TECHNISCHE UNIVERSITÄT DARMSTADT

UBIQUITOUS KNOWLEDGE PROCESSING

HOW CAN WE AVOID GETTING CARRIED AWAY IN AN OCEAN OF QA MODELS AND INSTEAD BENEFIT FROM THEM?

Huawei Semantic Search Academic Workshop

21st November 2022

TECHNISCHE UNIVERSITÄT DARMSTADT

AGENDA

1 New Possibilities

2 Meta-QA

3 Relevance Feedback In Neural Re-Ranking

4 UKP-SQuARE

5 Future Work

TECHNISCHE UNIVERSITÄT DARMSTADT

CHAPTER

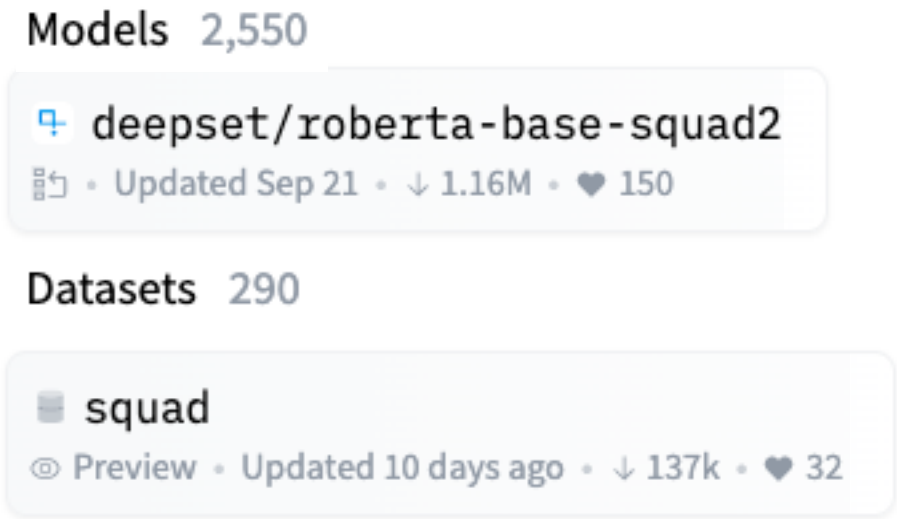
NEW POSSIBILITIES

TECHNISCHE UNIVERSITÄT DARMSTADT

EXPLOSION OF QA DATASETS AND MODELS

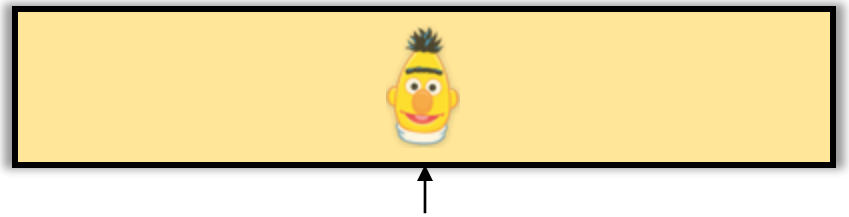
How to leverage all this collective effort?

How to study all these models and datasets?

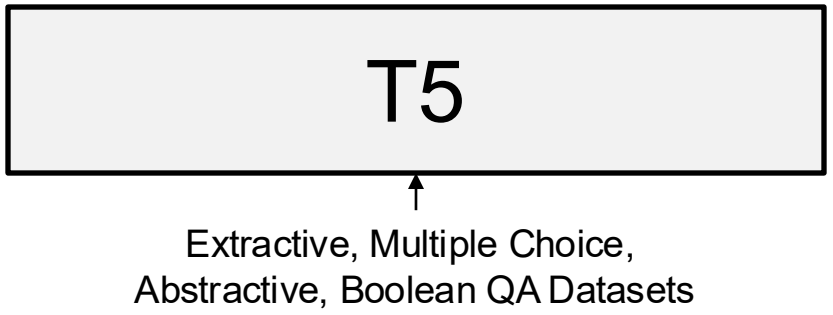


TECHNISCHE UNIVERSITÄT DARMSTADT

MULTI-DATASET MODELS

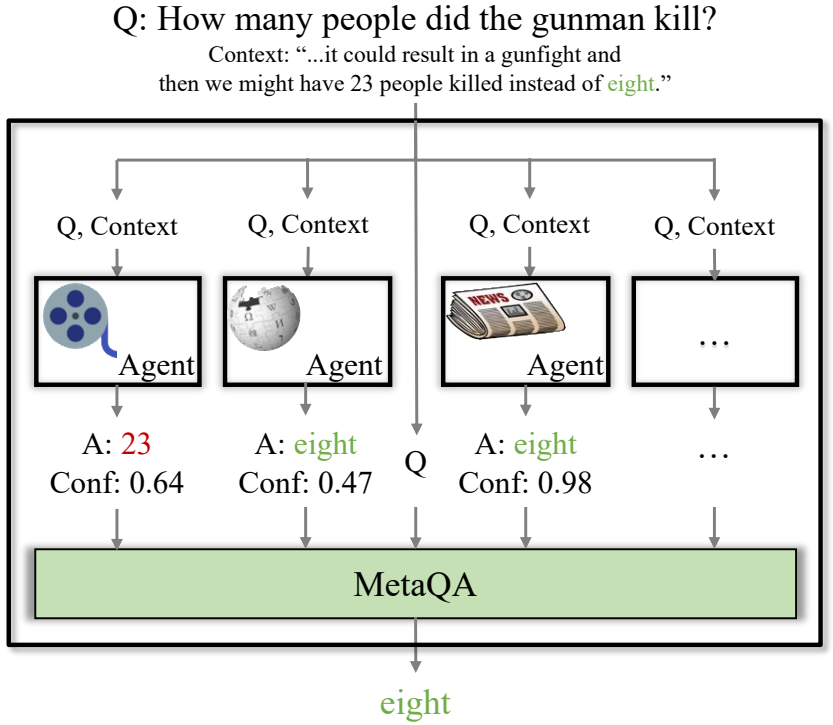


MultiQA (Talmor & Berant, ACL 2019)



UNIFIEDQA (Khashabi et al., Findings 2020)

MULTI-AGENT MODELS



MetaQA (Puerto et al., arXiv 2021)

