THE UNIVERSITY OF QUEENSLAND AUSTRALIA | CREATE CHANGE

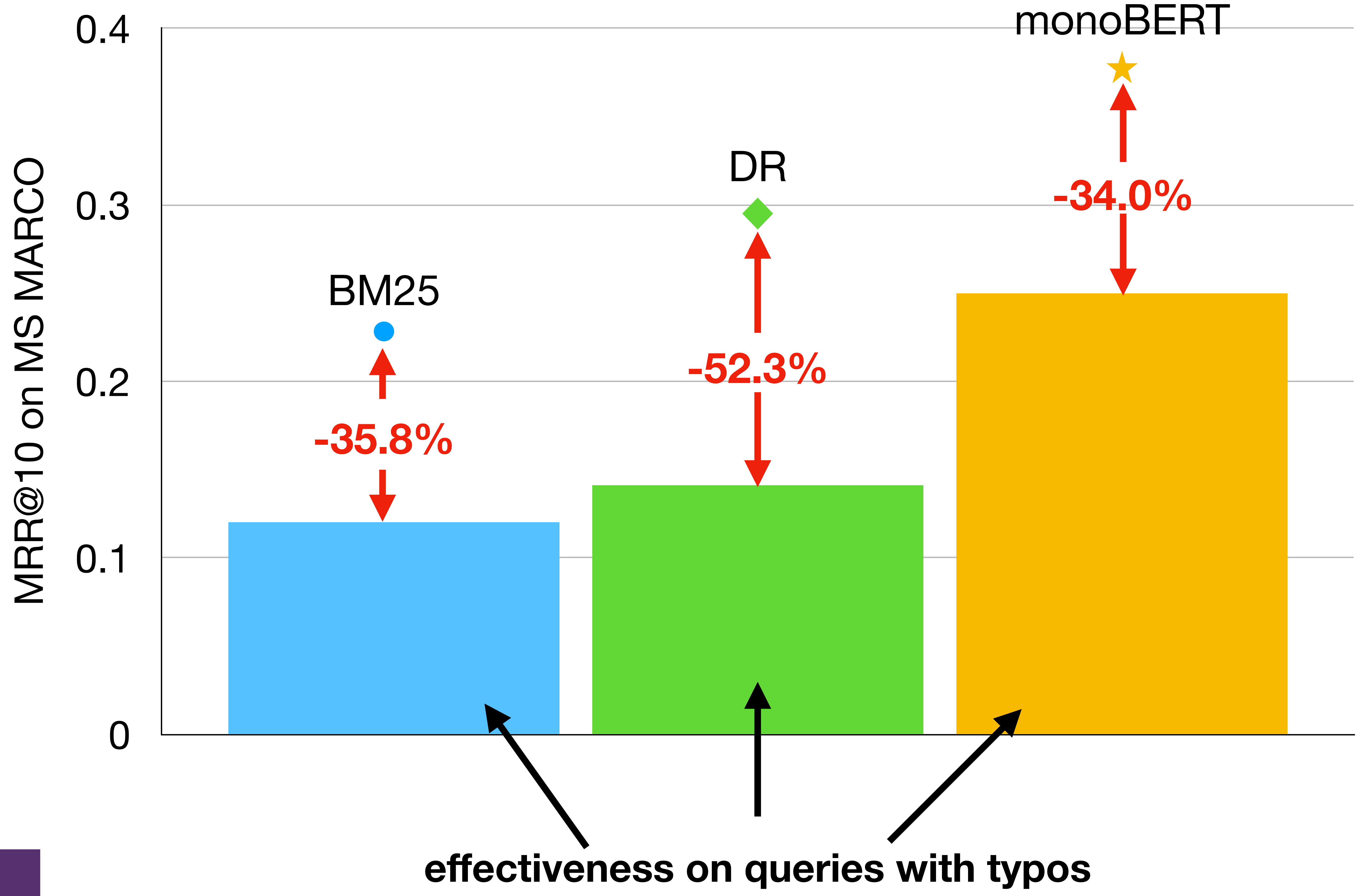
CharacterBERT and Self-Teaching for Improving the Robustness of Dense Retrievers on Queries with Typos

Shengyao Zhuang & Guido Zuccon

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

ie lab

BERT-based rankers cannot deal with queries with typos

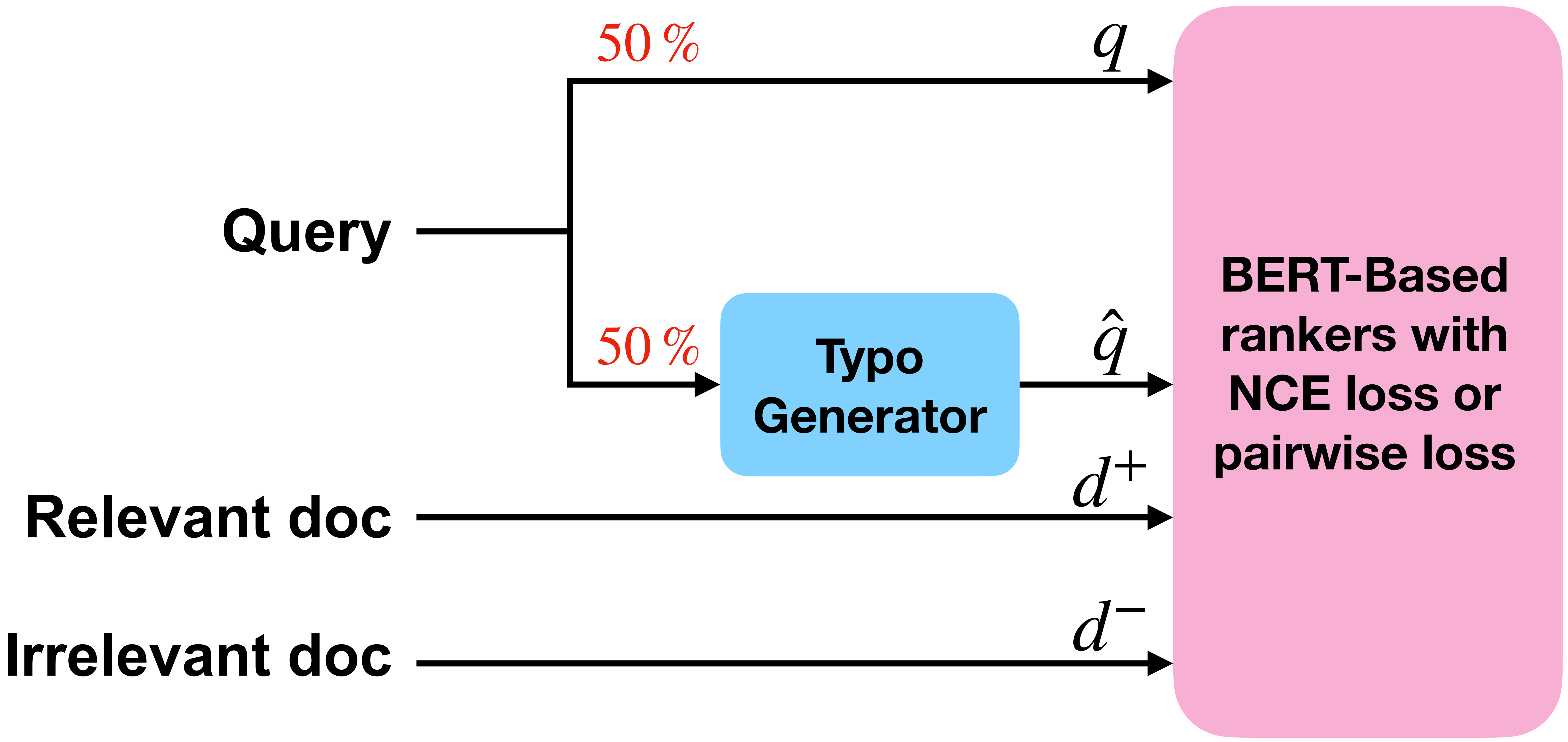


Both DR and monoBERT perform poorly on queries with typos

Zhuang, Zuccon, *"Dealing with Typos for BERT-based Passage Retrieval and Ranking*", EMNLP 2021

