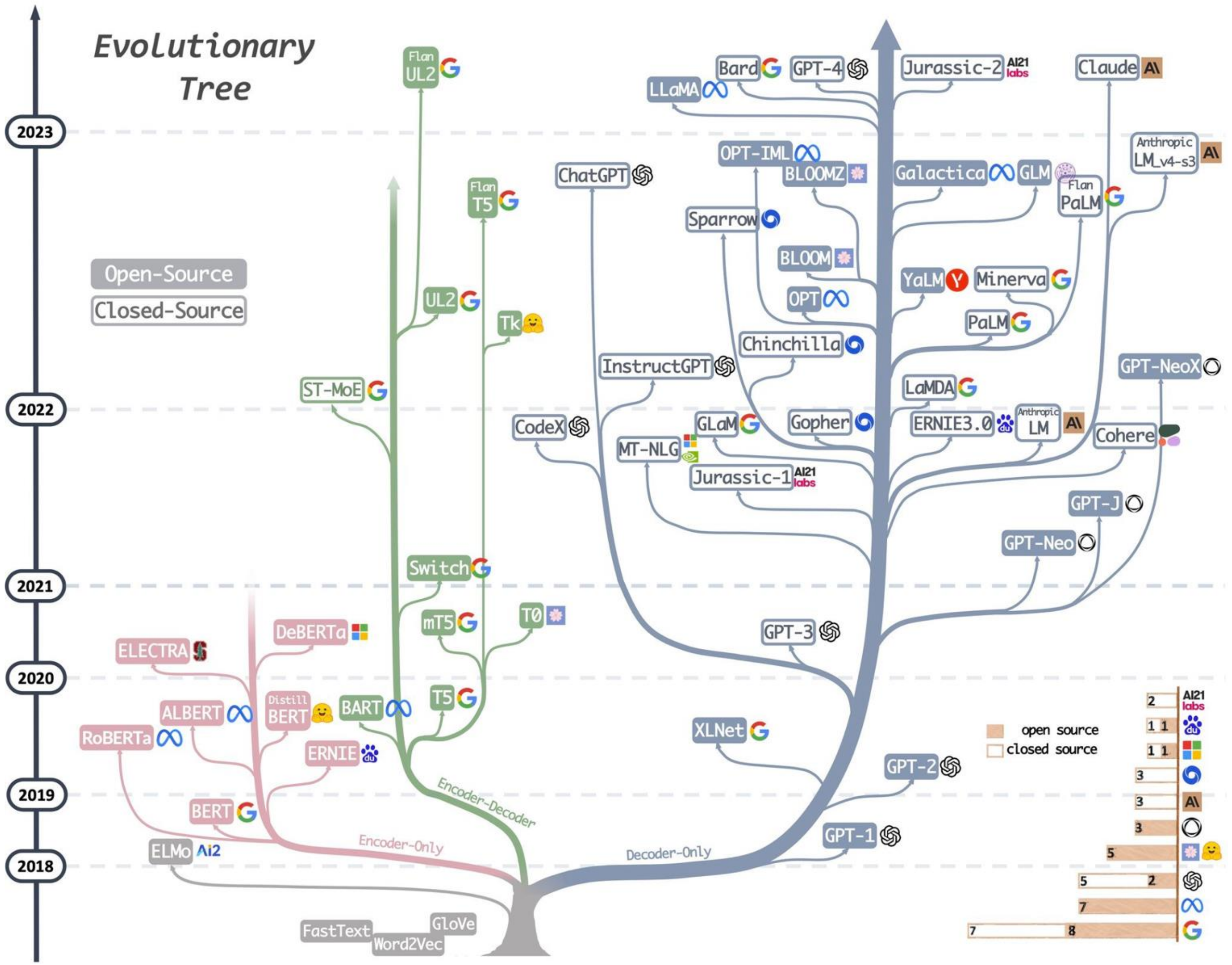
BROKEN TELEPHONE

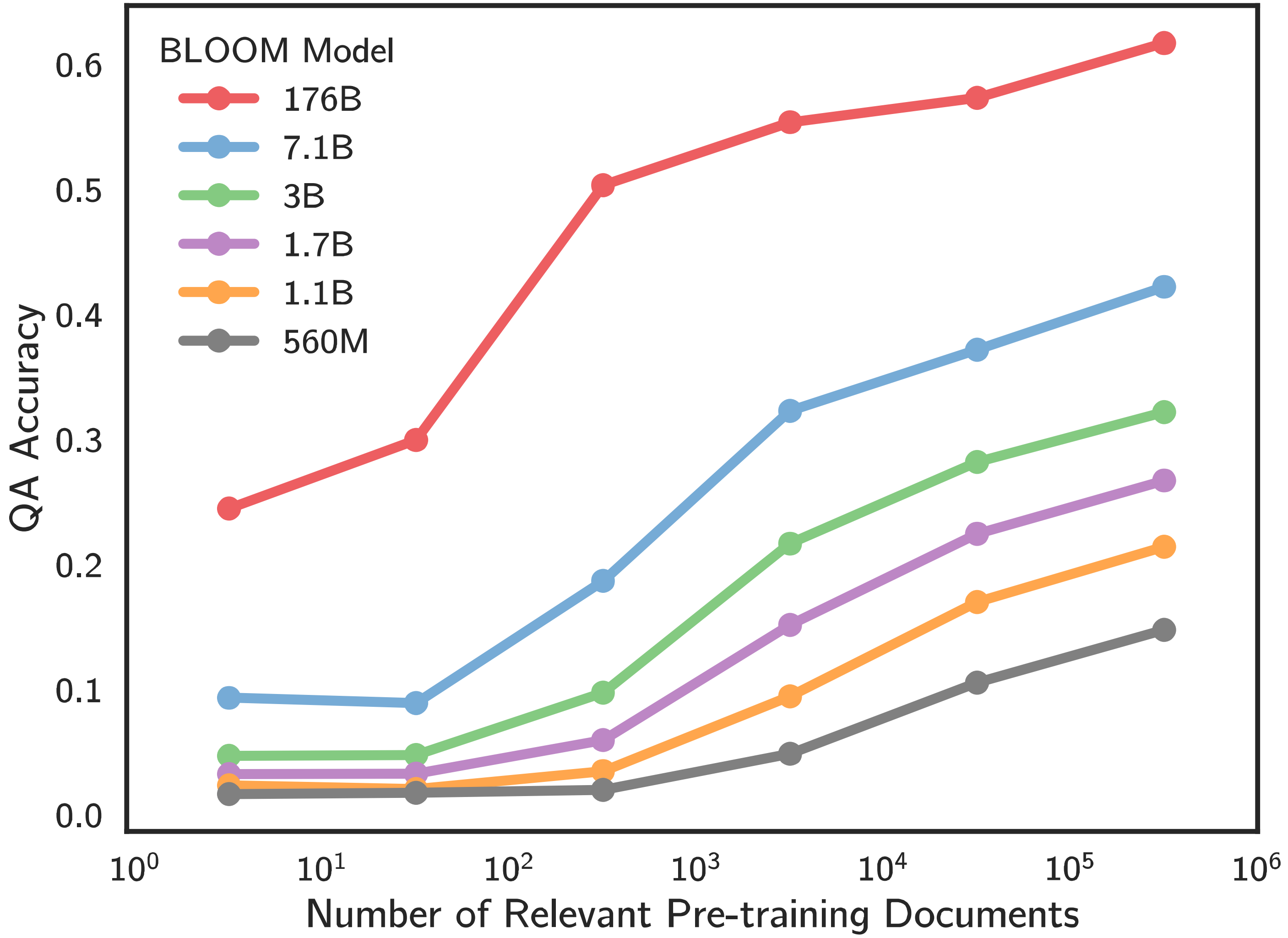
Evangelos Kanoulas

University of Amsterdam

INFORMATION ACCESS IN THE ERA OF LLMS



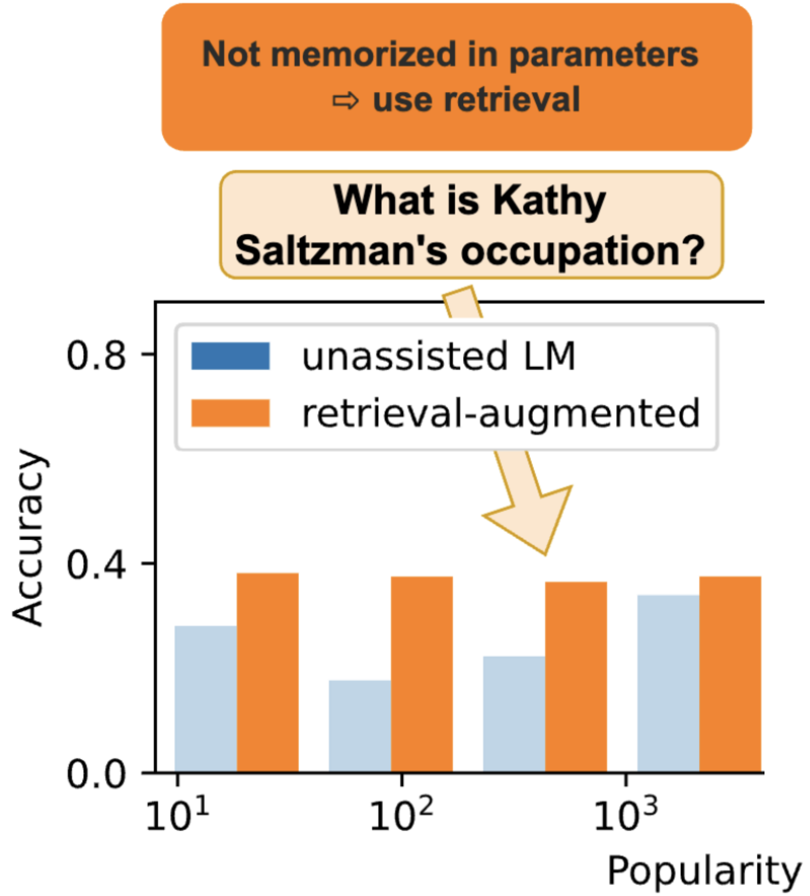
WHAT LLMS DO NOT KNOW



Kandpal, Nikhil, et al. "Large language models struggle to learn long-tail knowledge." International Conference on Machine Learning. PMLR, 2023.

RETRIEVAL AUGMENTATION

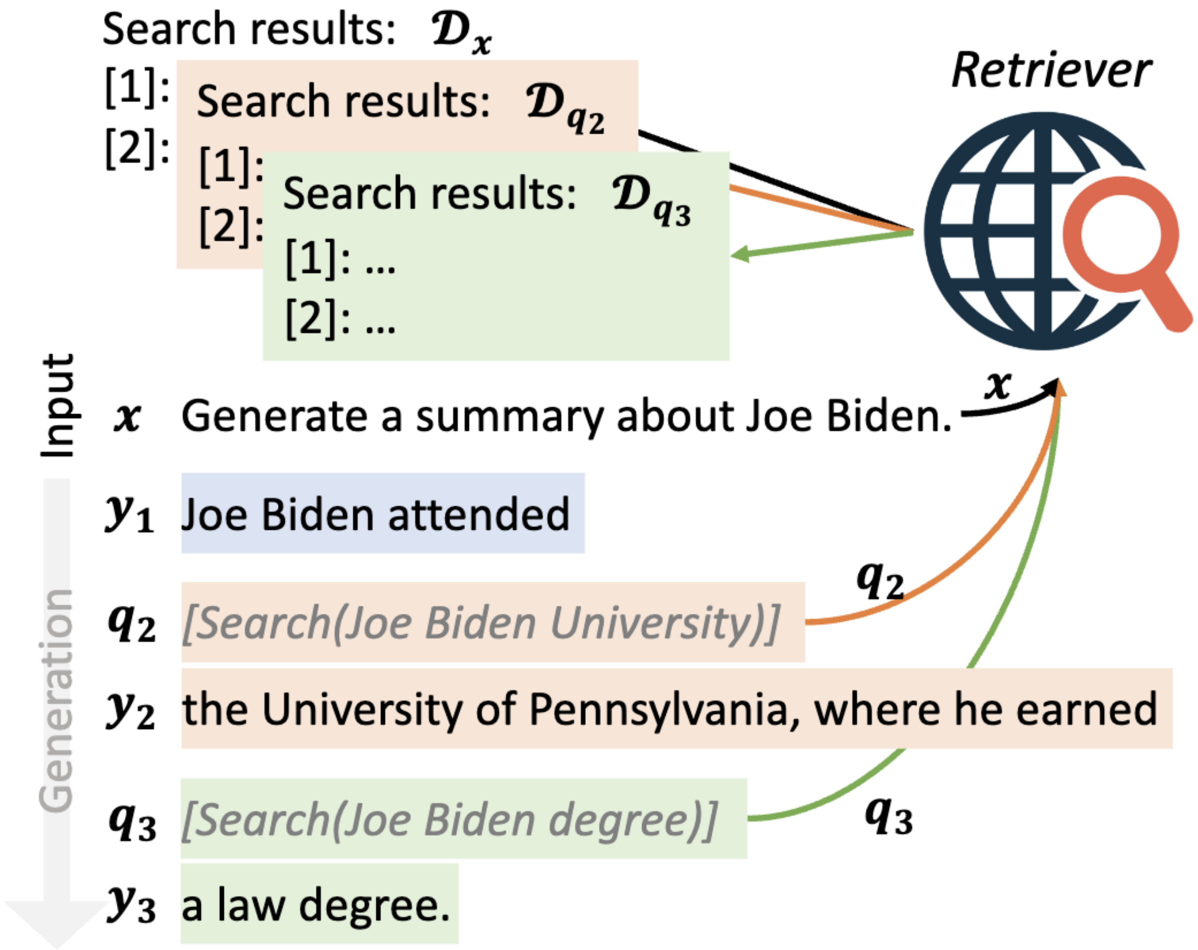
Retrieval augmentation helps when the LLM has not seen enough …



Mallen, Alex, et al. "When not to trust language models: Investigating effectiveness of parametric and non-parametric memories." In ACL 2023.

