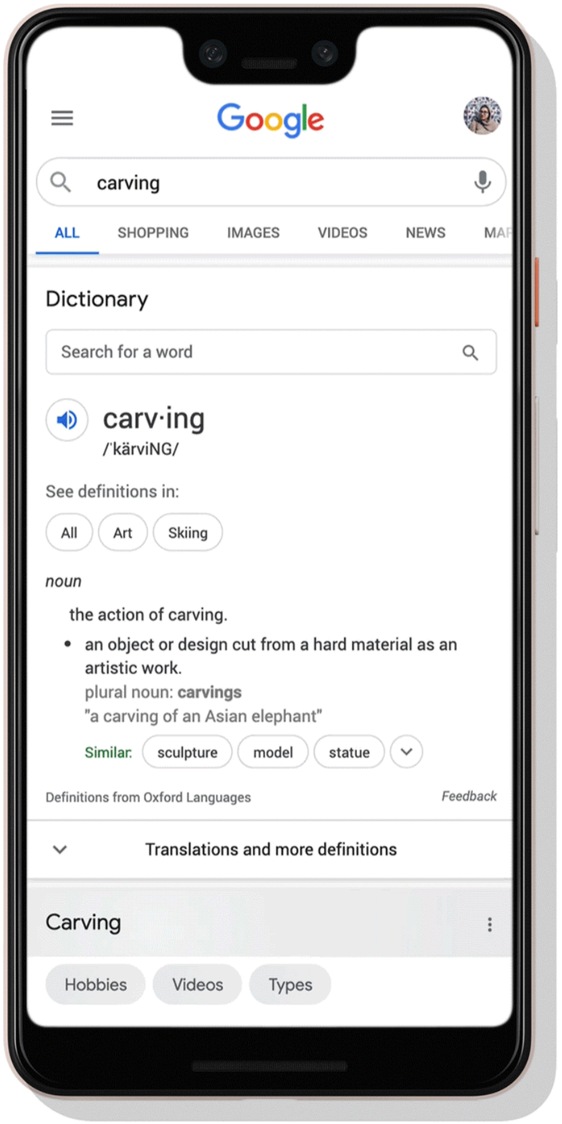
Contextualized Query Embeddings for Conversational Search

Presenter: Sheng-Chieh Lin


Google
carving
ALL
SHOPPING
IMAGES
VIDEOS
NEWS
Dictionary
Search for a word
carv·ing
/ˈkärviNG/
See definitions in:
All
Art
Skiing
noun
the action of carving.
an object or design cut from a hard material as an artistic work.
plural noun: carvings
"a carving of an Asian elephant"
Similar:
sculpture
model
statue
Definitions from Oxford Languages
Feedback
Translations and more definitions
Carving
Hobbies
Videos
Types