ie
lab

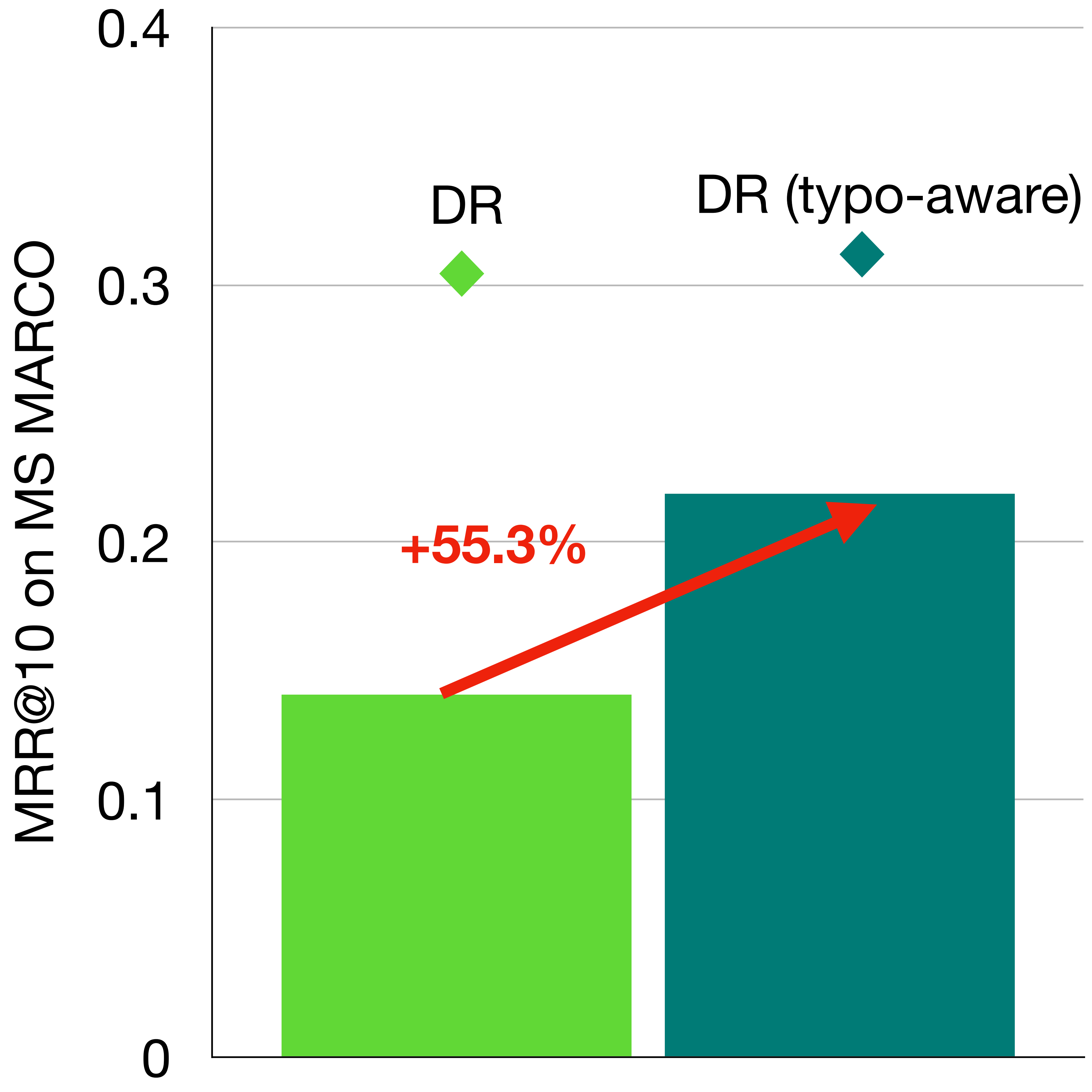
Prior work: Typos-aware training



Data Augmentation strategy: inject typos in training queries

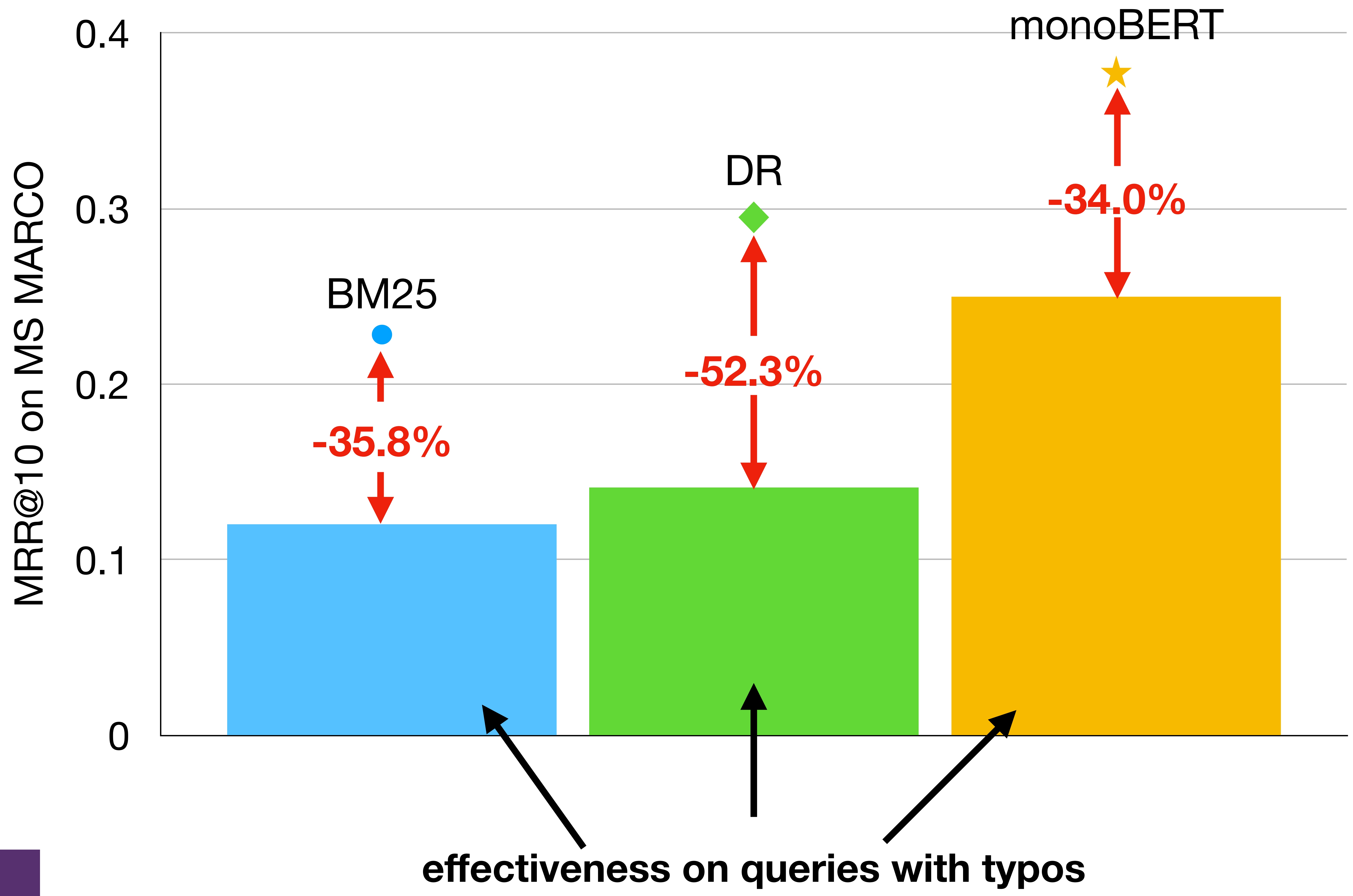
Zhuang, Zuccon, *"Dealing with Typos for BERT-based Passage Retrieval and Ranking*", EMNLP 2021

Prior work: Typos-aware training



- Data Augmentation improves effectiveness
  - Typo-aware training as effective as traditional training on queries without typos
  - Typo-aware training more effective on queries with typos

In this paper…



- Focus on **Dense Retrievers**
- **Why** Dense Retrievers cannot deal with typos in queries?
- **How** can we **improve** Dense Retriever's robustness to typos in 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]**

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

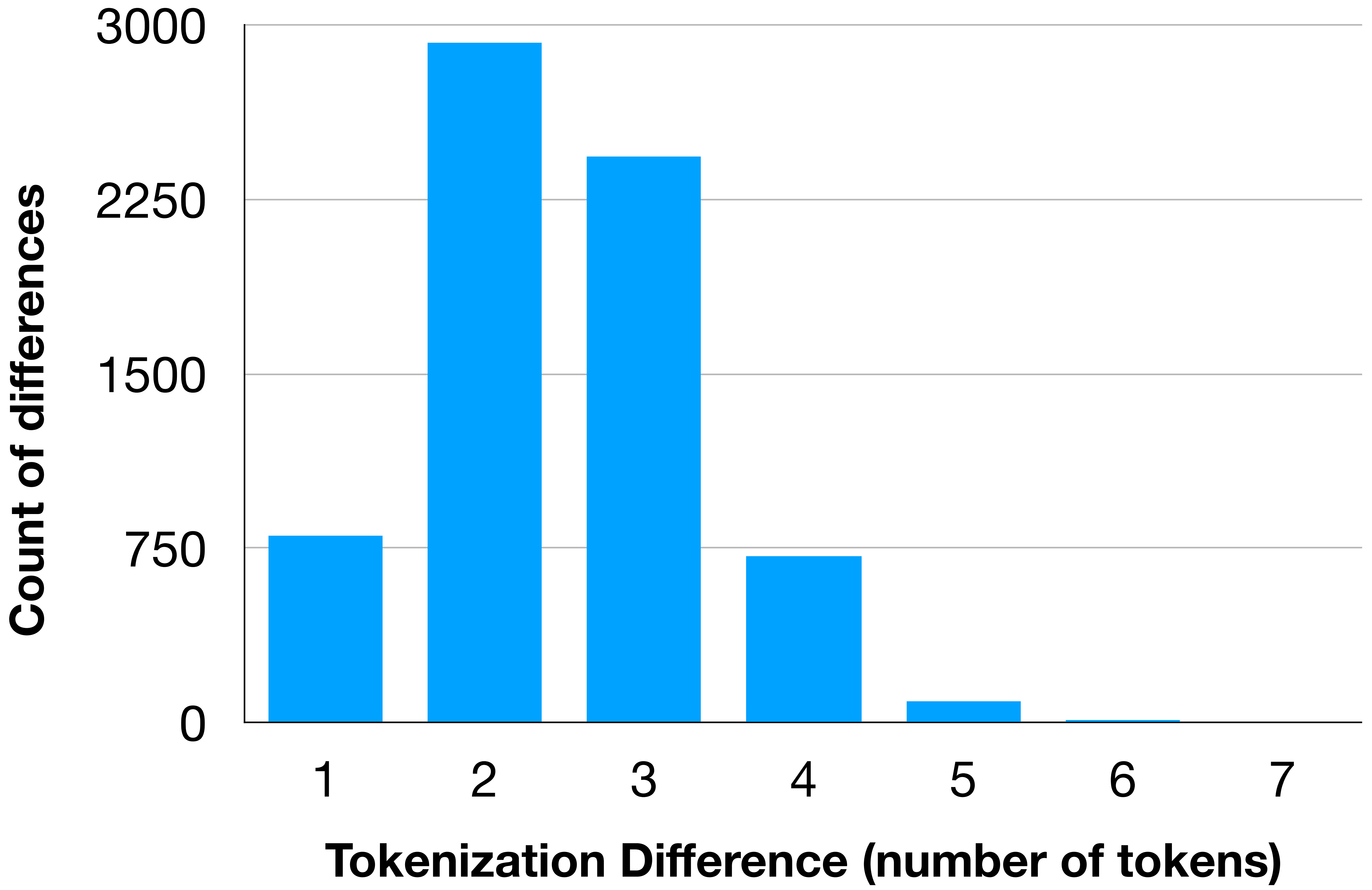
- What is the difference in output from the WordPiece Tokenization in presence of a typo?

‘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?