RETRIEVAL AUGMENTATION

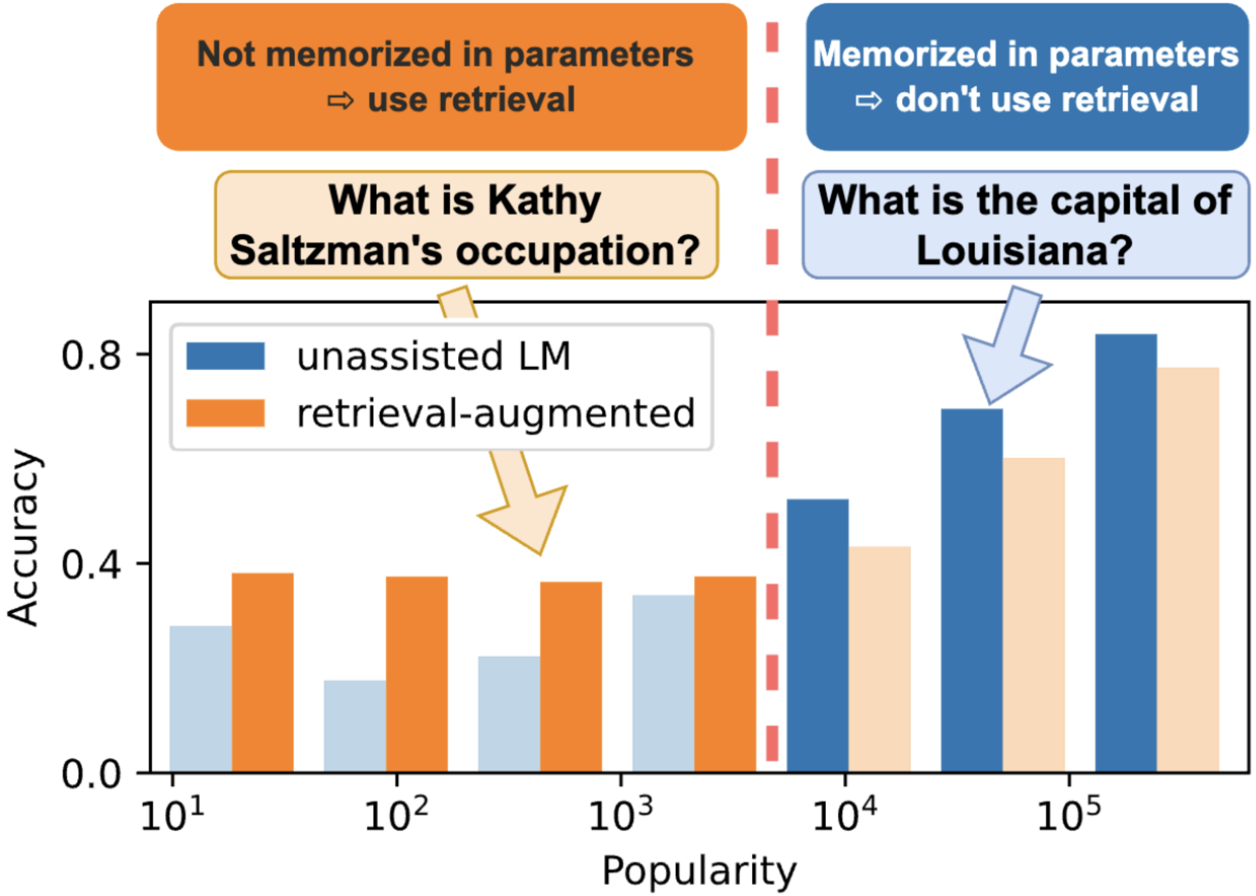
Retrieve when the LLM is not confident enough



Jiang, Zhengbao, et al. "Active retrieval augmented generation." arXiv preprint arXiv:2305.06983 (2023).

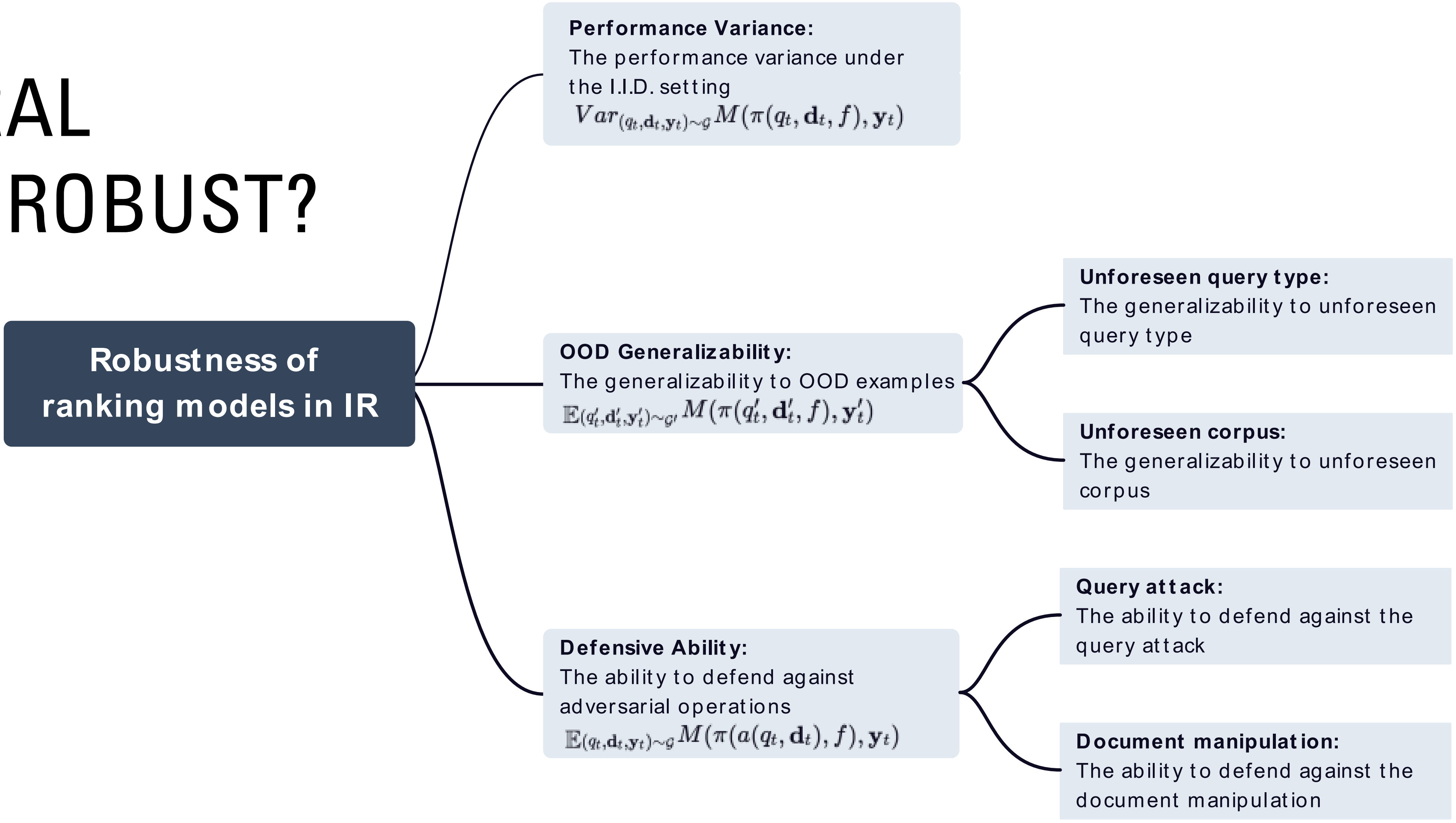
RETRIEVAL AUGMENTATION

Retrieval augmentation helps when the LLM has not seen enough … but if the LLM has seen enough it may even harm performance.



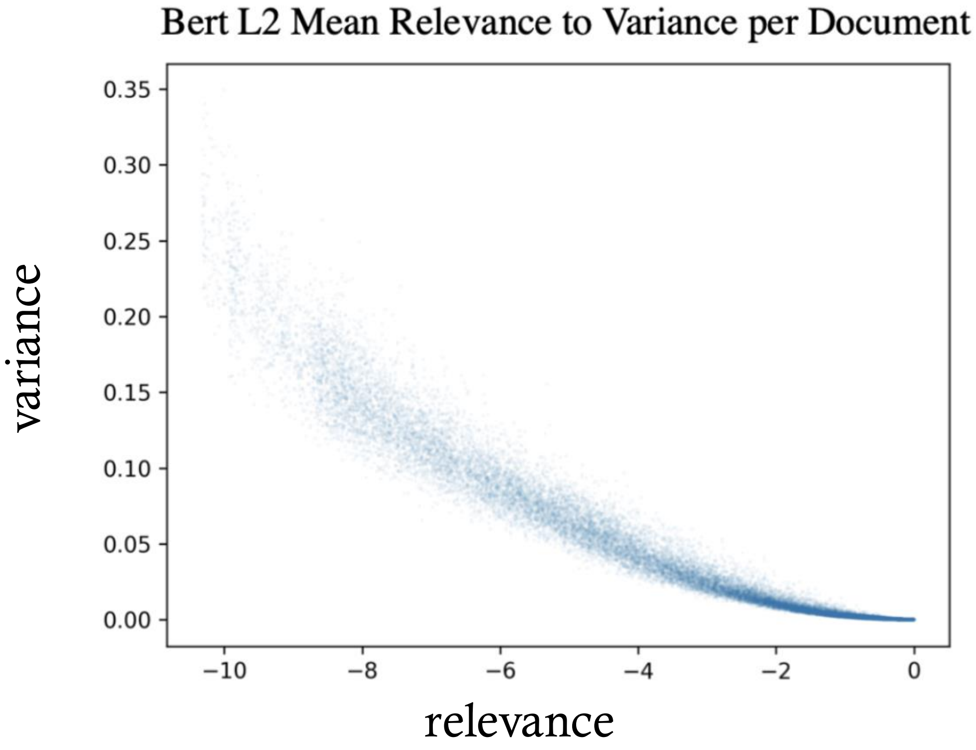
Mallen, Alex, et al. "When not to trust language models: Investigating effectiveness of parametric and non-parametric memories." In ACL2023.

ARE NEURAL RANKERS ROBUST?

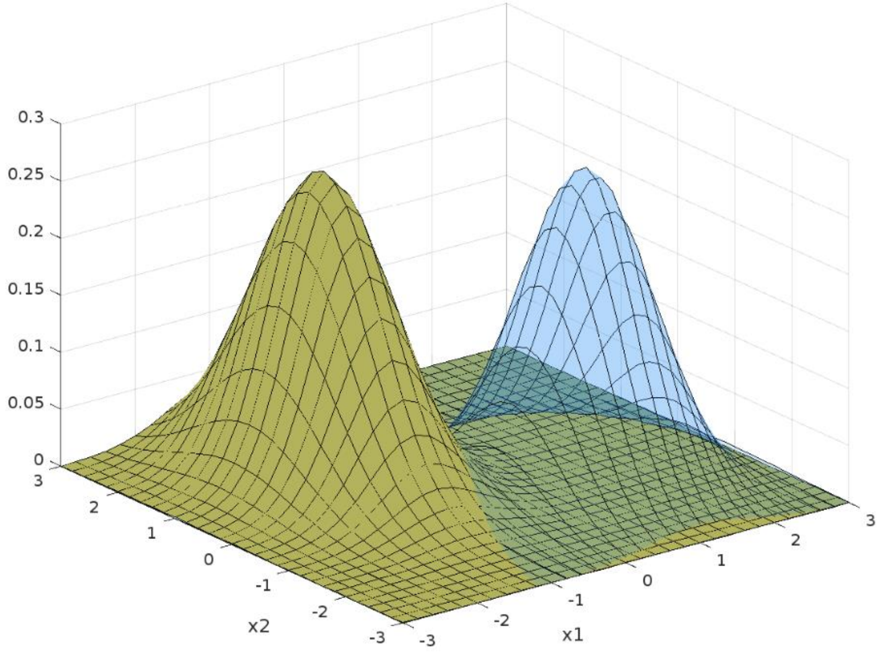


Wu, Chen, et al. "Are neural ranking models robust?." ACM Transactions on Information Systems 41.2 (2022): 1-36.