TREC: Conversational Assistant Track

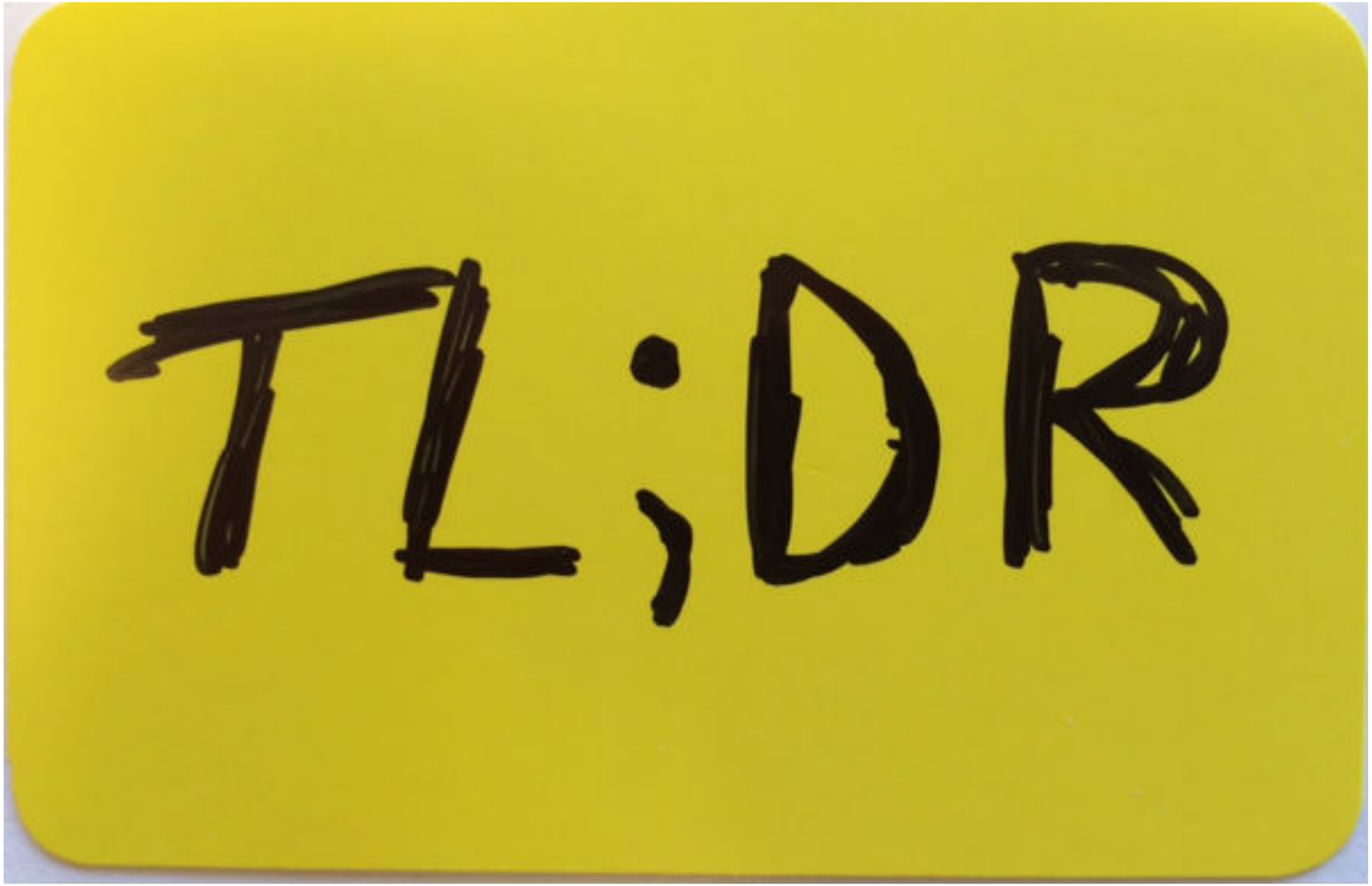
CAsT Example

q_1 **What is throat cancer?**

q_2 **Is it treatable?**

q_3 **Tell me about lung cancer.**

q_4 **What are its symptoms?**

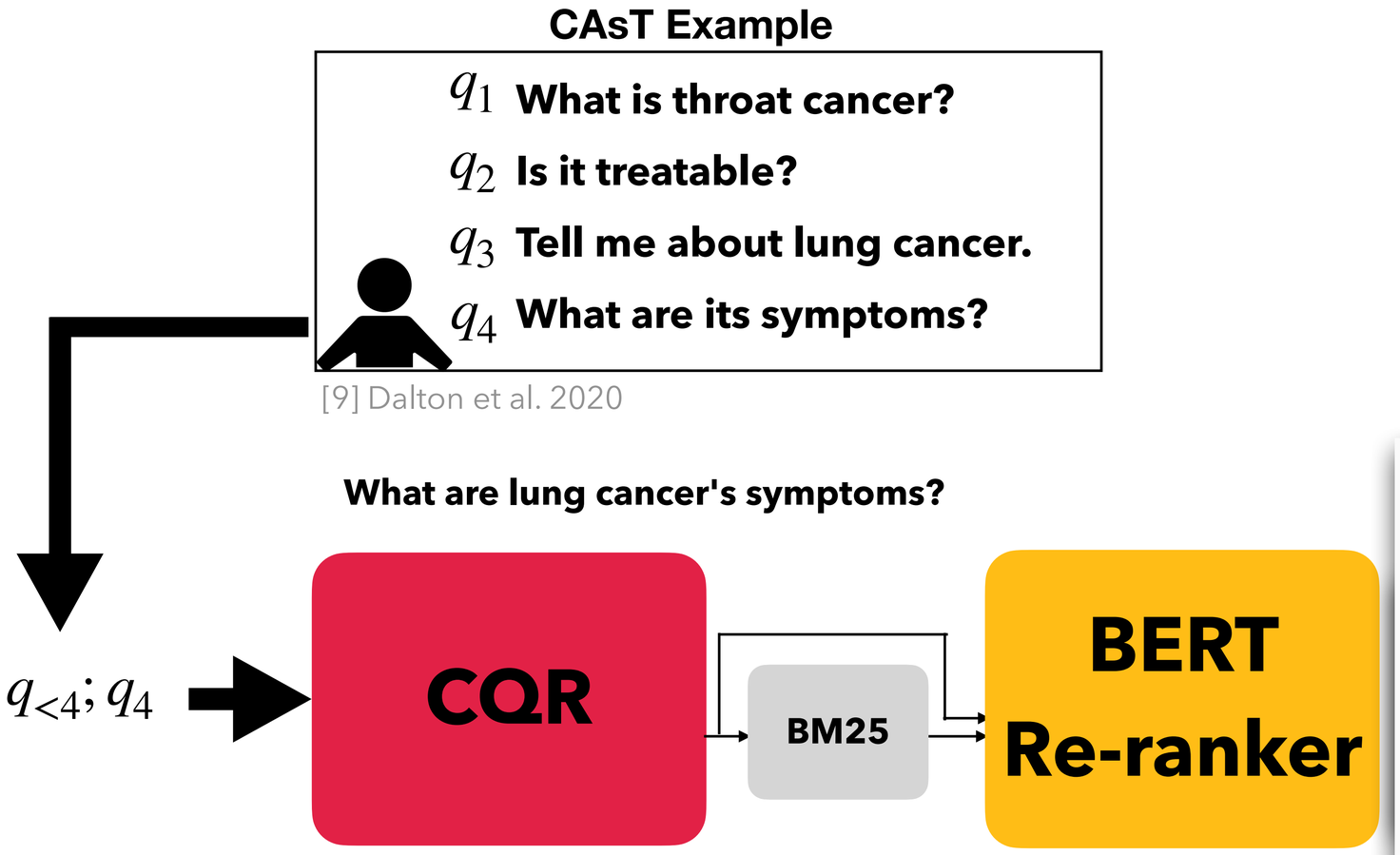

TL;DR

Successfully Apply DPR to Conversational Search

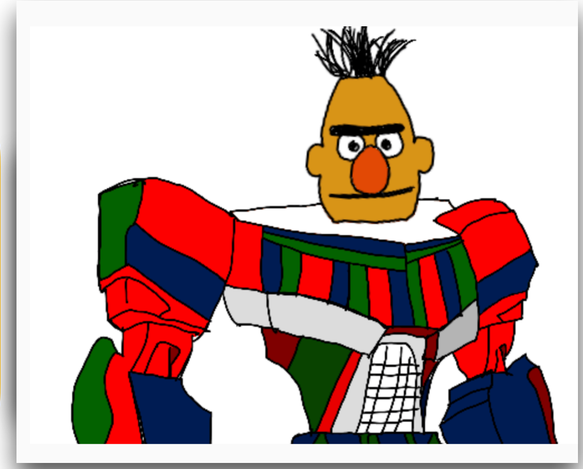
Interpret how DPR Understand Conversational Query