TECHNISCHE UNIVERSITÄT DARMSTADT

CHAPTER

META-QA

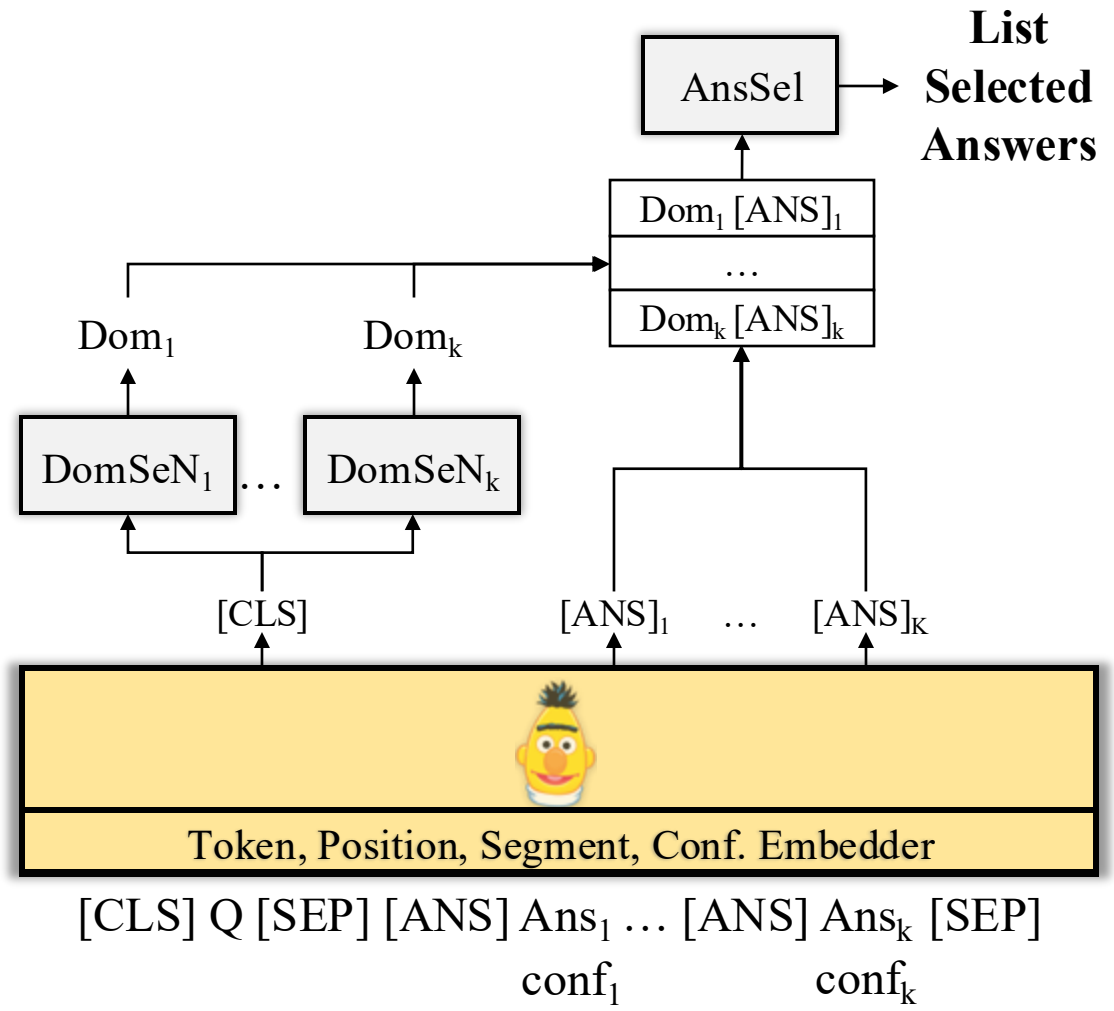
TECHNISCHE UNIVERSITÄT DARMSTADT

Goals

- Identify the best answer (even if it comes from an OOD agent)

Multi-Task

- Domain Selection
- Answer Selection



TECHNISCHE UNIVERSITÄT DARMSTADT

EXPERIMENTS

1 agent for each dataset

	Dataset
Extractive	SQuAD (Rajpurkar et al., 2016) NewsQA (Trischler et al., 2017) HotpotQA (Yang et al., 2018) SearchQA (Dunn et al., 2017) NQ (Kwiatkowski et al., 2019) TriviaQA-web (Joshi et al., 2017) QAMR (Michael et al., 2018) DuoRC (Saha et al., 2018)
Multiple-Choice	RACE (Lai et al., 2017) CSQA (Talmor et al., 2019) BoolQ (Clark et al., 2019) HellaSWAG (Zellers et al., 2019) SIQA (Sap et al., 2019)
Abs.	DROP (Dua et al., 2019) NarrativeQA (Kočiský et al., 2018)
MM	HybridQA (Chen et al., 2020)

TECHNISCHE UNIVERSITÄT DARMSTADT

RESULTS

- If many and very different datasets/tasks

  → multi-agent systems are better

Model	Avg. F1
MetaQA	**76.00**
UnifiedQA	65.90

Dataset	MetaQA	UnifiedQA
NewsQA	**71.71**	65.57
TriviaQA-web	**80.63**	72.34
NQ	**81.20**	75.58
DuoRC	**51.24**	34.65
CSQA	**78.66**	58.43
HellaSWAG	**73.19**	36.01
RACE	**84.71**	69.65
SIQA	**74.17**	61.62
DROP	**73.04**	42.45

TECHNISCHE UNIVERSITÄT DARMSTADT

OOD RESULTS

- OOD MetaQA outperforms OOD UnifiedQA by 8.45 in F1
- OOD MetaQA outperforms in-domain UnifiedQA in 4 datasets

Dataset	TriviaQA	DuoRC	CSQA	HellaSWAG
MetaQA	80.65	51.01	78.40	72.14
UnifiedQA	72.34	34.65	58.43	36.01
OOD MetaQA	**77.26**	**50.64**	**58.75**	**51.94**
OOD UnifiedQA	69.33	32.84	**50.57**	29.35

Leave-one-out evaluation. SQuAD F1 metric

TECHNISCHE UNIVERSITÄT DARMSTADT

AGENT COLLABORATION

Dataset	In-Domain Agent	OOD Agent
DuoRC	48%	NewsQA (18%)
TriviaQA	80%	DuoRC (3%)
SearchQA	86%	TriviaQA (5%)

Dataset	Question	In-domain Agent	OOD Agent
DuoRC	Who does Rocky Balboa work for as an enforcer?	Adrian	Tony Gazzo (NewsQA Agent)
TriviaQA-web	Who played the character Mr Chips in the 2002 TV adaptation of Goodbye Mr Chips?	Timothy Carroll	MartinClunes (DuoRC Agent)
SearchQA	This short story, written around 1820, contains the line "If I can but reach that bridge... I am safe"	Legend	Legend of Sleepy Hollow (TriviaQA Agent)

Table 3: Examples of questions where our MetaQA system disregard the in-domain agent due to their incorrect predictions (in red) and selects and an out-of-domain (OOD) agent that returns the right answer (in green).

TECHNISCHE UNIVERSITÄT DARMSTADT

MetaQA: Combining Expert Agents for Multi-Skill Question Answering

Haritz Puerto, Gözde Gül Şahin, Iryna Gurevych
Ubiquitous Knowledge Processing Lab (UKP-TUDA)
Department of Computer Science, Technical University of Darmstadt
https://www.ukp.tu-darmstadt.de
puerto@ukp.informatik.tu-darmstadt.de

TECHNISCHE UNIVERSITÄT DARMSTADT

CHAPTER

RELEVANCE FEEDBACK IN NEURAL RE-RANKING

TECHNISCHE UNIVERSITÄT DARMSTADT

QUERY TYPES [1]

	Navigational / Transactional	Information-Seeking
Example query	“arxiv sentence bert” “acl paper submission”	“origin coronavirus” “neural sentence representation”
# relevant documents	1-few	Many
Scenarios	Navigation Known-Item Retrieval	Scientific Literature Review News Background Case Law Factoid QA

MOTIVATION

Challenges

- Unknown Search Domain
- Difficult Query Formulation

Observation

- Judging a document is easier than formulating good queries
- Some relevant documents might be already known to the user