MODEL UNCERTAINTY

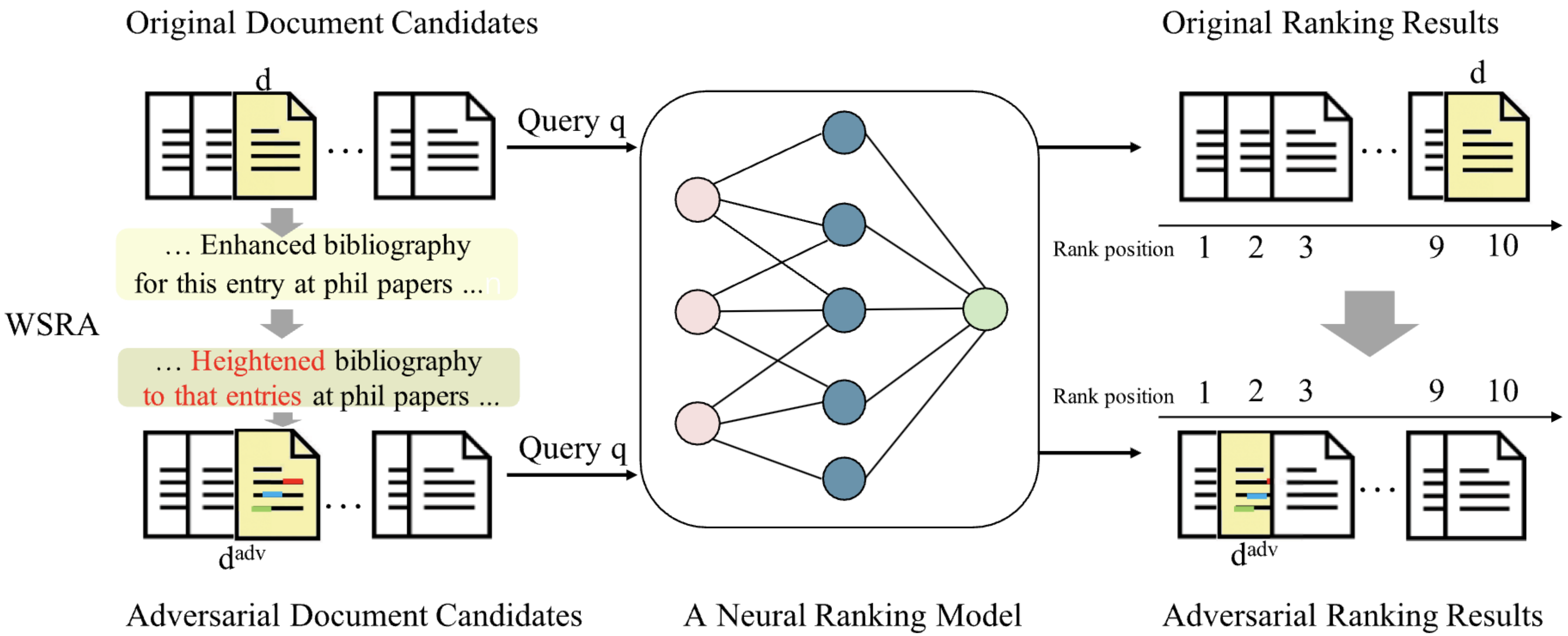


Cohen, Daniel, et al. "Not all relevance scores are equal: Efficient uncertainty and calibration modeling for deep retrieval models." In SIGIR 2021



Hamed Zamani and Michael Bendersky. 2023. Multivariate Representation Learning for Information Retrieval. In SIGIR ’23.

ADVERSARIAL ATTACKS

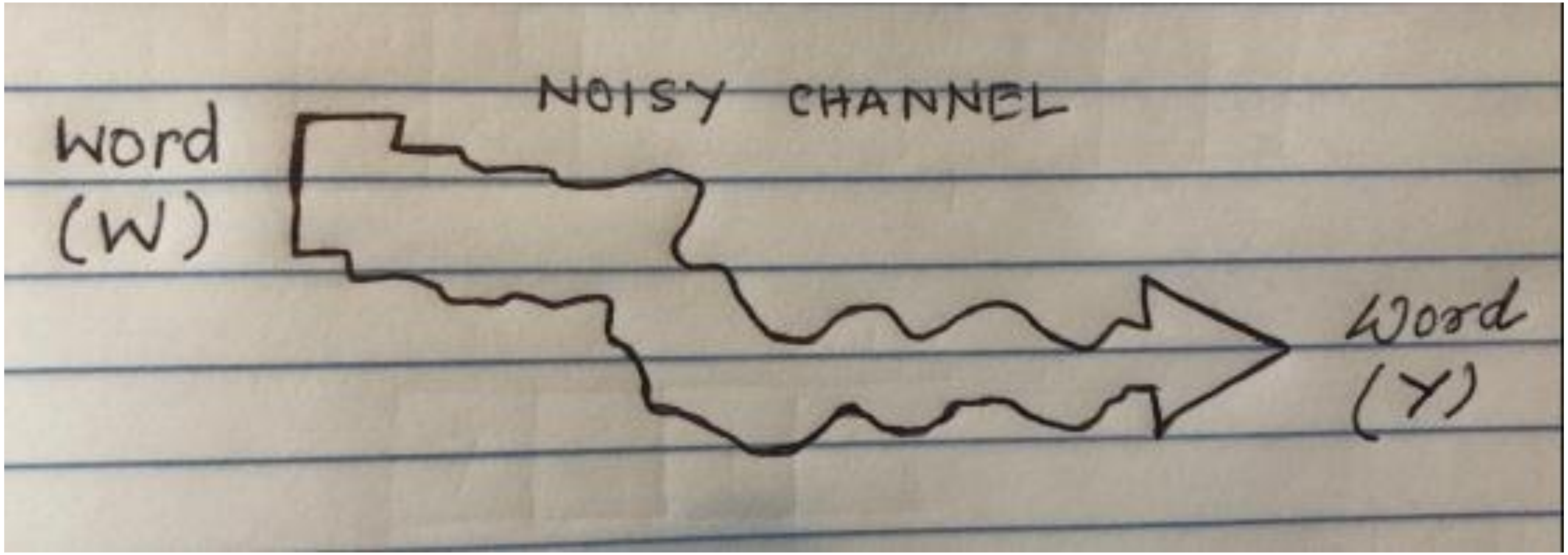


Wu, Chen, et al. "Prada: practical black-box adversarial attacks against neural ranking models." ACM Transactions on Information Systems 41.4 (2023): 1-27.

UNCERTAINTY

- Estimate uncertainty/confidence
- Attribute uncertainty
- Develop remedies
  - Augment data
  - Transfer knowledge across domains
  - Learn to defer
- Express/pass uncertainty

DALL*E



(Attack) Typos and ASR mistakes → Contrastive learning and data augmentation.

(Attack) False memories → Uncertainty estimation and attribution.

(OOD) Conversational questions → Domain transfer and data generation.

QUESTIONS WITH TYPOS

Random

- Inserts, deletes, swaps, or substitutes a random character

Keyboard

- Swaps a random character with those close to each other on the QWERTY keyboard

Common misspellings

- Replaces words with misspelled ones, defined in a dictionary of common user-generated misspellings

Sidiropoulos, Georgios, and Evangelos Kanoulas. "Analysing the robustness of dual encoders for dense retrieval against misspellings." In SIGIR 2022.

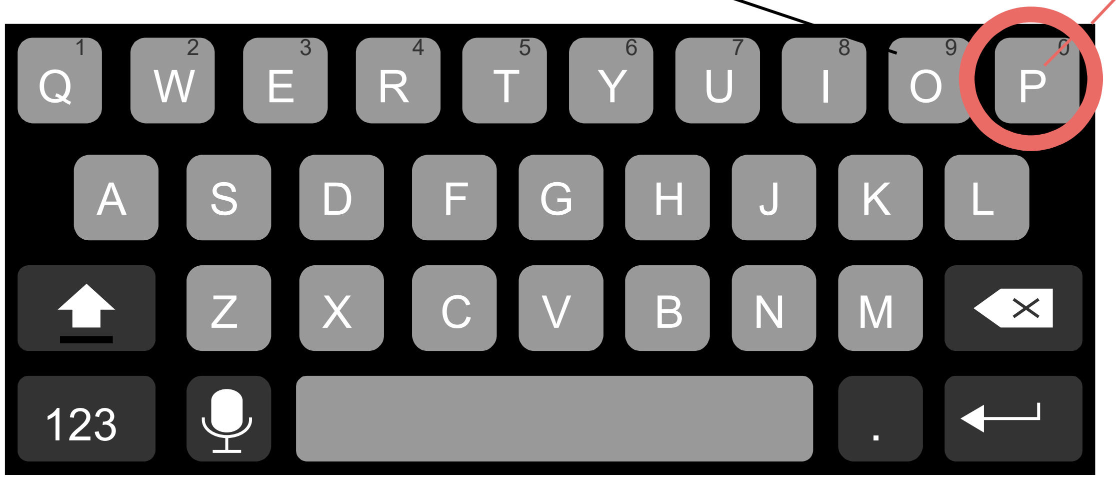
METHODS

- **Dense Retriever (DR)** is a dual-encoder BERT-based model used for scoring question-passage pairs:

$$L_1(q_i, p_i^+, p_{i,1}^-, \cdots, p_{i,n}^-) = -\log \frac{e^{\mathrm{sim}(q_i, p_i^+)}}{e^{\mathrm{sim}(q_i, p_i^+)} + \sum_{j=1}^{n} e^{\mathrm{sim}(q_i, p_{i,j}^-)}}. \quad (1)$$

Sidiropoulos, Georgios, and Evangelos Kanoulas. "Analysing the robustness of dual encoders for dense retrieval against misspellings." In SIGIR 2022.

Original: Where was Obama born? ***Typo: Where was Obama bprn?***



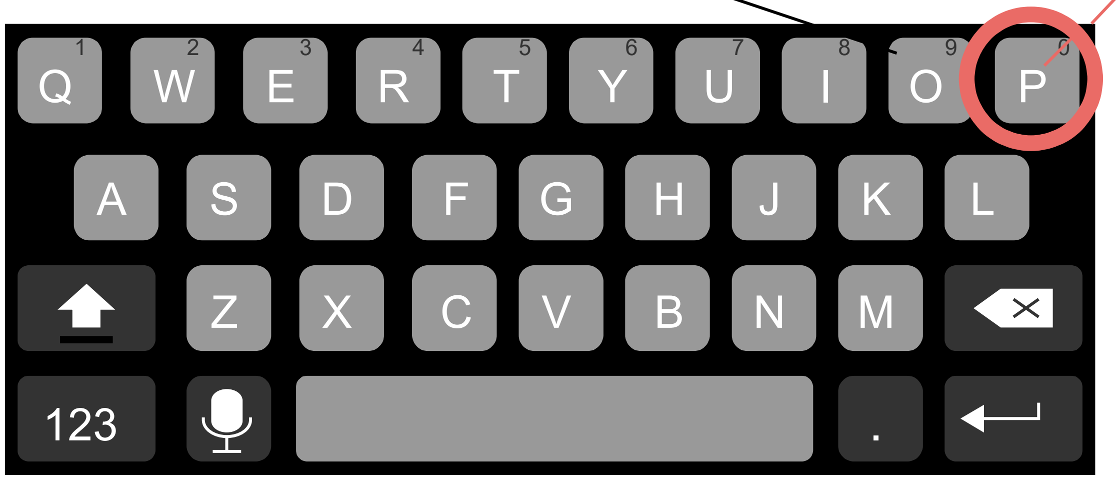
	Natural Questions (Test)	
	Original	Typos in Random Words
	AR@5	AR@5
Dense Retrieval	67.31	49.52

26% drop

Sidiropoulos, Georgios, and Evangelos Kanoulas. "Analysing the robustness of dual encoders for dense retrieval against misspellings." In SIGIR 2022.

Original: Where was Obama born?

Typo: Where was Obama bprn?



	MSMarco (Dev)	
	Original	Typos in Random Words
	MRR@10	MRR@10
Dense Retrieval	$\boxed{28.11}$	$\boxed{15.11}$

46% drop

Sidiropoulos, Georgios, and Evangelos Kanoulas. "Analysing the robustness of dual encoders for dense retrieval against misspellings." In SIGIR 2022.

METHODS

- **Dense Retriever (DR)** is a dual-encoder BERT-based model used for scoring question-passage pairs:

$$L_1(q_i, p_i^+, p_{i,1}^-, \cdots, p_{i,n}^-) = -\log \frac{e^{\mathrm{sim}(q_i, p_i^+)}}{e^{\mathrm{sim}(q_i, p_i^+)} + \sum_{j=1}^{n} e^{\mathrm{sim}(q_i, p_{i,j}^-)}}. \quad (1)$$

- **Data Augmentation (DR + Data augm.)** is one of the traditional approaches for robustifying neural models. By exposing DR on questions with and without typos, the model learns to be invariant to typos.

Sidiropoulos, Georgios, and Evangelos Kanoulas. "Analysing the robustness of dual encoders for dense retrieval against misspellings." In SIGIR 2022.

	Natural Questions (Test)		MSMarco (Dev)	
	Original	Typos in Random Words	Original	Typos in Random Words
	AR@5	AR@5	MRR@10	MRR@10
DR	$\boxed{67.31}$	$\boxed{49.52}$	$\boxed{28.11}$	$\boxed{15.11}$
DR + Data augm.	66.45	$\boxed{60.69}$	28.26	$\boxed{22.00}$

26% → 10% drop 46% → 21% drop

Sidiropoulos, Georgios, and Evangelos Kanoulas. "Analysing the robustness of dual encoders for dense retrieval against misspellings." In SIGIR 2022.

METHODS

- **Contrastive learning (DR + CL)** of representations works by maximizing the agreement between differently augmented views of the same object. We propose a contrastive loss that compares the similarity between a question and its typoed variations and other distinct questions:

$$L_2(q_i, q_i^+, q_{i,1}^{\neq}, \cdots, q_{i,n}^{\neq}) = \neq \log \frac{e^{\mathrm{sim}(q_i, q_i^+)}}{e^{\mathrm{sim}(q_i, q_i^+)} + \mathrm{P}_{j=1}^{n} e^{\mathrm{sim}(q_i, q_{i,j}^{\neq})}}. \quad (2)$$

The final loss is a weighted average of the two losses:

$$L = w_1 \cdot L_1 + w_2 \cdot L_2. \quad (3)$$

Sidiropoulos, Georgios, and Evangelos Kanoulas. "Analysing the robustness of dual encoders for dense retrieval against misspellings." In SIGIR 2022.

