The University Of Queensland
AUSTRALIA
CREATE CHANGE

A Reproducibility Study of Question Retrieval for Clarifying Questions

Sebastian Cross, **Guido Zuccon**, and Ahmed Mourad
g.zuccon@uq.edu.au
ielab, The University of Queensland, Australia
www.ielab.io

ie
lab

Clarifying Questions

- Creating a single query that is complex and detailed enough to retrieve the required information accurately is a difficult task
- Systems designed to assist the user with query formulation
    - Clarifying questions is such an approach
- Goal: identify a user's information-seeking intent by posing a clarifying question to the user, expecting their answer to clarify aspects of their query.

Clarifying Questions Increasingly Useful Feature for Conversational Search (and beyond)

From: Aliannejadi, et al., "Asking clarifying questions in open-domain information-seeking conversations.", SIGIR 2019

Useful

Zamani et al. WWW'20: asking clarifying questions is useful in **web search**

Zou et al. CIKM'20: question-based systems helpful towards **completing tasks**

Signals

Lotze et al. ECIR'21: exploit predicted **user engagement** with clarification pane

Bi et al. SIGIR'21: clarifying questions from **negative feedback**

Zhao et al. SIGIR'22: Generate clarifying questions from **web search results**

How to Generate

Sekulić et al. ICTIR'21: **GPT-2** to generate clarifying questions with respect to query and facets

Wang&Li CIKM'22: **Template-guided** clarifying question generation

ie lab

The University of Queensland Australia
Create Change

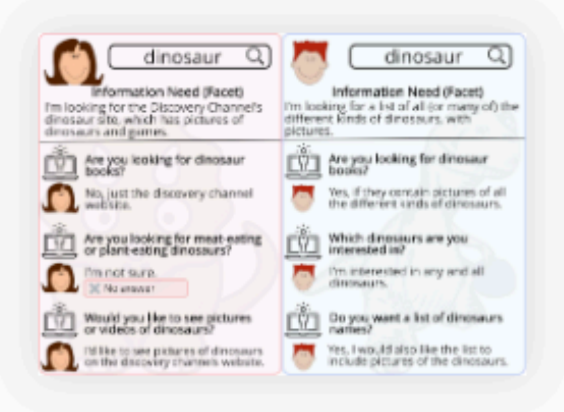
Asking Clarifying Questions in Open-Domain Information-Seeking Conversations

Authors: Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, W. Bruce Croft Authors Info & Claims

SIGIR'19: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval • July 2019 • Pages 475–484 • https://doi.org/10.1145/3331184.3331265

Published: 18 July 2019 Publication History

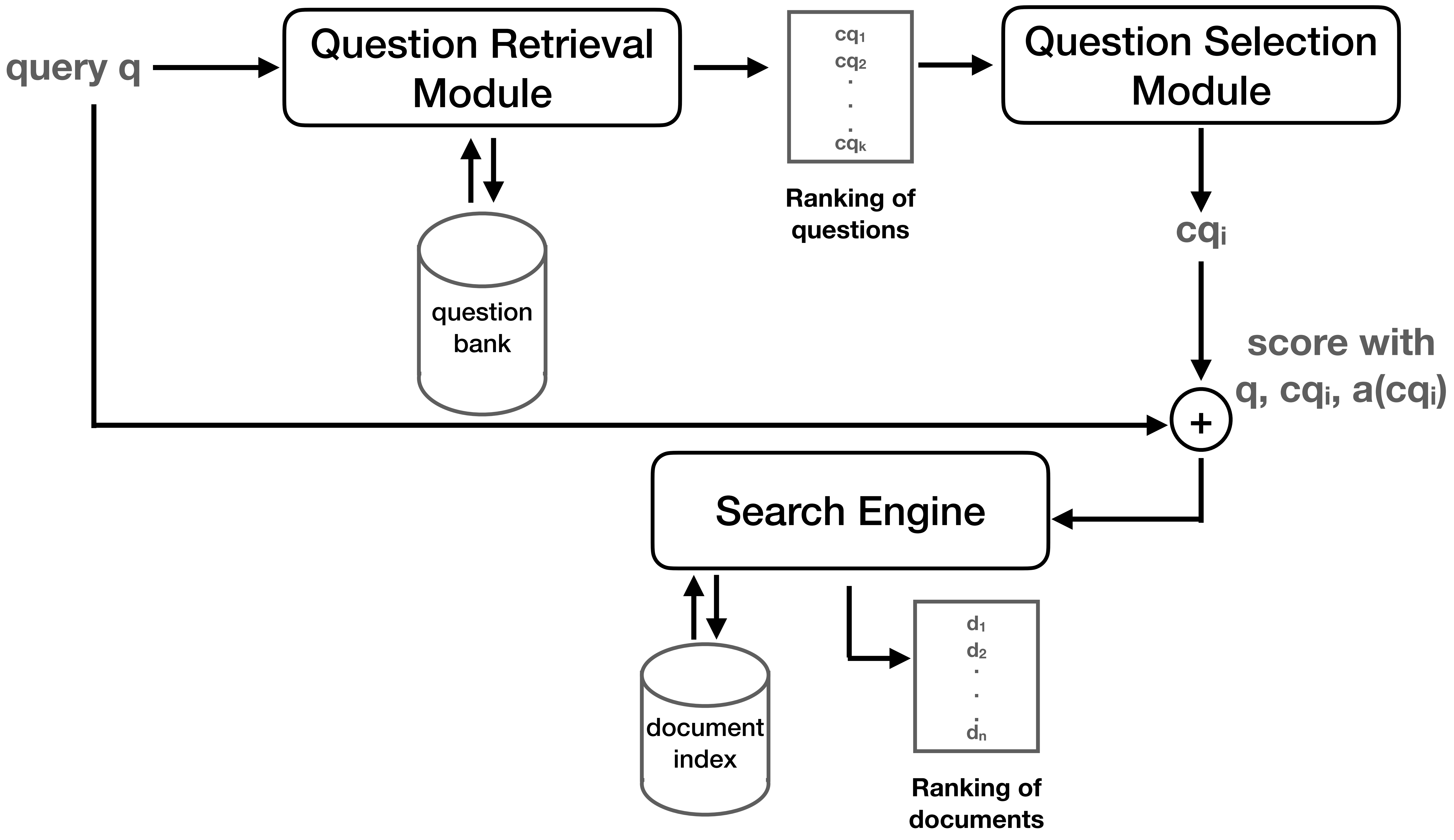
ABSTRACT



Users often fail to formulate their complex information needs in a single query. As a consequence, they may need to scan multiple result pages or reformulate their queries, which may be a frustrating experience. Alternatively, systems can

- Key milestone for research in methods for asking clarifying questions
- Provided a blue-print architecture for the task
  - Not just in terms of pipeline components, but also sub-tasks, evaluation
- Contributed a rich dataset (Qulac)
- Evaluated common baselines for components, developed new methods

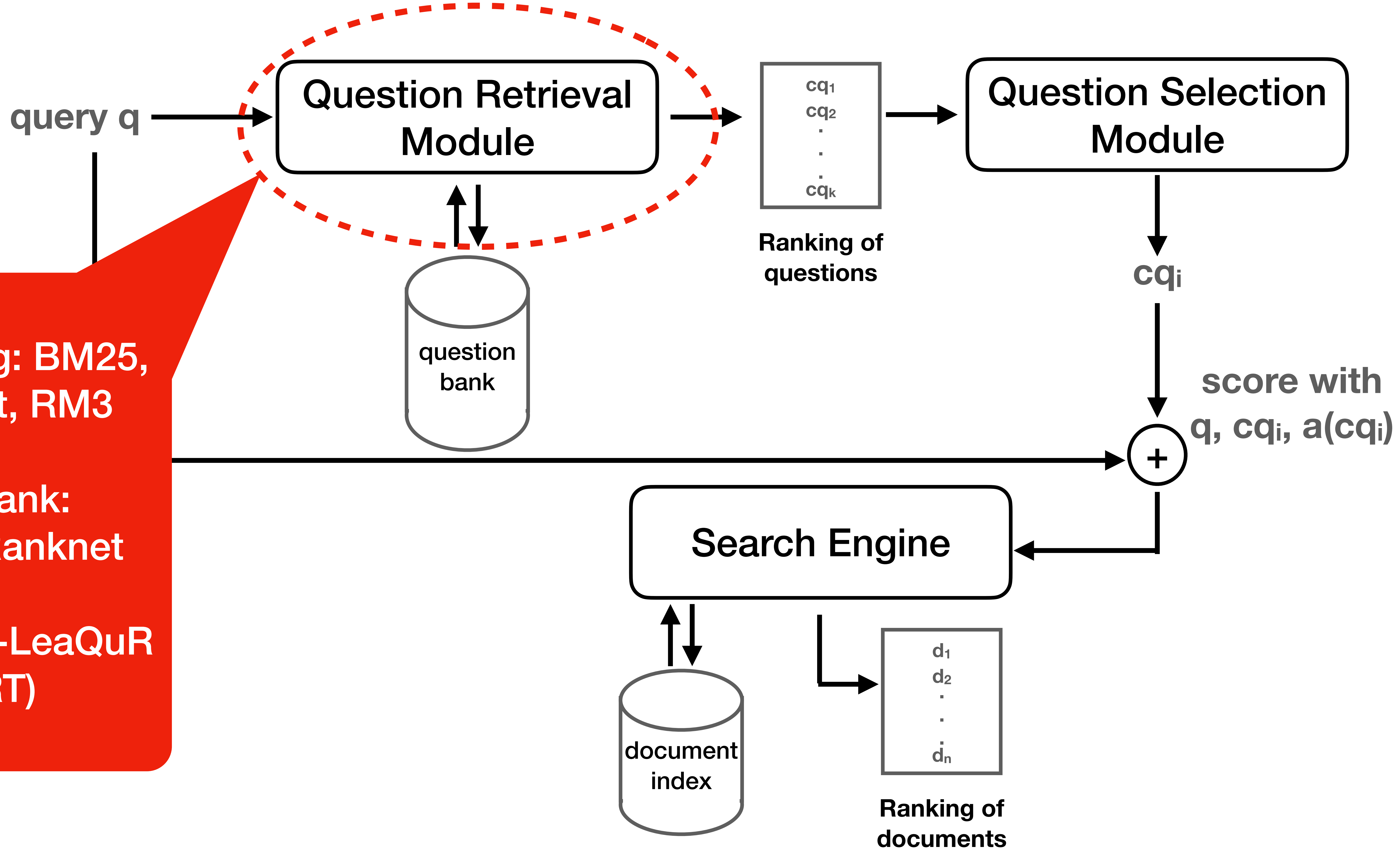
SIGIR 2019's
Retrieval with Clarifying Questions