WHY

Previous Solution



[9] Dalton et al. 2020



Our Idea: Conversational DPR

CAsT Example

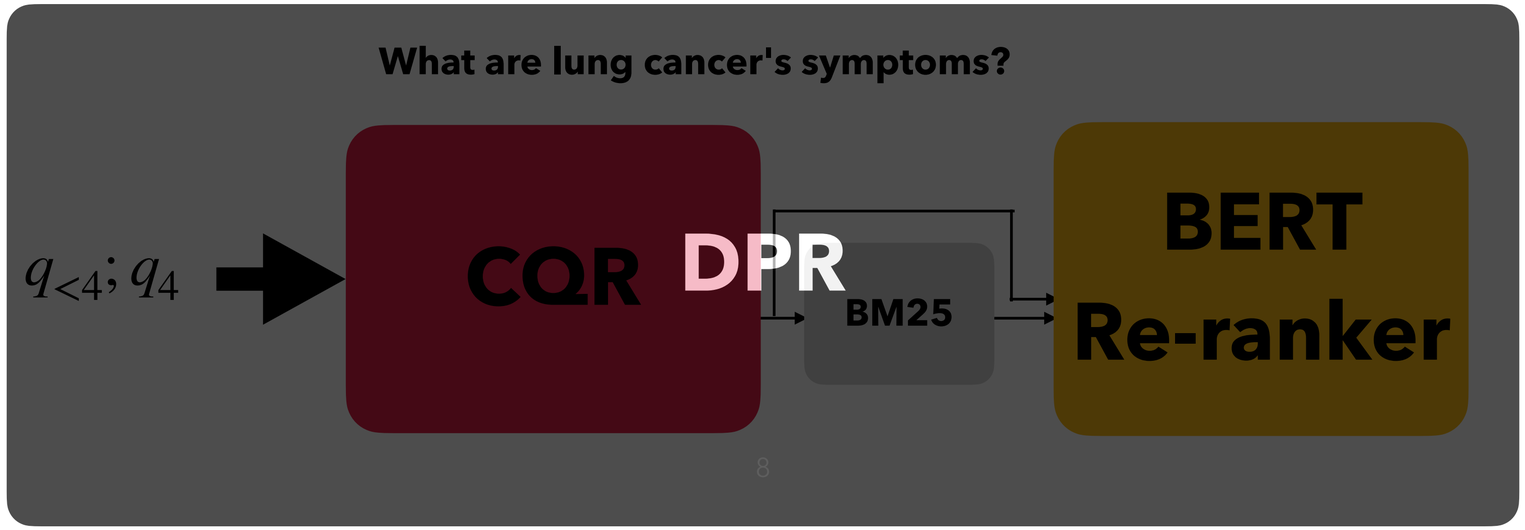
q_1 **What is throat cancer?**

q_2 **Is it treatable?**

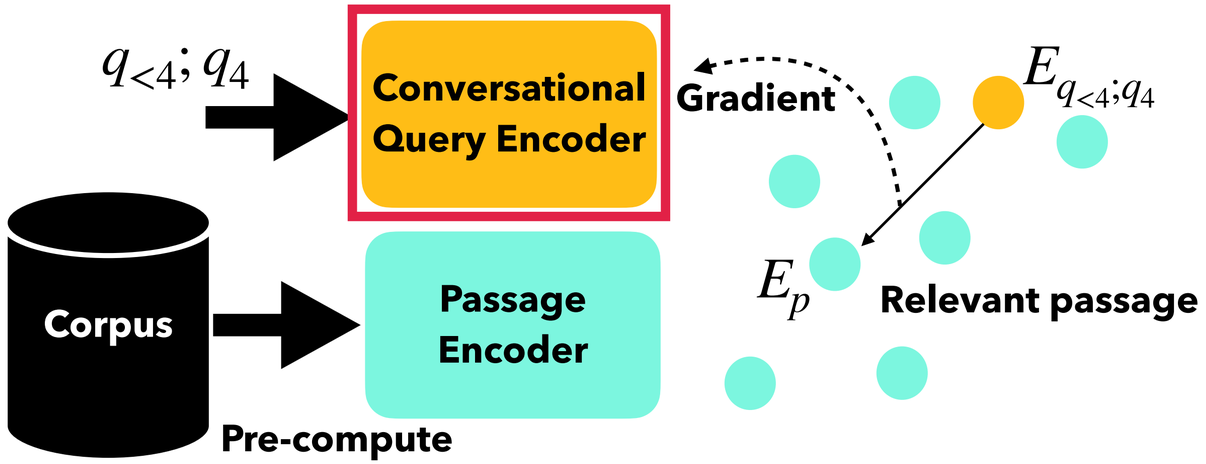
q_3 **Tell me about lung cancer.**

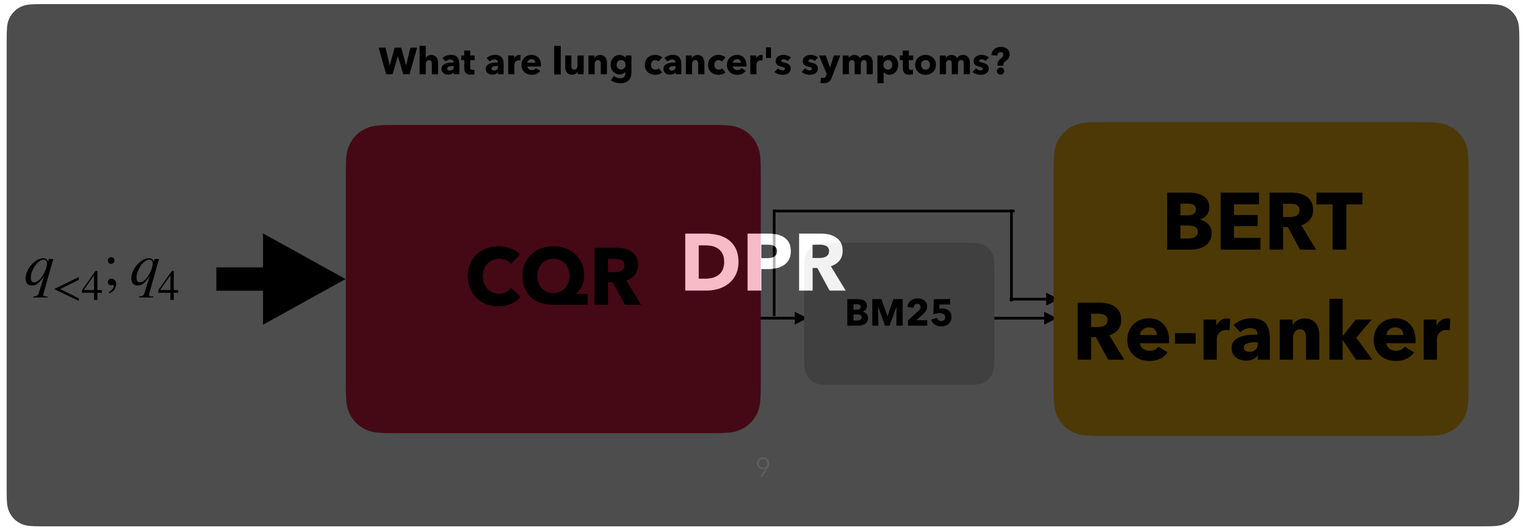
q_4 **What are its symptoms?**

[9] Dalton et al. 2020



Our Idea: Conversational DPR



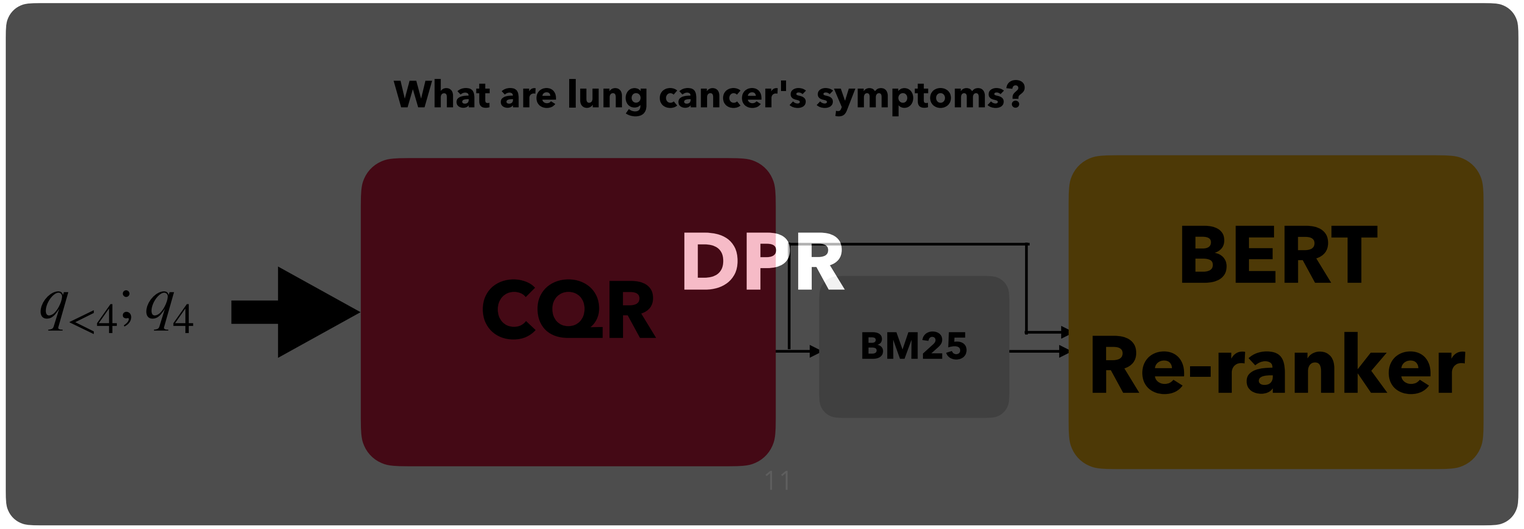



Kevin
160229

Challenge

Table 2: CAsT dataset statistics.

	CAsT19		CAsT20
	Training	Eval	Eval
# Queries	108	173	208
# Dialogues	13	20	25
# Passages		38M	



Successfully Apply DPR to Conversational Search

Interpret how DPR Understand Conversational Query

Fine-tuning with Pseudo Relevance Judgement

Table 1: CANARD dataset statistics.

CANARD	Training	Dev	Test
# Queries	31,526	3,430	5,571
# Dialogues	4,383	490	771

[4] Elgohary et al. 2020

Fine-tuning with Pseudo Relevance Judgement

Question Rewriting

Q1:	What happened in 1983?	What happened to Anna Vissi in 1983?

A1: In May 1983, she marries Nikos Karvelas, a composer

Q2:	Did they have any children?	Did Anna Vissi and Nikos Karvelas have any children together?

A2: In November, she gave birth to her daughter Sofia

Q3:	Did she have any other children?	Did Anna Vissi have any other children than her daughter Sofia?

[4] Elgohary et al. 2020

Fine-tuning with Pseudo Relevance Judgement

Question Rewriting

Q1:	What happened in 1983?	What happened to Anna Vissi in 1983?

A1: In May 1983, she marries Nikos Karvelas, a composer

Q2:	Did they have any children?	Did Anna Vissi and Nikos Karvelas have any children together?