	Natural Questions (Test)		MSMarco (Dev)	
	Original AR@5	Typos in Random Words AR@5	Original MRR@10	Typos in Random Words MRR@10
DR	67.31	49.52	28.11	15.11
DR + Data augm.	66.45	60.69	28.26	22.00
DR + Data augm. + CL (ours)	67.47	**62.13**	29.14	**22.84**

26% → 7% drop

46% → 18% drop

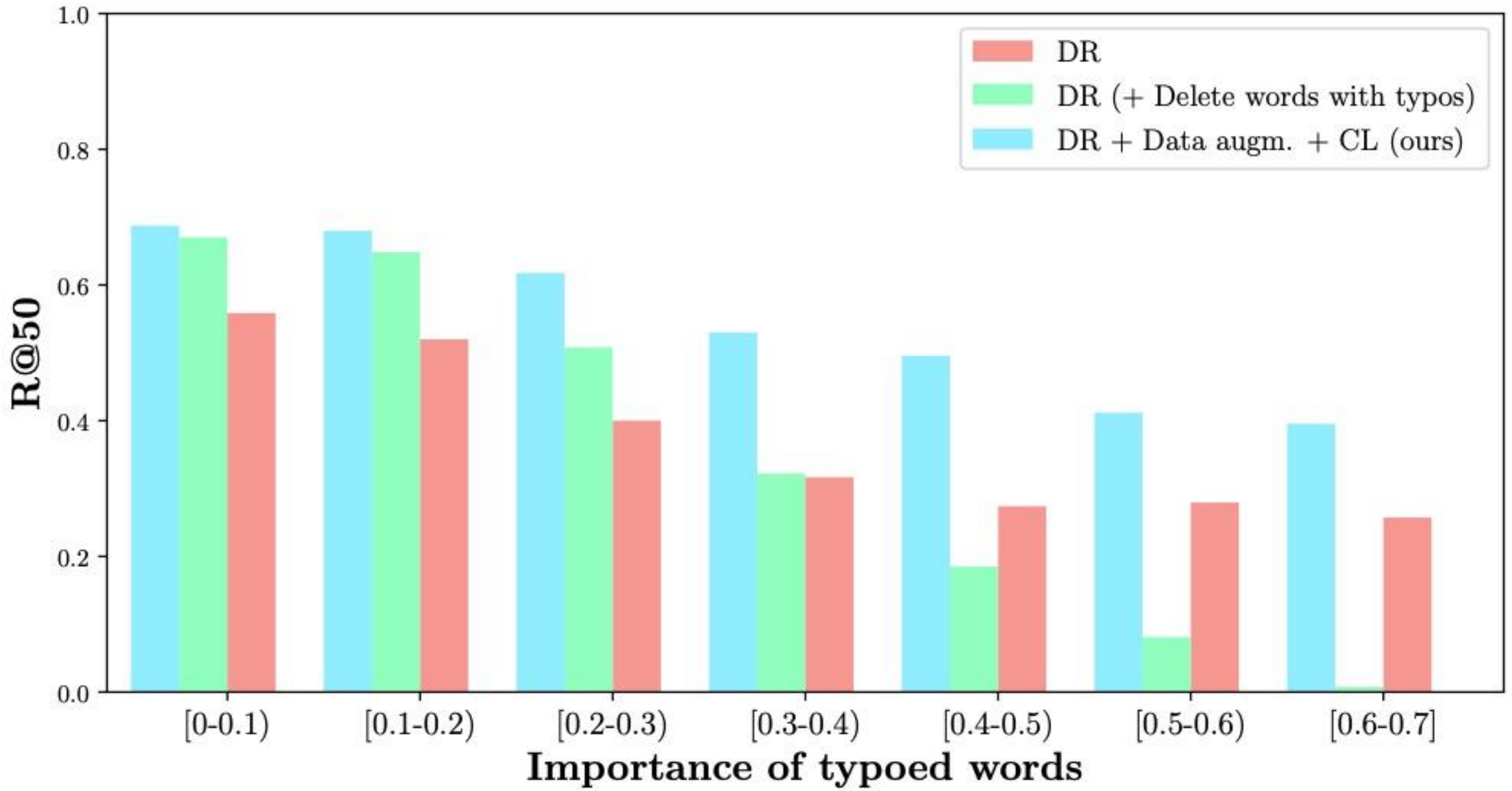
Sidiropoulos, Georgios, and Evangelos Kanoulas. "Analysing the robustness of dual encoders for dense retrieval against misspellings." In SIGIR 2022.

DO ALL TYPOS MATTER EQUALLY?

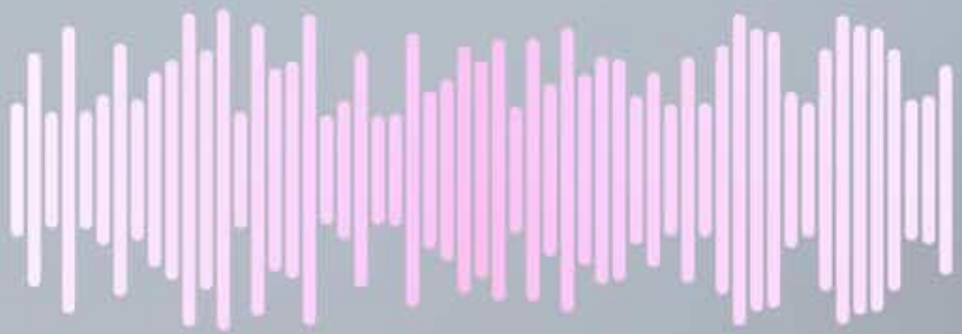
	Natural Questions (Test)		
	Original AR@5	Typos in Random Words AR@5	Typos in Discriminative Utterances AR@5
DR	67.31	49.52	38.89
DR + Data augm.	66.45	60.69	51.68
DR + Data augm. + CL (ours)	67.47	**62.13**	**53.15**

Sidiropoulos, Georgios, and Evangelos Kanoulas. "Analysing the robustness of dual encoders for dense retrieval against misspellings." In SIGIR 2022.

WHAT DOES THE MODEL LEARN?



Sidiropoulos, Georgios, and Evangelos Kanoulas. "Analysing the robustness of dual encoders for dense retrieval against misspellings." In SIGIR 2022.



SPEECH-BASED QUESTIONS

Prof. dr. E. Kanoulas – University of Amsterdam

SYNTHETIC QUESTIONS

Sidiropoulos, Georgios, Svitlana Vakulenko, and Evangelos Kanoulas. "On the impact of speech recognition errors in passage retrieval for spoken question answering." In CIKM 2022

SIGNIFICANT DROP IN PERFORMANCE

		Natural Questions (Test)	
		Original	ASR
		AR@5	AR@5
DR		66.26	41.91

37% drop

Sidiropoulos, Georgios, Svitlana Vakulenko, and Evangelos Kanoulas. "On the impact of speech recognition errors in passage retrieval for spoken question answering." In CIKM 2022

DATA AUGMENTATION FOR THE WIN

		Natural Questions (Test)	
	Noise	Original AR@5	ASR AR@5
DR	--	$\boxed{66.26}$	$\boxed{41.91}$
DR + Data augm.	Typos	67.47	$\boxed{46.75}$

29% drop

Sidiropoulos, Georgios, Svitlana Vakulenko, and Evangelos Kanoulas. "On the impact of speech recognition errors in passage retrieval for spoken question answering." In CIKM 2022

DATA AUGMENTATION FOR THE WIN

		Natural Questions (Test)	
	Noise	Original AR@5	ASR AR@5
DR	--	66.26	41.91
DR + Data augm.	Typos	67.47	46.75
DR + Data augm.	ASR	66.67	54.84

17% drop

Sidiropoulos, Georgios, Svitlana Vakulenko, and Evangelos Kanoulas. "On the impact of speech recognition errors in passage retrieval for spoken question answering." In CIKM 2022

… BUT HOW MUCH AND WHAT KIND

# Additional training questions with ASR noise	AR@5
40K	54.84
4K	48.69
400	44.29
0	41.91

Data hungry

Sidiropoulos, Georgios, Svitlana Vakulenko, and Evangelos Kanoulas. "On the impact of speech recognition errors in passage retrieval for spoken question answering." In CIKM 2022

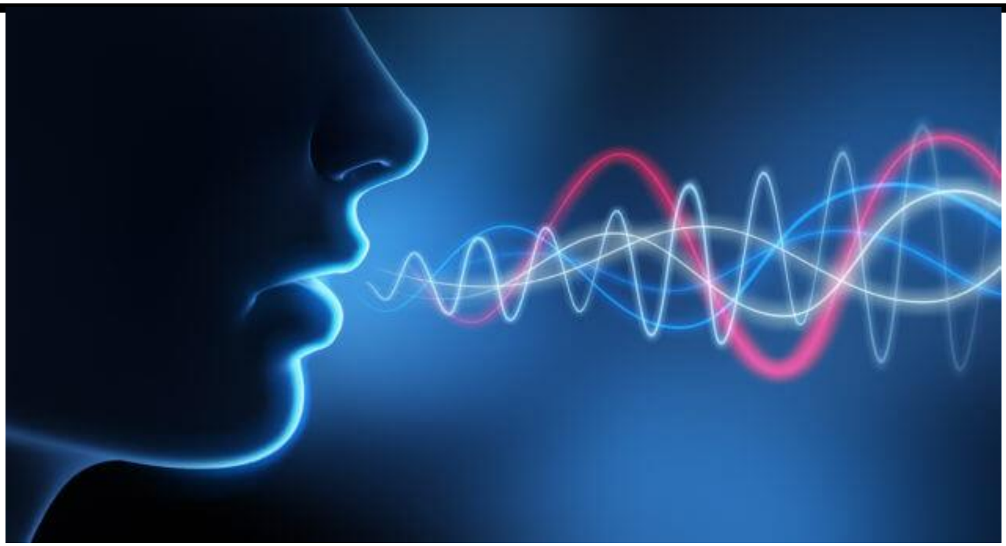
... BUT HOW MUCH AND WHAT KIND

Data augm.	test	AR@5
en-US	en-US	54.84
en-US	en-AU	53.54
en-US	en-IN	46.37

Does not generalize well

Sidiropoulos, Georgios, Svitlana Vakulenko, and Evangelos Kanoulas. "On the impact of speech recognition errors in passage retrieval for spoken question answering." In CIKM 2022

NATURAL QUESTIONS

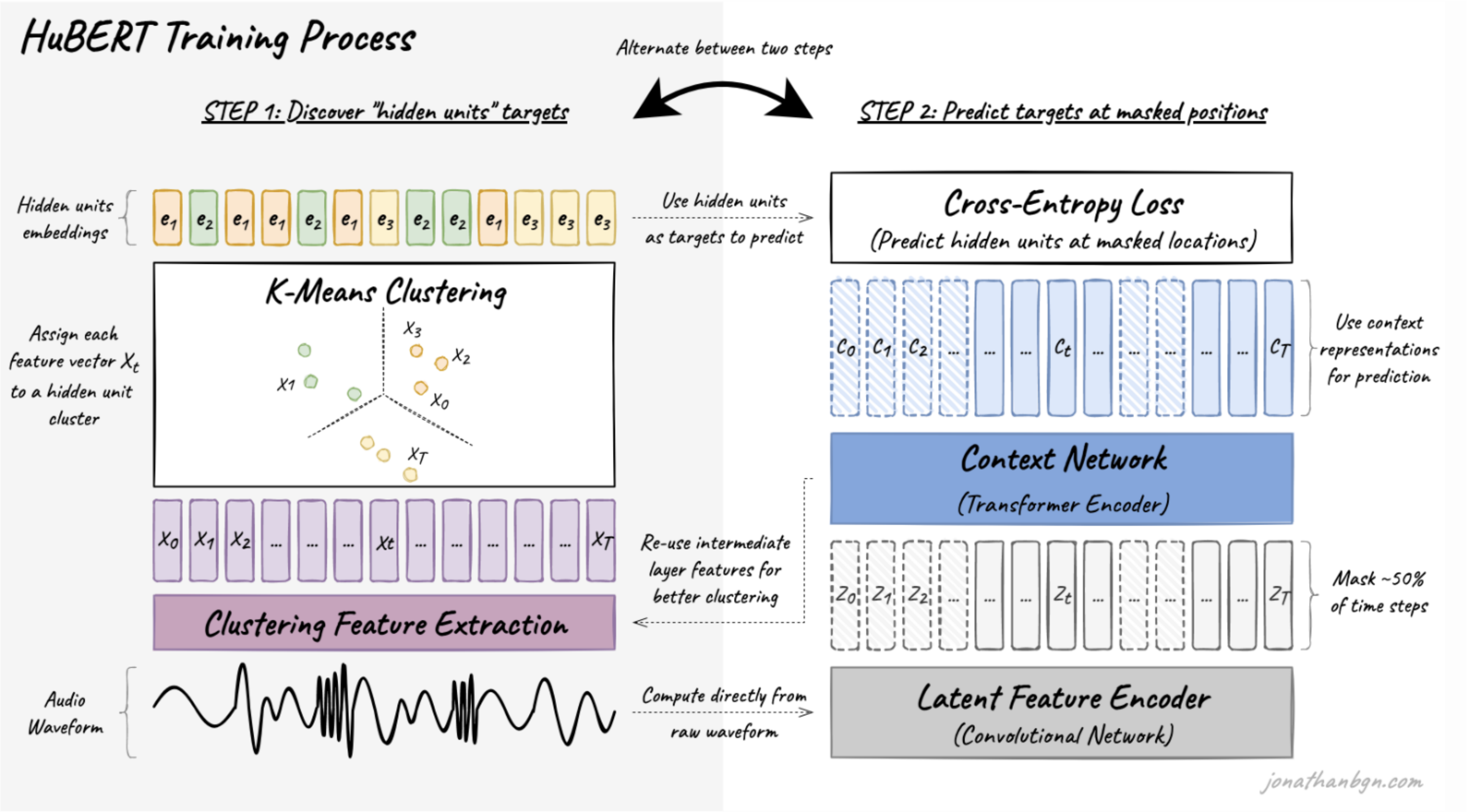


		Natural Questions (Test)		
	Noise	Original AR@5	ASR Synthetic AR@5	ASR NAtural AR@5
DR	--	66.26	41.91	16.64
DR + Data augm.	Typos	67.47	46.75	24.82
DR + Data augm.	ASR Synthetic	66.67	54.84	29.98

