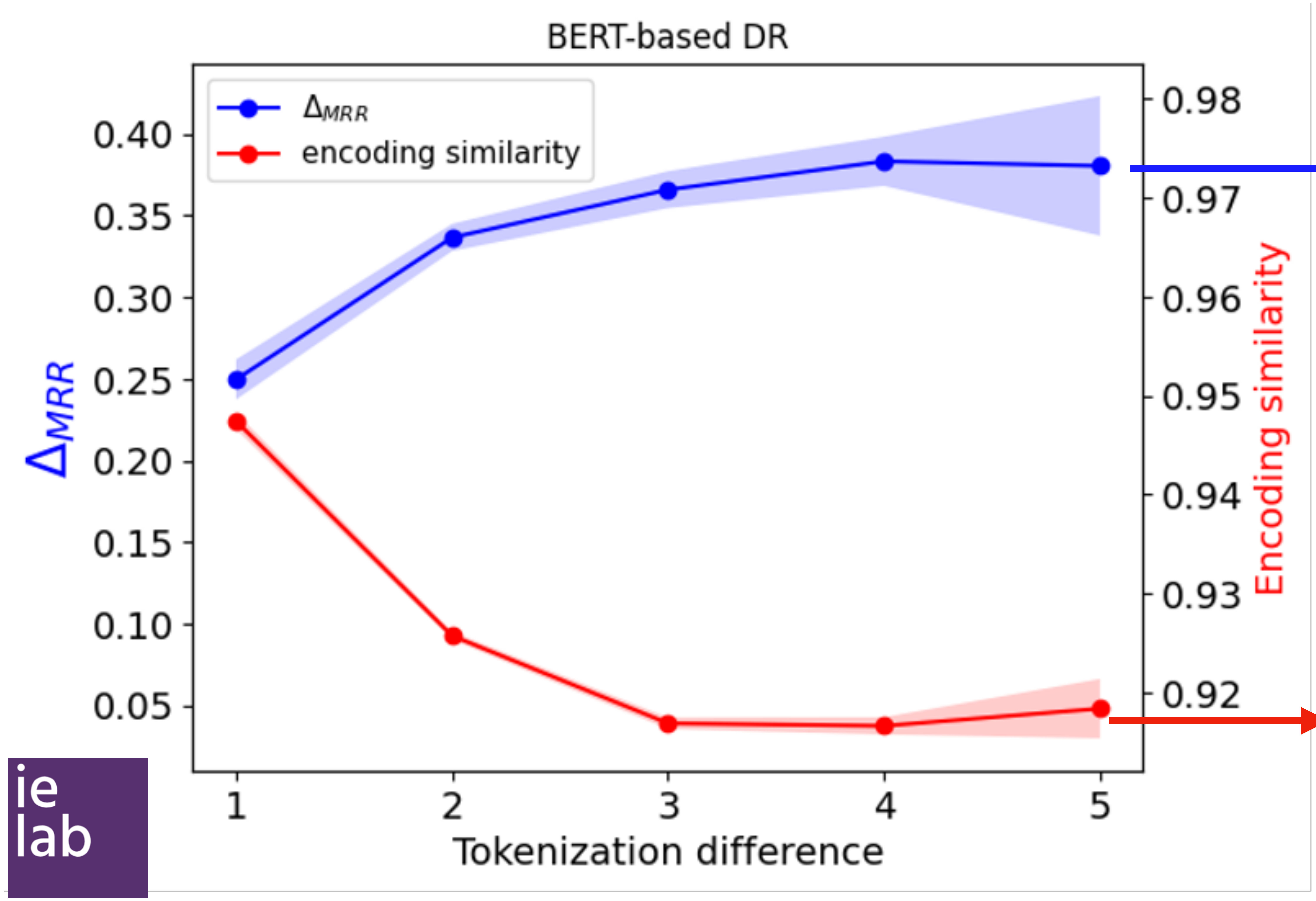
**Representation with typos diverges from original**