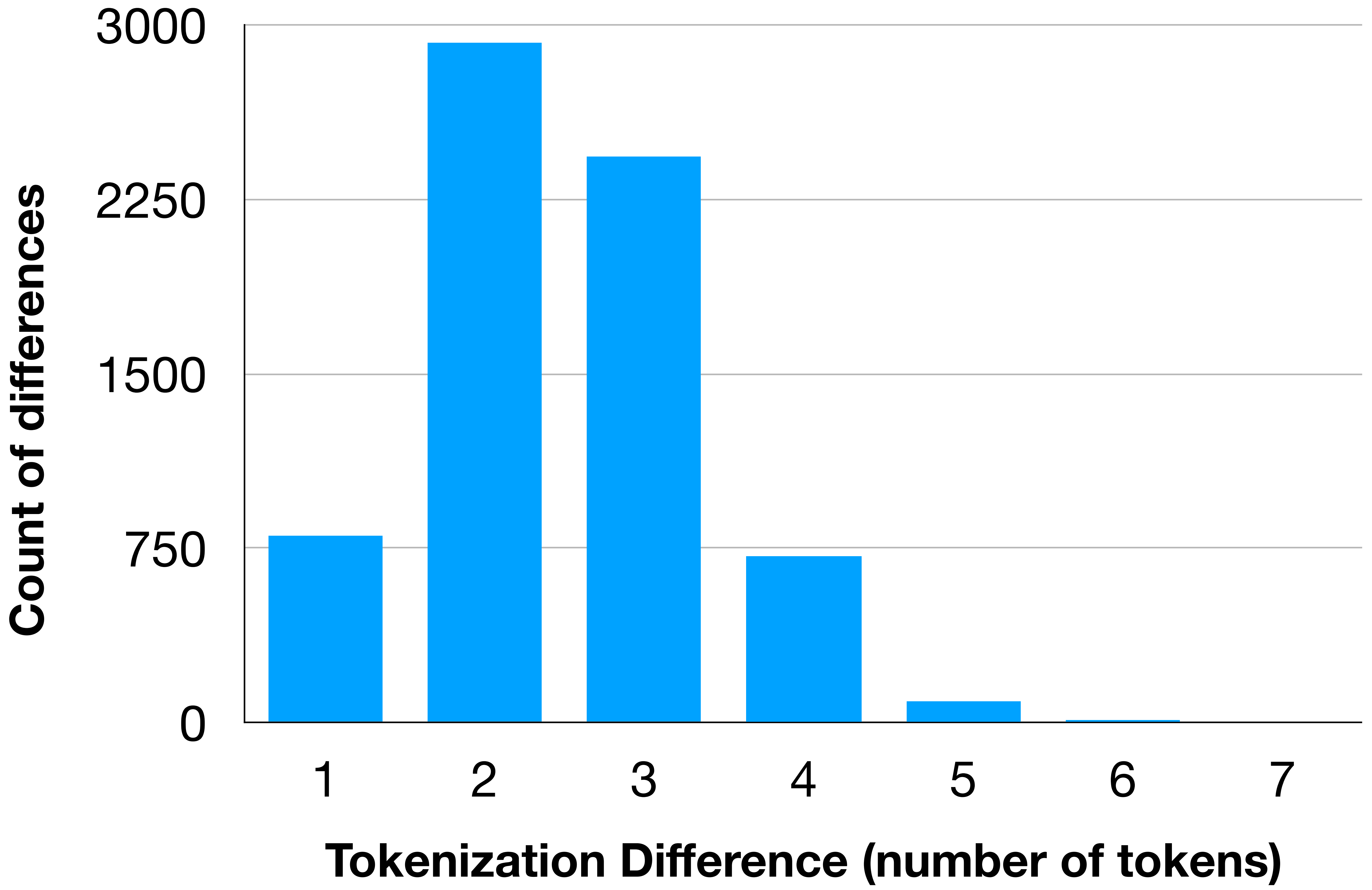


ie lab

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

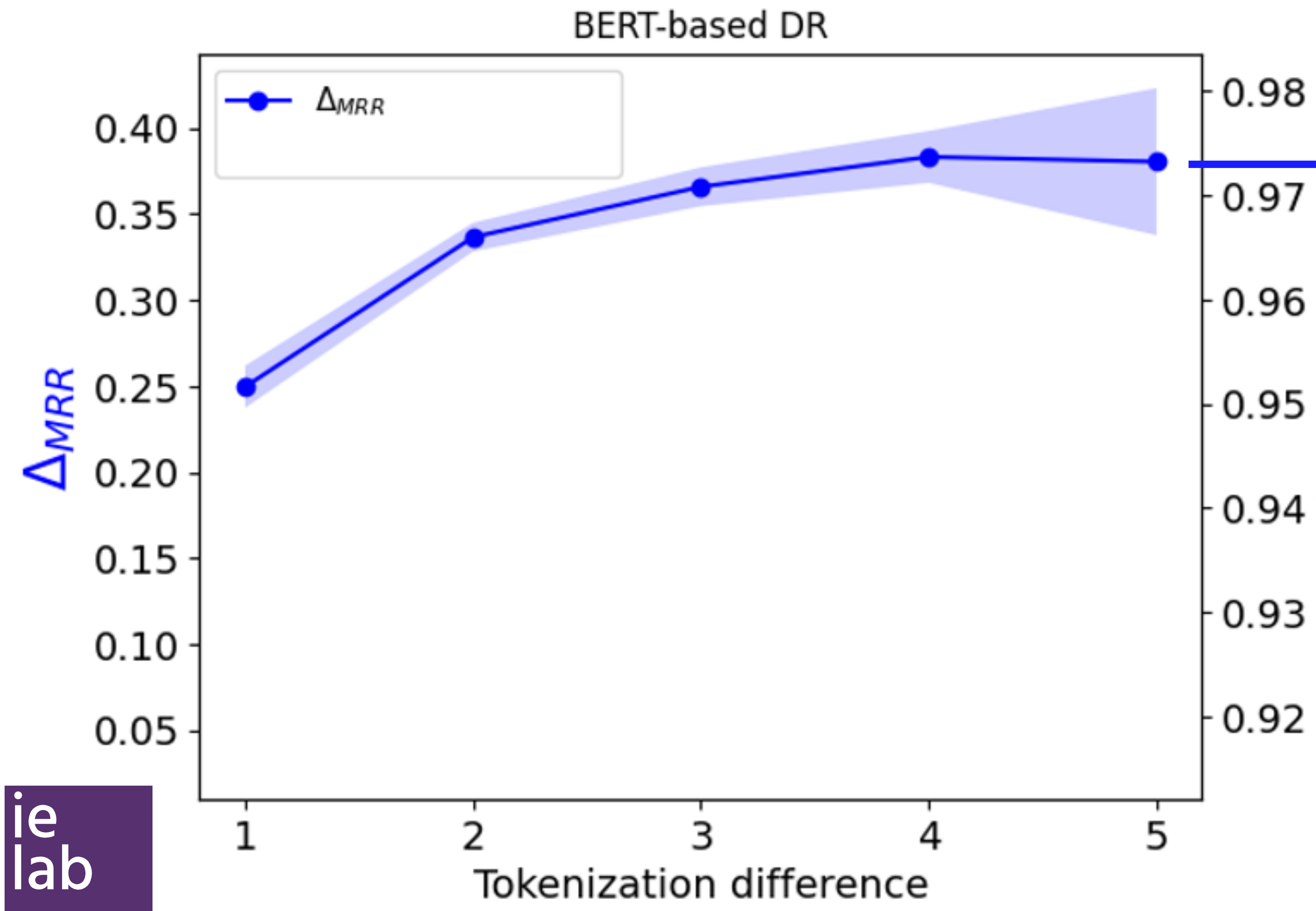
CREATE CHANGE

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



What are the effects of such tokenization difference?

Tokenization Difference produces effectiveness losses



Effectiveness drops

=

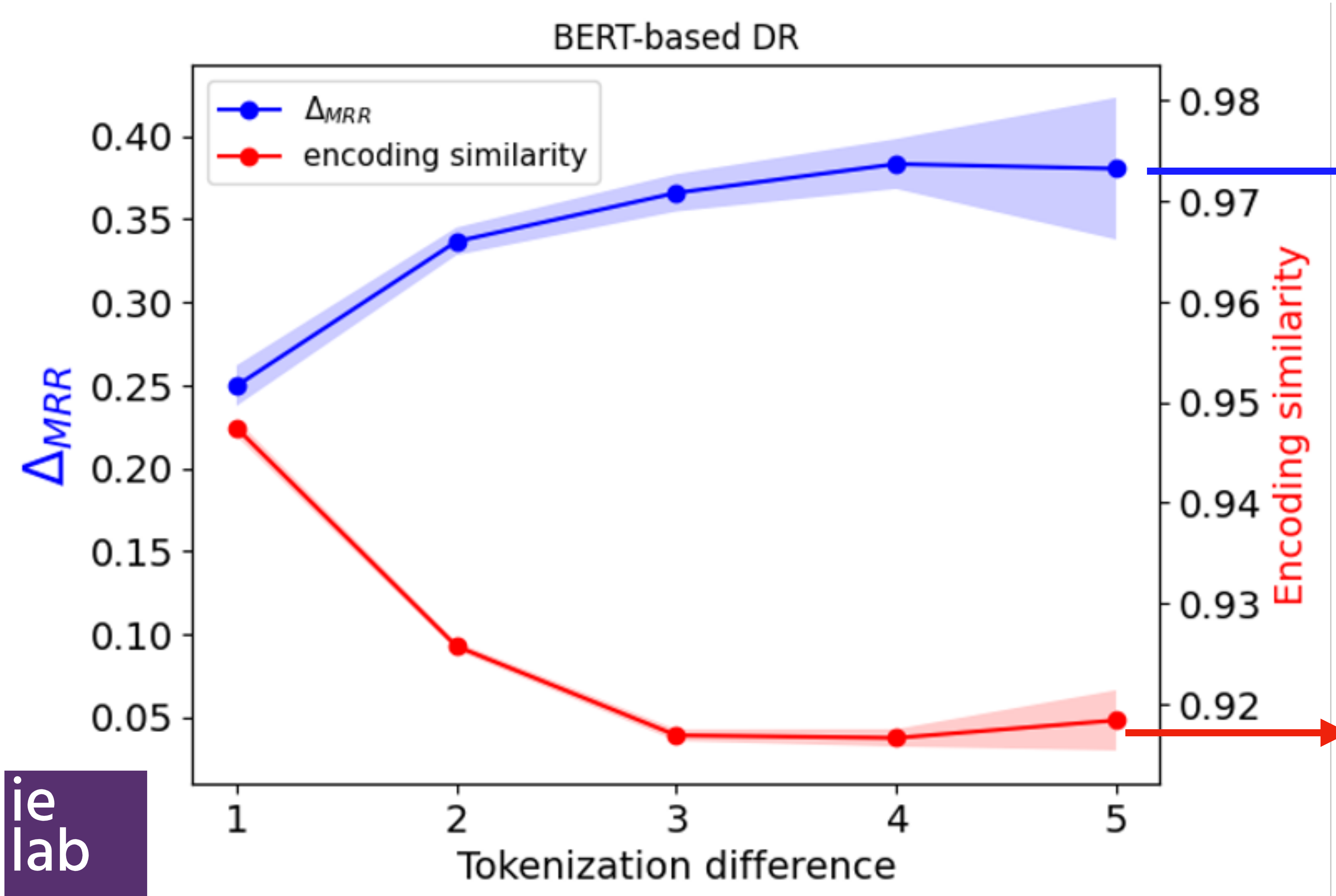
The higher the loss, the worse the DR does on typos