ie lab

THE UNIVERSITY OF QUEENSLAND AUSTRALIA
CREATE CHANGE

SIGIR 2019's
Retrieval with Clarifying Questions



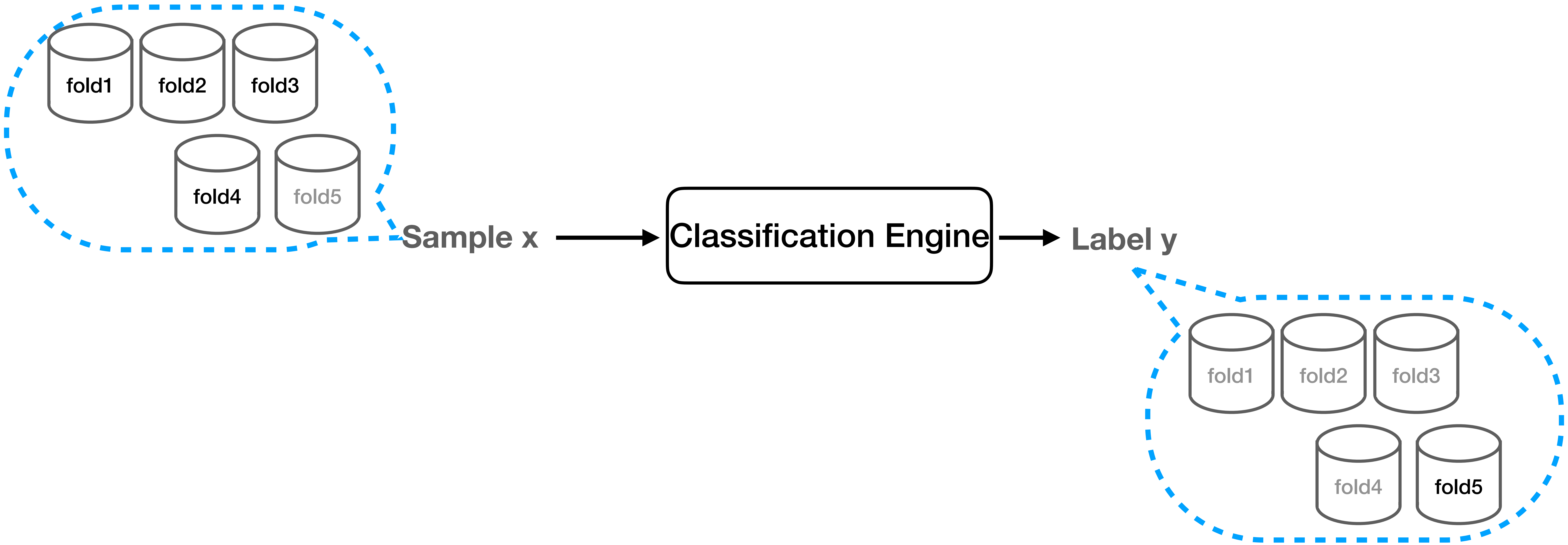
- Keyword matching: BM25, QL with Dirichlet, RM3
- Learning to Rank: LambdaMART, Ranknet
- PLM based: BERT-LeaQuR (~monoBERT)

ie lab

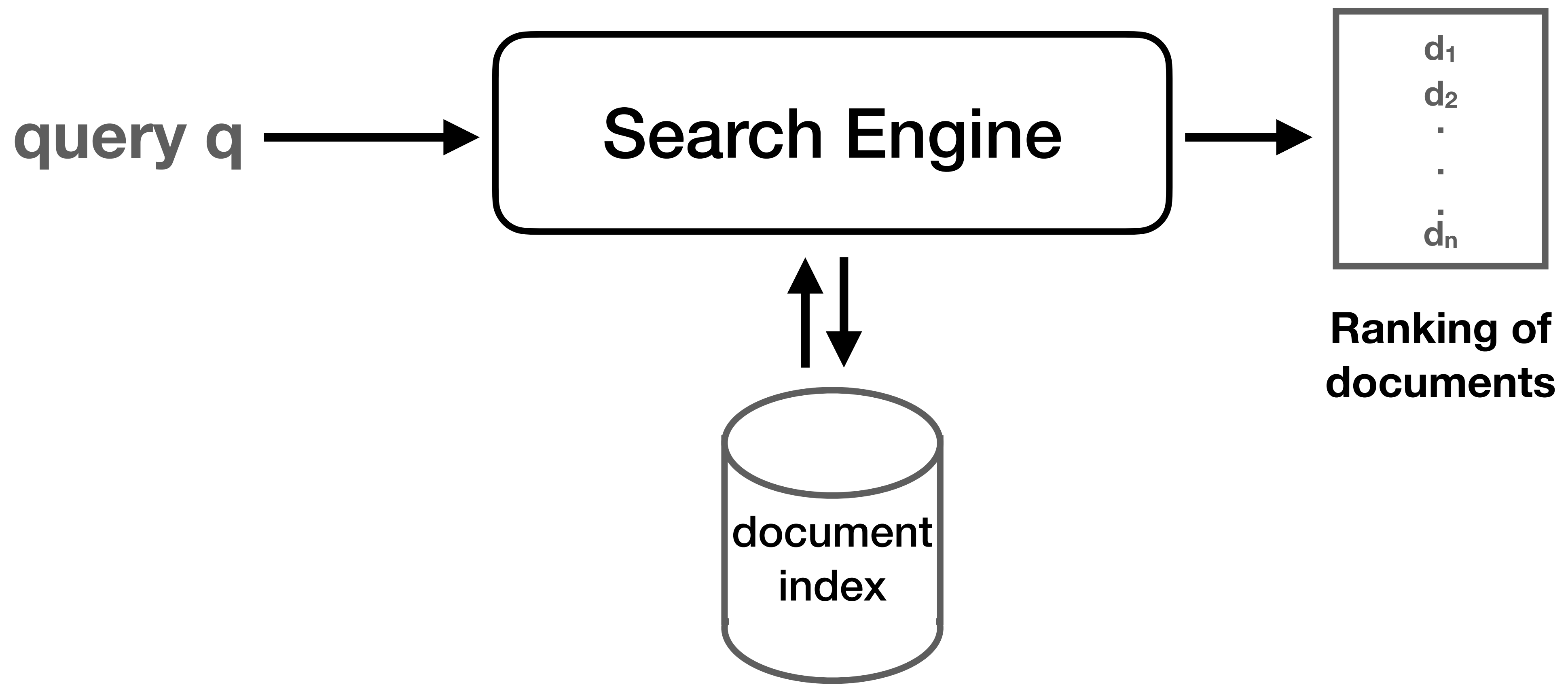
THE UNIVERSITY OF QUEENSLAND
AUSTRALIA
CREATE CHANGE

K-fold Cross-Validation in Machine Learning

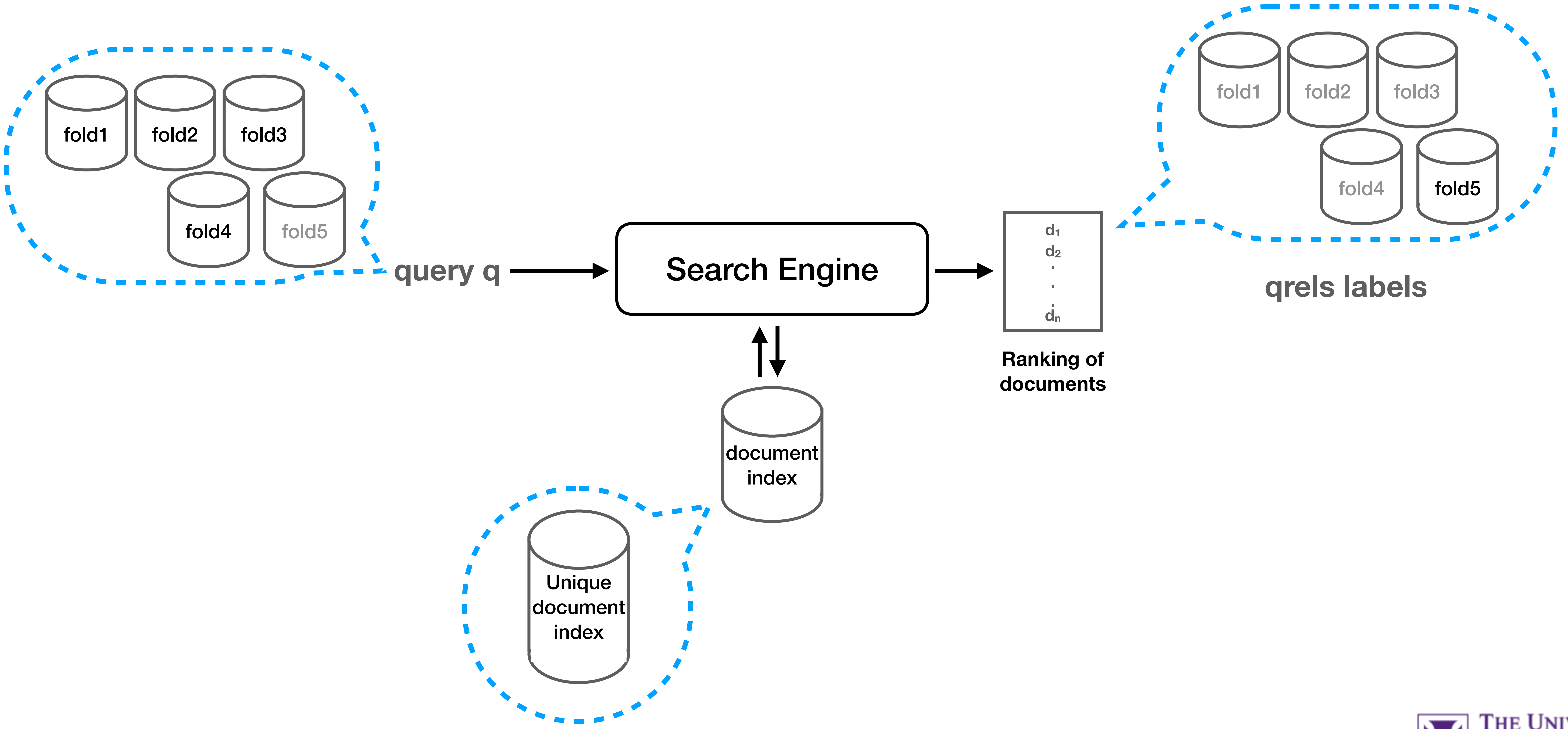
K-fold Cross-Validation in Machine Learning