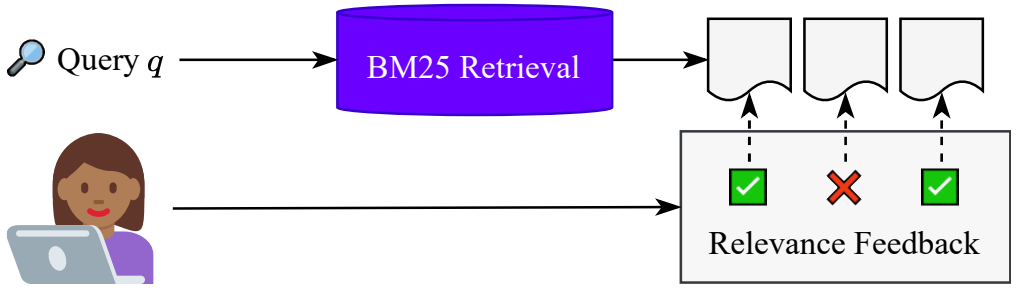
→ **Use relevance feedback to improve search**

TECHNISCHE UNIVERSITÄT DARMSTADT

BACKGROUND

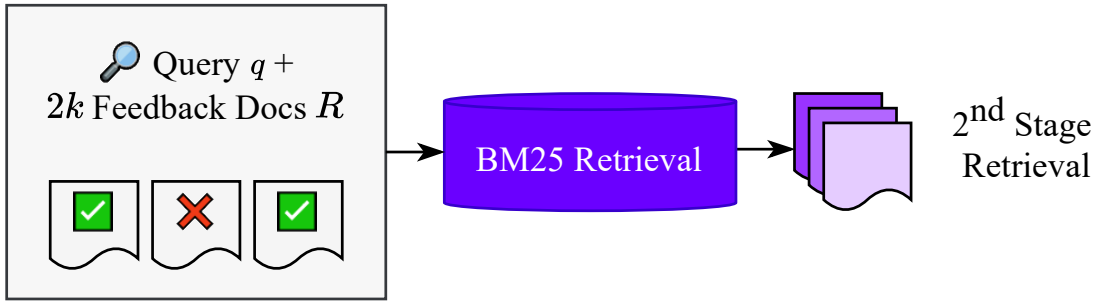
1) Retrieve Documents using the query

2) Obtain feedback on retrieved documents



3) Extract ”expansion terms” from documents

4) Retrieve with query + expansion terms



How to integrate relevance feedback into neural retrieval?

TECHNISCHE UNIVERSITÄT DARMSTADT

TASK SETUP

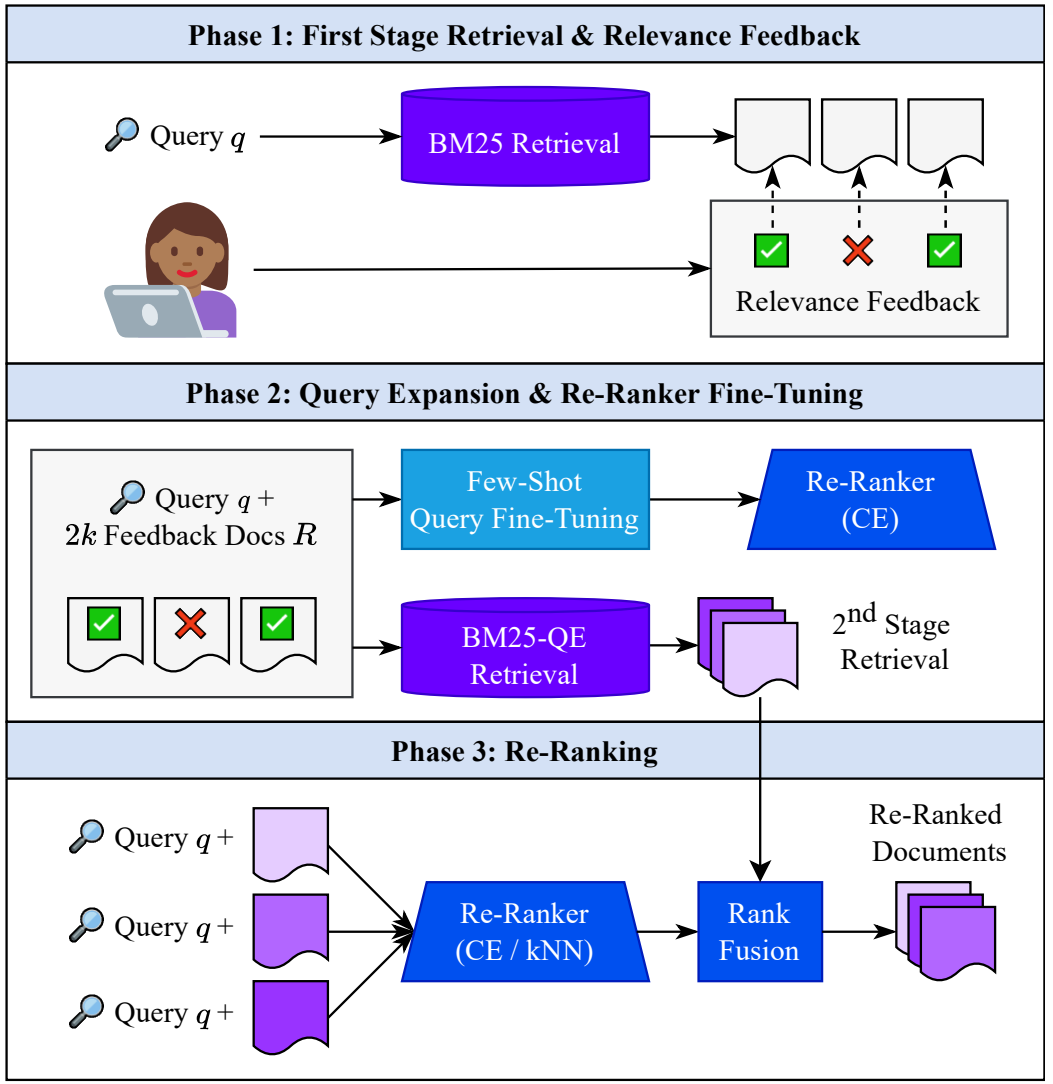
Goal: Re-rank documents from 2. Stage Retrieval with neural re-rankers incorporating relevance feedback

Input

- Query q
- k relevant and non-relevant feedback documents, where $k \in \{2, 4, 8\}$
- 1000 documents from 2. stage retrieval

Evaluation

- Re-ranking [nDCG@20]