ielab

The University Of Queensland Australia

Create Change

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



Effectiveness drops

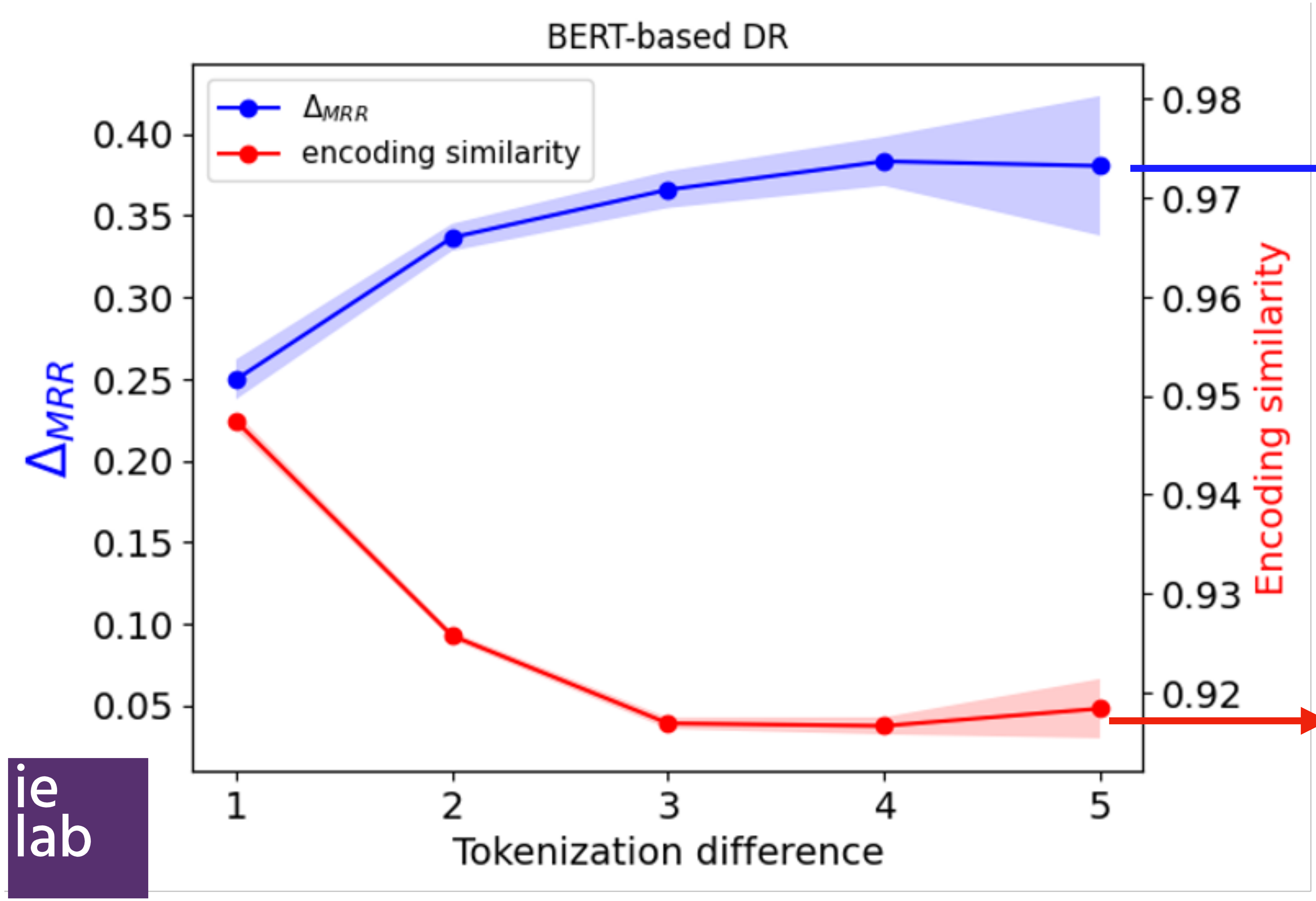
Representation with typos diverges from original