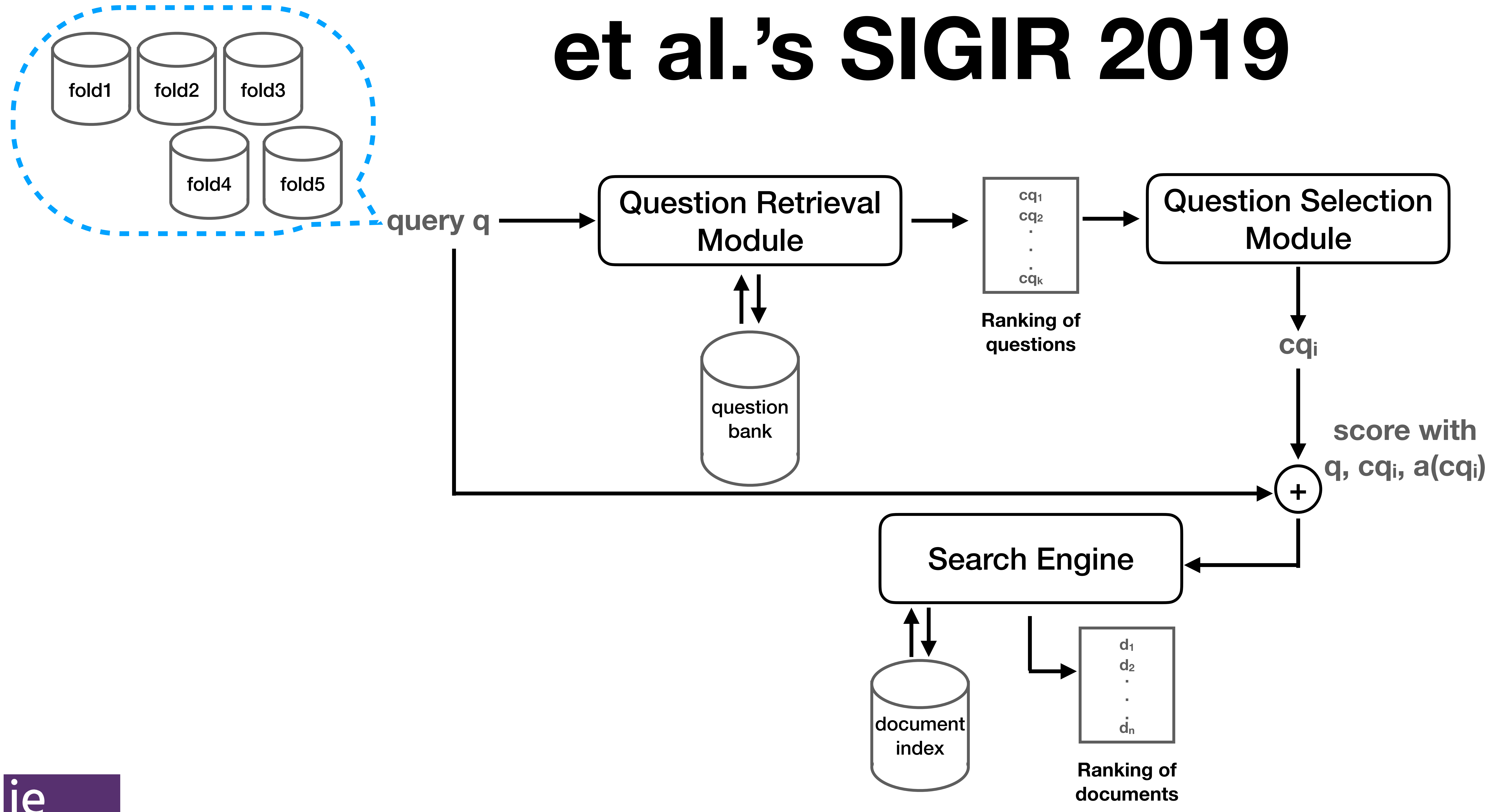
K-fold Cross-Validation in IR



K-fold Cross-Validation in IR



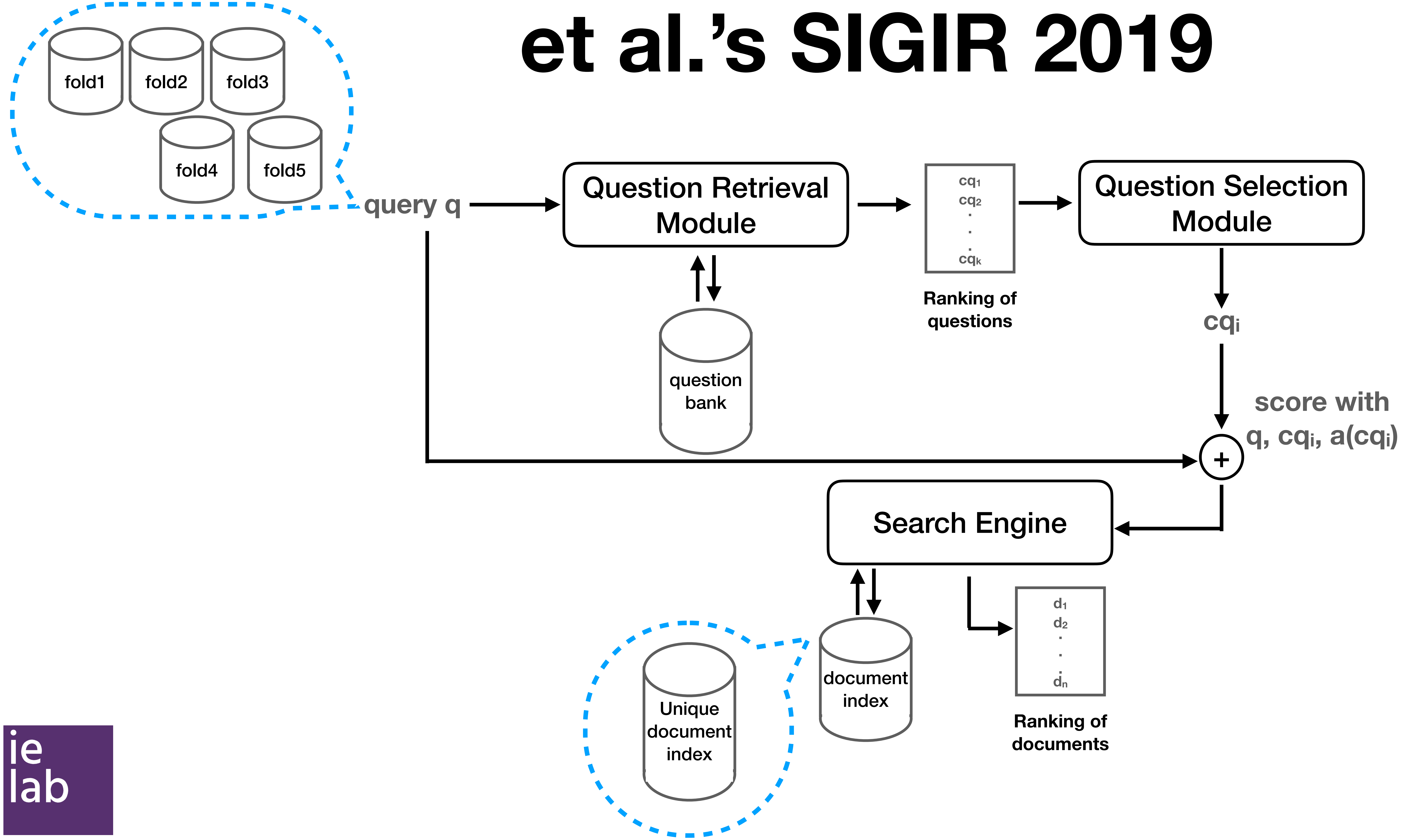
The Key Experimental Issue with Aliannejadi et al.'s SIGIR 2019