54% drop

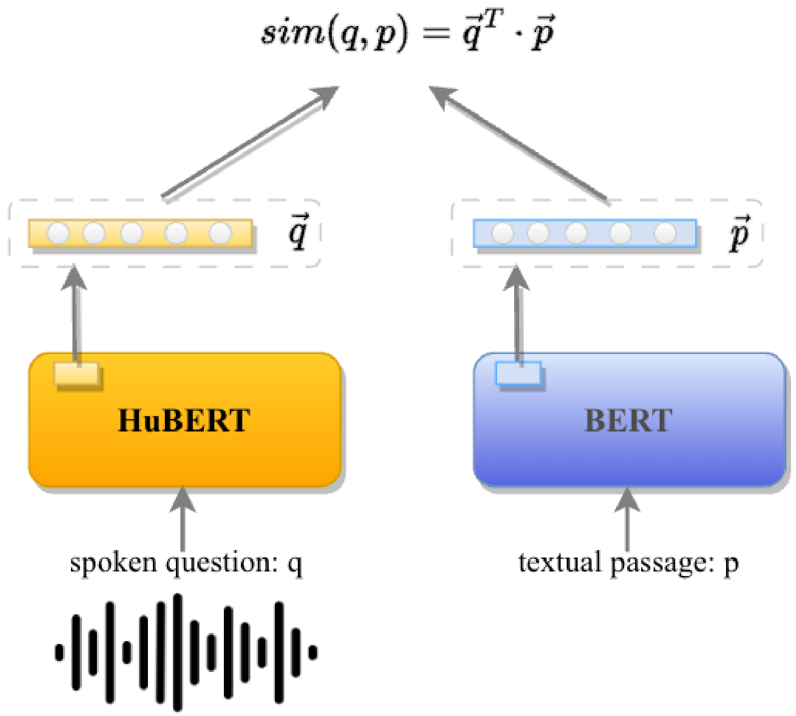
Sidiropoulos, Georgios, Svitlana Vakulenko, and Evangelos Kanoulas. "On th
speech recognition errors in passage retrieval for spoken question answering." in CIKM 2022

END-TO-END SPEECH-TO-TEXT RETRIEVAL



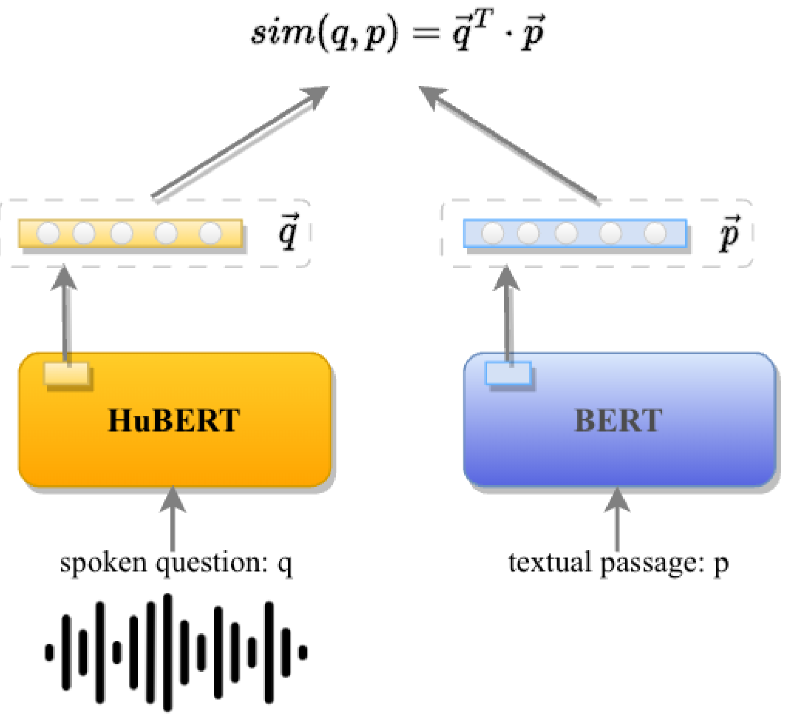
Hsu, Wei-Ning, et al. "Hubert: Self-supervised speech representation learning by masked prediction of hidden units." IEEE/ACM Transactions on Audio, Speech, and Language Processing 29 (2021

END-TO-END SPEECH-TO-TEXT RETRIEVAL



Sidiropoulos, Georgios and Evangelos Kanoulas, Multimodal Dense Passage Retrieval For Speech-based Open-Domain Question Answering. Preprint. Under Review.

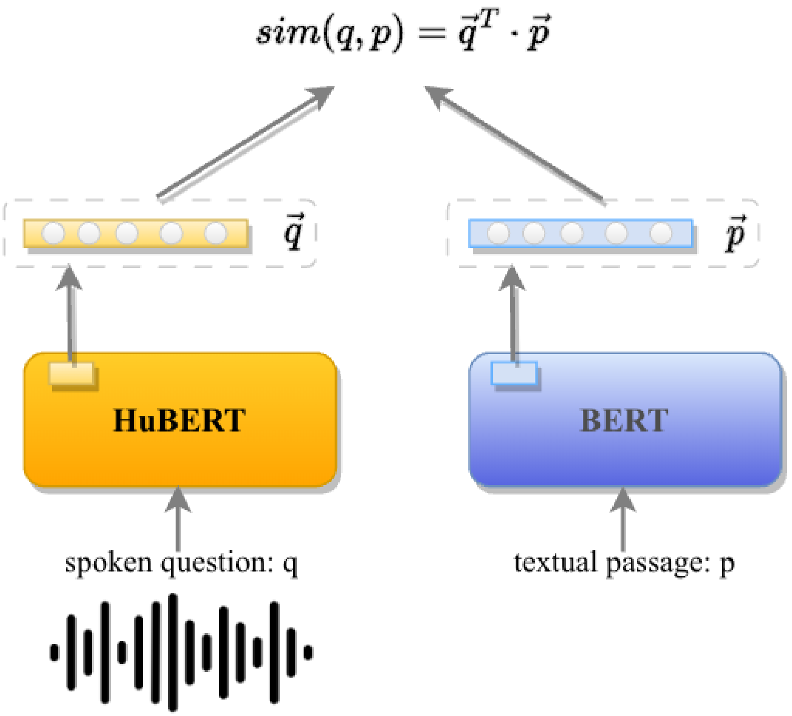
END-TO-END SPEECH-TO-TEXT RETRIEVAL



	Natural Question
	AR@5
DR + ASR	68,36
MDR	57,25

Sidiropoulos, Georgios and Evangelos Kanoulas, Multimodal Dense Passage Retrieval For Speech-based Open-Domain Question Answering. Preprint. Under Review.

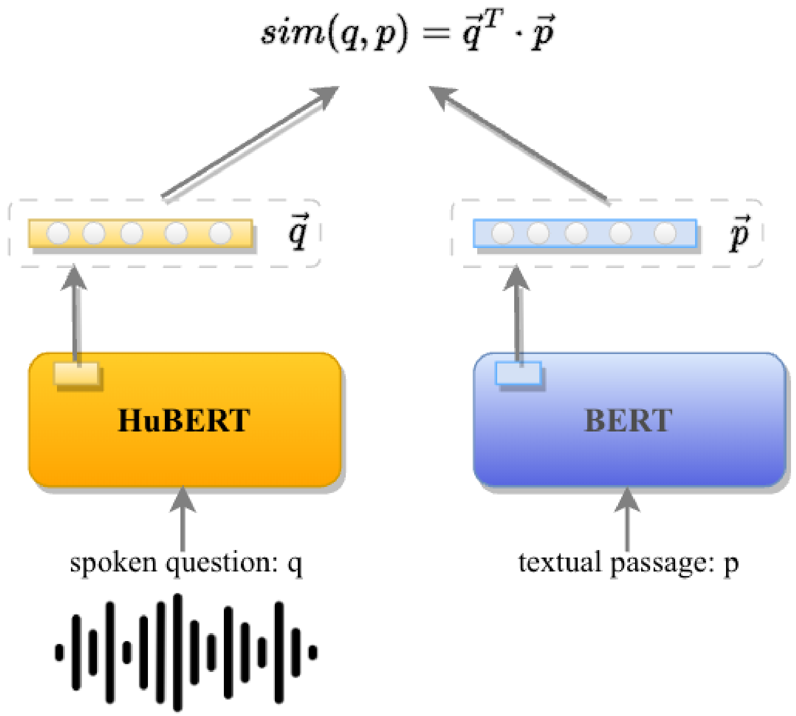
END-TO-END SPEECH-TO-TEXT RETRIEVAL



	MS Marco
	MMR
DR + ASR	17.74
MDR	15.77

Sidiropoulos, Georgios and Evangelos Kanoulas, Multimodal Dense Passage Retrieval For Speech-based Open-Domain Question Answering. Preprint. Under Review.

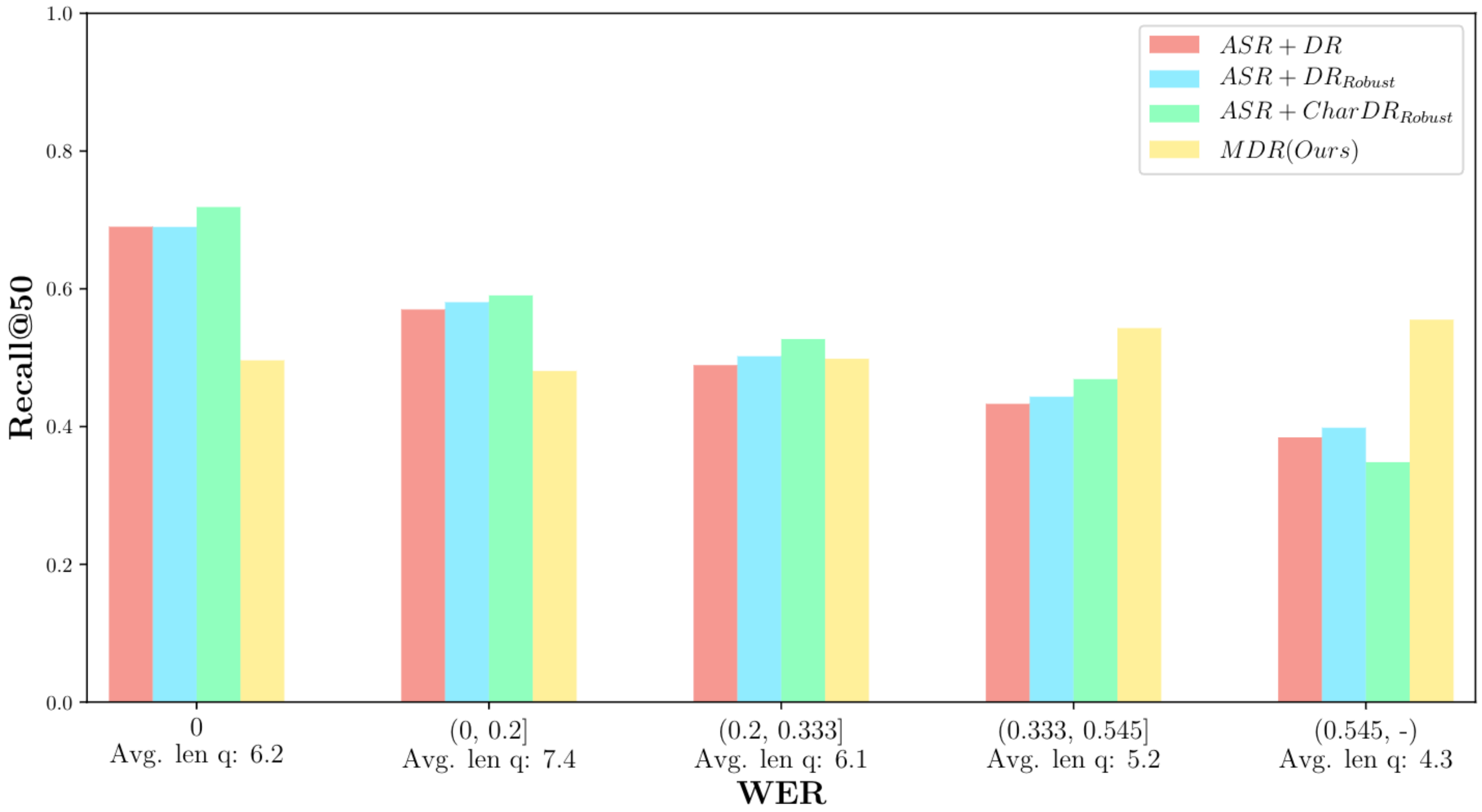
END-TO-END SPEECH-TO-TEXT RETRIEVAL



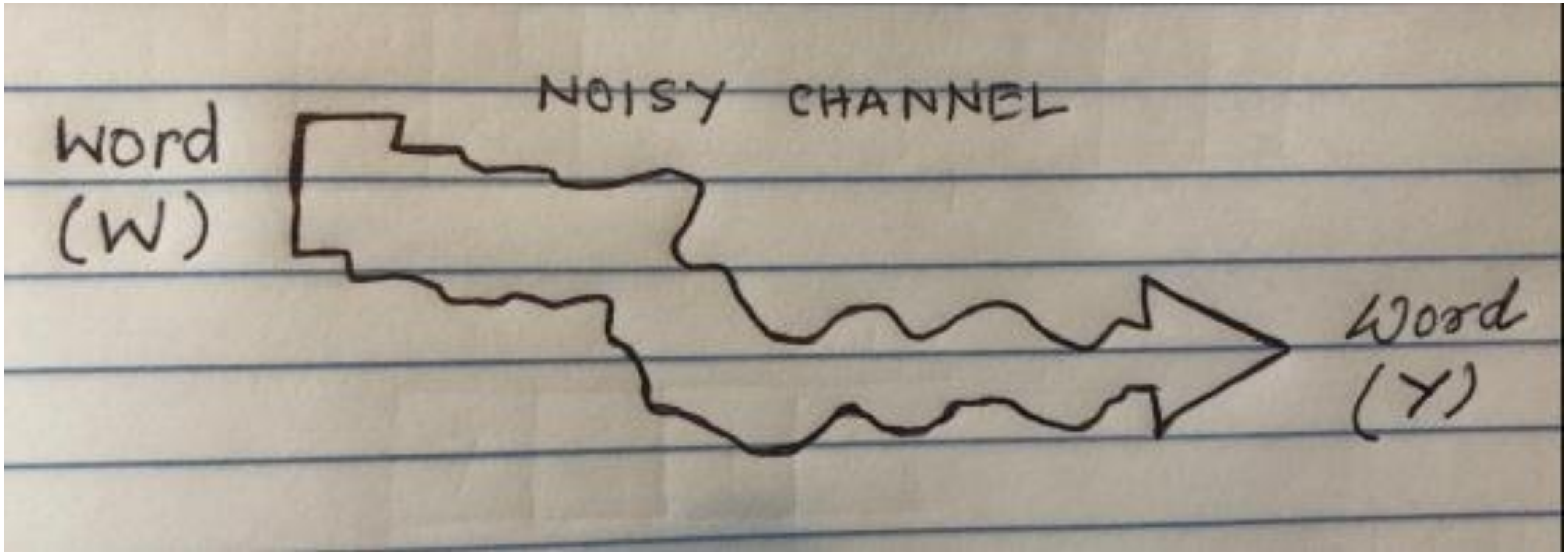
	MS Marco Seen	MS Marco Unseen
	MMR	MMR
DR + ASR	16.69	13.45
MDR	15.47	18.52

Sidiropoulos, Georgios and Evangelos Kanoulas, Multimodal Dense Passage Retrieval For Speech-based Open-Domain Question Answering. Preprint. Under Review.