ie lab

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

CREATE CHANGE

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

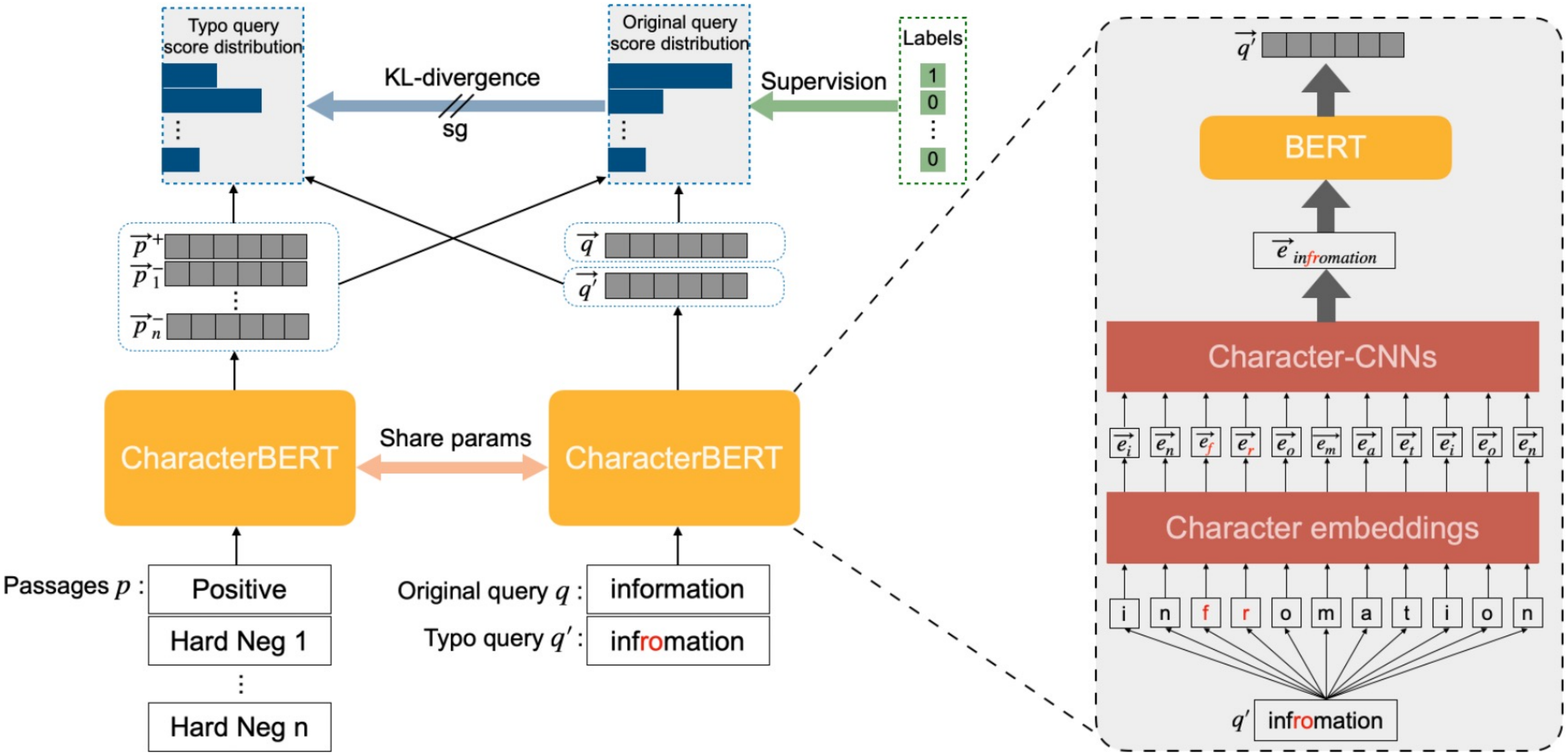


Effectiveness drops

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

Representation with typos diverges from original

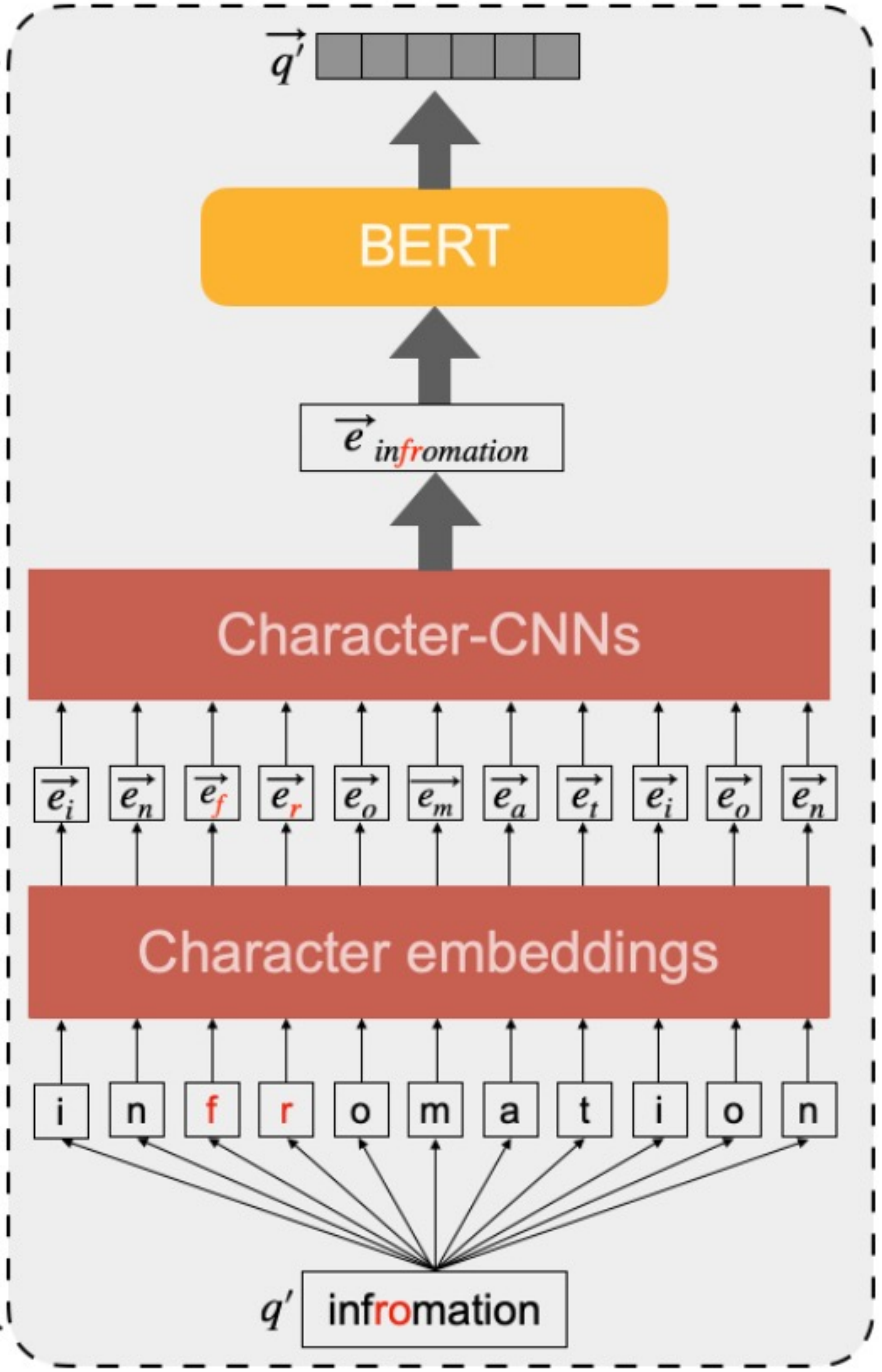
CharacterBERT-DR + Self-Teaching



CharacterBERT-DR + Self-Teaching

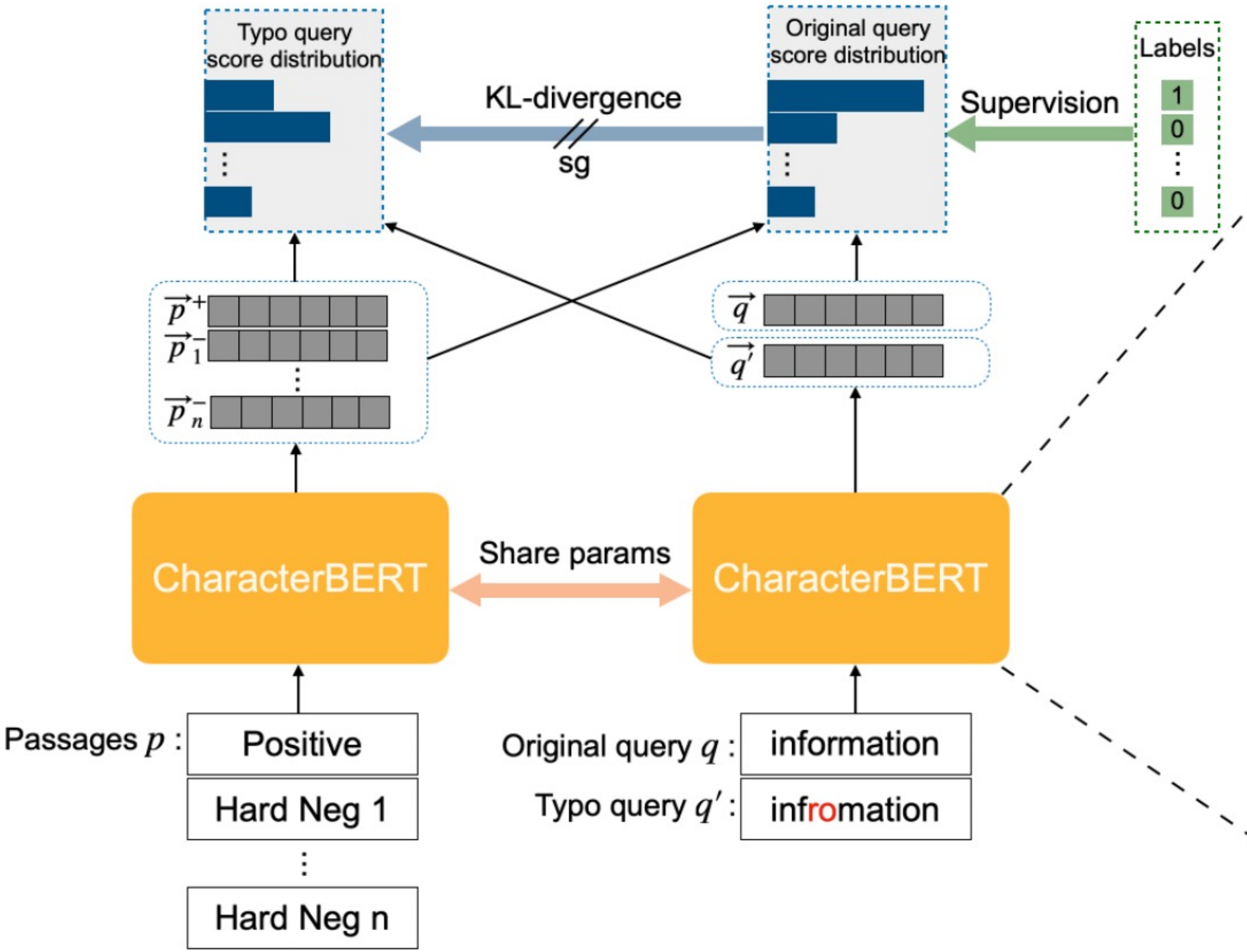
- Replace BERT WordPiece Tokenizer with CharacterBERT [2] to create query and passages embeddings.

- 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.

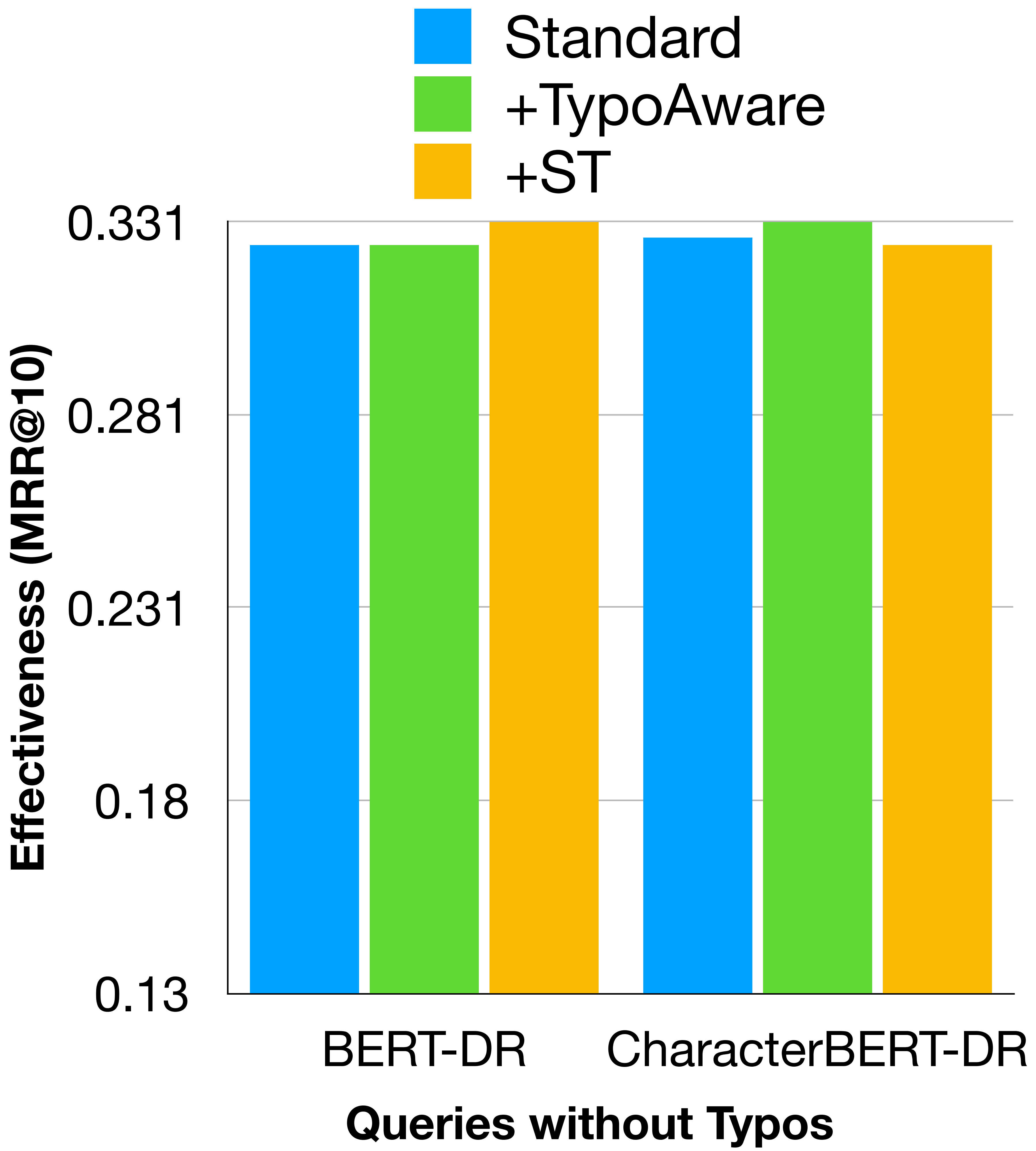
2. Self-Teaching: minimize score distribution difference b/w original query & query with typos:

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

3. Supervised contrastive loss:

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

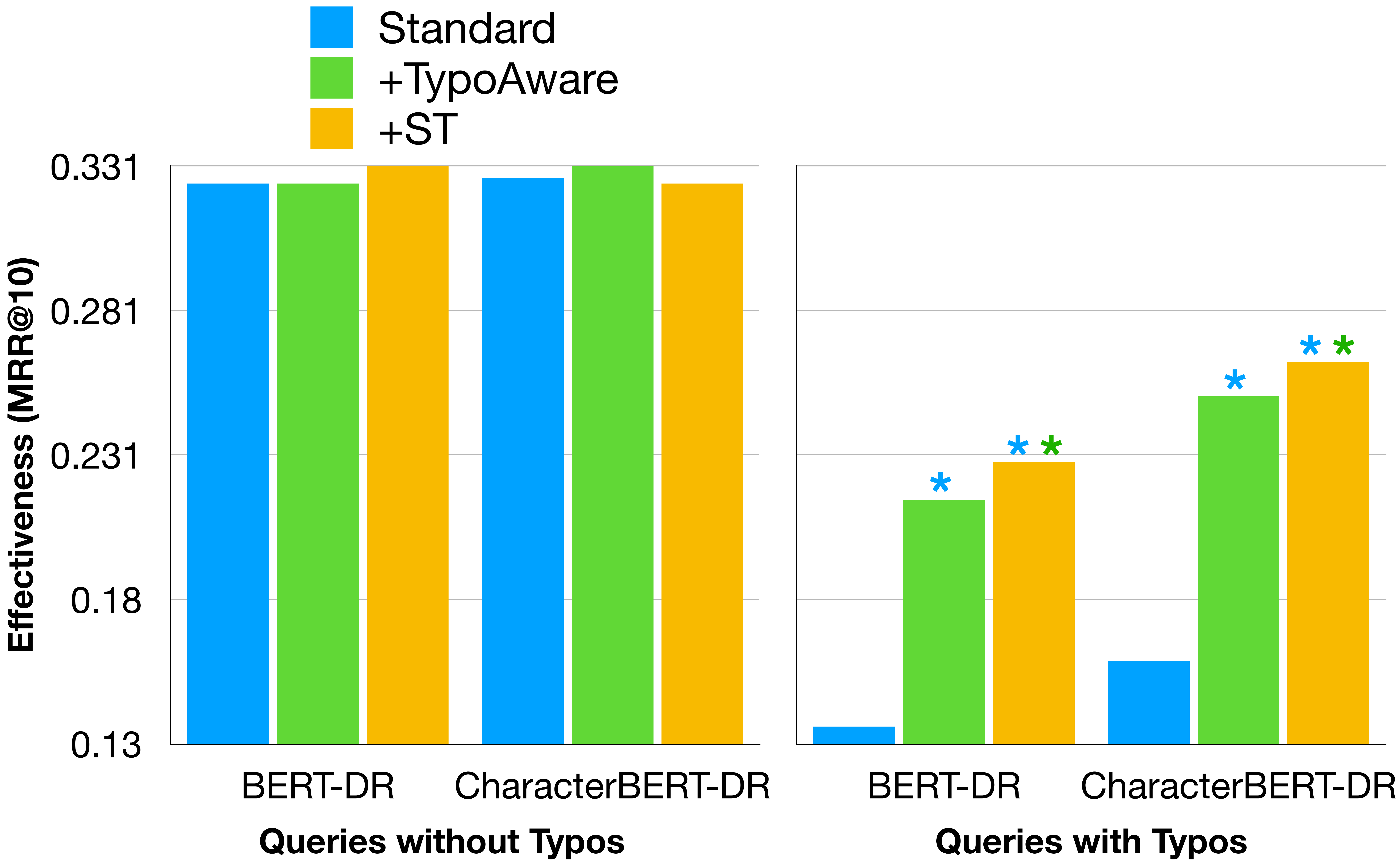
Does CharacterBERT+ST produce unwanted effect on queries w/o typos?



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

THE UNIVERSITY OF QUEENSLAND AUSTRALIA
CREATE CHANGE

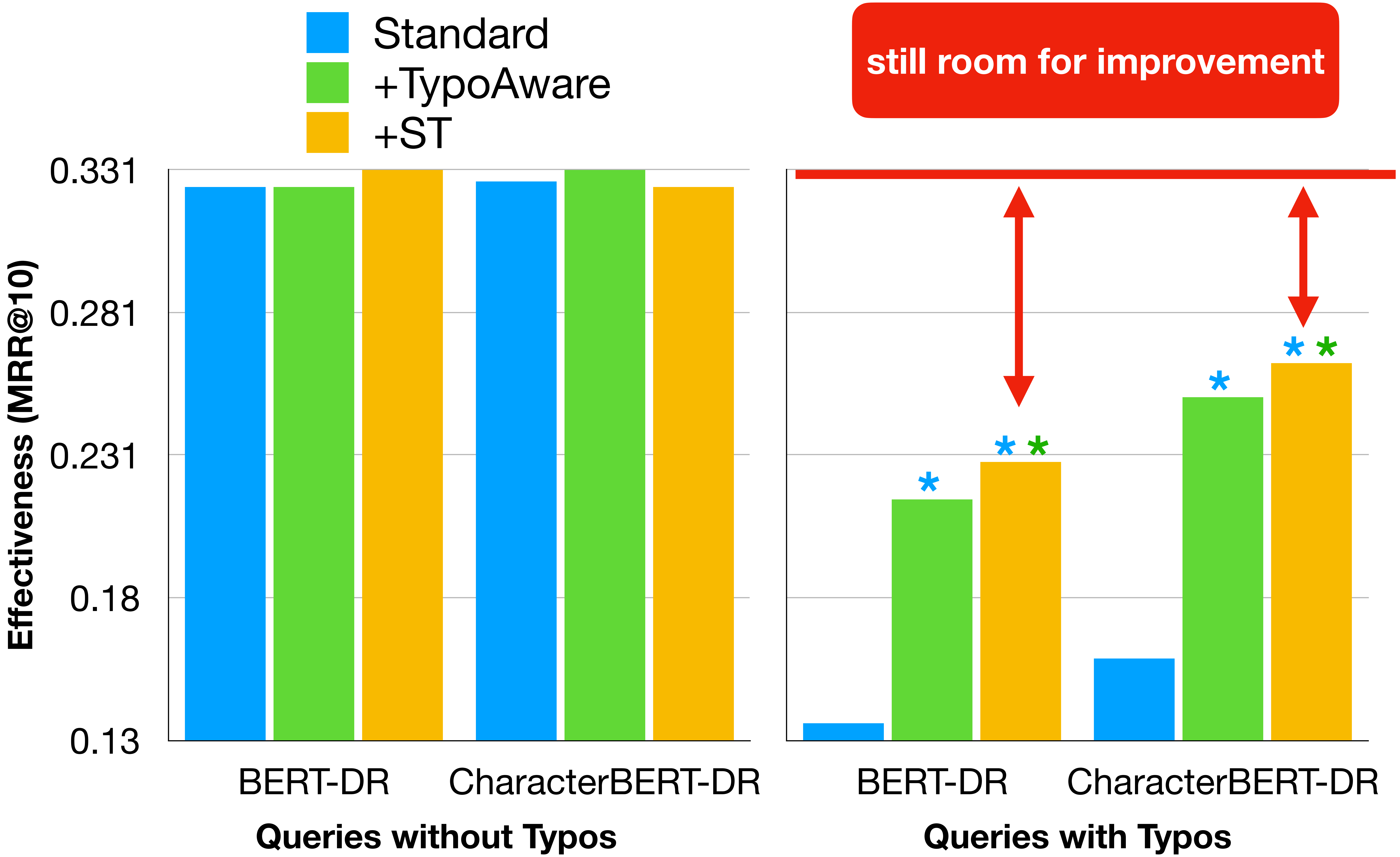
Does CharacterBERT+ST produce improvements on queries with typos?



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

THE UNIVERSITY OF QUEENSLAND AUSTRALIA
CREATE CHANGE

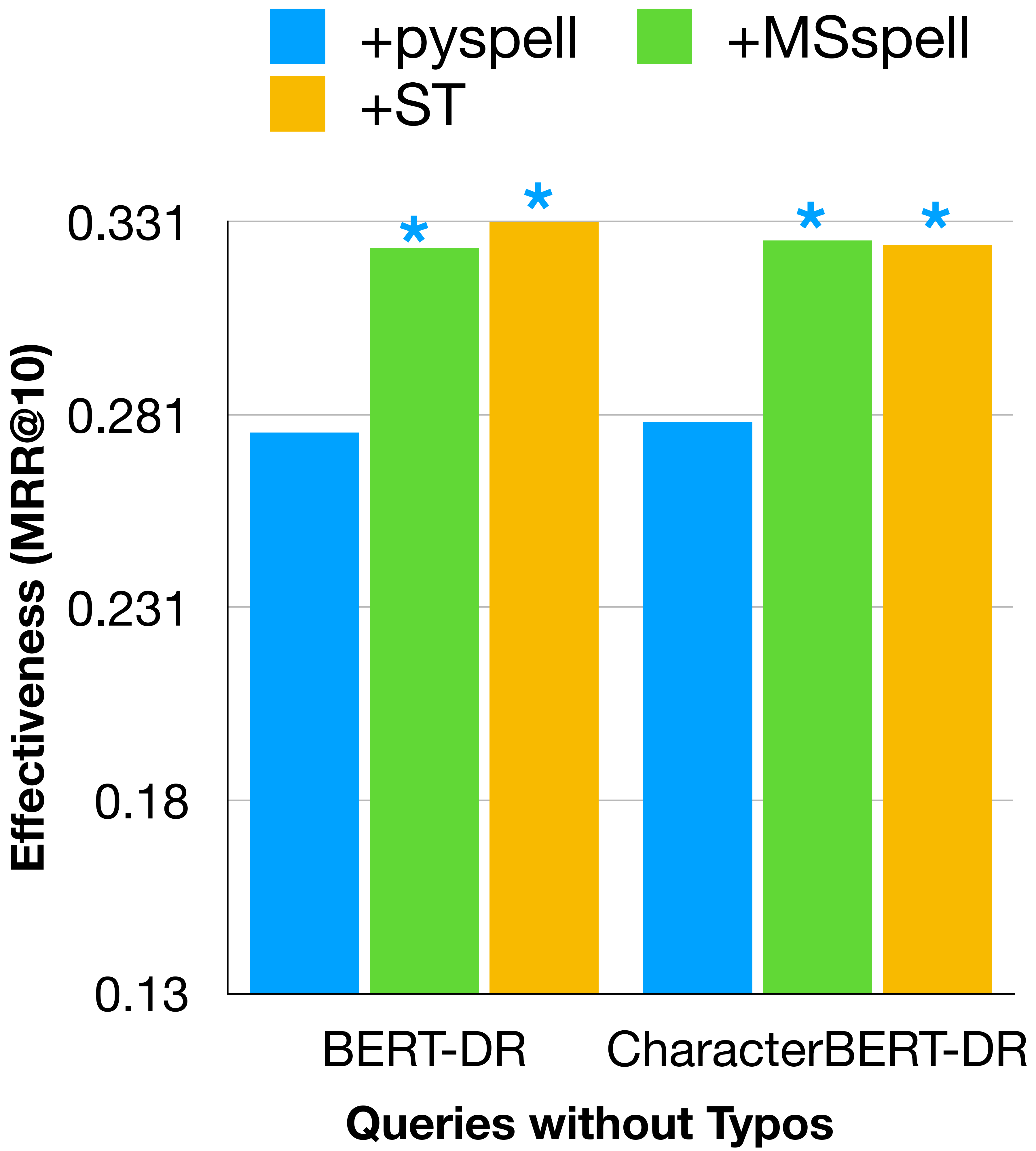
Does CharacterBERT+ST produce improvements on queries with typos?