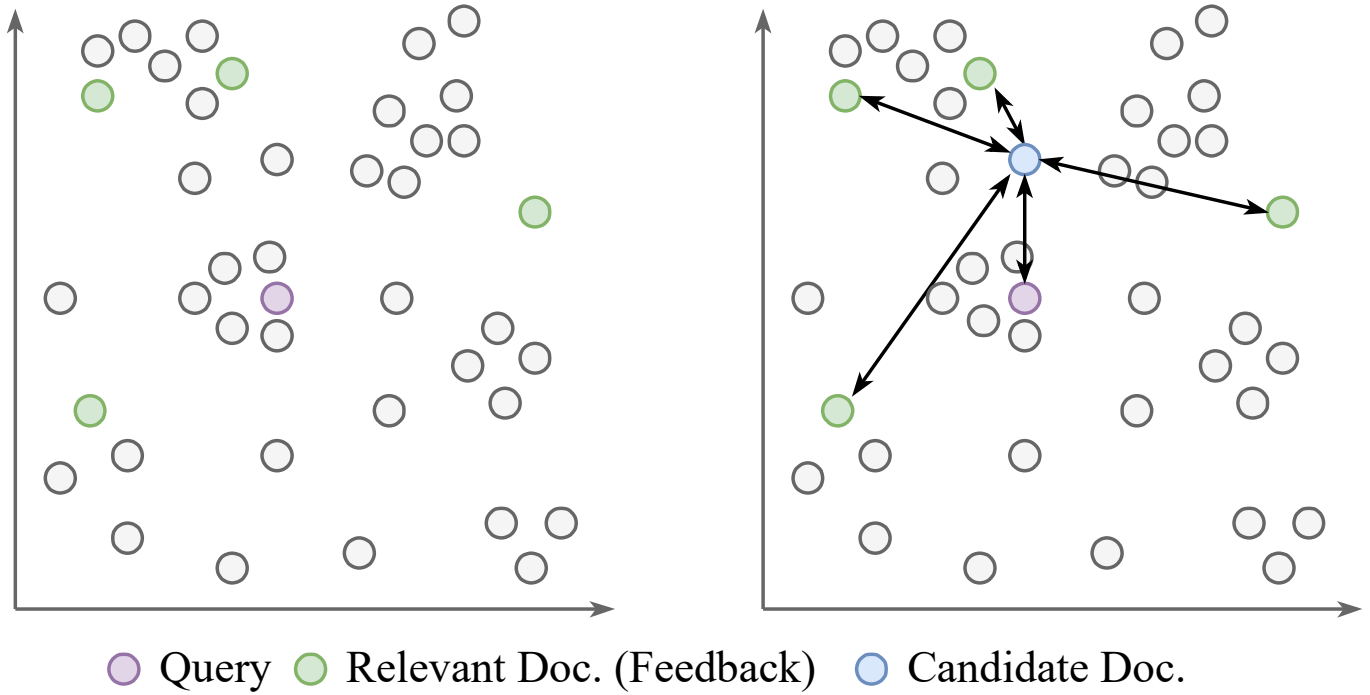


TECHNISCHE UNIVERSITÄT DARMSTADT

METHOD: KNN

- Compute document representations $d_i \in D$
- Score documents by summing the similarities between the candidate document d_i, query q and relevant feedback documents $d_j \in R^+$

$$s_i = f(d_i, q) + \sum_{d_j \in R^+} f(d_i, d_j)$$



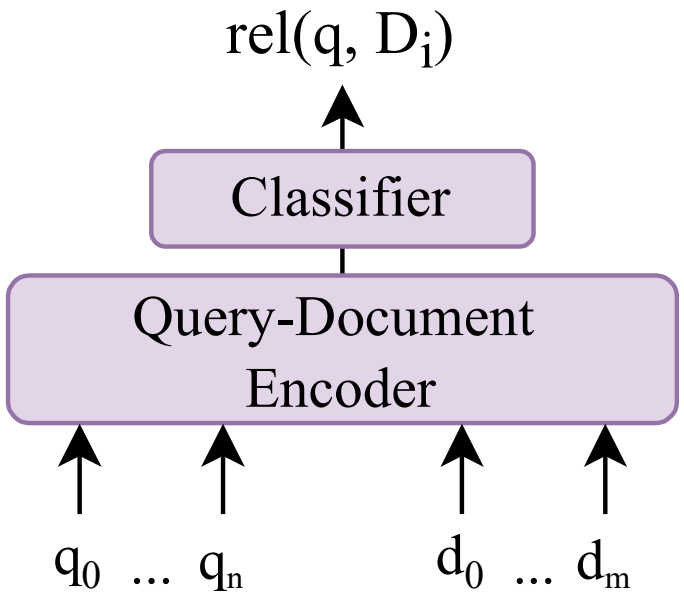
METHOD: CROSS-ENCODER [2]

CE Query Fine-Tuning

Fine-Tune bias layers per query on $2k$ Relevance Feedback Documents

CE MAML + Query Fine-Tuning

1. Fine-Tune bias layers on in-domain annotations with Meta-Learning to obtain ”fast parameters”
2. Fine-Tune per query on on $2k$ Feedback Documents



METHOD: RANK-FUSION [3]

Idea: Merge rankings of different ranking functions $h \in H$

Problem: Different methods produce different scores, simple adding the scores is biased

=> Use ranks instead of raw scores

$$s_i = \sum_{h \in H} \frac{1}{c + h(d_i)}$$

$h(d_i)$ returns the integer rank that method h assigns document d_i

c is constant smoothing the impact of top ranked documents

DATASETS

Dataset	Domain	Docs.	Queries	Judgments
Robust04	News	528k	148	1287 (±501)
TREC-Covid	Biomedical	191k	50	1370 (±323)
TREC-News	News	595k	34	259 (±82)
Touché-2020	Debates	383k	49	50 (±7)

Includes only queries with at least 32 relevant documents.

RESULTS

Method	Robust	Covid	News	Touché	Avg
BM25-QE (2. Stage Retrieval)	**0.496**	0.610	**0.392**	**0.271**	**0.442**
kNN	0.443	0.686	0.365	0.174	0.417
CE Zero-Shot	0.415	**0.702**	0.314	0.176	0.402

=> BM25-QE is hard to beat!

RESULTS

Method	Robust	Covid	News	Touché	Avg
BM25-QE (2. Stage Retrieval)	0.496	0.610	**0.392**	**0.271**	**0.442**
CE Query Fine-Tune	0.484	0.723	0.335	0.198	0.435
CE MAML + Query FT	**0.506**	**0.735**	0.314	0.223	**0.445**

=> Fine-tuning per query helps
=> CE MAML + Query FT is on par with BM25-QE

RESULTS

Method	Robust	Covid	News	Touché	Avg
BM25-QE (2. Stage Retrieval)	0.496	0.610	0.392	**0.271**	0.442
kNN	0.443	0.686	0.365	0.174	0.417
CE Zero-Shot	0.415	0.702	0.314	0.176	0.402
CE Query Fine-Tune	0.484	0.723	0.335	0.198	0.435
CE MAML + Query FT	0.506	0.735	0.314	0.223	0.445
BM25-QE ∩ kNN	0.507	0.707	**0.412**	0.248	0.468
BM25-QE ∩ CE MAML + Query FT	**0.570**	**0.740**	0.405	**0.272**	**0.497**

=> Fusing BM25-QE and neural model is highly effective!