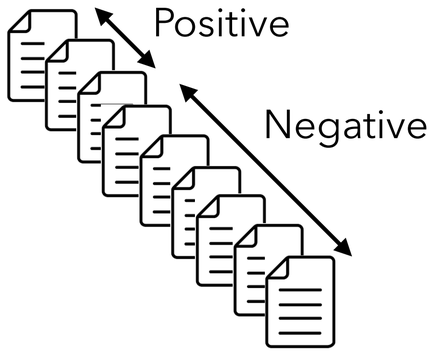
A2: In November, she gave birth to her daughter Sofia

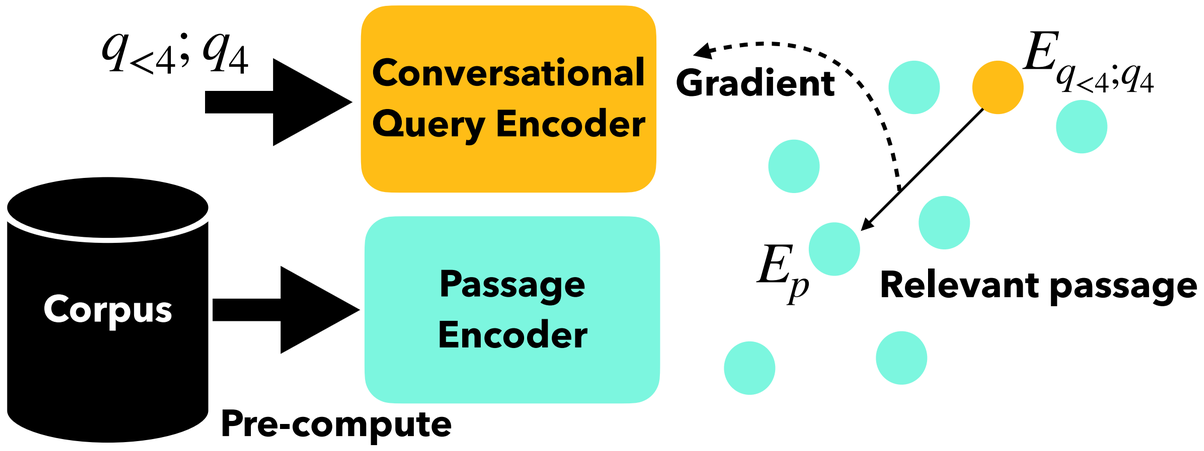
Q3:	Did she have any other children?	Did Anna Vissi have any other children than her daughter Sofia?

[4] Elgohary et al. 2020

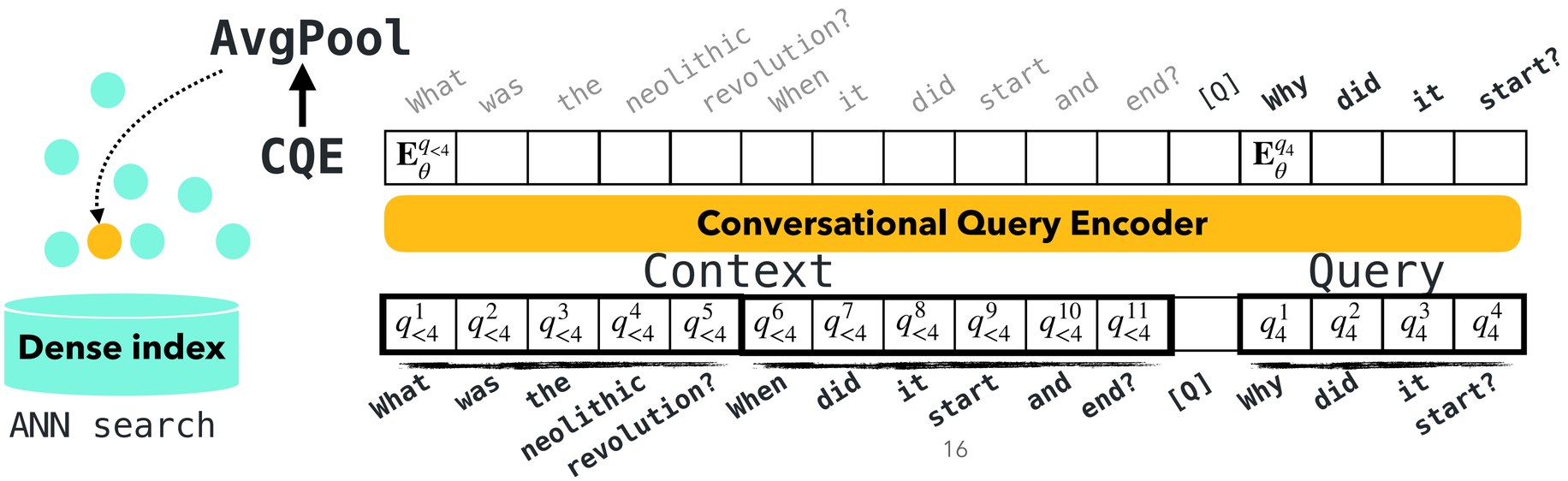
Query

[CLS] What happened in 1983 ? | Did they have any children? [Q] Did she have any other children?





Contextualized Query Embeddings



Successfully Apply DPR to Conversational Search

- Reformulate conversational query directly in dense space
- Create training data with pseudo relevance judgement

Interpret how DPR Understand Conversational Query

- Text compression (or filtering)

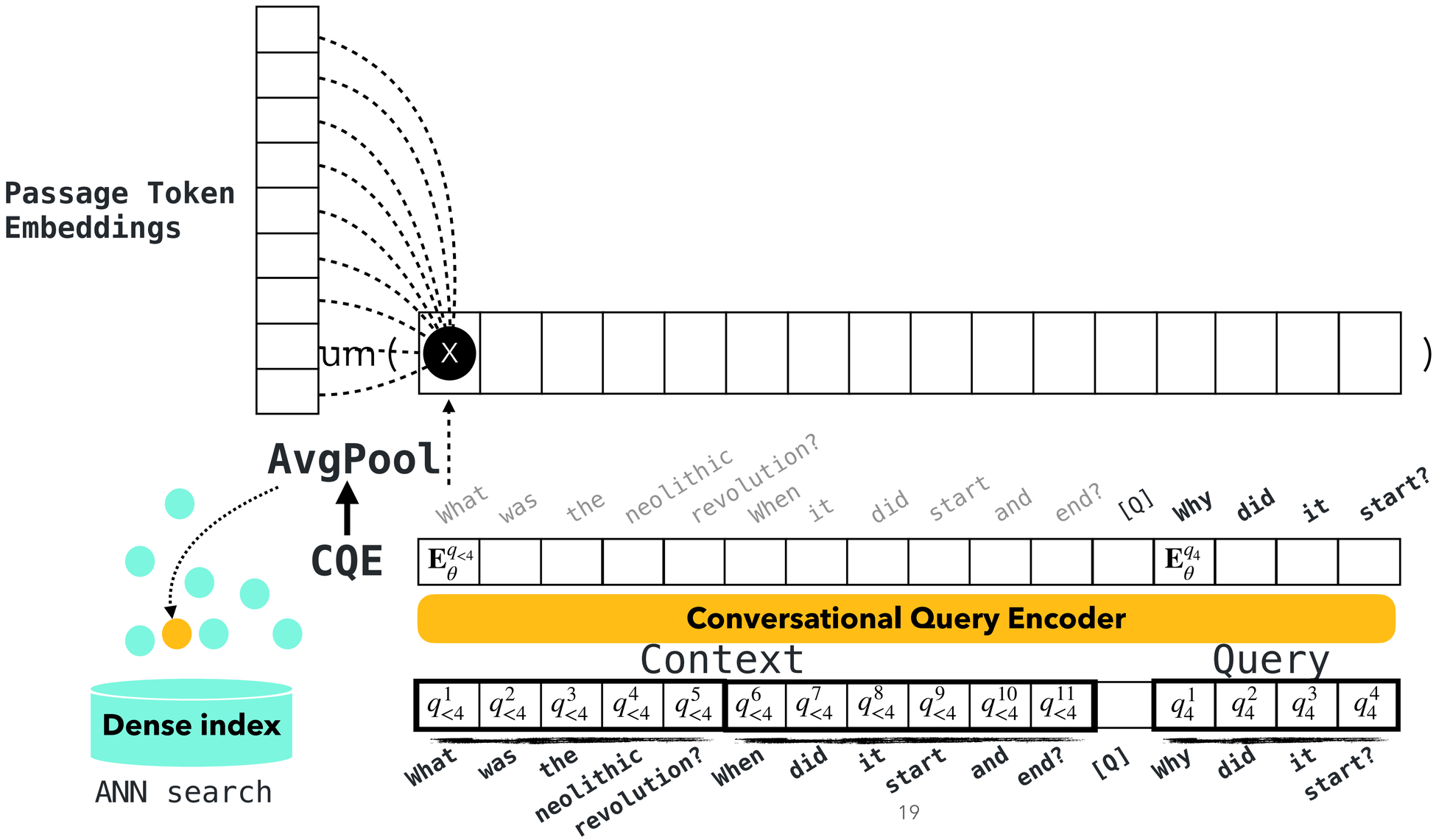
Successfully Apply DPR to Conversational Search