ie lab

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

CREATE CHANGE

# Tokenization Difference produces encodings different from query without typo & effectiveness losses




Effectiveness drops

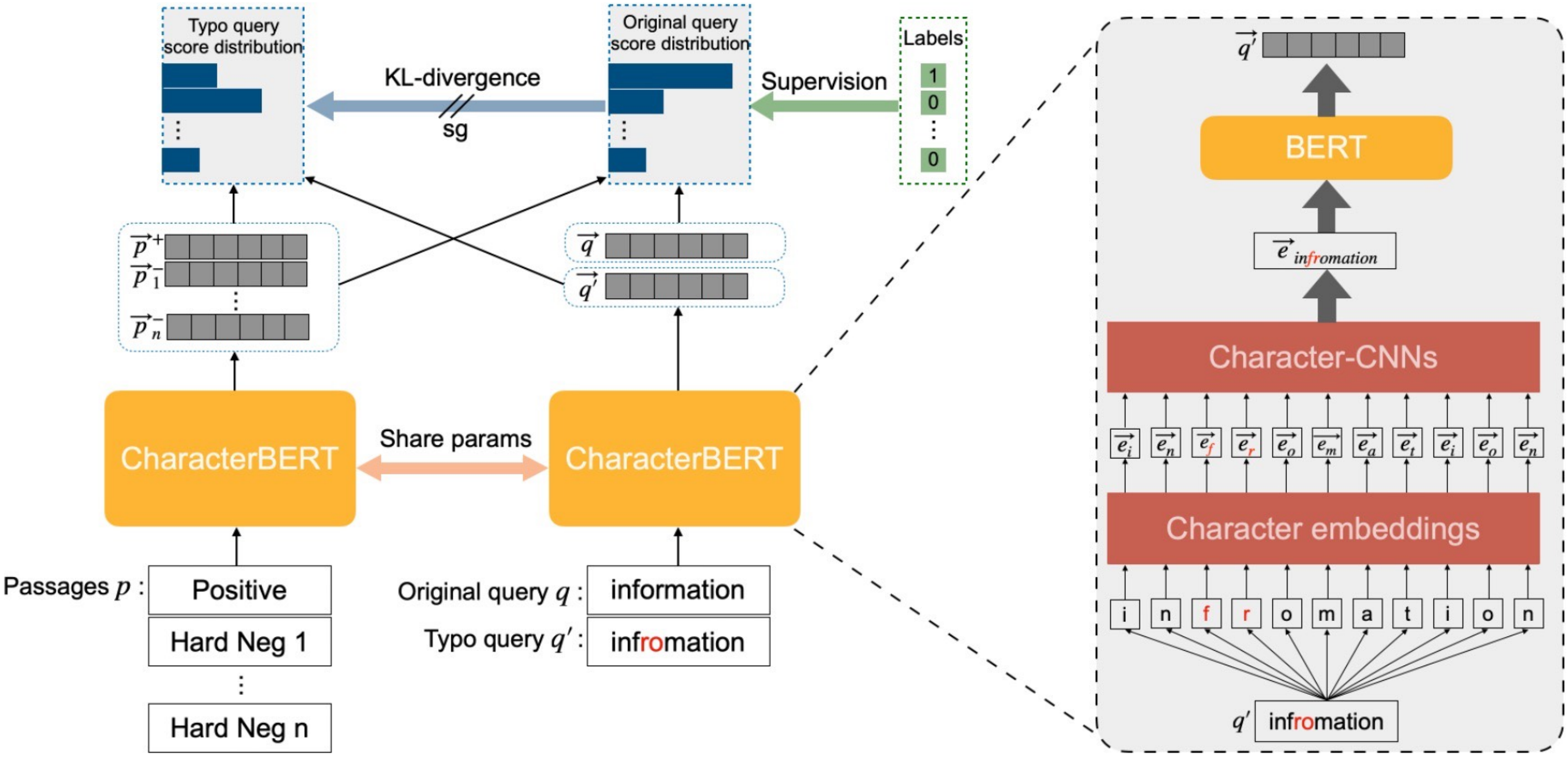
***Can we make representations that are invariant to the presence of typos?***