TECHNISCHE UNIVERSITÄT DARMSTADT

Incorporating Relevance Feedback for Information-Seeking Retrieval using Few-Shot Document Re-Ranking

Tim Baumgärtner,[1] Leonardo F. R. Ribeiro,[1*] Nils Reimers,[2] Iryna Gurevych[1]
[1]Ubiquitous Knowledge Processing Lab (UKP Lab),
Department of Computer Science and Hessian Center for AI (hessian.AI),
Technical University of Darmstadt
[2]cohere.ai
www.ukp.tu-darmstadt.de

EMNLP 2022

TECHNISCHE UNIVERSITÄT DARMSTADT

CHAPTER

UKP-SQUARE

AN ONLINE PLATFORM FOR QA RESEARCH

TECHNISCHE UNIVERSITÄT DARMSTADT

AN ECOSYSTEM FOR QA RESEARCH

- Common **interface**
- Easy **analysis** of the strengths, weaknesses, and biases
- Common **explainability** methods
- Common **adversarial** attacks
- Graph **visualizations**

ACL 2022 AACL 2022

- Running on the **browser** (no configurations, no installations)
- Re-use **data sources**
- Easy to **deploy** new **models**
- Dozens of **models available**

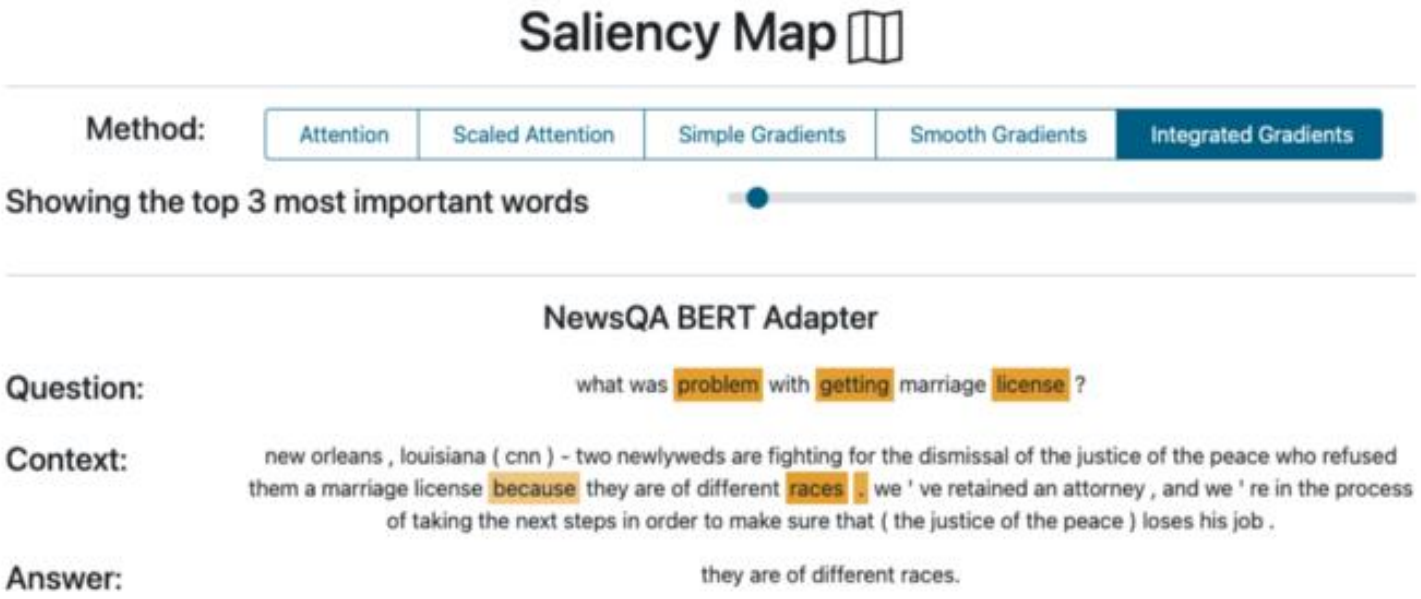
Data icons created by Freepik - Flaticon

TECHNISCHE UNIVERSITÄT DARMSTADT

EXPLAINABILITY

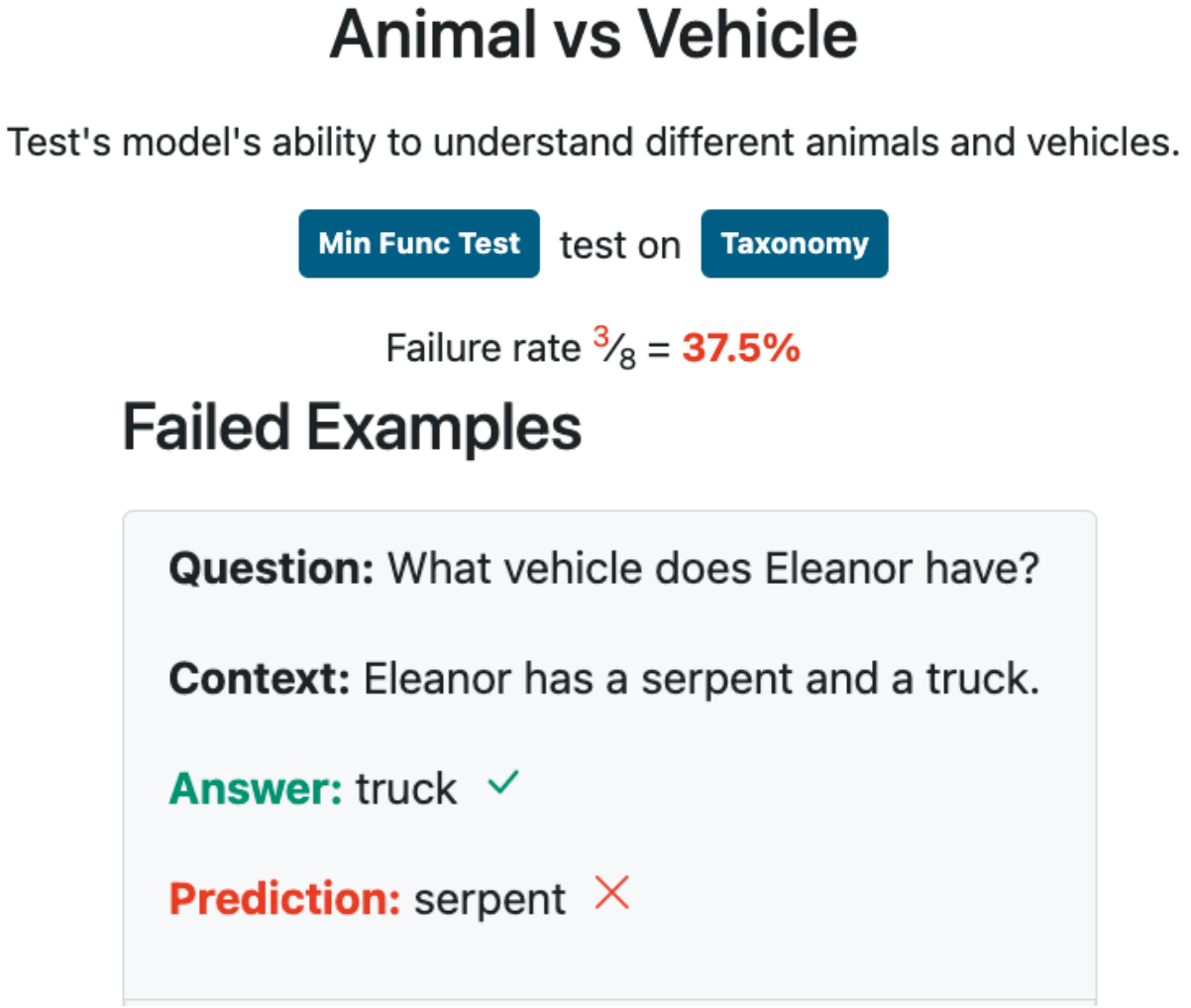
Saliency Maps

- Attention (Jain et al., ACL 2020)
- Scaled Attention (Serrano & Smith, ACL 2019)
- Simple Gradient (Simonyan et al., arXiv 2013)
- Smooth Gradients (Smilkov et al., arXiv 2017)
- Integrated Gradients (Sundararajan et al., PMLR 2017)