END-TO-END SPEECH-TO-TEXT RETRIEVAL



Sidiropoulos, Georgios and Evangelos Kanoulas, Multimodal Dense Passage Retrieval For Speech-based Open-Domain Question Answering. Preprint. Under Review.

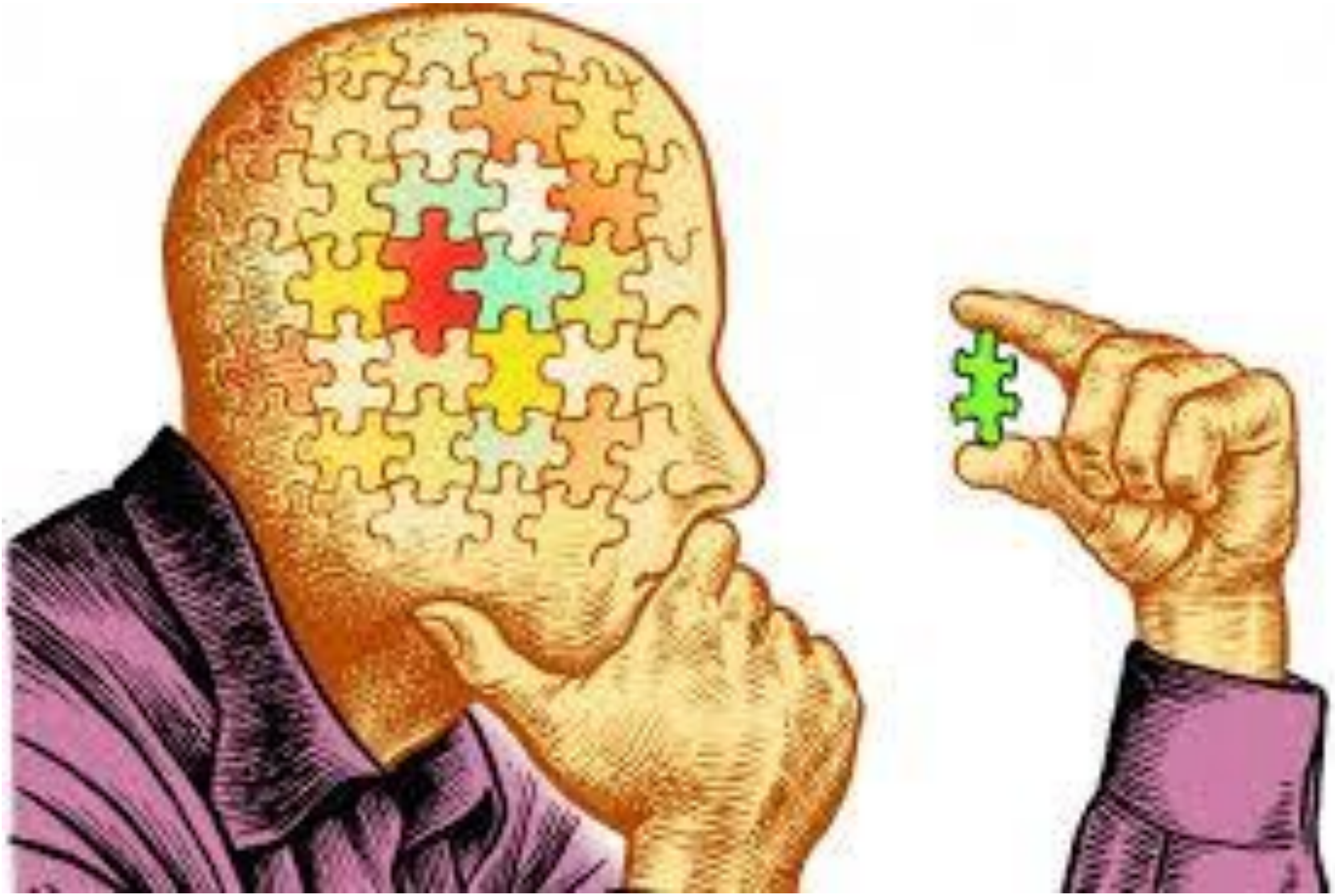


Typos and ASR mistakes → Contrastive learning and data augmentation.

False memories → Uncertainty estimation and attribution.

Conversational questions → Domain transfer.

RIGHT ON THE TIP OF MY TONGUE



IT'S ON THE TIP OF MY TONGUE

Ok so I don't really remember anything but one scene, so I will try to give as much detail as I can. I saw this movie when I was very young and only remember this scene because (from what I remember) I don't think I actually watched most of the movie because I found it pretty scary, so it is probably not a children/family movie. I think I watched it in 2006 (mid-end of the year probably) and it was on a tv in someone's house, so it was old enough to be released on tv/dvd (not still in cinemas). It was in English and colour I'm pretty sure. So this scene: There was a (or multiple) giant robot-like things and I think they were sort of sphere shaped. It was destroying a city and going around picking people up (possible killing them?) in giant net-like things i think. There was a father and daughter (i think it was a daughter) and they ended up getting separated because the father got picked up by the robot thing. When the robot thing was picking people up there was a lot of red liquid stuff (quite possibly blood, but maybe something else?). In the end I think the father made it back to the daughter. And that is all I remember, sorry if it is vague, but really hope someone can help.

COMPLEXITY

- Social nieces
- Contextual episodic memories
- Semantic memories about metadata, scenes, plots, etc.
- Exclusion criteria
- Multi-hop reasoning (comparisons)

Ok so I don't really remember anything but one scene, so I will try to give as much detail as I can. I saw this movie when I was very young and only remember this scene because (from what I remember) I don't think I actually watched most of the movie because I found it pretty scary, so it is probably not a children/family movie. I think I watched it in 2006 (mid-end of the year probably) and it was on a tv in someone's house, so it was old enough to be released on tv/dvd (not still in cinemas). It was in English and colour I'm pretty sure. So this scene: There was a (or multiple) giant robot-like things and I think they were sort of sphere shaped. It was destroying a city and going around picking people up (possible killing them?) in giant net-like things i think.

COMPLEXITY

- Social nieces
- Contextual episodic memories
- Semantic memories about metadata, scenes, plots, etc.
- Exclusion criteria
- Multi-hop reasoning (comparisons)

Ok so I don't really remember anything but one scene, so I will try to give as much detail as I can. I saw this movie when I was very young and only remember this scene because (from what I remember) I don't think I actually watched most of the movie because I found it pretty scary, so it is probably not a children/family movie. I think I watched it in 2006 (mid-end of the year probably) and it was on a tv in someone's house, so it was old enough to be released on tv/dvd (not still in cinemas). It was in English and colour I'm pretty sure. So this scene: There was a (or multiple) giant robot-like things and I think they were sort of sphere shaped. It was destroying a city and going around picking people up (possible killing them?) in giant net-like things i think.

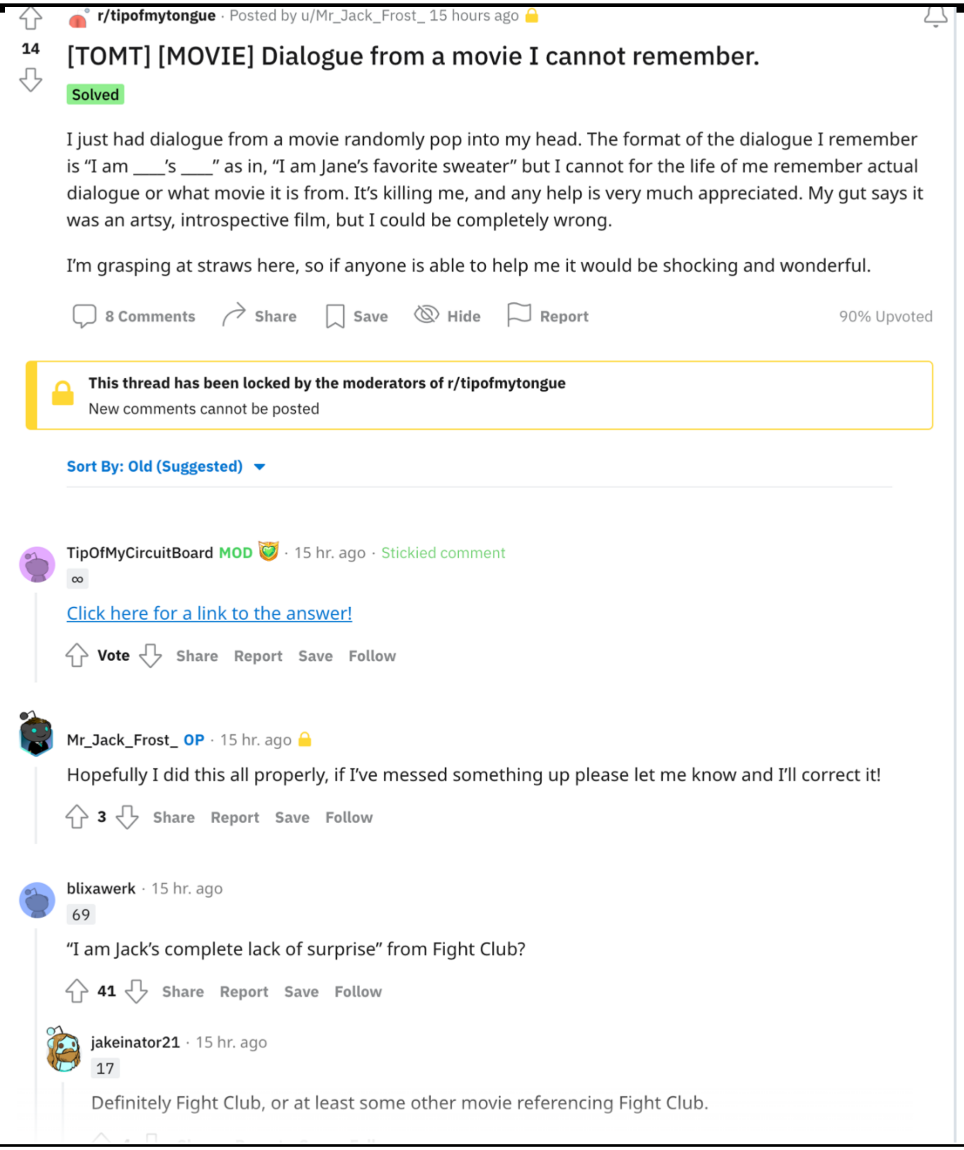
COMPLEXITY

- Social nieces
- Contextual episodic memories
- Semantic memories about metadata, scenes, plots, etc.
- Exclusion criteria
- Multi-hop reasoning (comparisons)

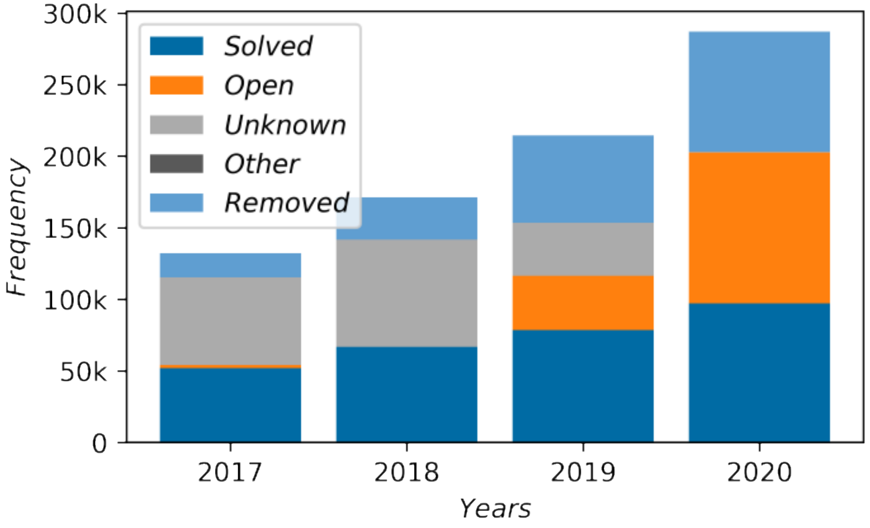
Ok so I don't really remember anything but one scene, so I will try to give as much detail as I can. I saw this movie when I was very young and only remember this scene because (from what I remember) I don't think I actually watched most of the movie because I found it pretty scary, so it is probably not a children/family movie. I think I watched it in 2006 (mid-end of the year probably) and it was on a tv in someone's house, so it was old enough to be released on tv/dvd (not still in cinemas). It was in English and colour I'm pretty sure. So this scene: There was a (or multiple) giant robot-like things and I think they were sort of sphere shaped. It was destroying a city and going around picking people up (possible killing them?) in giant net-like things i think.