Representation with typos diverges from original

ie lab

The University of Queensland
Australia
Create Change

# CharacterBERT-DR + Self-Teaching

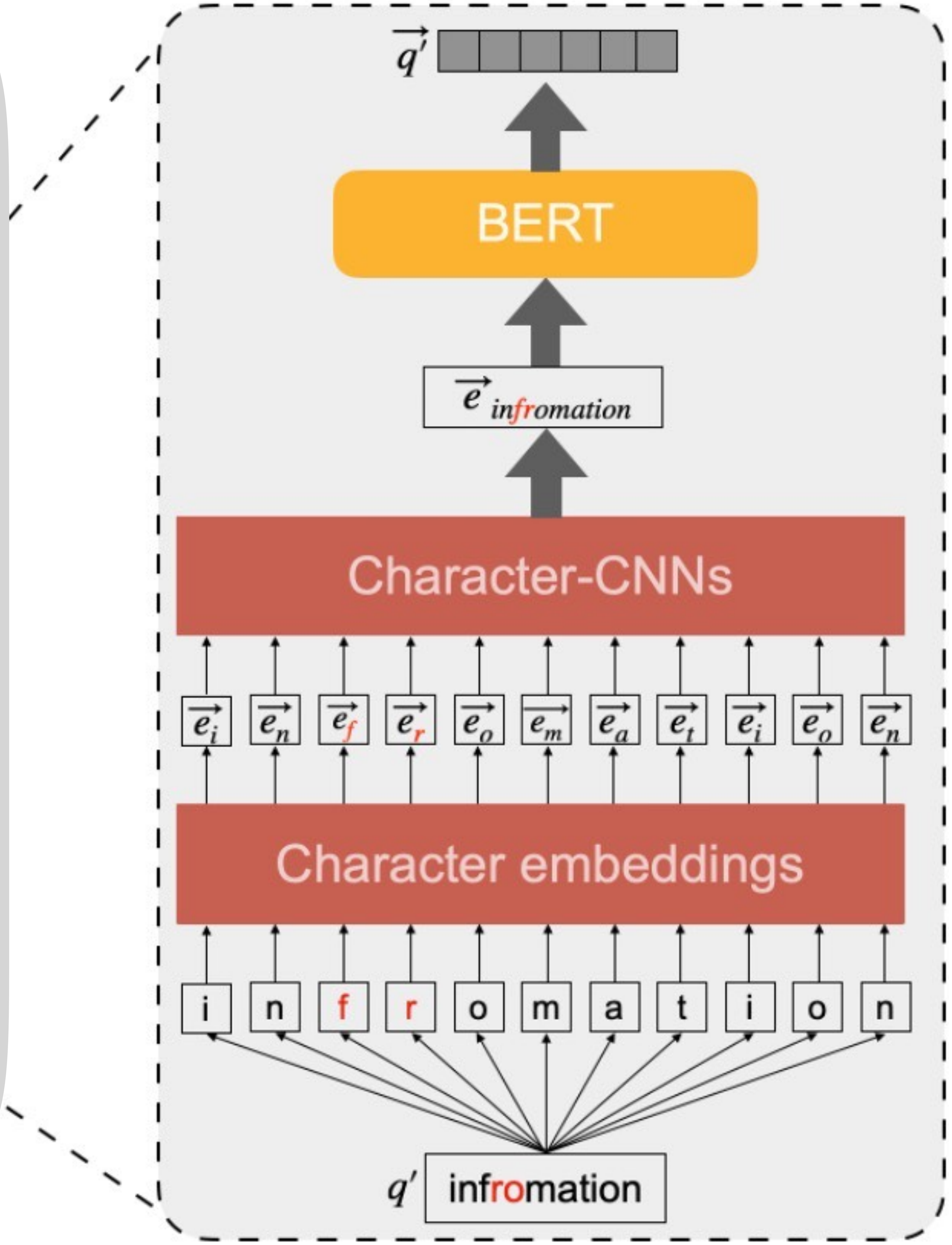


# CharacterBERT-DR + self-teaching

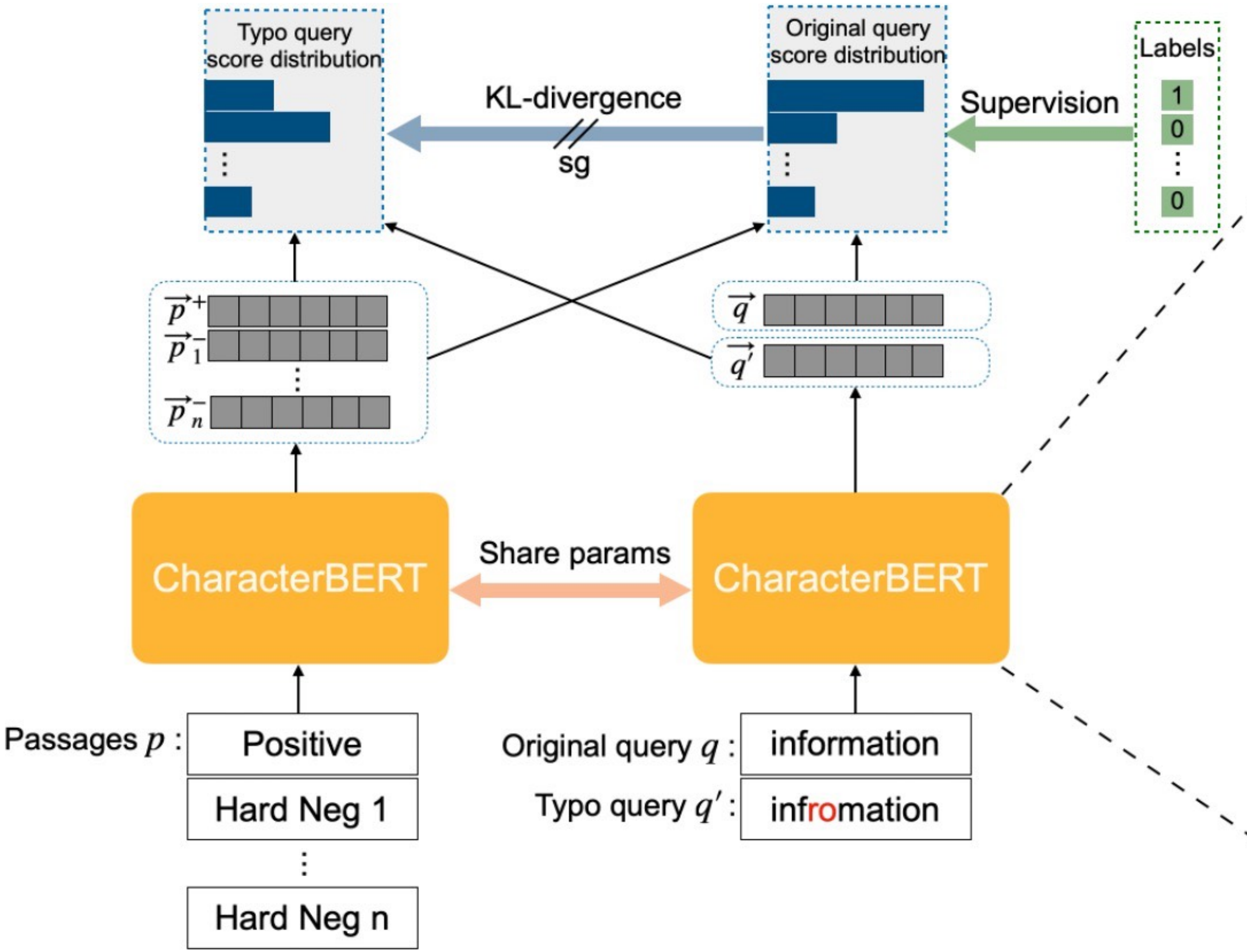
- Replace BERT WordPiece Tokenizer with CharacterBERT [2] to create query and passages embeddings.

- Does not rely on WordPiece vocabulary: any word will be represented by a single word embedding.

[2] CharacterBERT: Reconciling ELMo and BERT for Word-Level Open-Vocabulary Representations From Characters, Hicham et al, COLING 2020