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

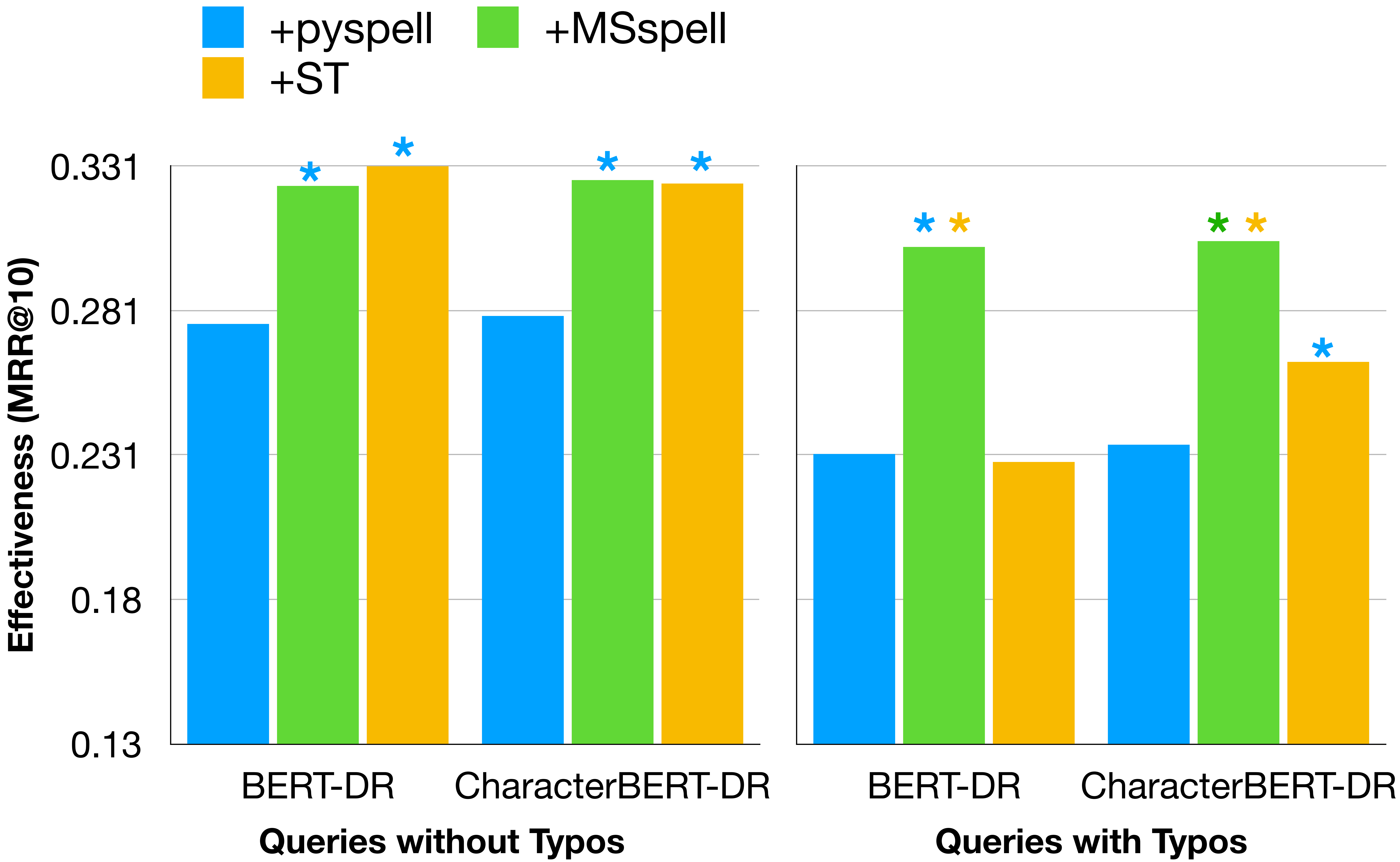
What about using Spell Checkers instead?

What about using Spell Checkers instead?



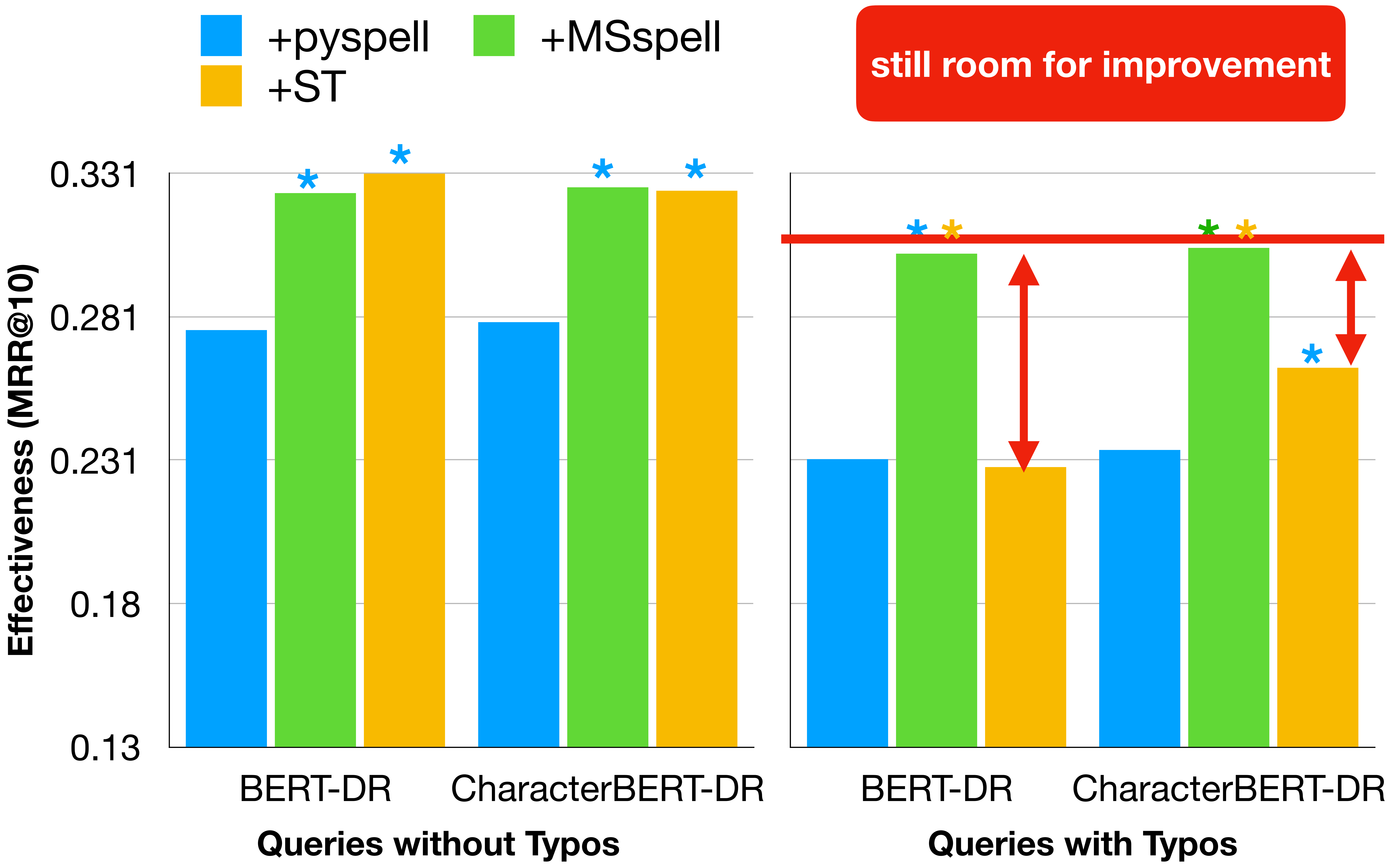
THE UNIVERSITY OF QUEENSLAND
AUSTRALIA
CREATE CHANGE

What about using Spell Checkers instead?



- MSspell provides significantly higher effectiveness
  - Most likely leverages extensive training data
- CharacterBERT+ST better than rule-based spell checker (pyspell)
- Engineering advantages in end-to-end DR pipeline rather than additional spell checker

What about using Spell Checkers instead?