REDDIT

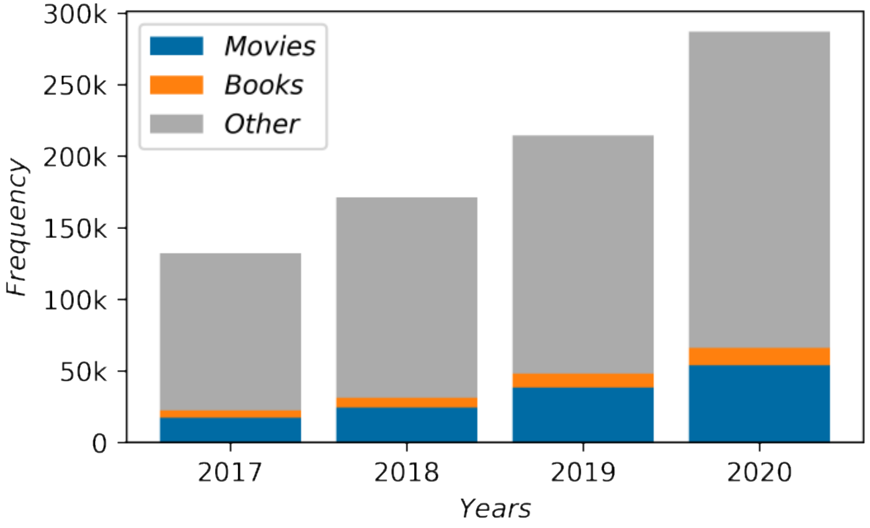
- /r/TipOfMyTongue Reddit community



Bhargav, Samarth, Georgios Sidiropoulos, and Evangelos Kanoulas. "'It's on the tip of my tongue' A new Dataset for Known-Item Retrieval." In WSDM 2022.



(a) Submission Status



(b) Category

Subset	No. queries	No. positive documents	No. other candidates	No. negatives	Total documents
Books	2319	1910	710	-	2620
Movies	13469	8845	4797	1221	14863
Total	15788	10755	5507	1221	17483

Table 2: Overall statistics of *Reddit-TOMT* .

Bhargav, Samarth, Georgios Sidiropoulos, and Evangelos Kanoulas. "'It's on the tip of my tongue' A new Dataset for Known-Item Retrieval." In WSDM 2022.

COMPLEXITY

Dataset	Query Length	Lexical Overlap
MSMarco (Campos et al., 2016)	7.68	0.55
Natural Questions (Kwiatkowski et al., 2019)	10.35	0.52
BioASQ (Tsatsaronis et al., 2015)	14.82	0.58
TREC-COVID (Roberts et al., 2020)	15.94	0.41
SciFact (Wadden et al., 2022)	19.52	0.50
HotPotQA (Yang et al., 2018b)	22.78	0.45
TOMT (Bhargav et al., 2022)	136.50	0.25

Kevin Lin, Kyle Lo, Joseph E. Gonzalez, Dan Klein. Decomposing Complex Queries for Tip-of-the-tongue Retrieval. In EMNLP Findings 2023

COMPLEXITY

Dataset	Query Length	Lexical Overlap
MSMarco (Campos et al., 2016)	7.68	0.55
Natural Questions (Kwiatkowski et al., 2019)	10.35	0.52
BioASQ (Tsatsaronis et al., 2015)	14.82	0.58
TREC-COVID (Roberts et al., 2020)	15.94	0.41
SciFact (Wadden et al., 2022)	19.52	0.50
HotPotQA (Yang et al., 2018b)	22.78	0.45
TOMT (Bhargav et al., 2022)	136.50	0.25
WhatsThatBook	156.20	0.19

Kevin Lin, Kyle Lo, Joseph E. Gonzalez, Dan Klein. Decomposing Complex Queries for Tip-of-the-tongue Retrieval. In EMNLP Findings 2023

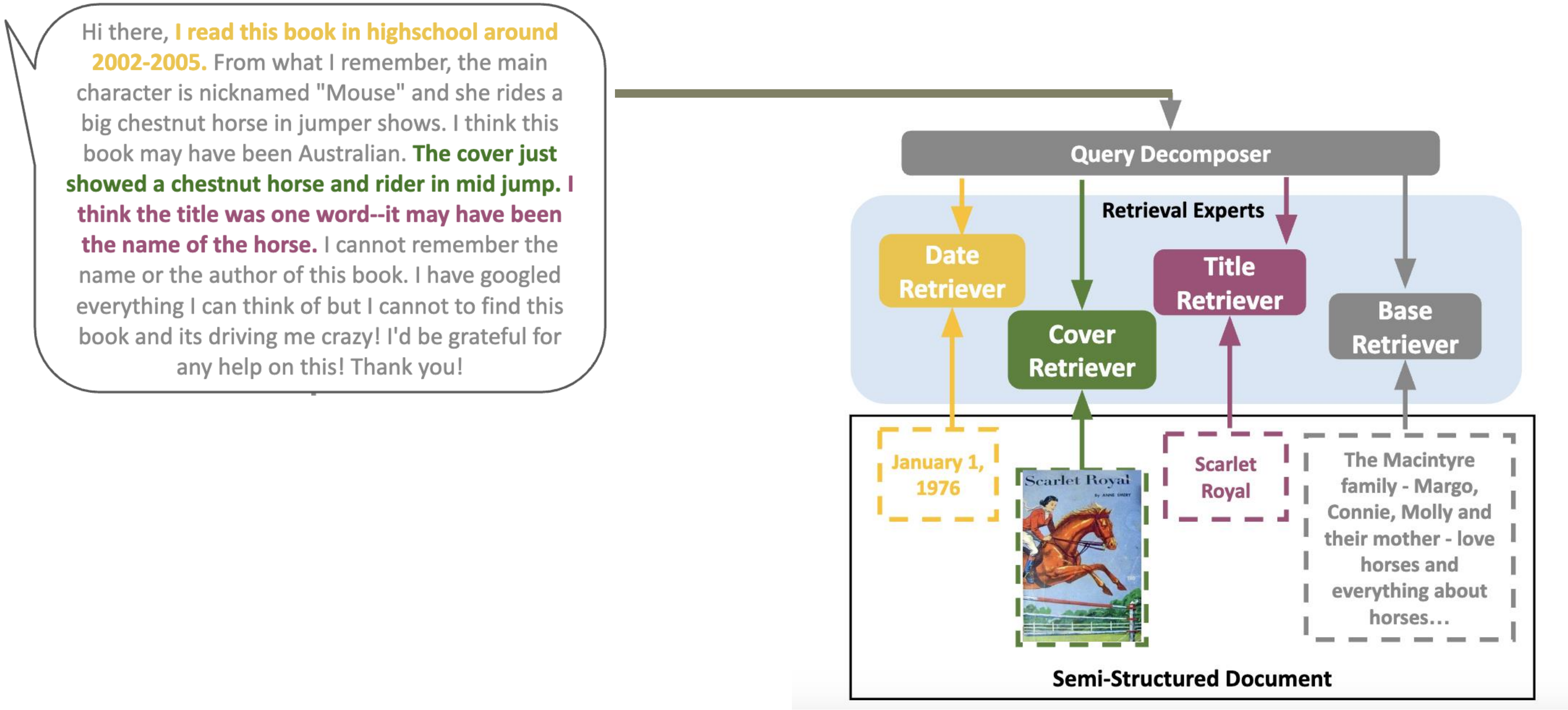
Dataset	Method	MRR@10
Books	BM25	$\boxed{0.197}$
	DR	0.278
Movies	BM25	$\boxed{0.167}$
	DR	0.193
MSMarco	BM25	$\boxed{0.240}$
	DR	0.311

Bhargav, Samarth, Georgios Sidiropoulos, and Evangelos Kanoulas. "'It's on the tip of my tongue' A new Dataset for Known-Item Retrieval." In WSDM 2022.

Dataset	Method	MRR@10
Books	BM25	0.197
	DR	$\boxed{0.278}$
Movies	BM25	0.167
	DR	$\boxed{0.193}$
MSMarco Dev	BM25	0.240
	DR	$\boxed{0.311}$

Bhargav, Samarth, Georgios Sidiropoulos, and Evangelos Kanoulas. "'It's on the tip of my tongue' A new Dataset for Known-Item Retrieval." In WSDM 2022.

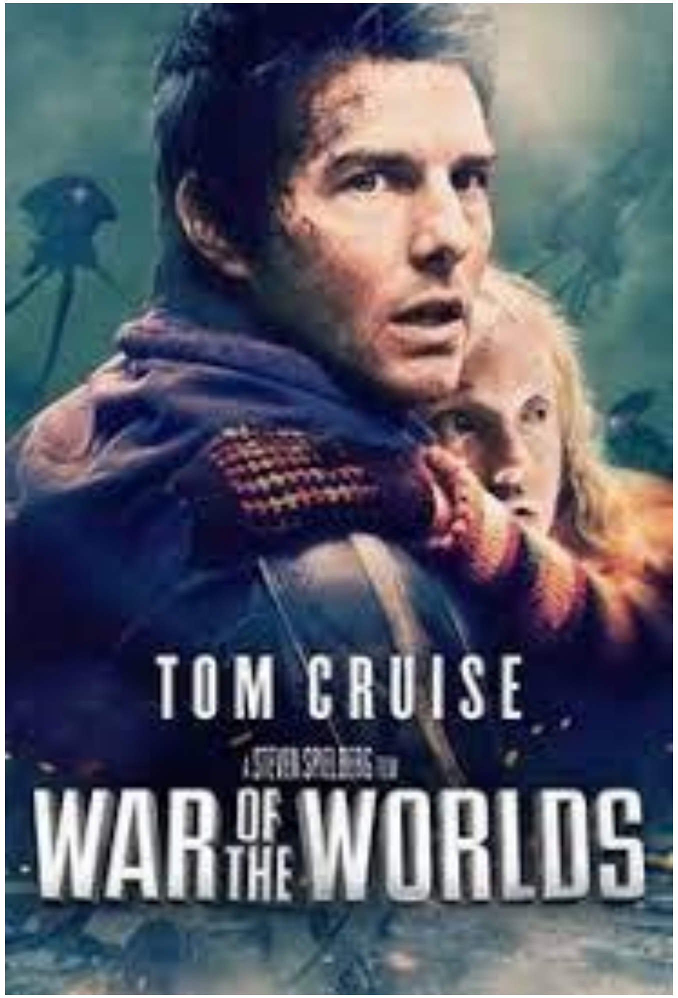
COMPLEXITY



Decomposing Complex Queries for Tip-of-the-tongue Retrieval Kevin Lin, Kyle Lo, Joseph E. Gonzalez, Dan Klein EMNLP Findings 2023

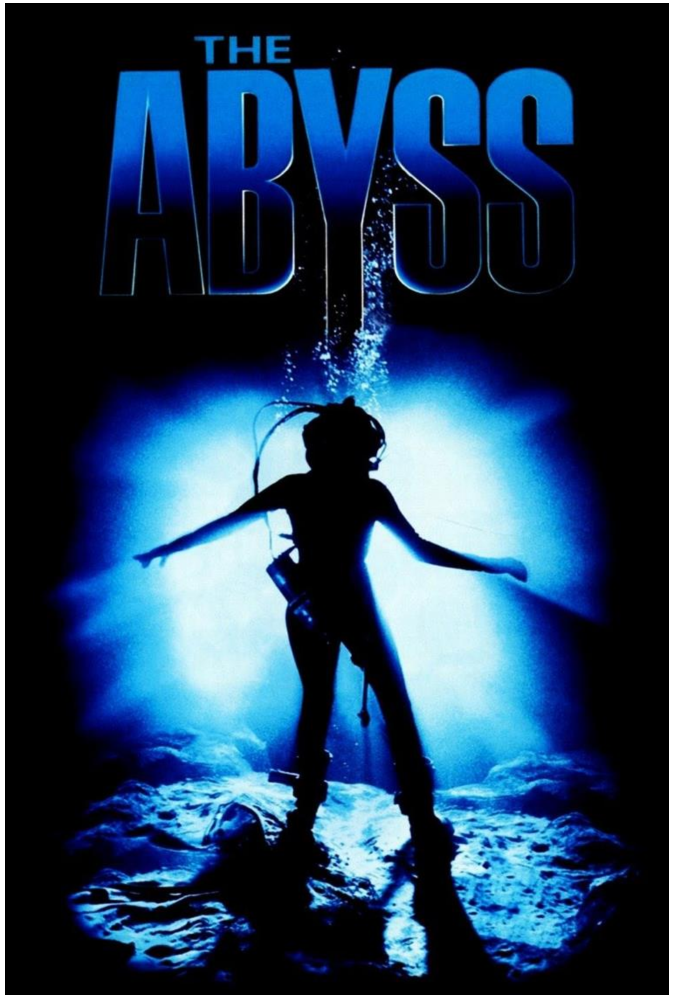
HEDGING

I don't think I actually watched most of the movie because I found it pretty scary, so it is probably not a children/family movie. I think I watched it in 2006 (mid-end of the year probably) and it was on a tv in someone's house, so it was old enough to be released on tv/dvd (not still in cinemas). It was in English and colour I'm pretty sure. So this scene: There was a (or multiple) giant robot-like things and I think they were sort of sphere shaped. It was destroying a city and going around picking people up (possible killing them?) in giant net-like things i think.