# CharacterBERT-DR + Self-Teaching




1. Make a typo augmentation during training.

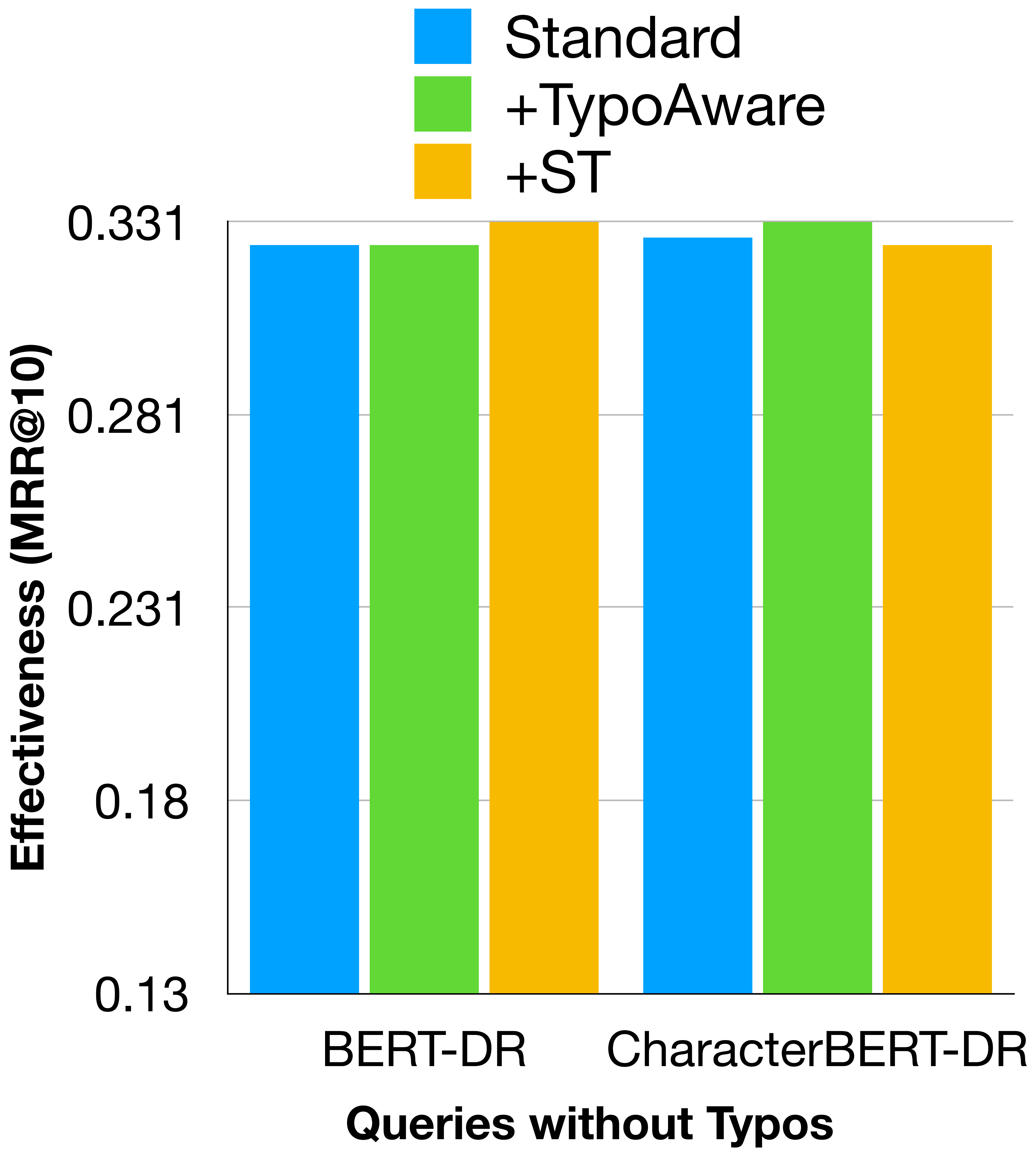
2. Self-Teaching: minimize score distribution difference b/w original query & query with typos:

$$\mathcal{L}_{KL}(\tilde{s}_{q'}, \tilde{s}_q) = \tilde{s}_{q'}(q', p) \cdot \log \frac{\tilde{s}_{q'}(q', p)}{\tilde{s}_q(q, p)}$$

3. Supervised contrastive loss:

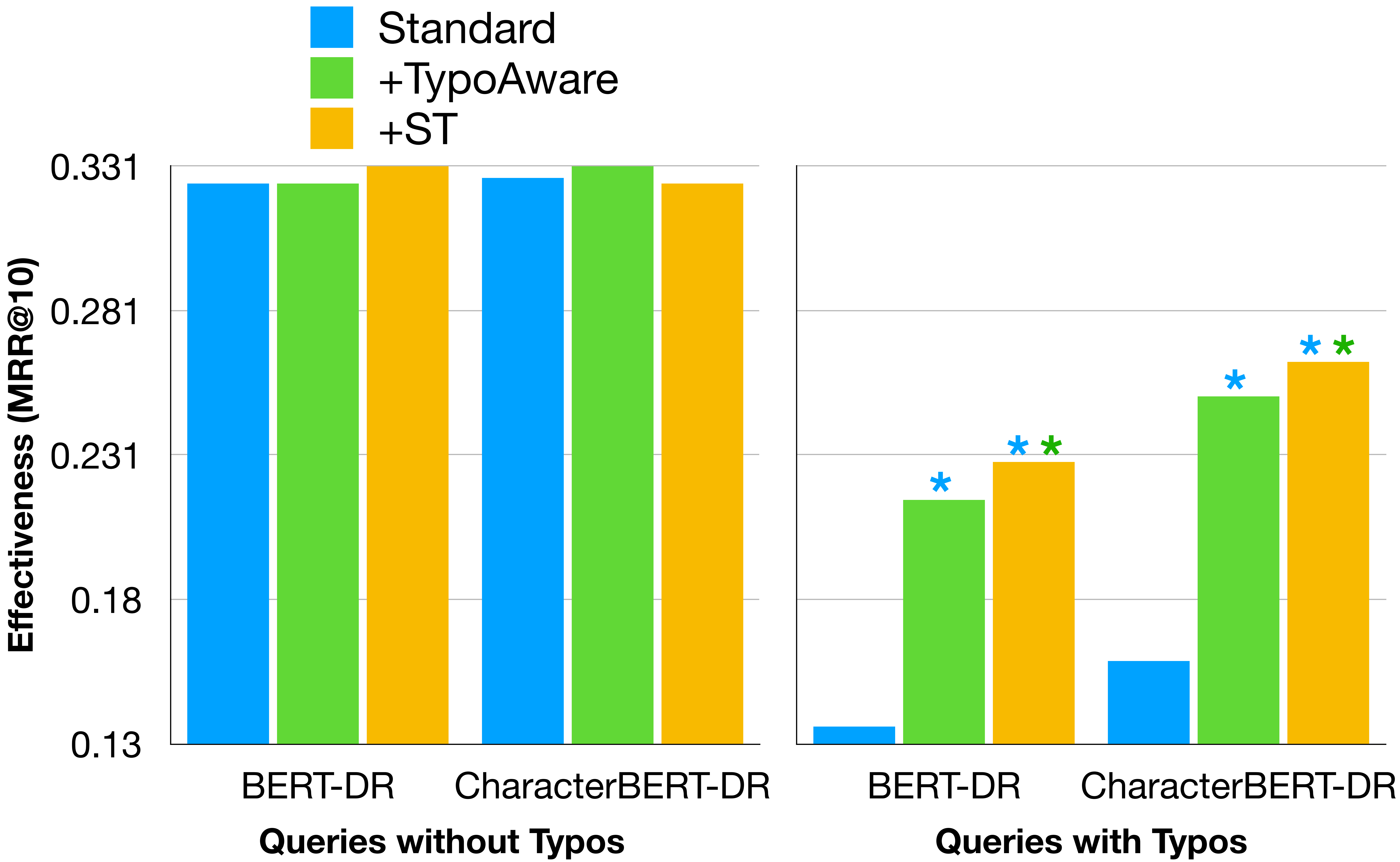
$$\mathcal{L}_{CE}(s_q) = -\log \frac{e^{s_q(q,p^+)}}{e^{s_q(q,p^+)} + \sum_{p^-} e^{s_q(q,p^-)}}$$