- Reformulate conversational query directly in dense space
- Create training data with pseudo relevance judgement

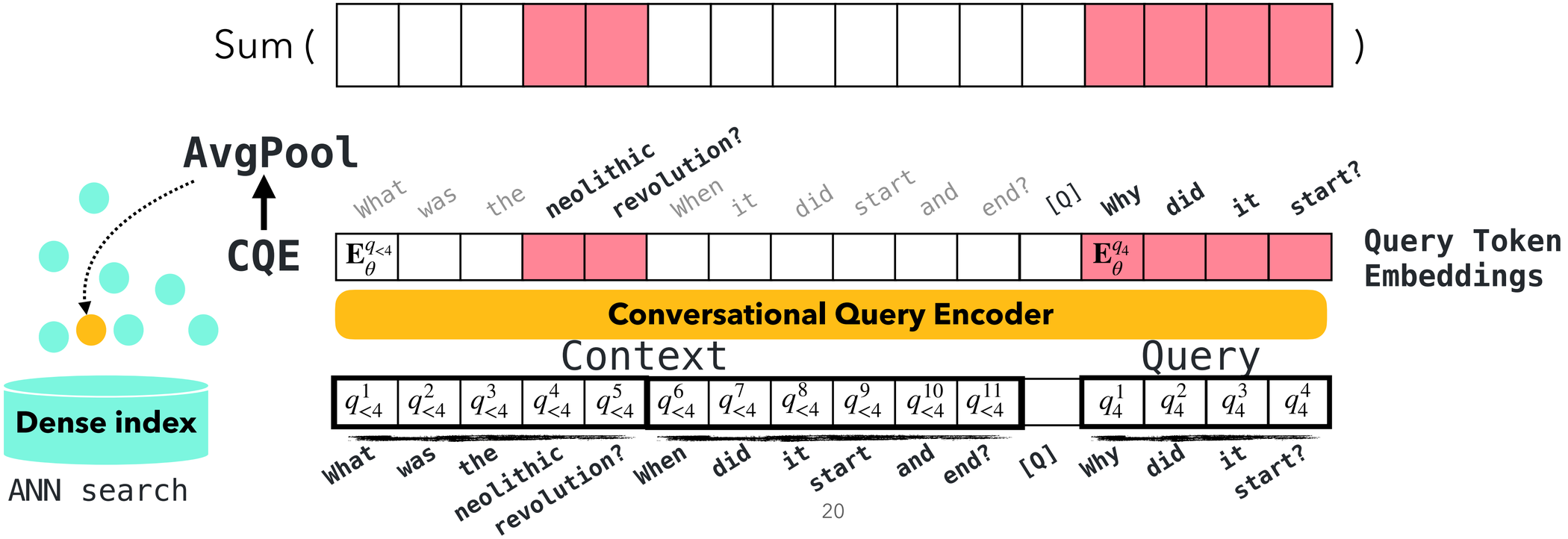
Interpret how DPR Understand Conversational Query

- Text compression (or filtering)