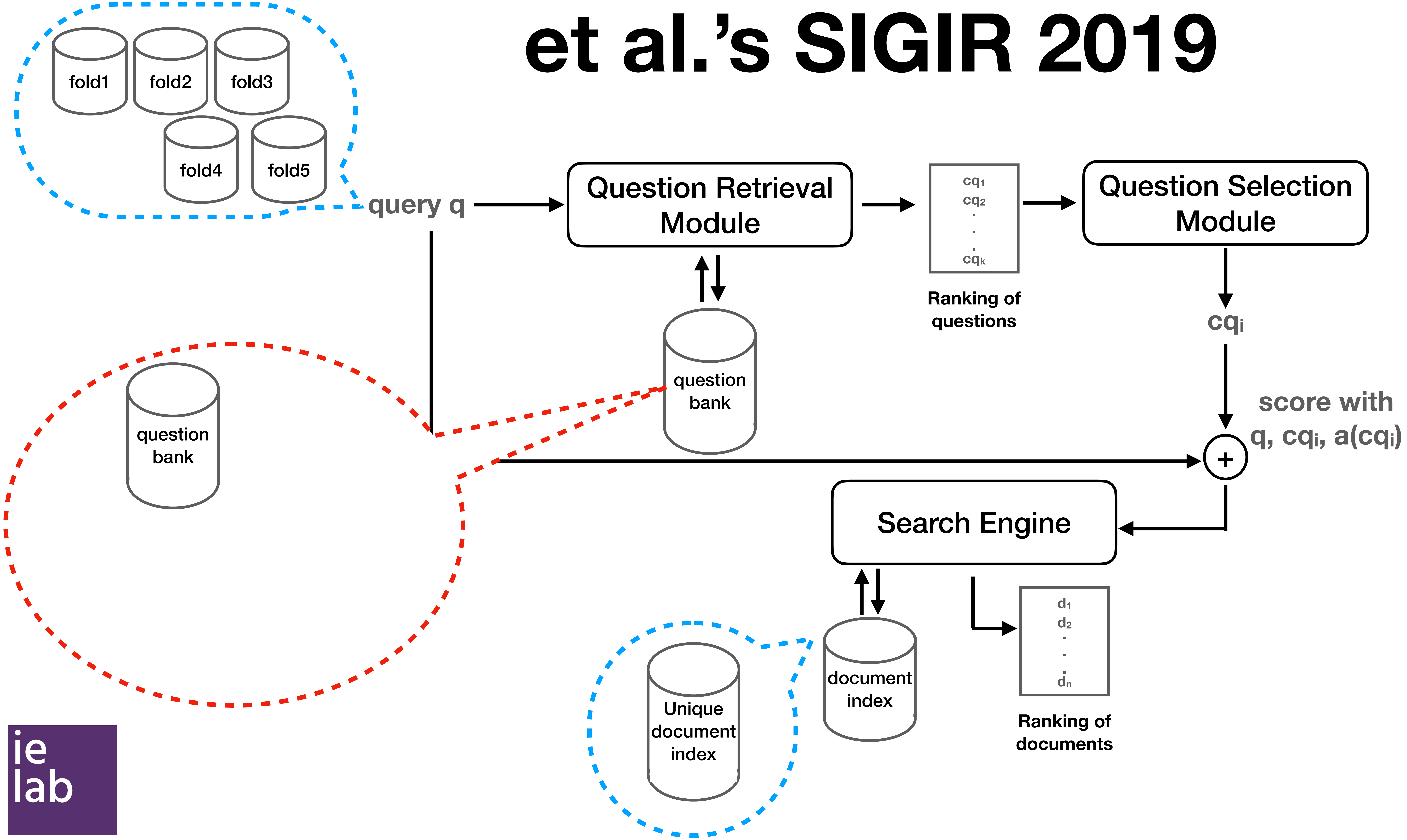
ie lab

THE UNIVERSITY OF QUEENSLAND AUSTRALIA
CREATE CHANGE

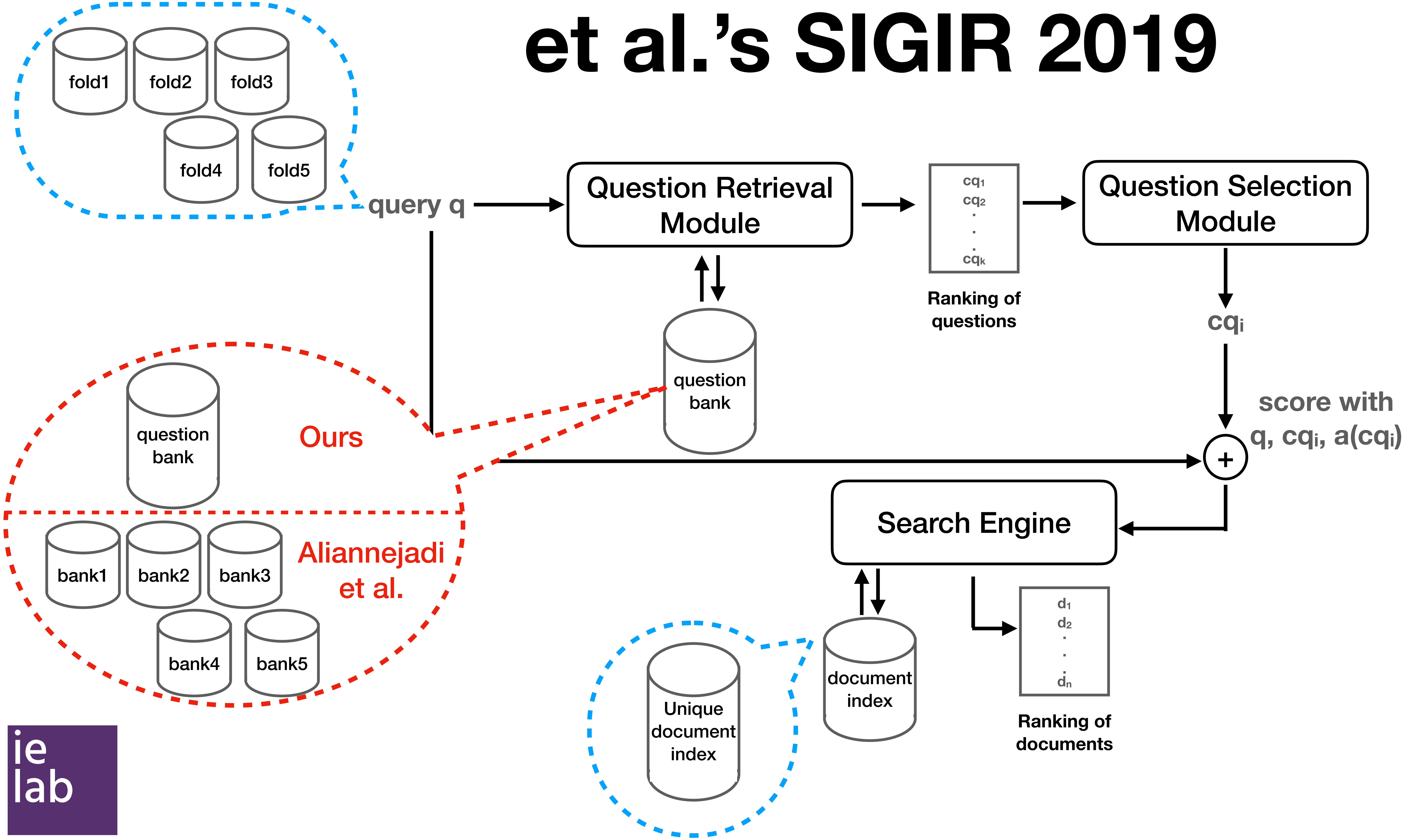
The Key Experimental Issue with Aliannejadi et al.'s SIGIR 2019



The Key Experimental Issue with Aliannejadi et al.'s SIGIR 2019



The Key Experimental Issue with Aliannejadi et al.'s SIGIR 2019

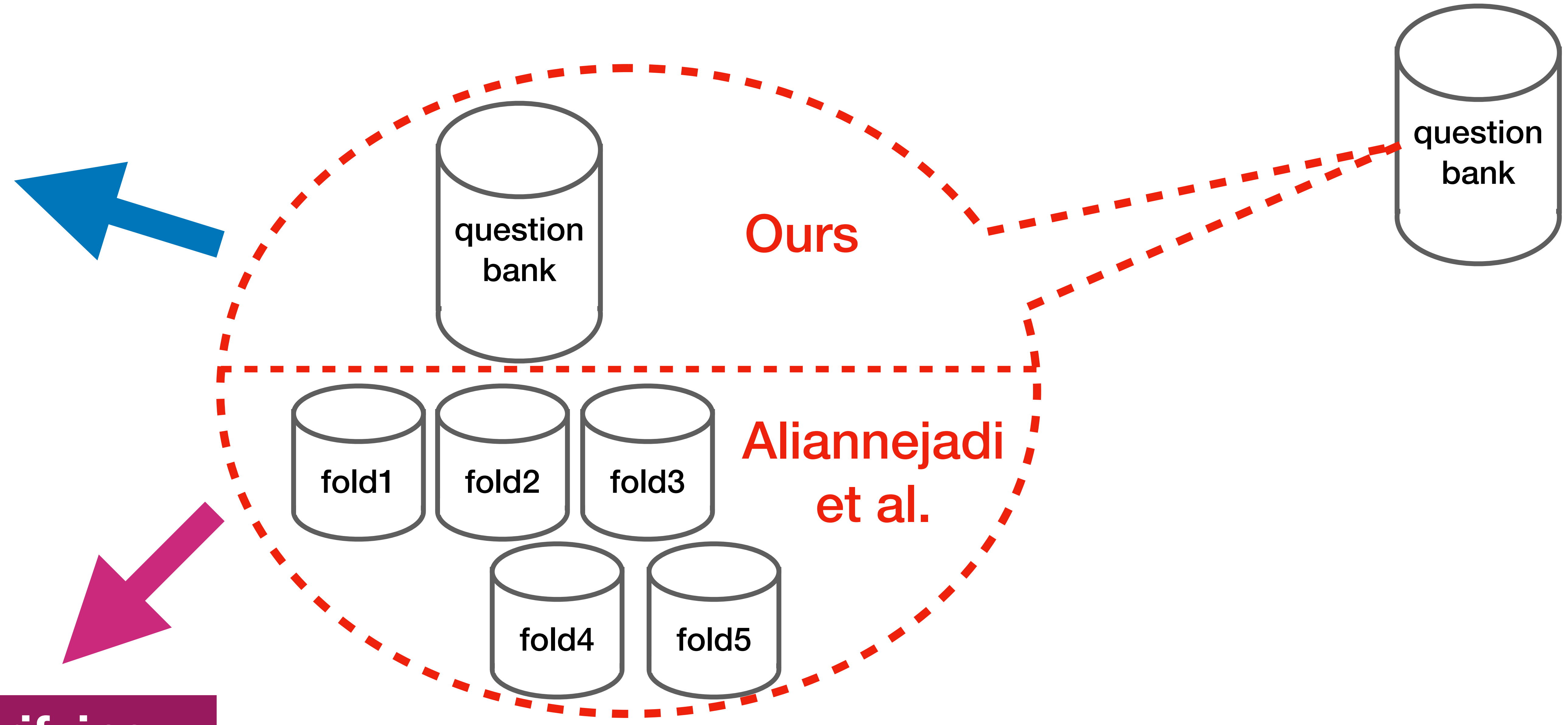


Fold Formation in Aliannejadi et al.’s SIGIR 2019

- Each fold contained a subset of topics and a subset of all candidate clarifying questions
- Each clarifying questions subset always contained all the relevant questions for a given topic
- Each clarifying questions subset contained far less non-relevant clarifying questions than those present in the question-bank

Differences in Data Preparation

	Avg # of topics per fold	Avg # clarifying questions per topic
Train	118.8	2,593
Validation	39.6	2,593
Test	39.6	2,593



	Avg # of topics per fold	Avg # clarifying questions per topic
Train	118.8	1,558.8
Validation	39.6	521.8
Test	39.6	521.8