# Does CharacterBERT+ST produce unwanted effect on queries w/o typos?



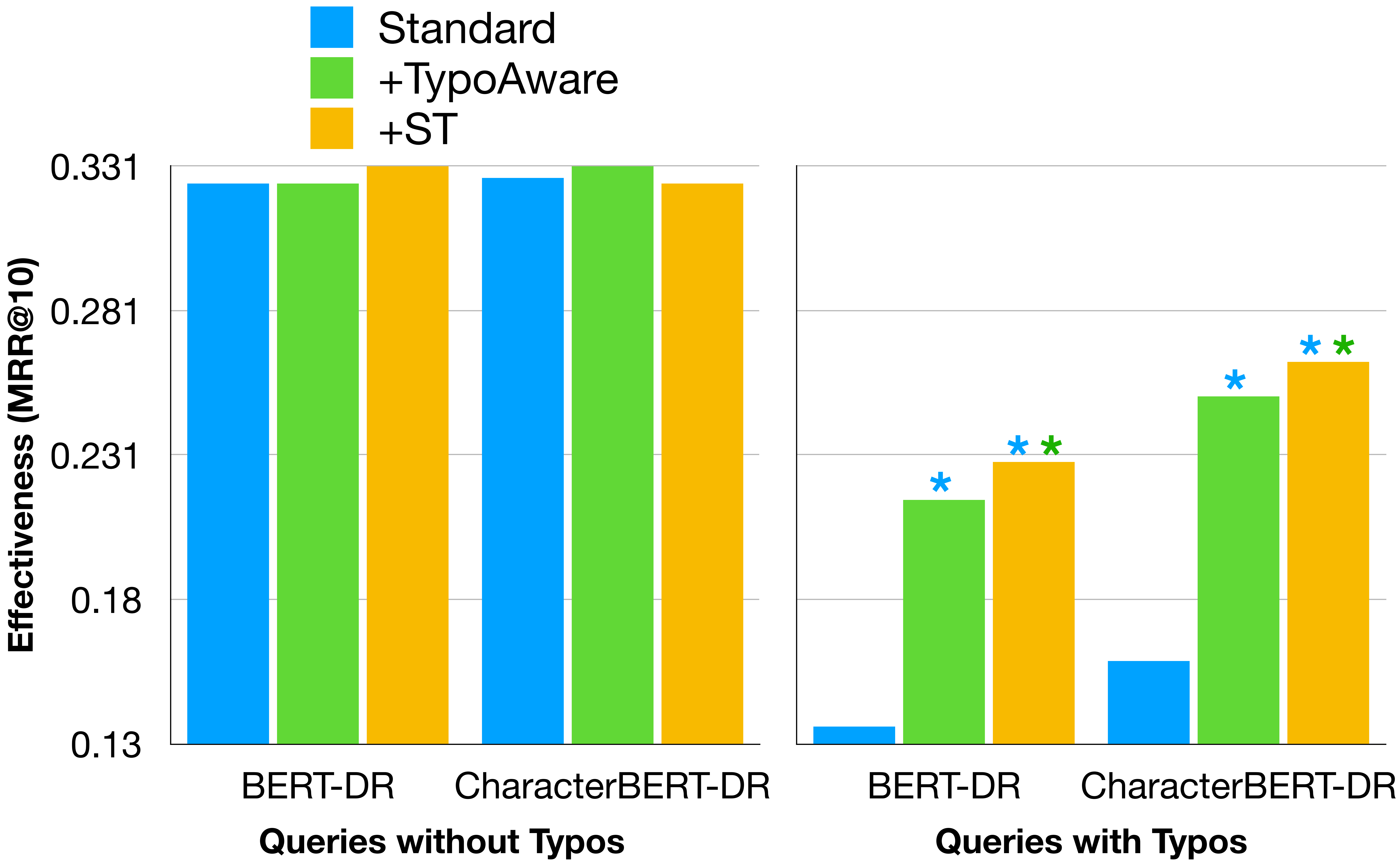

- TypoAware, ST do not provide significant differences on queries without typos: **no risk to use them**