BEHAVIORAL TESTS

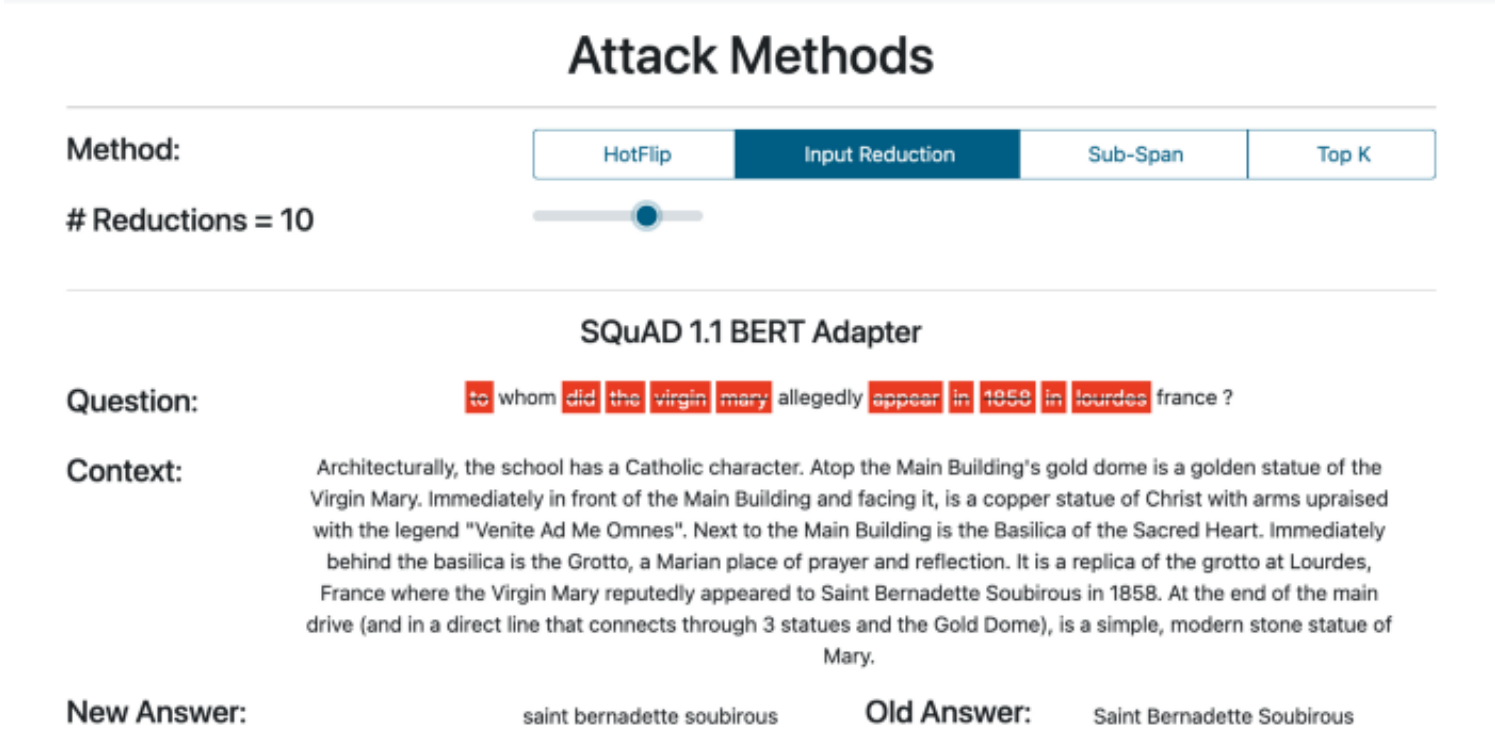
- **Unit Tests** of ML
- Useful to **detect shortcuts** to solve a type of questions
- Validates the input-output behavior w/o knowing the model internals



Beyond Accuracy: Behavioral Testing of NLP Models with CheckList (Ribeiro et al., ACL 2020)

ADVERSARIAL ATTACKS

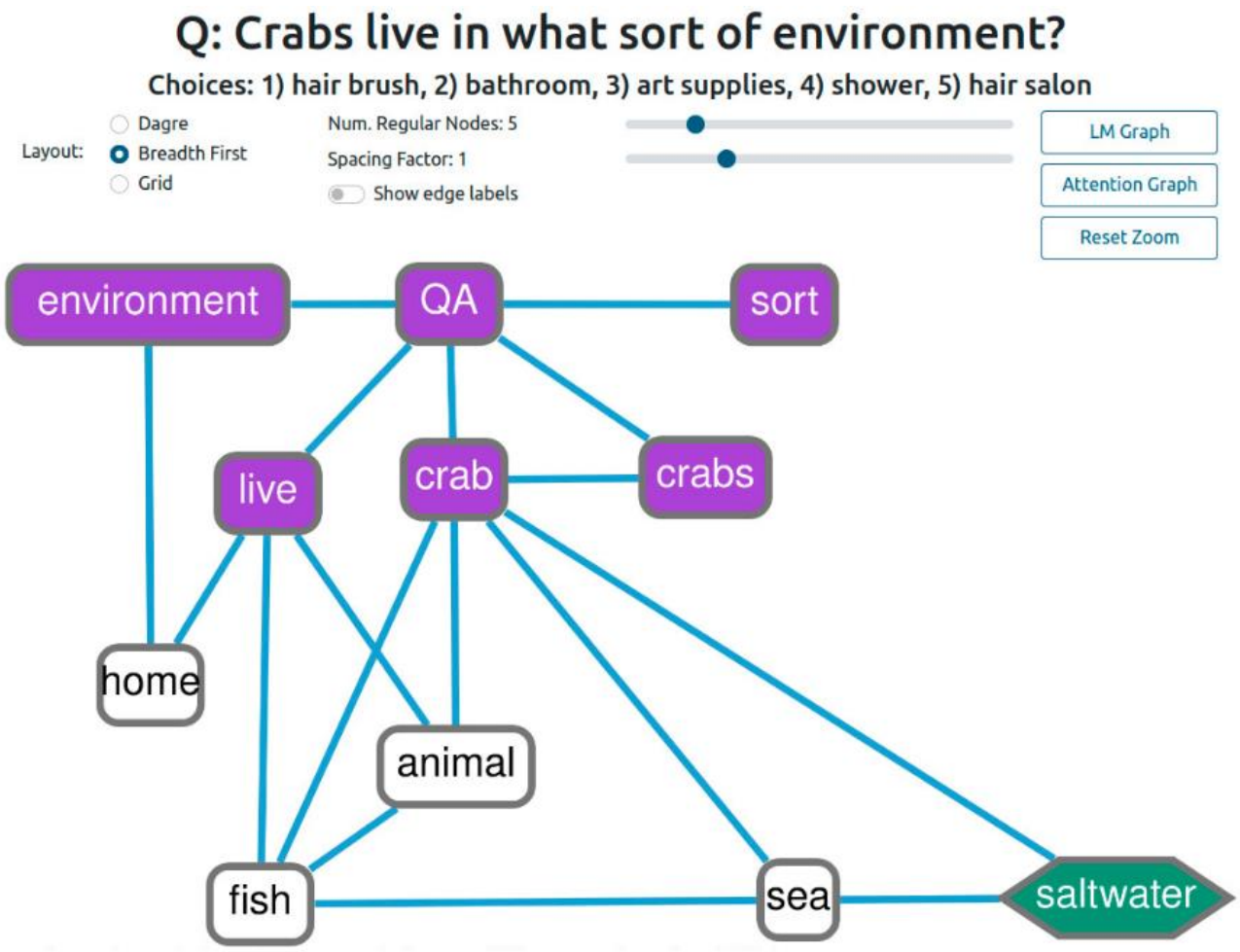
- HotFlip (Ebrahimi et al., ACL 2018)
- Input Reduction (Feng et al., EMNLP 2018)
- Sub-Span
- Top-K