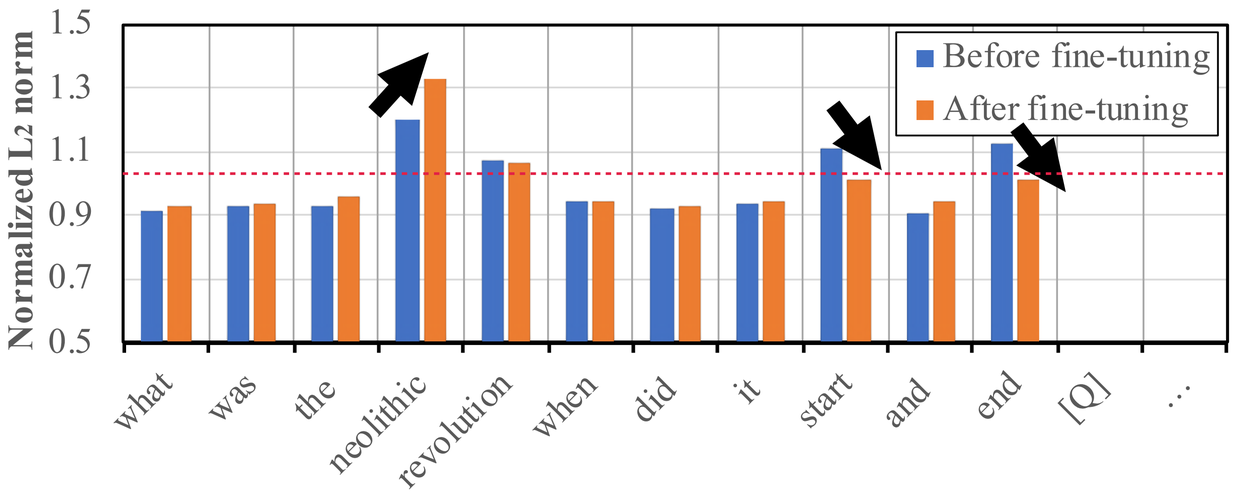
Interpretation

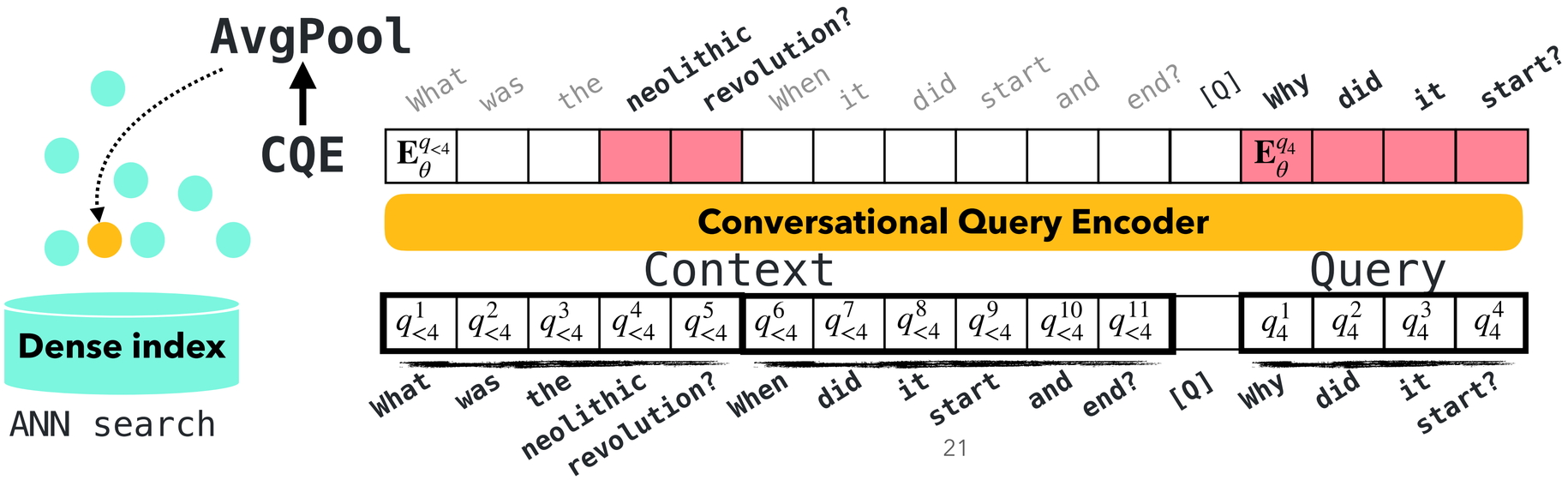


Interpretation

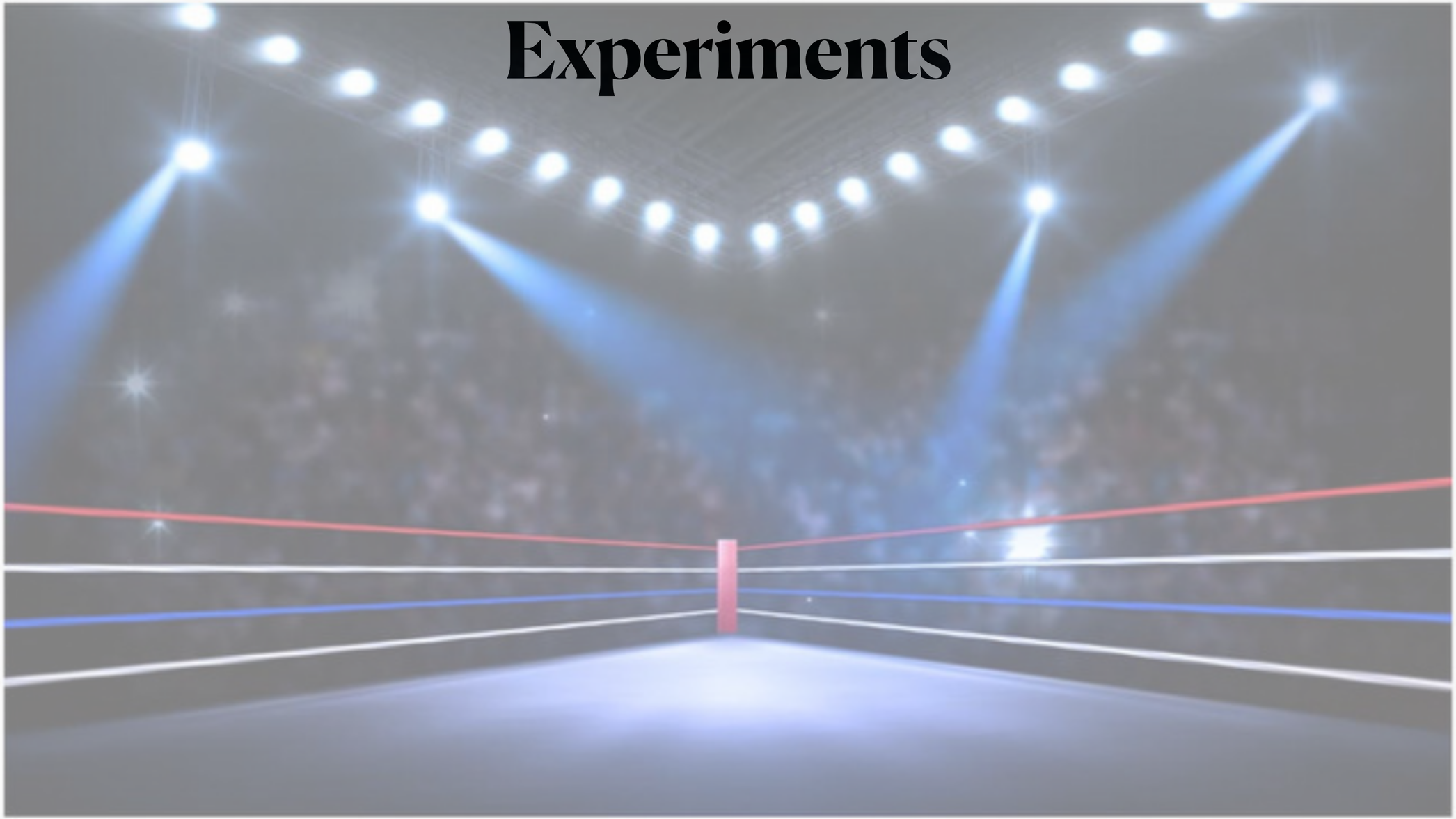


Interpretation





Experiments



Experiments

Models

CQE

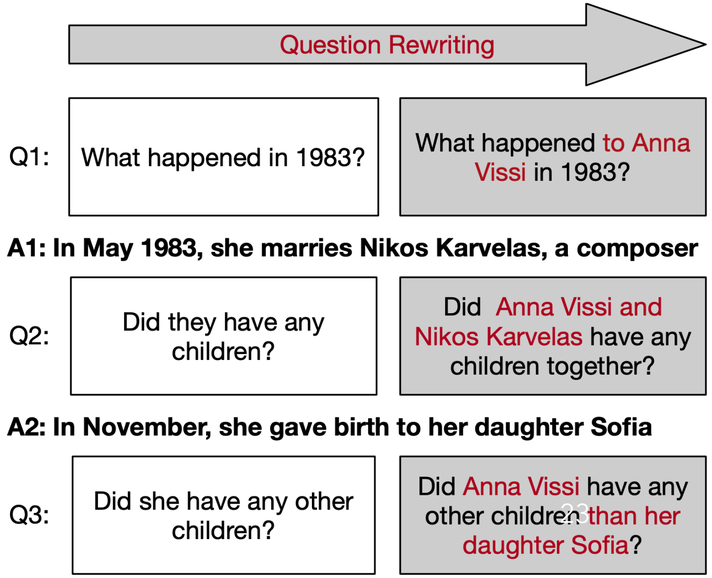
DPR

CQR

Training Data

Table 1: CANARD dataset statistics.

CANARD	Training	Dev	Test
# Queries	31,526	3,430	5,571
# Dialogues	4,383	490	771

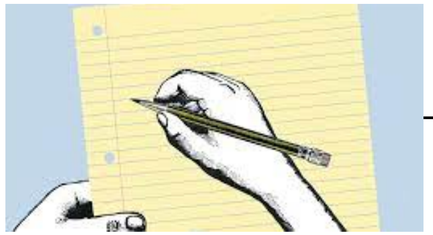


Eval Data

Table 2: CAsT dataset statistics.

	CAsT19		CAsT20
	Training	Eval	Eval
# Queries	108	173	208
# Dialogues	13	20	25
# Passages		38M	