# Does CharacterBERT+ST produce improvements on queries with typos?



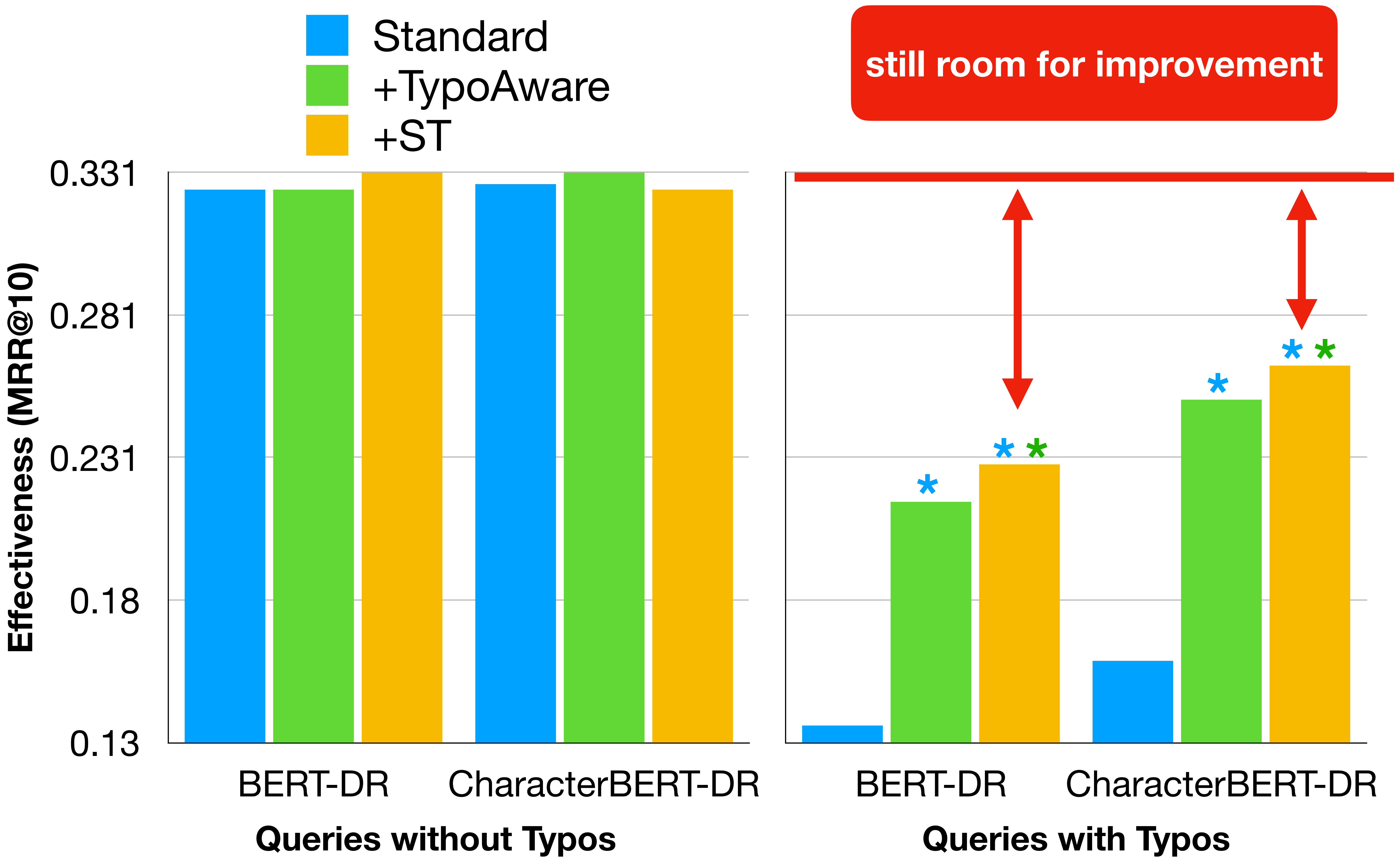

- TypoAware, ST do not provide significant differences on queries without typos: **no risk to use them**
- **ST provides largest gains** on queries with typos

# Does CharacterBERT+ST produce improvements on queries with typos?




- TypoAware, ST do not provide significant differences on queries without typos: **no risk to use them**
- **ST provides largest gains** on queries with typos
- Similar trend on **dataset we created with real typos**

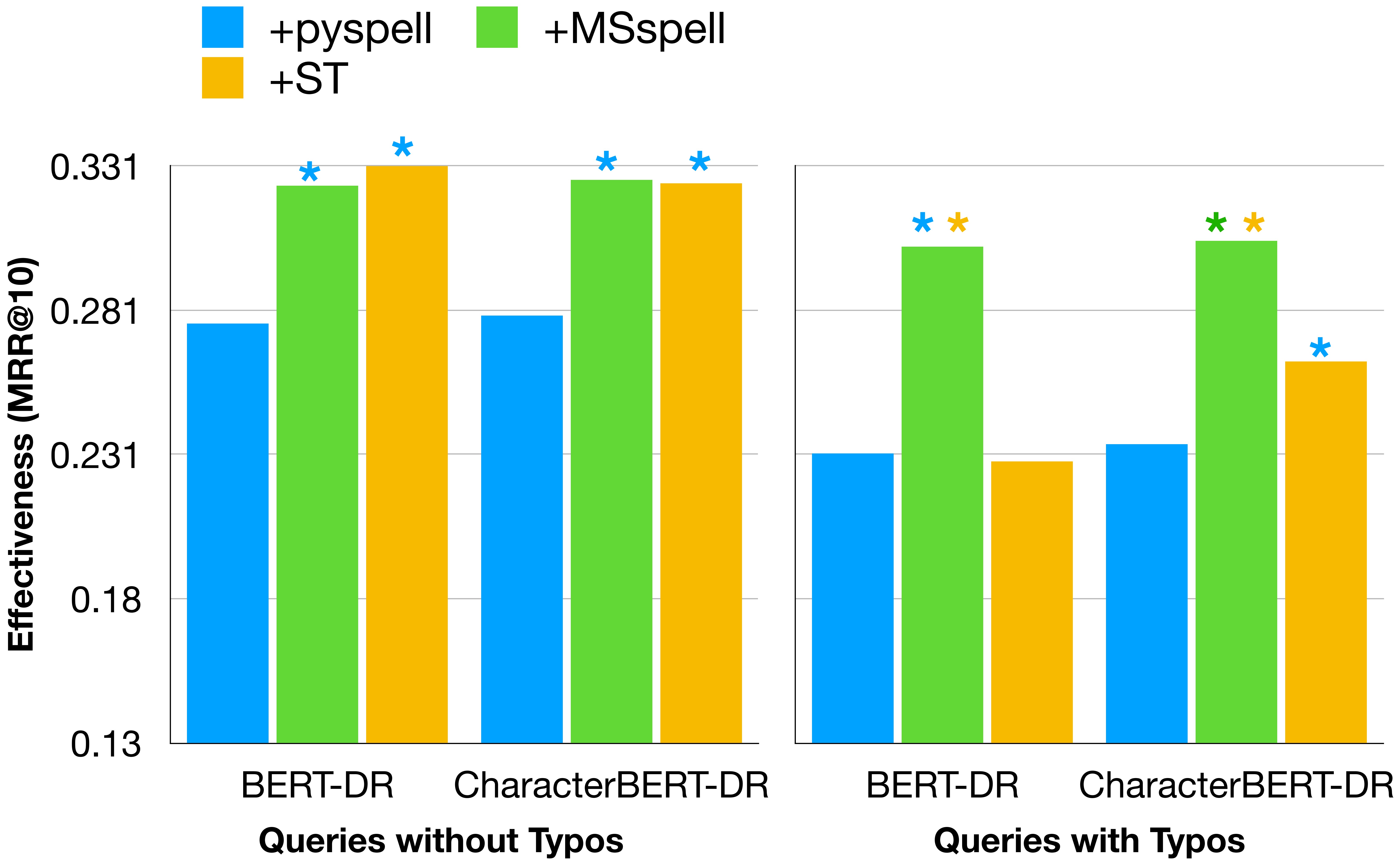
# Does CharacterBERT+ST produce improvements on queries with typos?