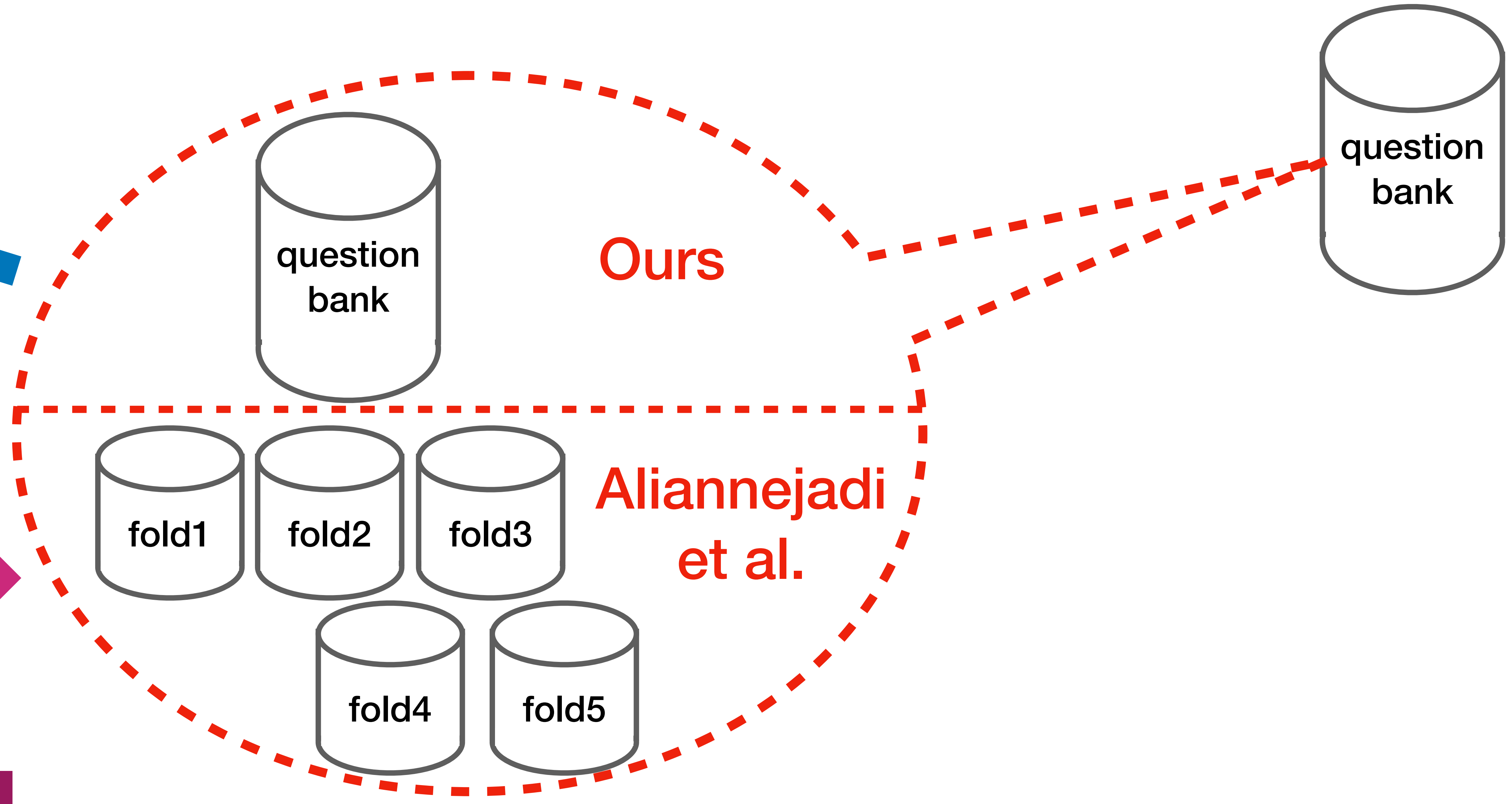
ie lab

The University Of Queensland
Australia
Create Change

Differences in Data Preparation

	Avg # of topics per fold	Avg # clarifying questions per topic
Train	118.8	2,593
Validation	39.6	2,593
Test	39.6	2,593

~13.1 are relevant



	Avg # of topics per fold	Avg # clarifying questions per topic
Train	118.8	1,558.8
Validation	39.6	521.8
Test	39.6	521.8

~13.1 are relevant

ie lab

THE UNIVERSITY OF QUEENSLAND AUSTRALIA
CREATE CHANGE

Content of testing data

	Our Data Preparation	SIGIR 2019 Data Preparation
	2,593	521.8
	13.1	13.1

ie lab

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

CREATE CHANGE

Our Data Preparation

SIGIR 2019 Data Preparation

Content of testing data

2,593

13.1

521.8

13.1

A relevant clarifying question X is ranked last in the ranking

Contribution made by X to MAP

Our Data Preparation

SIGIR 2019 Data Preparation

	Our Data Preparation	SIGIR 2019 Data Preparation
Content of testing data	2,593 13.1	521.8 13.1

A relevant clarifying question X is ranked last in the ranking

	Our Data Preparation	SIGIR 2019 Data Preparation
Contribution made by X to MAP	$g(X) = 3.8565 \times 10^{-4}$	$g(X) = 1.9164 \times 10^{-3}$

ie lab

The University Of Queensland
Australia
Create Change

Our Data Preparation

SIGIR 2019 Data Preparation

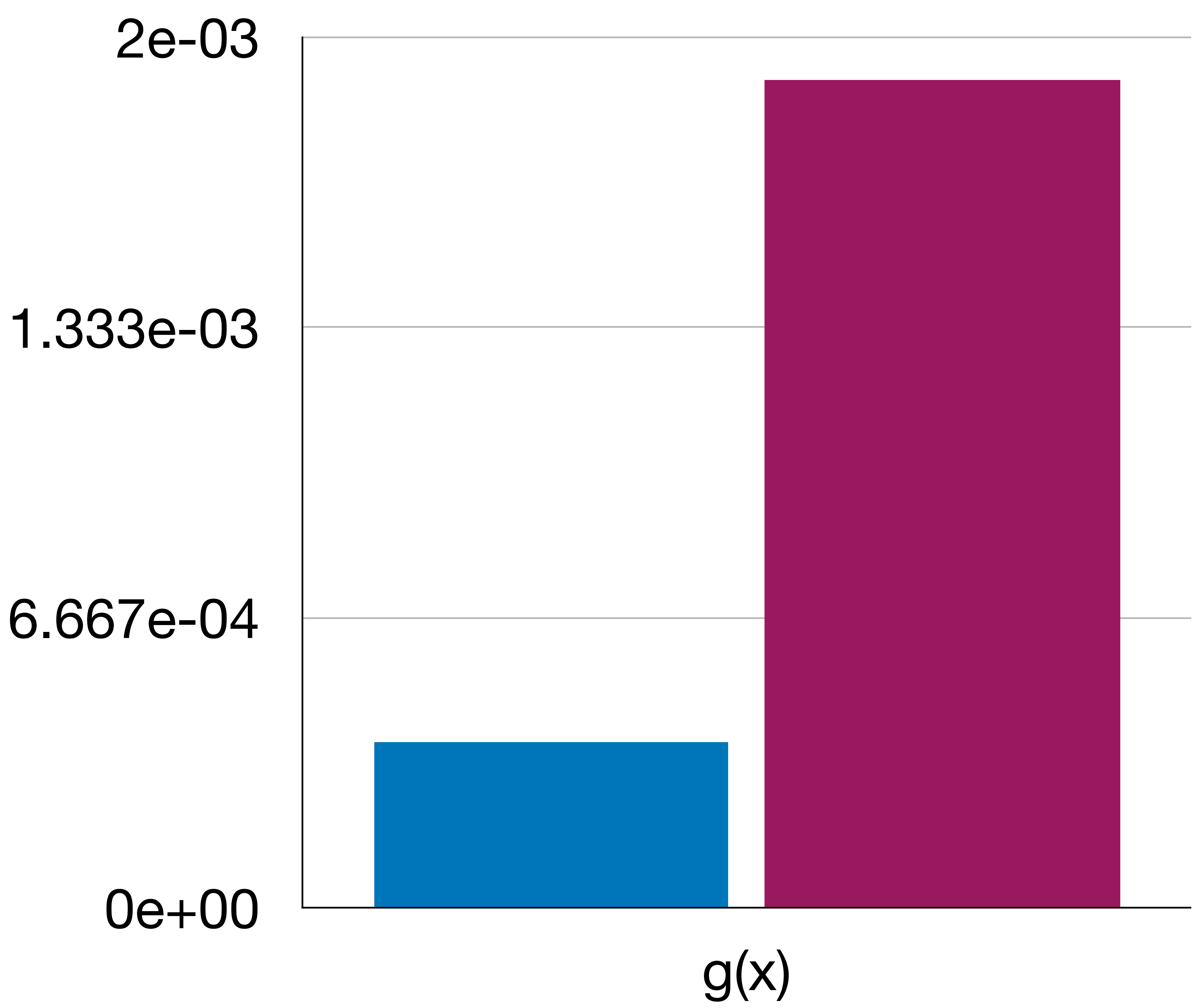
Content of testing data

2,593

521.8

A

ng



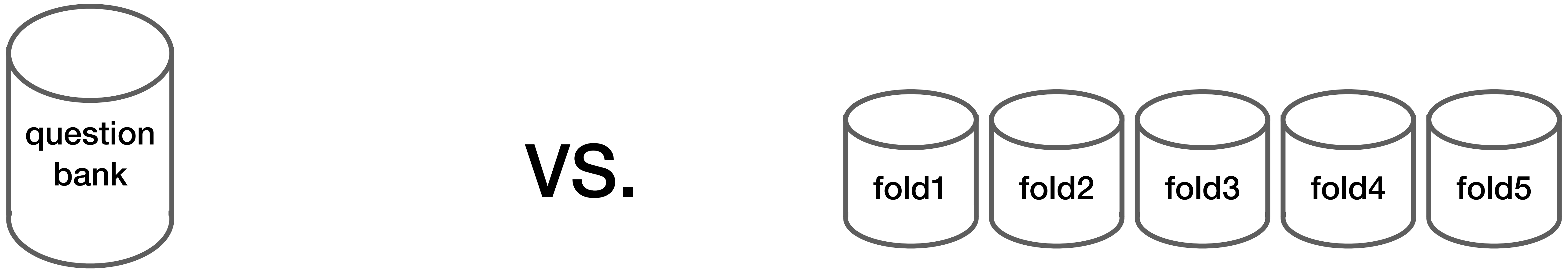
Contribution made by X to MAP

g

x 10^{-3}

What we do in this paper

- Replicate the methods of Aliannejadi et al. to investigate the **impact of difference in data preparation**