GRAPH-BASED MODELS

QA-GNN

- LM + KB for common-sense QA
- KB: ConceptNet
- Graph Viz can help us understand the behavior of the model



QA-GNN: Reasoning with Language Models and Knowledge Graphs for Question Answering (Yasunaga et al., NAACL 2021)

TECHNISCHE UNIVERSITÄT DARMSTADT

DEMO

HTTPS://SQUARE.UKP-LAB.DE

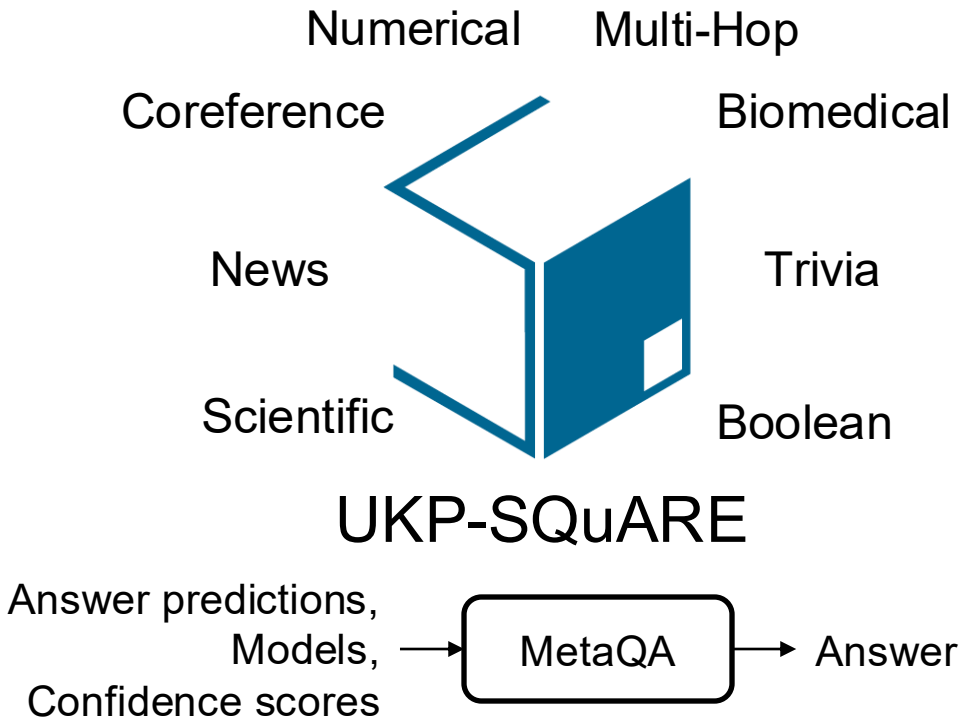
TECHNISCHE UNIVERSITÄT DARMSTADT

CHAPTER

FUTURE WORK

TECHNISCHE UNIVERSITÄT DARMSTADT

MULTI-AGENT QA SYSTEM



IR SYSTEM

- Downloadable & ready-to-use indices
- Relevance feedback incl. neural re-ranking
- Internet datastore using live Bing API

TECHNISCHE UNIVERSITÄT DARMSTADT

THANK YOU!

puerto@ukp.informatik.tu-darmstadt.de

https://square.ukp-lab.de

https://twitter.com/UKP_SQuARE

https://github.com/UKP-SQuARE

REFERENCES

TECHNISCHE UNIVERSITÄT DARMSTADT

[1] BRODER, A. (2002, SEPTEMBER). A TAXONOMY OF WEB SEARCH. IN ACM SIGIR FORUM (VOL. 36, NO. 2, PP. 3-10). NEW YORK, NY, USA: ACM.

[2] YATES, A., NOGUEIRA, R., & LIN, J. (2021, MARCH). PRETRAINED TRANSFORMERS FOR TEXT RANKING: BERT AND BEYOND. IN PROCEEDINGS OF THE 14TH ACM INTERNATIONAL CONFERENCE ON WEB SEARCH AND DATA MINING (PP. 1154-1156).

[3] CORMACK, G. V., CLARKE, C. L., & BUETTCHER, S. (2009, JULY). RECIPROCAL RANK FUSION OUTPERFORMS CONDORCET AND INDIVIDUAL RANK LEARNING METHODS. IN PROCEEDINGS OF THE 32ND INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL (PP. 758-759).

[4] FINN, C., ABBEEL, P., & LEVINE, S. (2017, JULY). MODEL-AGNOSTIC META-LEARNING FOR FAST ADAPTATION OF DEEP NETWORKS. IN INTERNATIONAL CONFERENCE ON MACHINE LEARNING (PP. 1126-1135). PMLR.

Dataset	MetaQA	TWEAC	Exp. Agent	UnifiedQA	MultiQA	Voting
SQuAD	91.98±0.11†	89.09±0.36	92.92	90.81	**93.14±0.18**	90.73
NewsQA	71.71±0.21†	66.86±0.75	**73.68**	65.57	73.59±0.60	66.60
HotpotQA	79.27±0.15†	74.96±0.59	80.60	77.92	**81.68±0.22**	71.71
SearchQA	**81.98±0.25**†‡	80.41±0.22	81.04	81.61	80.45±1.82	68.87
TriviaQA-web	**80.63±0.26**†‡	76.55±0.15	79.34	72.34	77.76±4.15	75.73
NQ	81.20±0.18†	78.06±0.37	81.97	75.58	**82.57±0.30**	72.25
DuoRC	**51.24±0.20**†‡	44.28±0.23	43.77	34.65	46.99±0.15	50.94
QAMR	83.78±0.14†	78.77±0.48	84.00	82.70	**84.62±0.14**	73.07
BoolQ	73.14±0.23†	72.20±0.03	72.17	**81.34**	n.a.	73.88
CSQA	**78.66±0.19**†	77.18±0.18	78.56	58.43	n.a.	68.41
HellaSWAG	73.19±1.01	77.12±0.30	**77.14**	36.01	n.a.	69.33
RACE	84.71±0.05†	83.02±0.27	**84.78**	69.65	n.a.	67.30
SIQA	74.17±0.64	75.39±0.05	**75.44**	61.62	n.a.	70.01
DROP	73.04±1.98†	69.12±0.36	**74.61**	42.45	n.a.	26.18
NarrativeQA	**67.19±0.00**	**67.19±0.00**	**67.19**	57.82	n.a.	**67.19**
HybridQA	**50.94±0.00**	**50.94±0.00**	**50.94**	n.a	n.a	**50.94**