- TypoAware, ST do not provide significant differences on queries without typos: **no risk to use them**
- **ST provides largest gains** on queries with typos
- Similar trend on **dataset we created with real typos**

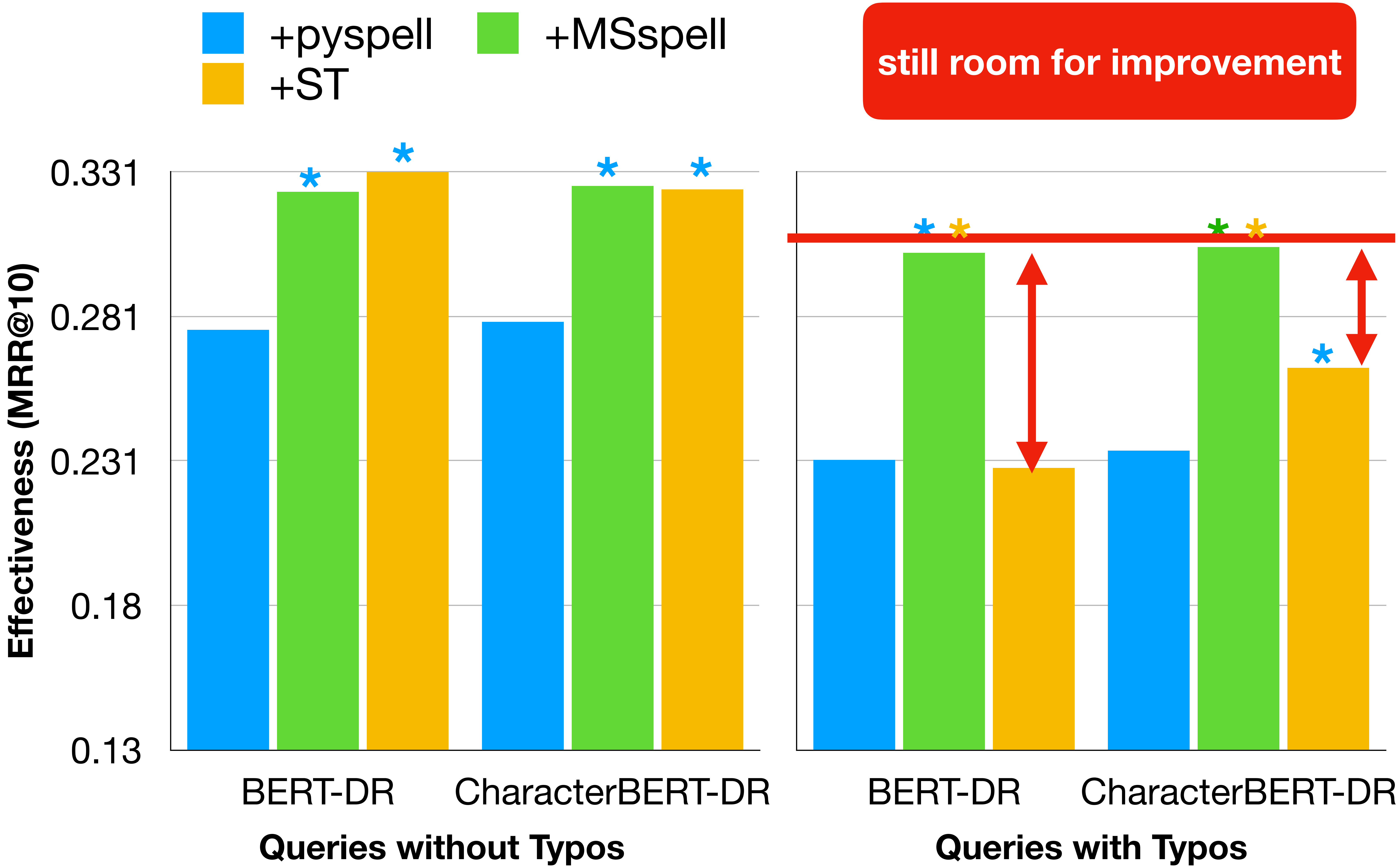
# What about using Spell Checkers instead?