- MSspell provides significantly higher effectiveness
    - Most likely leverages extensive training data
- CharacterBERT+ST better than rule-based spell checker (pyspell)
- Engineering advantages in end-to-end DR pipeline rather than additional spell checker

THE UNIVERSITY OF QUEENSLAND AUSTRALIA
CREATE CHANGE

Take-aways

- Typical Dense Retrievers do not perform well on queries with typos. Augmentation at training (EMNLP 2021) only goes so far
- The **key issue is in how a word with typos vs. without-typo is represented**
  - Bringing these two representations closer improves effectiveness
  - BERT’s Tokenizer major source for representation differences
- Replacing tokenizer with CharacterBERT encoder and using ST to further bring representations close drastically improves robustness of Dense Retrievers
- We also provide a **new dataset** for evaluation with real queries with typos

THE UNIVERSITY OF QUEENSLAND AUSTRALIA | CREATE CHANGE

Additional Material

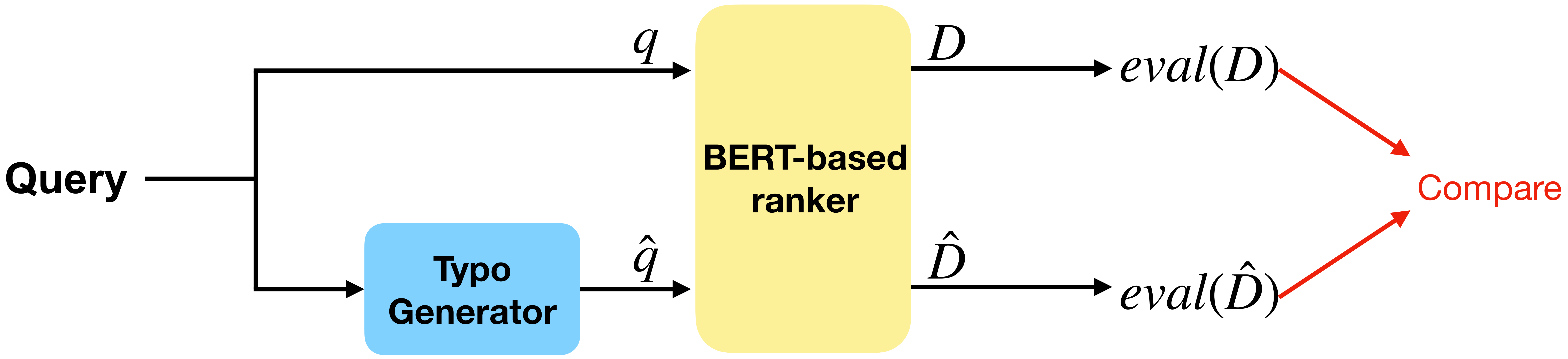
CharacterBERT and Self-Teaching for Improving the Robustness of Dense Retrievers on Queries with Typos

Shengyao Zhuang & Guido Zuccon

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

ie lab

Evaluation with typo generator



Typo query generation

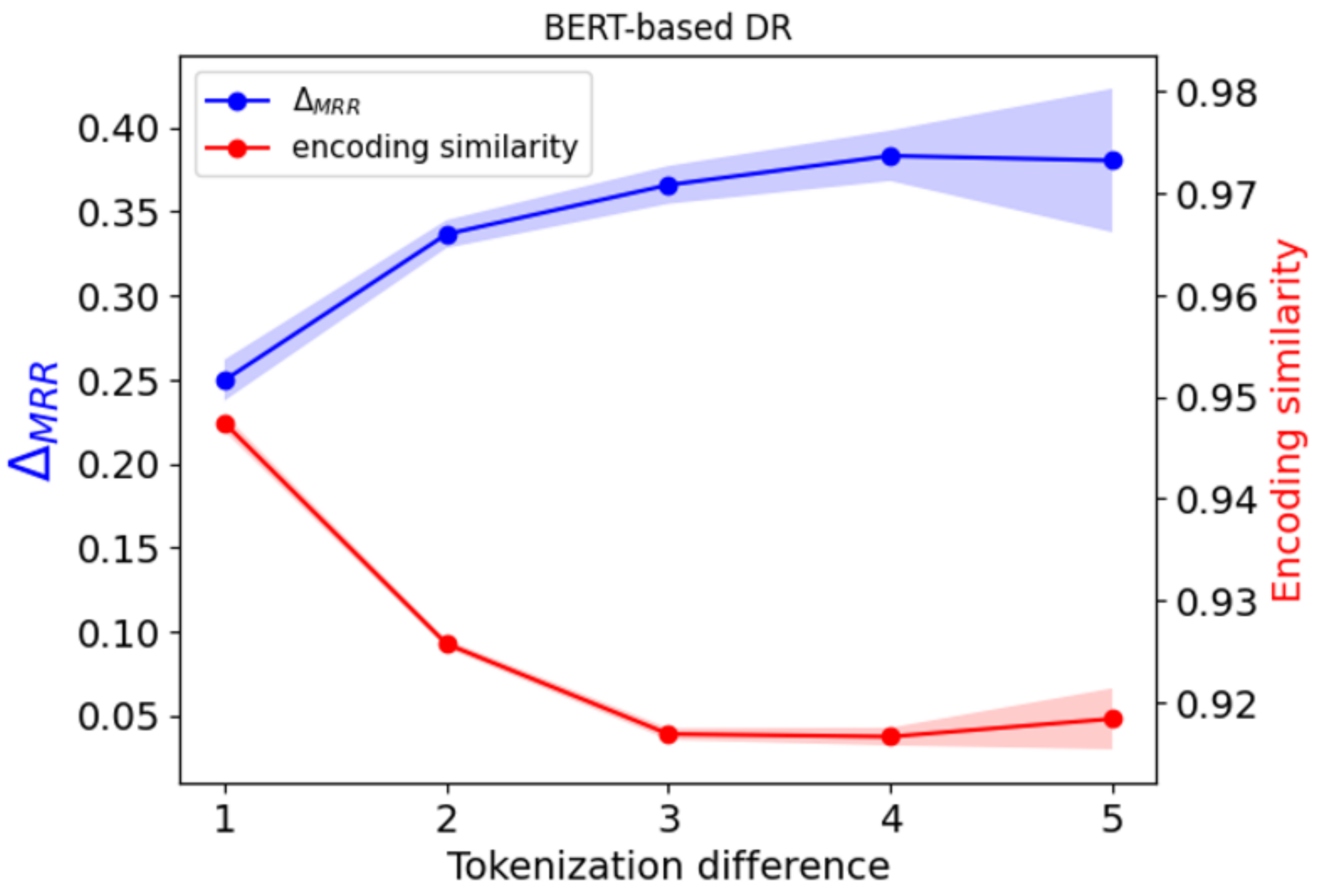
- Synthetic Typo generation for MS MARCO queries
  - Random character Insertion: ‘typo’ -> ‘tyapo’
  - Random character deletion: ‘typo’ -> ‘tyo’
  - Random character substitution: ‘typo’ -> ‘type’
  - Swap neighbour character: ‘typo’ -> ‘tyop’
  - Swap adjacent keyboard character: ‘typo’ -> ‘typi’
- These are common typos in real-world user queries [1]

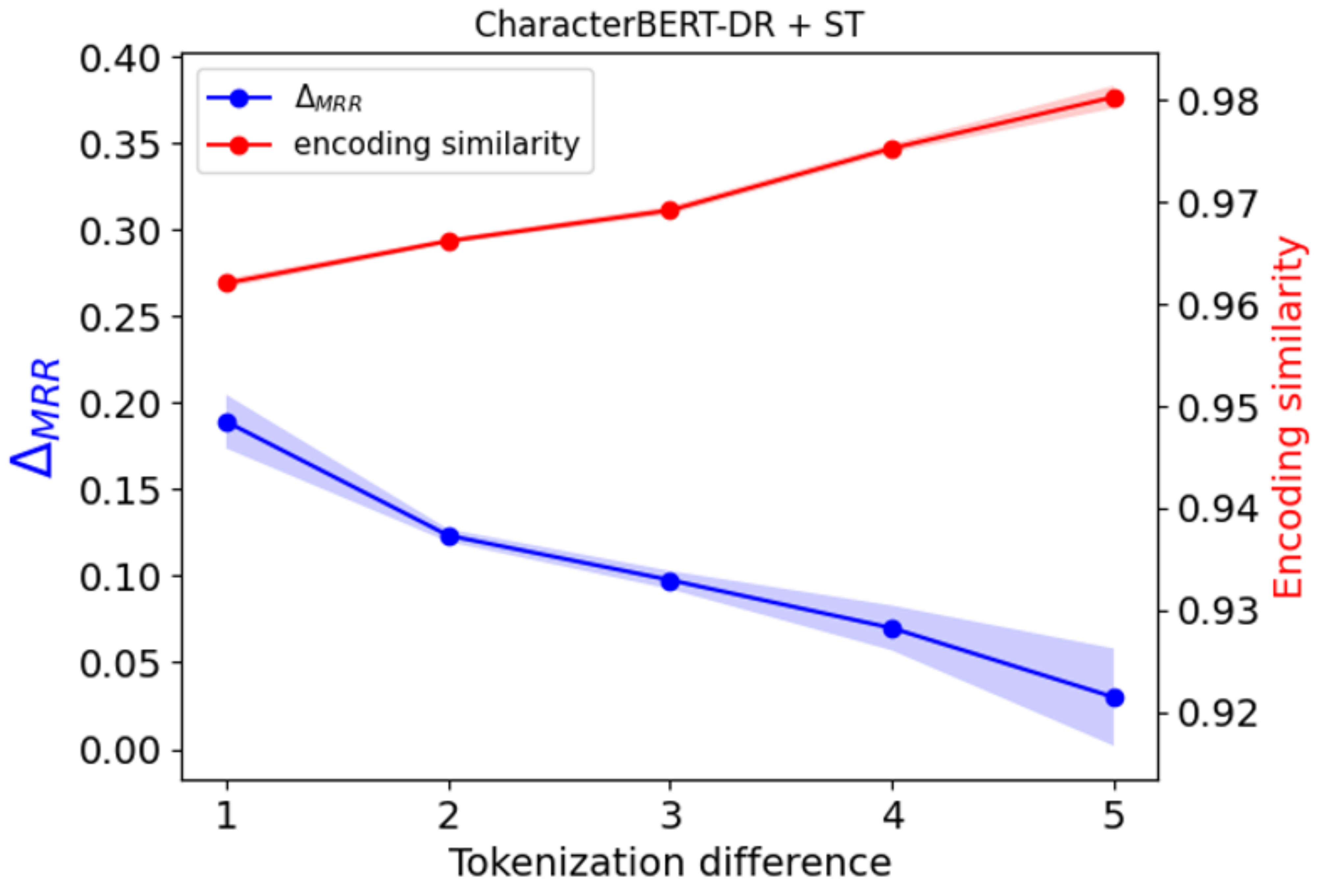
ie lab

[1] Matthias Hagen, Martin Potthast, Marcel Gohsen, Anja Rathgeber, and Benno Stein. 2017. A large-scale query spelling correction corpus. SIGIR

The University of Queensland
Australia
Create Change

CharacterBERT-based DR behaves differently





Similar trends in other datasets

- We experimented with **MS MARCO** (dev queries), **TREC 2019** and **2020**
- We created a **new dataset**, DL-typo
  - 60 queries with typos from AOL query log
    Human spelling corrections
  - Relevance assessments following TREC DL relevance criteria
    On average 63.52 judgements per query (relevant: 25.7)