Dense Retrieval

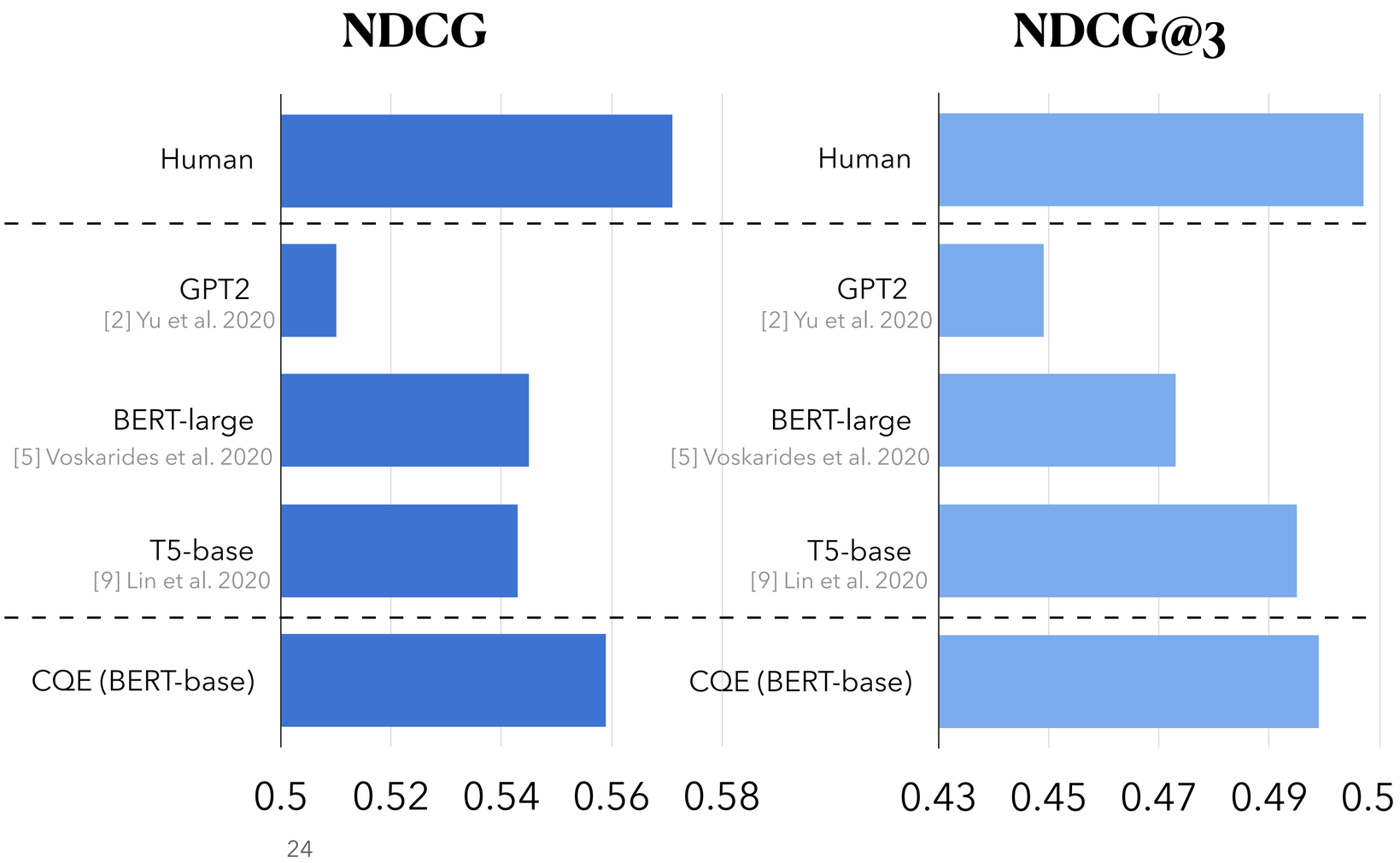


DPR

CQR

DPR

CQE
DPR



Sparse Retrieval

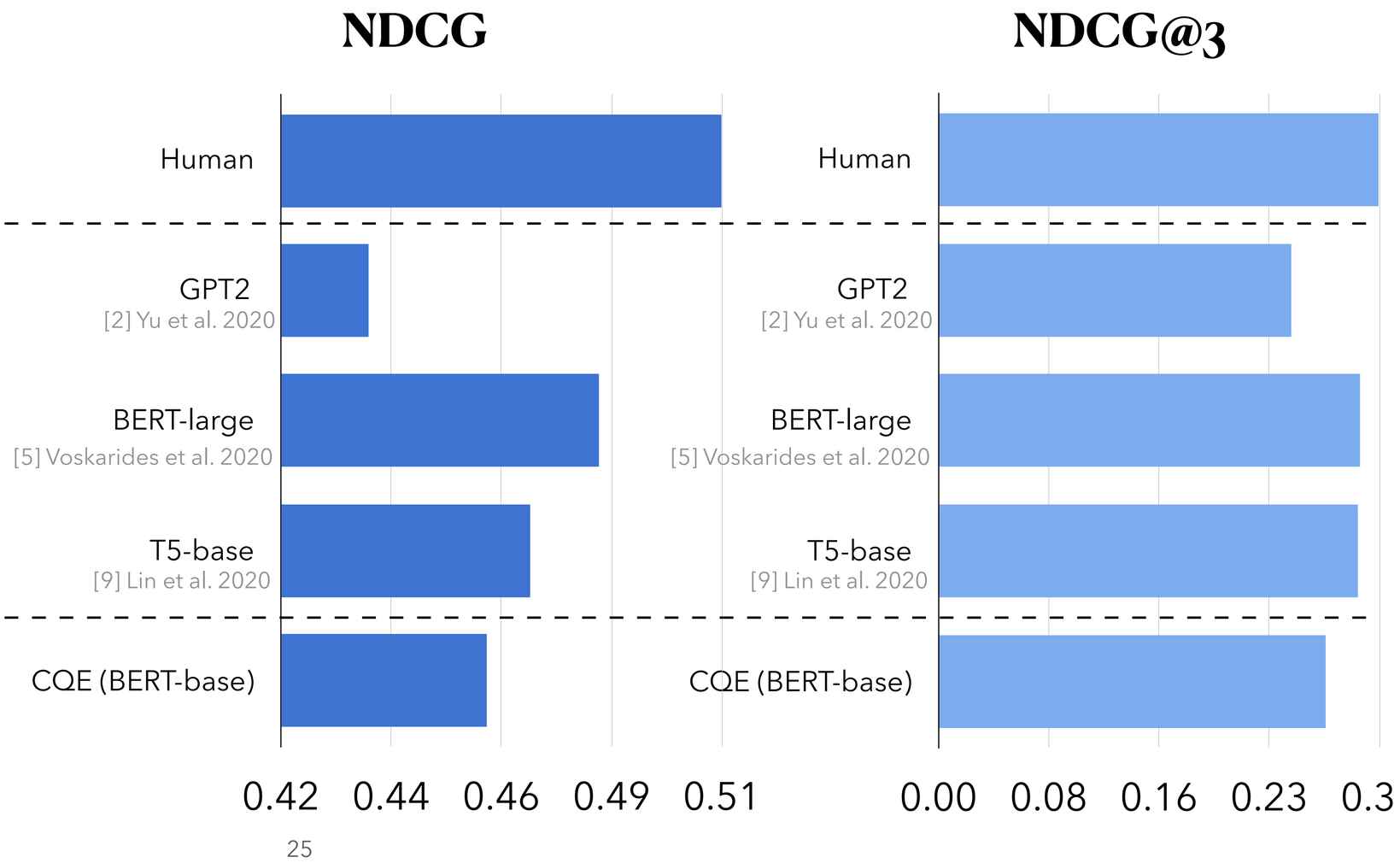
BM25

CQR

BM25

CQE
DPR

BM25



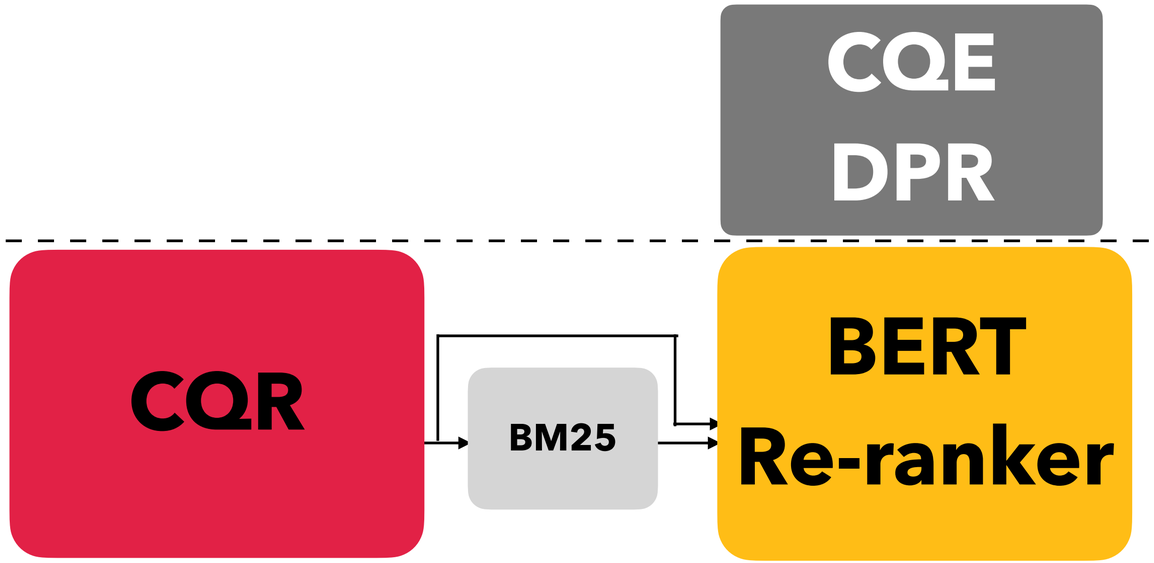
Comparison to Multi-stage Pipeline

CQE

DPR

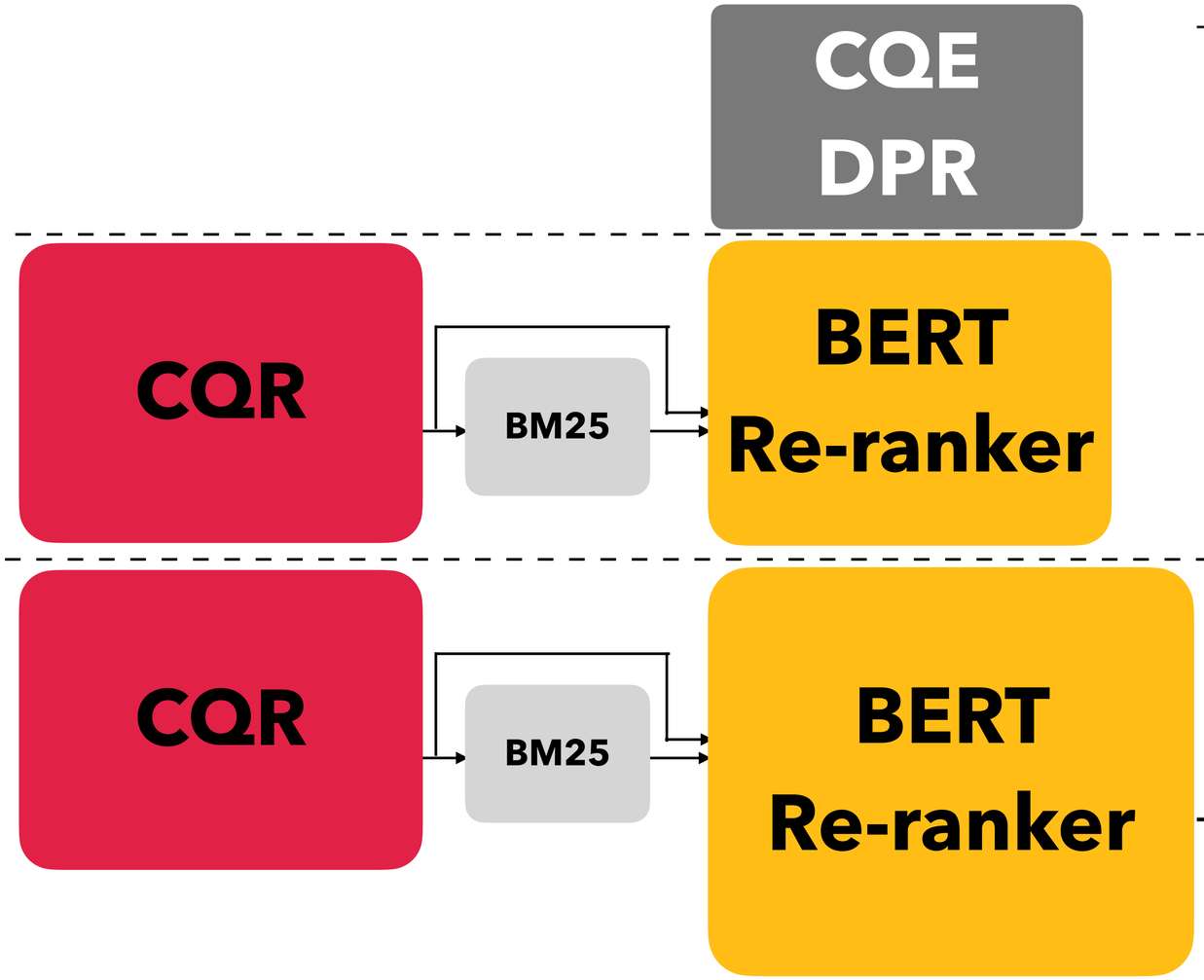
CAsT19 Eval	nDCG@3
BERT-base: latency = 314 ms	
CQE	.499
CQE-hybrid	.515

Comparison to Multi-stage Pipeline



CAsT19 Eval	nDCG@3
BERT-base: latency = 314 ms	
CQE	.499
CQE-hybrid	.515
CQR + BM25 + BERT-base: latency = 5,350 ms	
QuReTec (Voskarides et al., 2020)	.476
Few-Shot Rewriter (Yu et al., 2020)	.492
3CQR + BM25 + BERT-base: latency = 8,025 ms (est.)	
MVR (Kumar and Callan, 2020)	.565

Comparison to Multi-stage Pipeline



CAsT19 Eval	nDCG@3
BERT-base: latency = 314 ms	
CQE	.499
CQE-hybrid	.515
CQR + BM25 + BERT-base: latency = 5,350 ms	
QuReTec (Voskarides et al., 2020)	.476
Few-Shot Rewriter (Yu et al., 2020)	.492
3CQR + BM25 + BERT-base: latency = 8,025 ms (est.)	
MVR (Kumar and Callan, 2020)	.565
CQR + BM25 + BERT-large: latency = 16,450 ms	
Transformer++ (Vakulenko et al., 2020)	.529
NTR (T5) (Lin et al., 2021c)	.556
HQE + NTR (T5) (Lin et al., 2021c)	.565

Conclusions and Future Work

- Create training data with pseudo relevance judgement
  - Reformulate conversational query directly in dense space
  - Create training data with pseudo relevance judgement
- Explain how DPR reformulates queries in embedding space
  - Text compression (or filtering)
- Future Work:
  - Add system responses as context

Reference

[1] https://blog.google/products/search/the-new-conversational-search-experience-were-thankful-for/

[2] Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2019. CAsT 2019: The conversational assistance track overview. In *Proc. TREC*.

[3] Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with BERT. *arXiv:1901.04085*.

[4] Ahmed Elgohary, Denis Peskov, and Jordan Boyd- Graber. 2019. Can you unpack that? Learning to rewrite questions-in-context. In *Proc. EMNLP*, pages 5917–5923.

[5] Nikos Voskarides, Dan Li, Pengjie Ren, Evangelos Kanoulas, and Maarten de Rijke. 2020. Query res- olution for conversational search with limited supervision. In *Proc. SIGIR*, pages 921–930.

[6] Shi Yu, Jiahua Liu, Jingqin Yang, Chenyan Xiong, Paul Bennett, Jianfeng Gao, and Zhiyuan Liu. 2020. Few- shot generative conversational query rewriting. In *Proc. SIGIR*, pages 1933–1936.

[7] Vaibhav Kumar and Jamie Callan. 2020. Making in- formation seeking easier: An improved pipeline for conversational search. In *Proc. EMNLP Findings*.

[8] Svitlana Vakulenko, Shayne Longpre, Zhucheng Tu, and Raviteja Anantha. 2020. Question rewriting for conversational question answering. *arXiv:2004.14652*.

[9] Sheng-Chieh Lin, Jheng-Hong Yang, Rodrigo Nogueira, Ming-Feng Tsai, Chuan-Ju Wang, and Jimmy Lin. 2021c. Multi-stage conversational passage retrieval: An approach to fusing term importance estimation and neural query rewriting. *ACM Trans. Inf. Syst.*, 39(4).