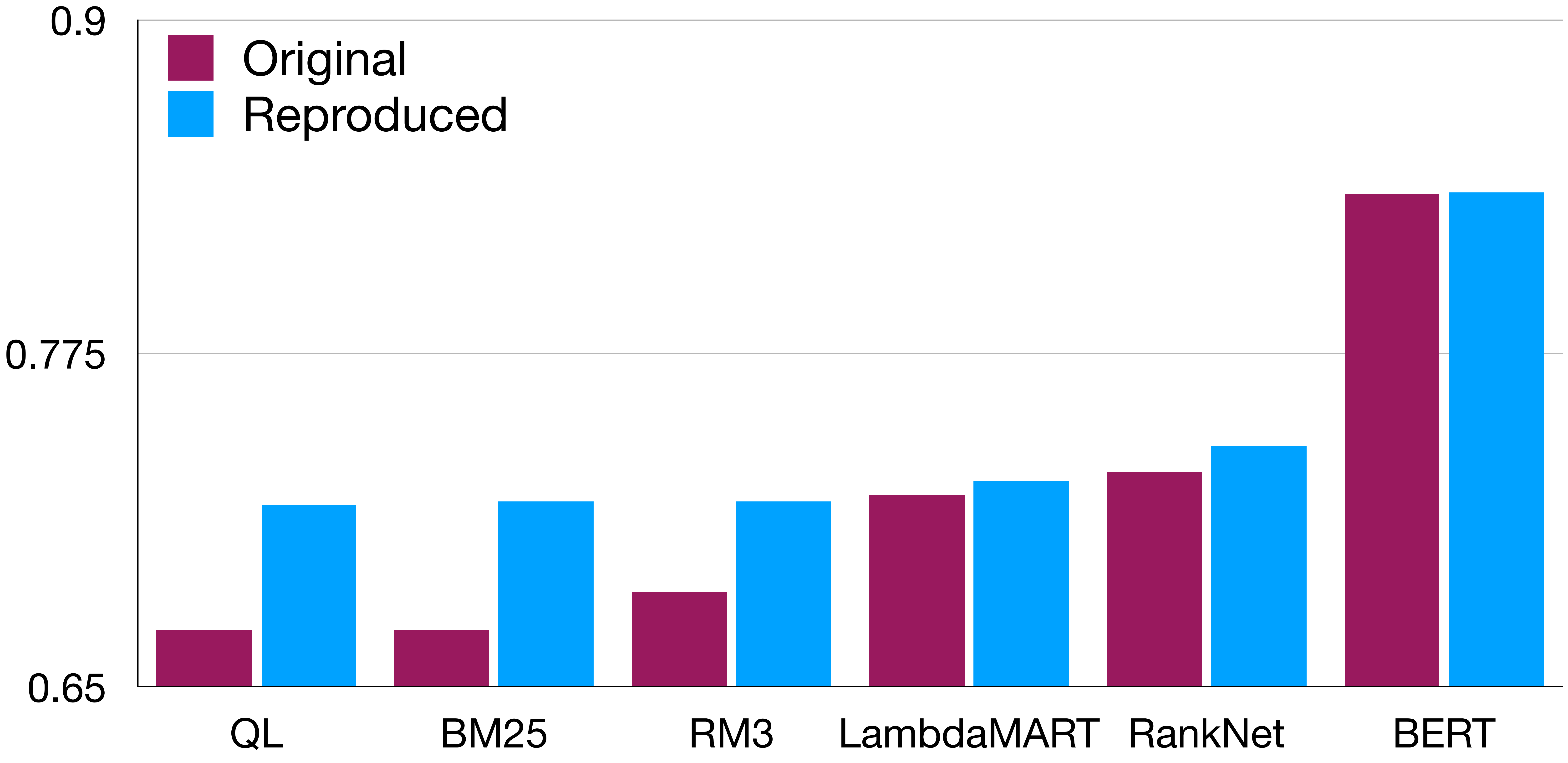


Ours

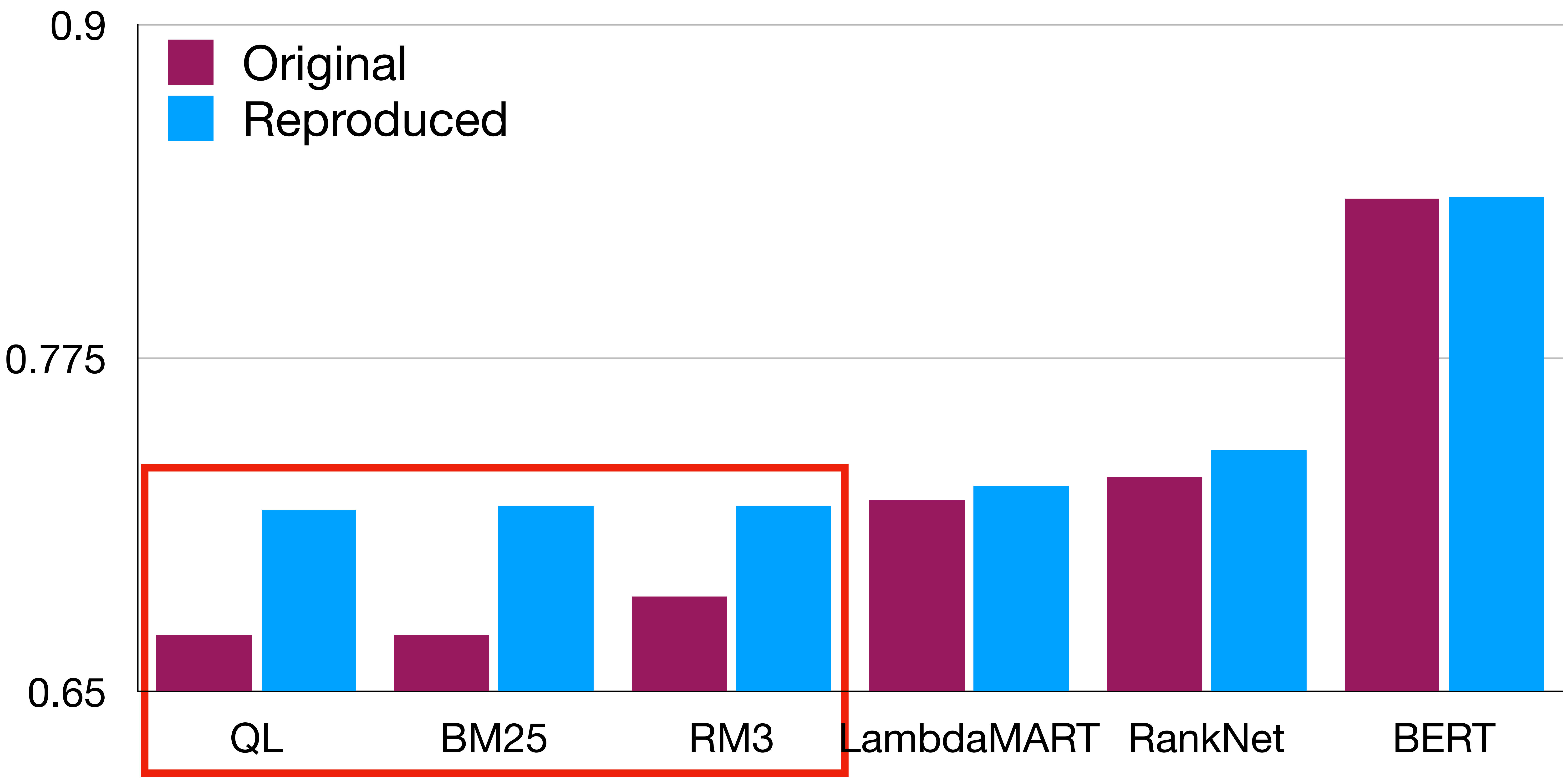
Aliannejadi et al.

- Further analyse results with respect to the use of **keyword matching scores** as only **features in learning to rank**
  - Zero-valued representations
  - Treatment of ties

Exp. 1: Aliannejadi et al.'s Data Preparation



Exp. 1: Aliannejadi et al.'s Data Preparation

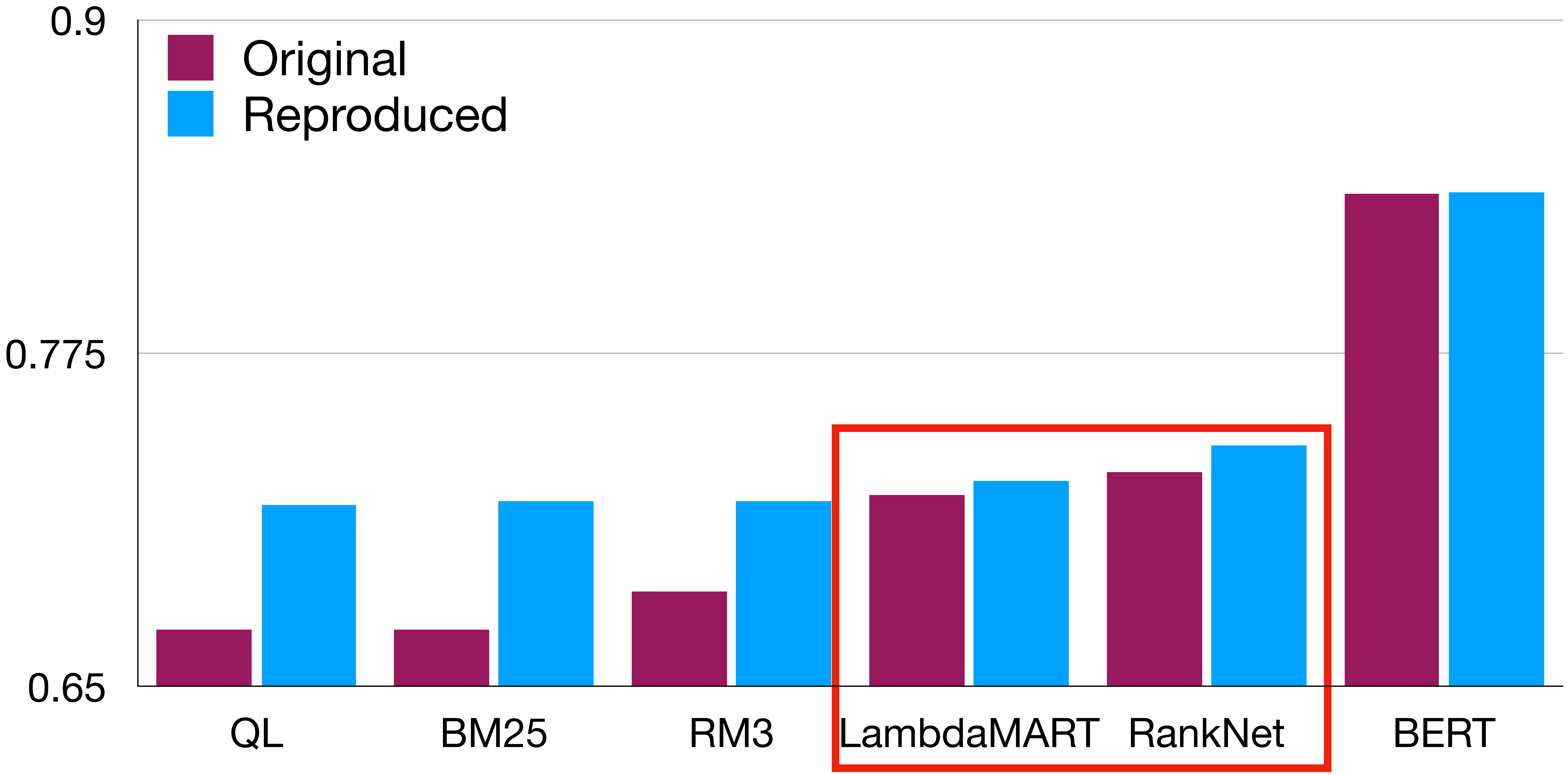


Keyword Matching

- could not reproduce results
- consistently higher effectiveness than reported

- *Hypothesis:* Aliannejadi et al. did not execute the keyword matching against the same data preparations used for learnt models.
- ?Results obtained against the whole question bank?