TECHNISCHE UNIVERSITÄT DARMSTADT

Dataset	NewsQA	HotpotQA	SearchQA	TriviaQA	NQ	DuoRC	QAMR	CSQA	HellaSWAG	SIQA	DROP	Δ
MetaQA	71.46	79.37	81.87	80.65	81.08	51.01	83.87	78.40	72.14	73.90	74.96	-
UnifiedQA	65.57	77.92	81.61	72.34	75.58	34.65	82.70	58.43	36.01	61.62	42.45	-
OOD MetaQA	62.26	69.41	66.59	75.02	67.51	**50.51**	72.20	58.59	52.13	59.28	22.14	-
OOD TWEAC	57.65	43.98	57.93	66.62	65.37	47.32	69.59	47.46	50.59	59.16	20.53	-6.31

TECHNISCHE UNIVERSITÄT DARMSTADT

MAML [4]

Ranking Function f_Θ
Task $T = \{q, D^+, D^-\}$
Binary Cross-Entropy Loss L

Update parameters on T_1 obtaining new parameters Θ'

$$\Theta' = \Theta - \alpha \nabla_\Theta \mathrm{L}(f_\Theta; T_1)$$

Update original parameters by evaluating $f_{\Theta'}$ on T_2

$$\Theta^* = \Theta - \alpha \nabla_\Theta \mathrm{L}(f_{\Theta'}; T_2)$$

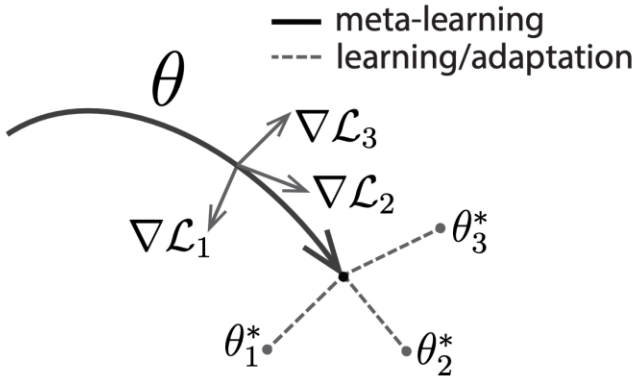


Figure 1. Diagram of our model-agnostic meta-learning algorithm (MAML), which optimizes for a representation θ that can quickly adapt to new tasks.

Intuitively, obtain parameters that are able to adapt to a new task quickly.

2. STAGE RETRIEVAL RESULTS

e	$k = 2$	$k = 4$	$k = 8$	Avg.
4	0.6187	0.6300	0.6566	0.6351
8	**0.6280**	**0.6414**	0.6721	**0.6472**
16	0.6195	0.6400	**0.6736**	0.6444
32	0.6039	0.6209	0.6477	0.6242
64	0.5597	0.5729	0.5843	0.5723
all	0.5723	0.5771	0.5828	0.5774

Table 2: Recall@1000 results on the test set with varying number of expansion terms e from each relevant document. Results are averaged over the shuffles.

RESULTS

	Robust	Covid	News	Touché	Avg.
BM25-QE					
$k = 2$	0.4480	0.5632	0.3846	0.2602	0.4140
$k = 4$	0.4843	0.6079	0.3877	0.2558	0.4339
$k = 8$	0.5568	0.6606	0.4049	**0.2982**	0.4801
Avg.	0.4964	0.6106	0.3924	0.2714	0.4427
kNN					
$k = 2$	0.4259	0.6736	0.3492	0.1646	0.4033
$k = 4$	0.4342	0.6789	0.3539	0.1697	0.4092
$k = 8$	0.4698	0.7069	0.3925	0.1904	0.4399
Avg.	0.4433	0.6865	0.3652	0.1749	0.4175
CE Zero-Shot					
$k = 2$	0.3937	0.6917	0.2955	0.1731	0.3885
$k = 4$	0.4185	0.7018	0.3189	0.1767	0.4040
$k = 8$	0.4335	0.7150	0.3285	0.1799	0.4142
Avg.	0.4152	0.7028	0.3143	0.1766	0.4022

	Robust	Covid	News	Touché	Avg.
CE Query-FT					
$k = 2$	0.4375	0.6833	0.2942	0.1887	0.4009
$k = 4$	0.4786	0.7182	0.3463	0.2080	0.4378
$k = 8$	0.5376	**0.7677**	0.3645	0.1975	0.4668
Avg.	0.4846	0.7231	0.3350	0.1981	0.4352
CE MAML + Query FT					
$k = 2$	0.4529	0.7129	0.2526	0.2212	0.4099
$k = 4$	0.5079	**0.7498**	0.3358	0.2292	0.4557
$k = 8$	0.5572	0.7449	0.3557	0.2201	0.4695
Avg.	0.5060	0.7359	0.3147	0.2235	0.4450
Rank Fusion: kNN & BM25-QE					
$k = 2$	0.4635	0.6903	0.3783	0.2263	0.4396
$k = 4$	0.5020	0.6858	**0.4228**	0.2438	0.4636
$k = 8$	0.5574	0.7470	**0.4359**	0.2744	0.5037
Avg.	0.5076	0.7077	**0.4123**	0.2482	0.4689
Rank Fusion: CE MAML + Query FT & BM25-QE					
$k = 2$	**0.5164**	**0.7269**	**0.3934**	**0.2670**	**0.4759**
$k = 4$	**0.5576**	0.7449	0.4084	**0.2701**	**0.4953**
$k = 8$	**0.6380**	0.7489	0.4148	0.2809	**0.5207**
Avg.	**0.5707**	**0.7402**	0.4055	**0.2727**	**0.4973**

ABLATIONS

	Robust	Covid	News	Touché	Avg.
BM25 without feedback documents					
	0.0459	0.1615	0.0551	0.1052	0.0919
kNN (Query Only)					
$k = 2$	0.3531	0.6611	0.2537	0.1637	0.3579
$k = 4$	0.3652	0.6486	0.2512	0.1649	0.3575
$k = 8$	0.3677	0.6854	0.2578	0.1687	0.3699
Avg.	0.3620	0.6650	0.2542	0.1658	0.3618
CE Query-FT (full)					
$k = 2$	0.4721	0.7168	0.3279	0.1797	0.4241
$k = 4$	0.5110	0.6872	0.3487	0.1858	0.4332
$k = 8$	0.5778	0.7644	0.3477	0.2021	0.4730
Avg.	0.5203	0.7228	0.3414	0.1892	0.4434
CE supervised + Query-FT (bias)					
$k = 2$	0.4540	0.7303	0.2716	0.2251	0.4203
$k = 4$	0.4896	0.7227	0.3657	0.2172	0.4488
$k = 8$	0.5353	0.7221	0.3390	0.2104	0.4517
Avg.	0.4930	0.7250	0.3254	0.2176	0.4403

TECHNISCHE UNIVERSITÄT DARMSTADT

LATENCY