# What about using Spell Checkers instead?




THE UNIVERSITY OF QUEENSLAND
AUSTRALIA
CREATE CHANGE

# What about using Spell Checkers instead?

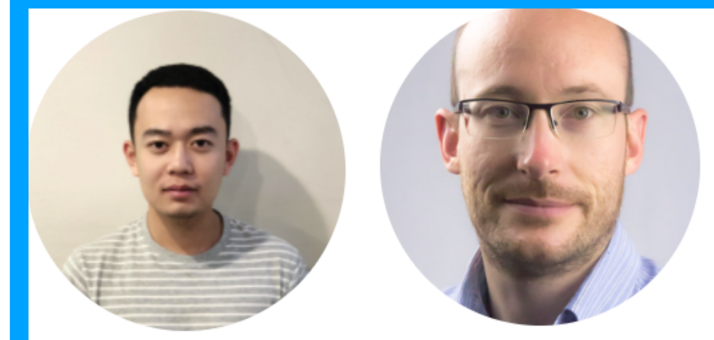



THE UNIVERSITY OF QUEENSLAND AUSTRALIA
CREATE CHANGE

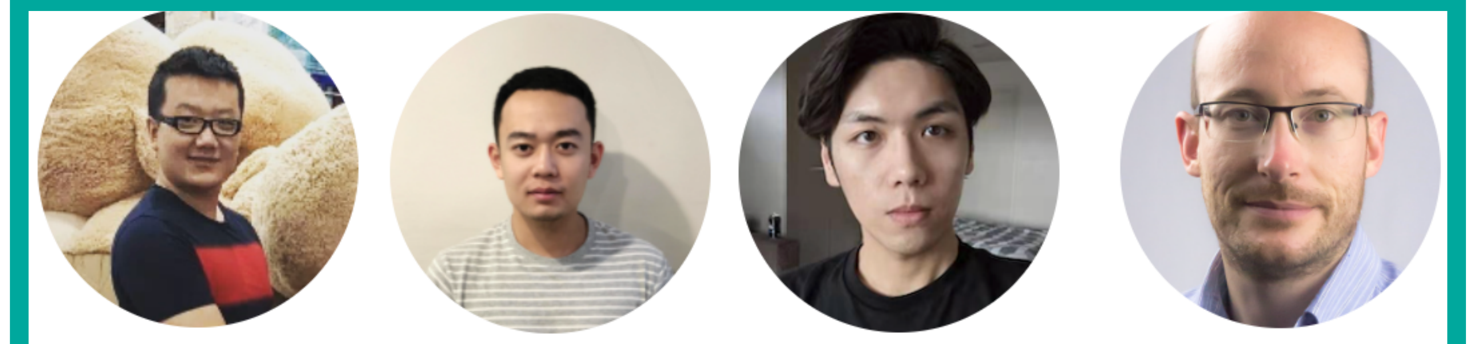
# Neural IR @ ielab

**Trade-off between effectiveness, efficiency and hardware**

**Robustness to out-of-distribution data**



**PRF integration, efficient PRF**



## Questions?

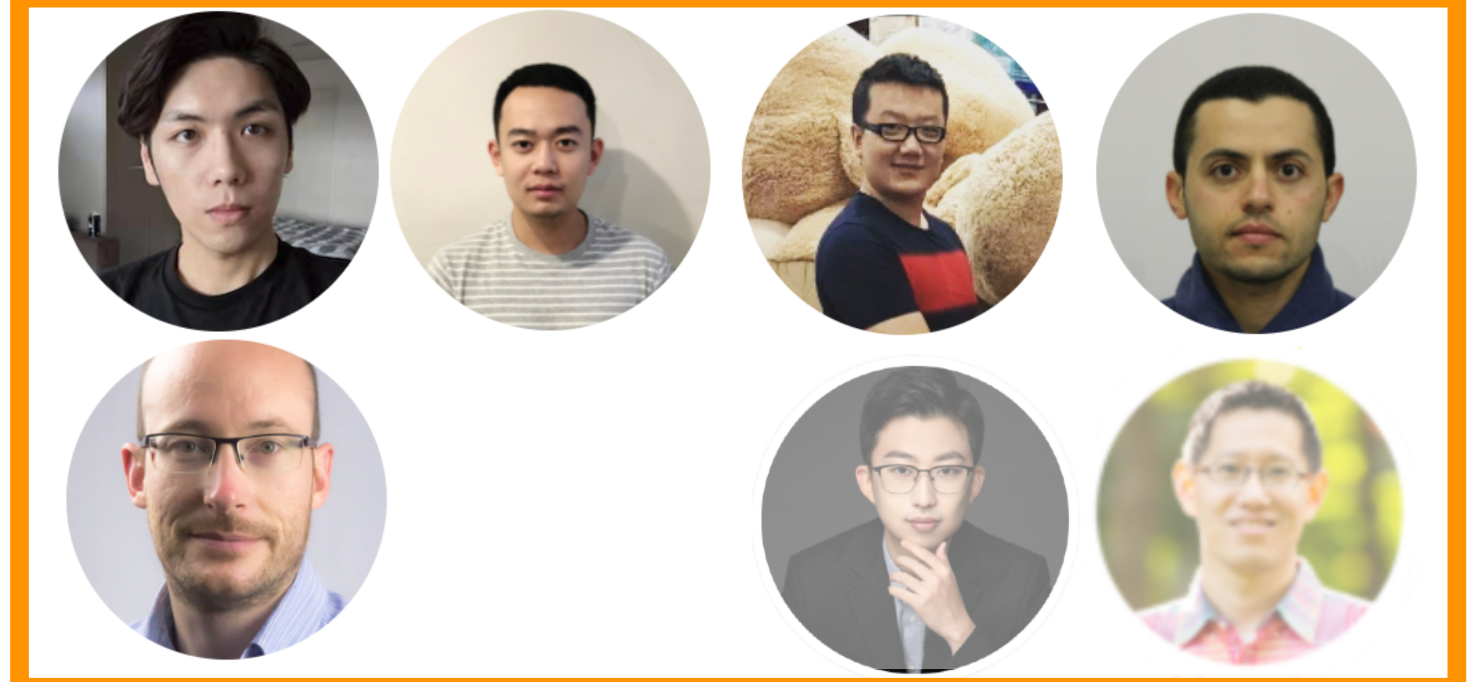
@guidozuc

https://ielab.io/guido

https://ielab.io/projects/transformers4ir.html

**Data & Training efficiency**

**Hybrid sparse-dense methods**



**BERT models in domain-specific tasks**