Exp. 1: Aliannejadi et al.'s Data Preparation

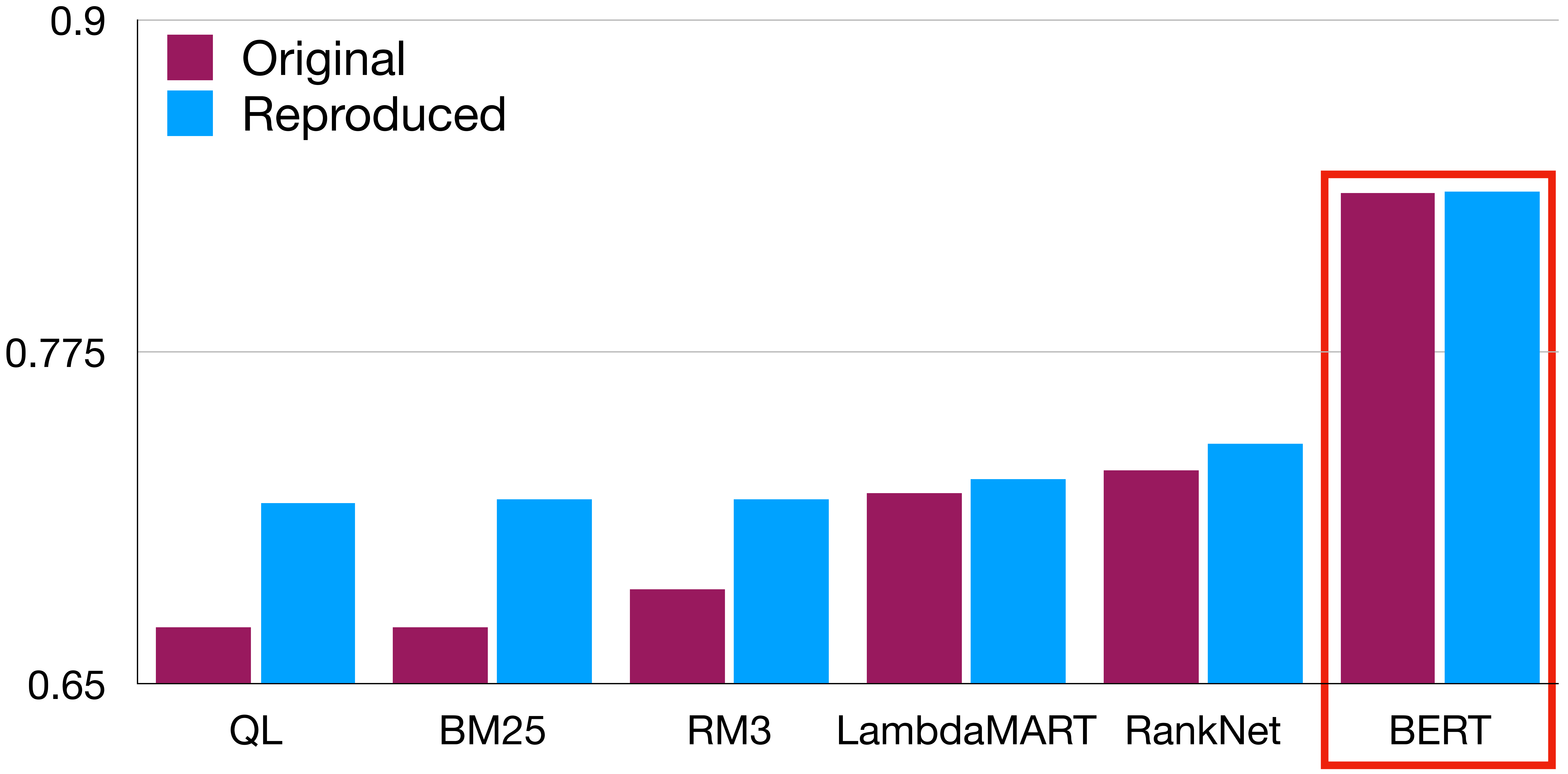


Learning to Rank

- could not obtain same results, but values very close

- *Hypothesis:* mismatch in original result reporting & data
- differences due to feature files they originally used containing more questions than the ones they gave us

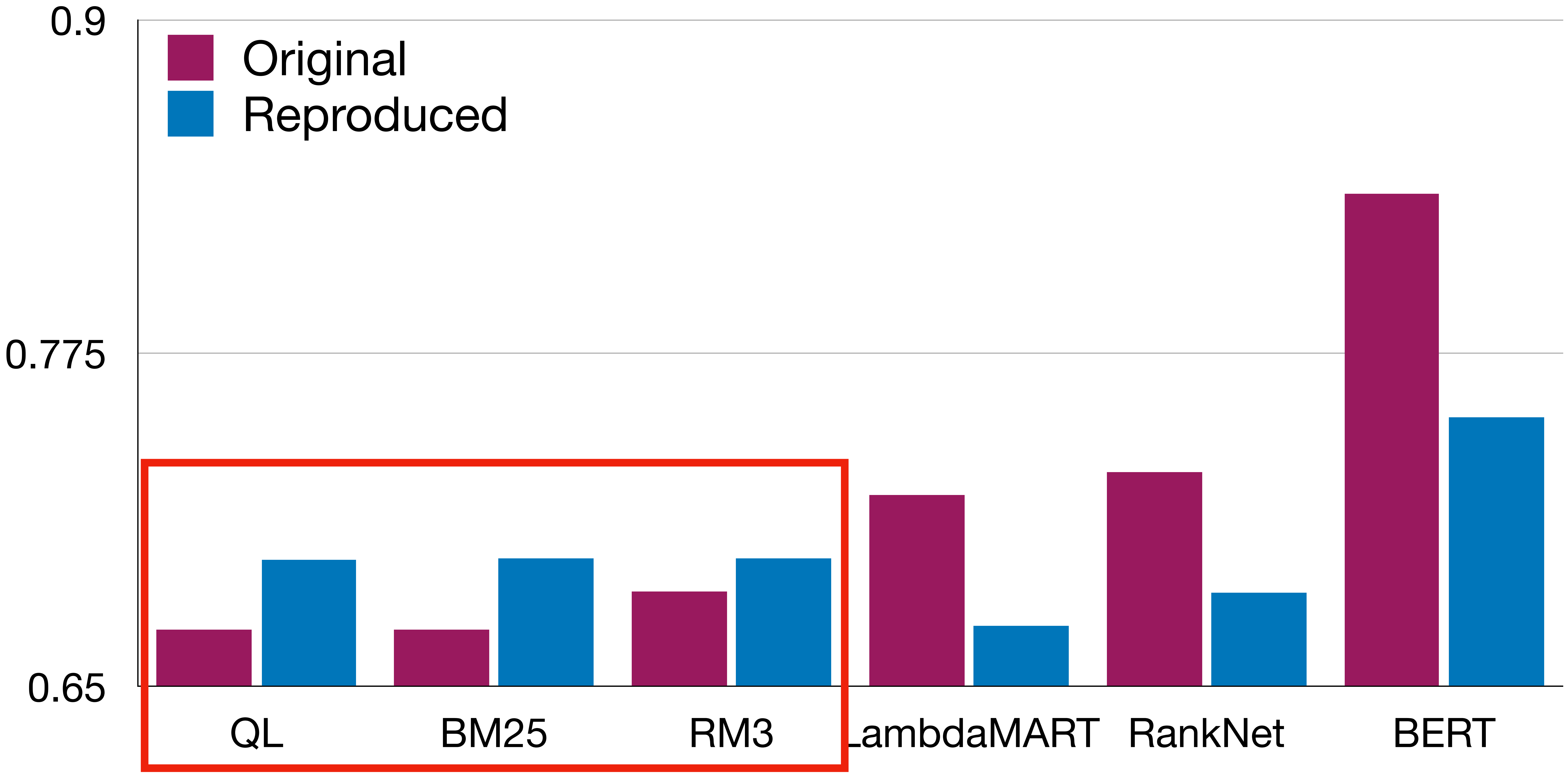
Exp. 1: Aliannejadi et al.'s Data Preparation



BERT

- we obtained values close to the ones reported

Exp. 2: Our Data Preparation



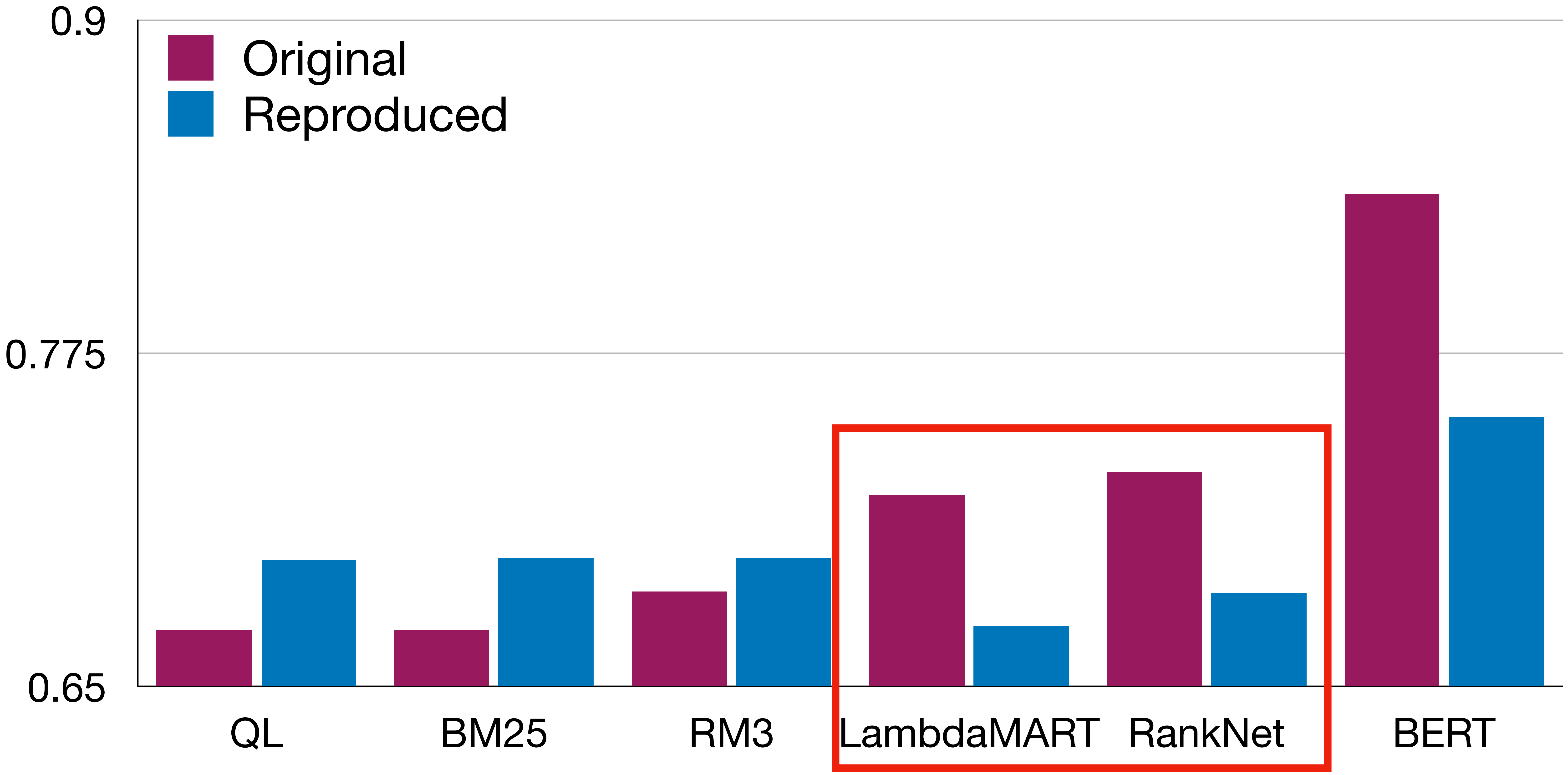
Keyword Matching

- we could not obtain same results, but values reasonably close (in context of setup)

- *Hypothesis:* Differences ascribed to tools (Anserini vs. Galago), model parameters, and question bank size.

ie lab

Exp. 2: Our Data Preparation

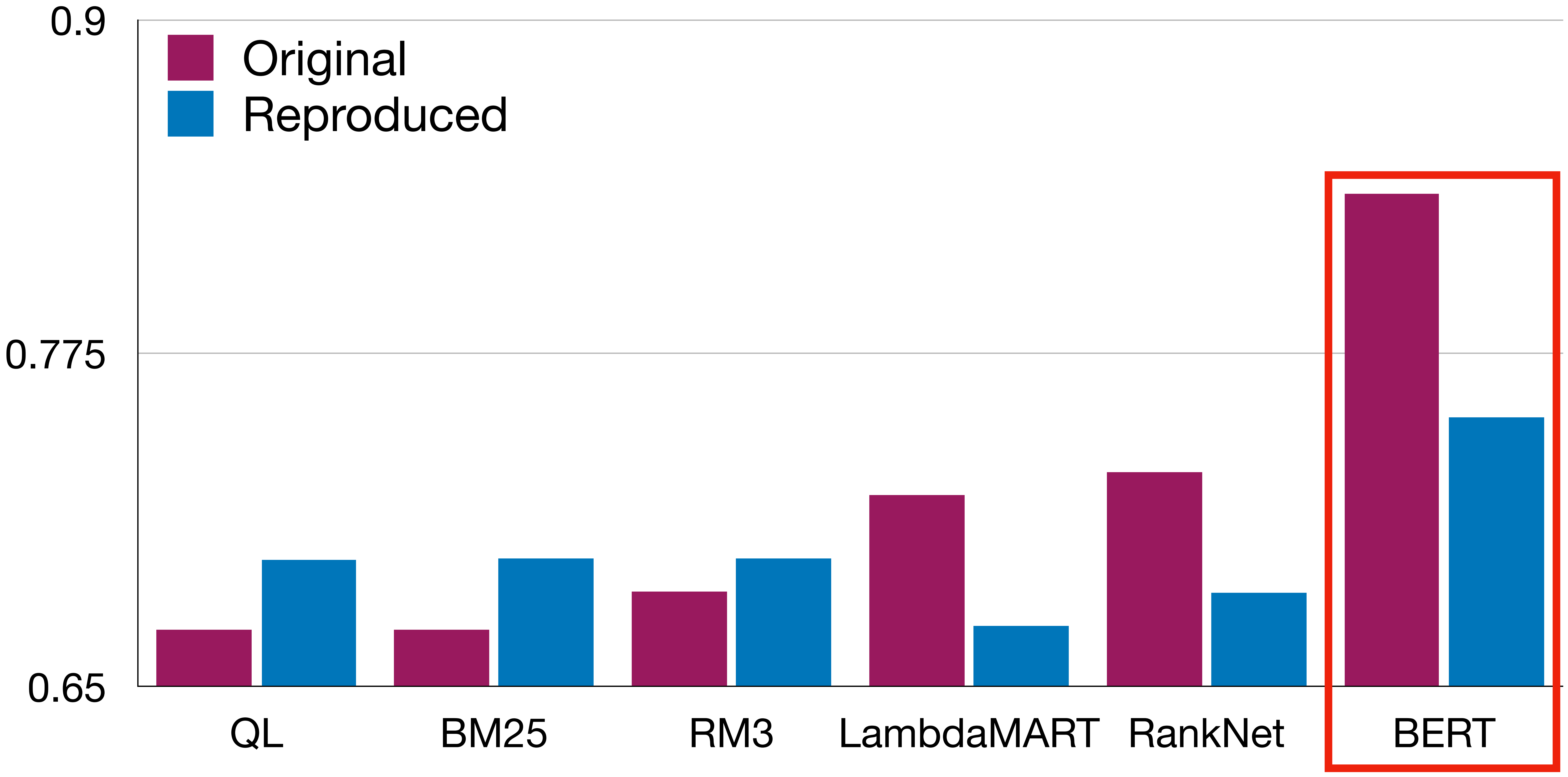


Learning to Rank

- could not obtain same results.
- Difference in trend: LTR lower effectiveness than keyword matching models

- Expected given they used only part of the available data for retrieval (i.e. fold VS. whole question bank)

Exp. 2: Our Data Preparation



BERT

- performs worst than in original work.
- BERT still best method, but gains over keyword matching sensibly lower
- e.g. +7.64% in ours vs. +24.33% in theirs.
- Gains not anymore significant

Take-aways

- We showed how data preparation affects the results reported in the original work
  - learning to rank cannot outperform keyword matching
  - BERT does outperform keyword matching, but much smaller gains (not statistically significant)
- We do not believe this is a generalisable result:
  (i) amount of training data is likely too little for those models (especially BERT)
  (ii) feature representation particularly poor for LTR, where most questions had identical representation.
- Data sharing and genuine collaboration b/w reproduction team and original team was fundamental to identify the data preparation aspect

BONUS SLIDES

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA
CREATE CHANGE

Zero-valued Representations

- LTR feature representation: 3 features — QL, BM25, RM3 scores
  - Many **relevant query-question** pairs share **same non-zero representation**
  - Many **query-question pairs** with **all features zero-valued**
    - often for **non-relevant** questions, sporadically for relevant questions

- At test time, LTR often ends up assigning to pairs one of two scores: 0 or 1 — thus, **ties**

ie
lab

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA
CREATE CHANGE

Treatment of Ties

Output of Ranker

d-A	0.9
d-C	0.8
d-B	0.8
d_D	0.7

Treatment of Ties

Output of Ranker

d-A	0.9
d-C	0.8
d-B	0.8
d_D	0.7

d-A	0.9
d-C	0.8
d-B	0.8
d_D	0.7

RankLib eval

Treatment of Ties

Output of Ranker

d-A	0.9
d-C	0.8
d-B	0.8
d_D	0.7

d-A	0.9
d-C	0.8
d-B	0.8
d_D	0.7

RankLib eval

trec_eval

d-A	0.9
d-B	0.8
d-C	0.8
d_D	0.7

Treatment of Ties

Output of Ranker

d-A	0.9
d-C	0.8
d-B	0.8
d_D	0.7

d-A	0.9
d-C	0.8
d-B	0.8
d_D	0.7

RankLib eval

trec_eval

d-A	0.9
d-B	0.8
d-C	0.8
d_D	0.7

RankLib eval	0.6728
`trec_eval` no ties	0.6728

ie lab

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

CREATE CHANGE

Treatment of Ties

Output of Ranker

d-A	0.9
d-C	0.8
d-B	0.8
d_D	0.7

d-A	0.9
d-C	0.8
d-B	0.8
d_D	0.7

RankLib eval

trec_eval

d-A	0.9
d-B	0.8
d-C	0.8
d_D	0.7

RankLib eval	0.6728
trec_eval	0.7233
trec_eval no ties	0.6728

ie lab

THE UNIVERSITY OF QUEENSLAND AUSTRALIA
CREATE CHANGE

Treatment of Ties

Output of Ranker

d-A	0.9
d-C	0.8
d-B	0.8
d_D	0.7

d-A	0.9
d-C	0.8
d-B	0.8
d_D	0.7

RankLib eval

trec_eval

d-A	0.9
d-B	0.8
d-C	0.8
d_D	0.7

RankLib eval	0.6728
trec_eval	0.7233
trec_eval no ties	0.6728

- Unsure what original study used
- In our experiments, we use trec_eval and break ties

ie lab

The University Of Queensland
Australia
Create Change