FALSE MEMORIES

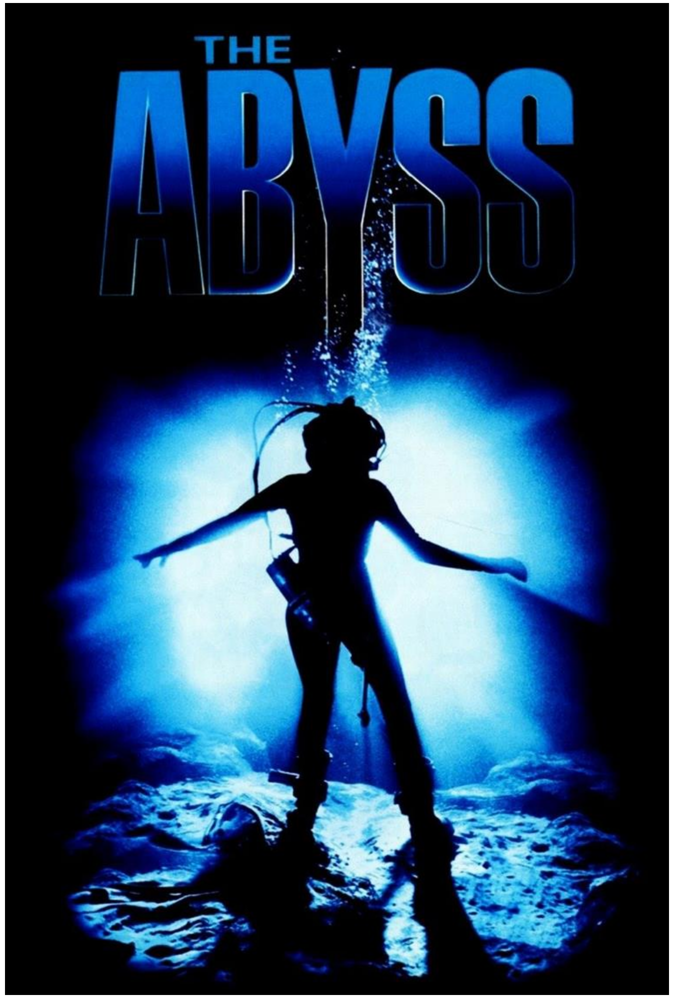
The only things that I remember are that there was something to do with water, and by this, I mean like an ocean, or maybe it was something like a submarine or an oil rig, and I think that the main character was a guy, who was either a ghost, or a silvery liquid metally alien/liquidy thing, that looks a bit like the T-1000 in Terminator 2. Sorry for the lack of detail, and please help me remember this!



FALSE MEMORIES

The only things that I remember are that there was something to do with water, and by this, I mean like an ocean, or maybe it was something like a submarine or an oil rig, and I think that the main character was a guy, who was either a ghost, or a silvery liquid metally alien/liquidy thing, that looks a bit like the T-1000 in Terminator 2. Sorry for the lack of detail, and please help me remember this!

The average rank of the correct item drops from position 10 to position 15.



52

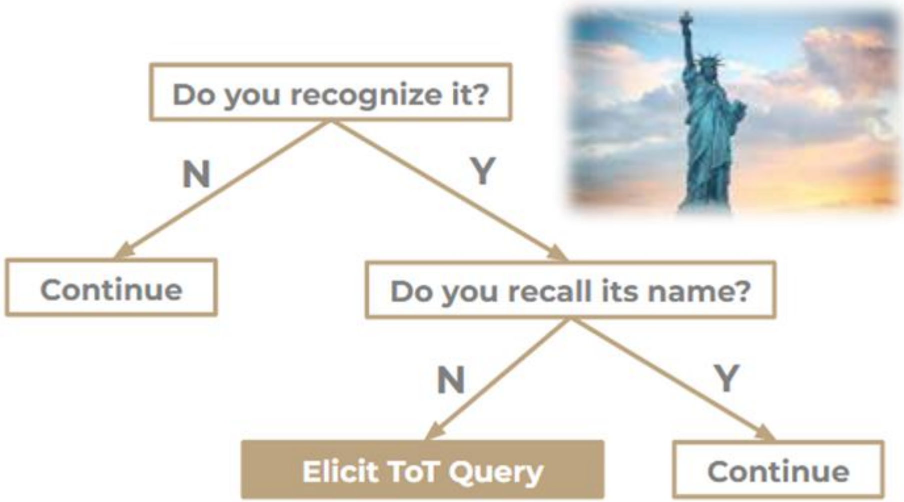
TREC 2024 TIP-OF-THE-TONGUE (TOT) TRACK

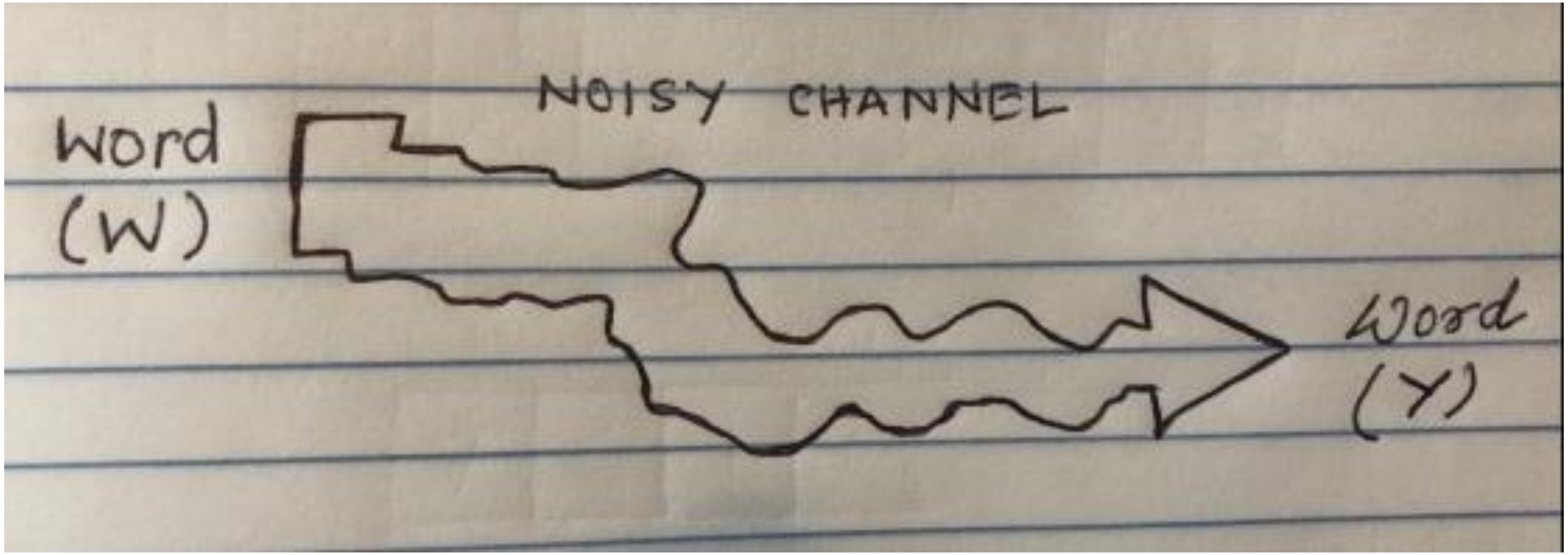
Task 1: Movie identification (same as 2023)

- **Test Queries:** 150 new queries sampled from MS-ToT dataset
- **Train/Dev Queries:** 450 queries from 2023 ToT Track
- **Corpus:** Wikipedia Corpus (~230K articles)
- **Task:** Given a ToT query, rank Wikipedia articles

Task 2: ToT known-item search for new domains with ToT query elicitation

- New ToT queries elicited from crowd workers
- **Domains:** movies, landmarks, celebrities, recipes, everyday objects, etc.



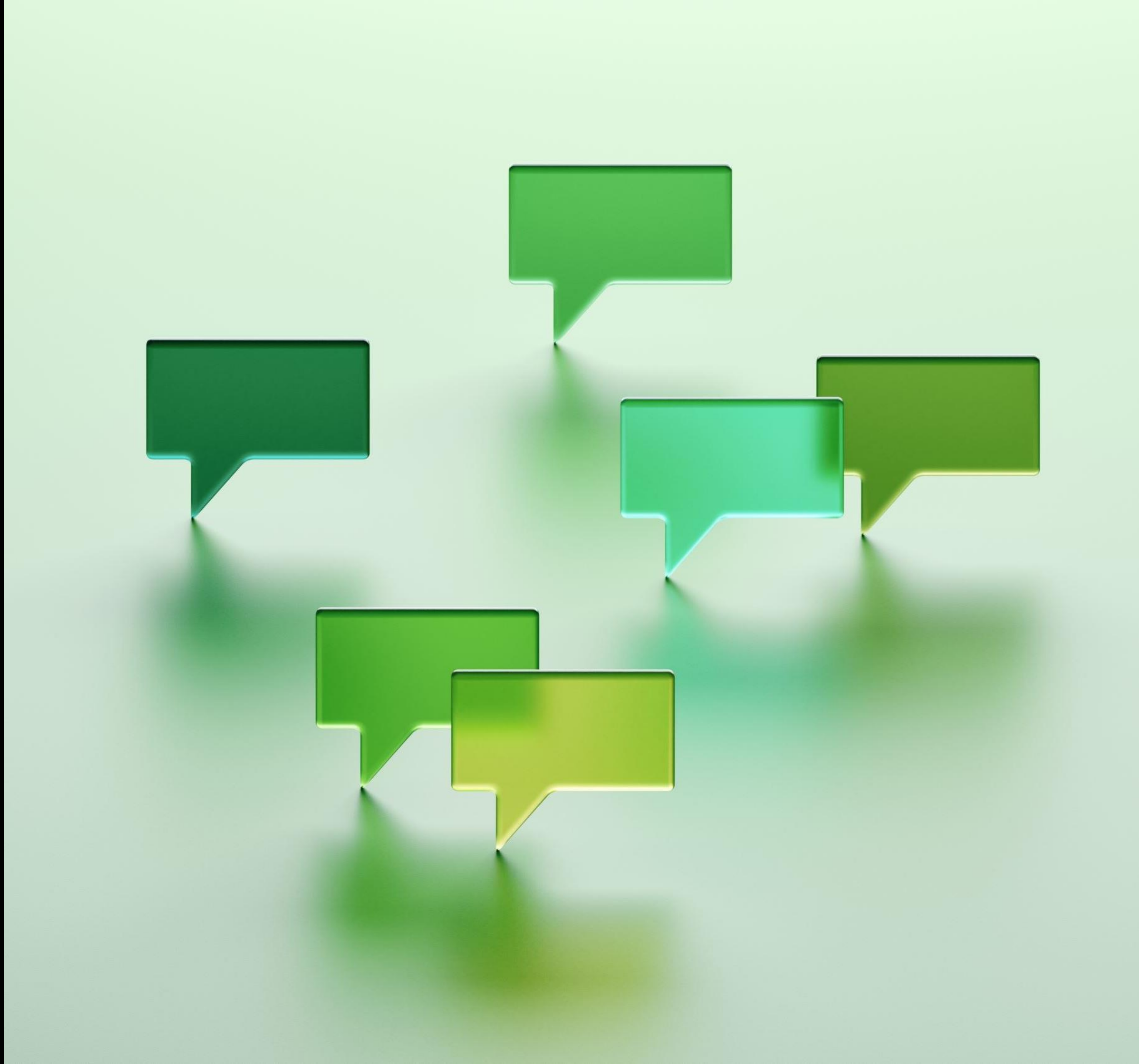


Typos and ASR mistakes → Contrastive learning and data augmentation.

False memories → Uncertainty estimation and attribution.

Conversational questions → Domain transfer.

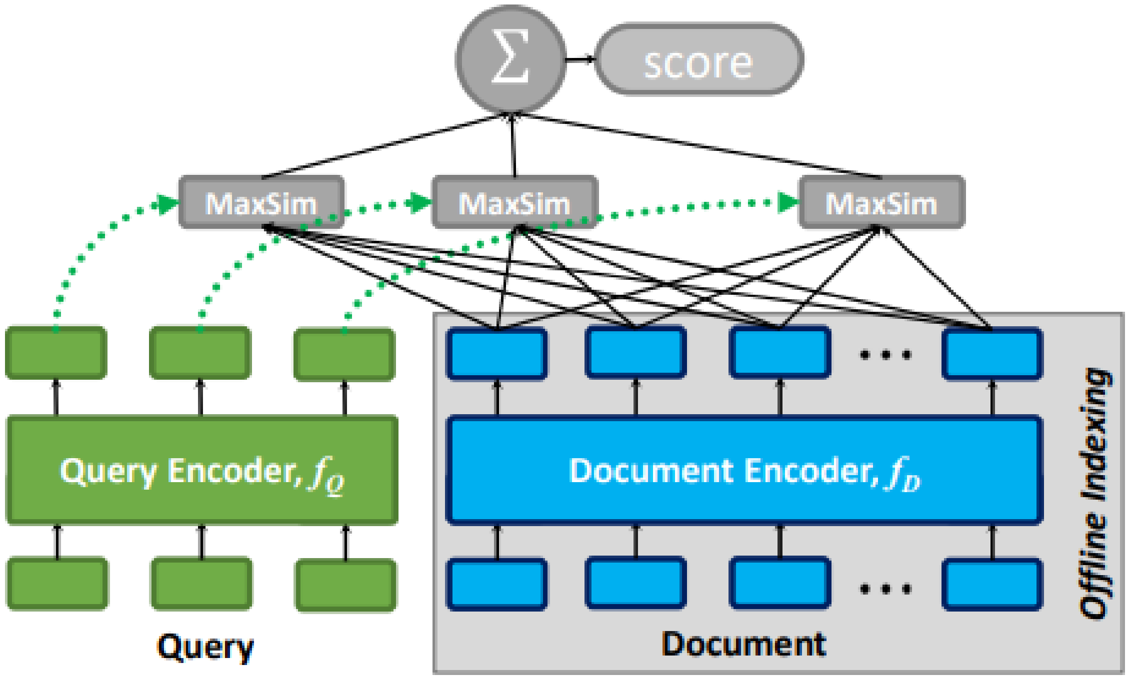
CONVERSATIONS



CONVERSATIONS

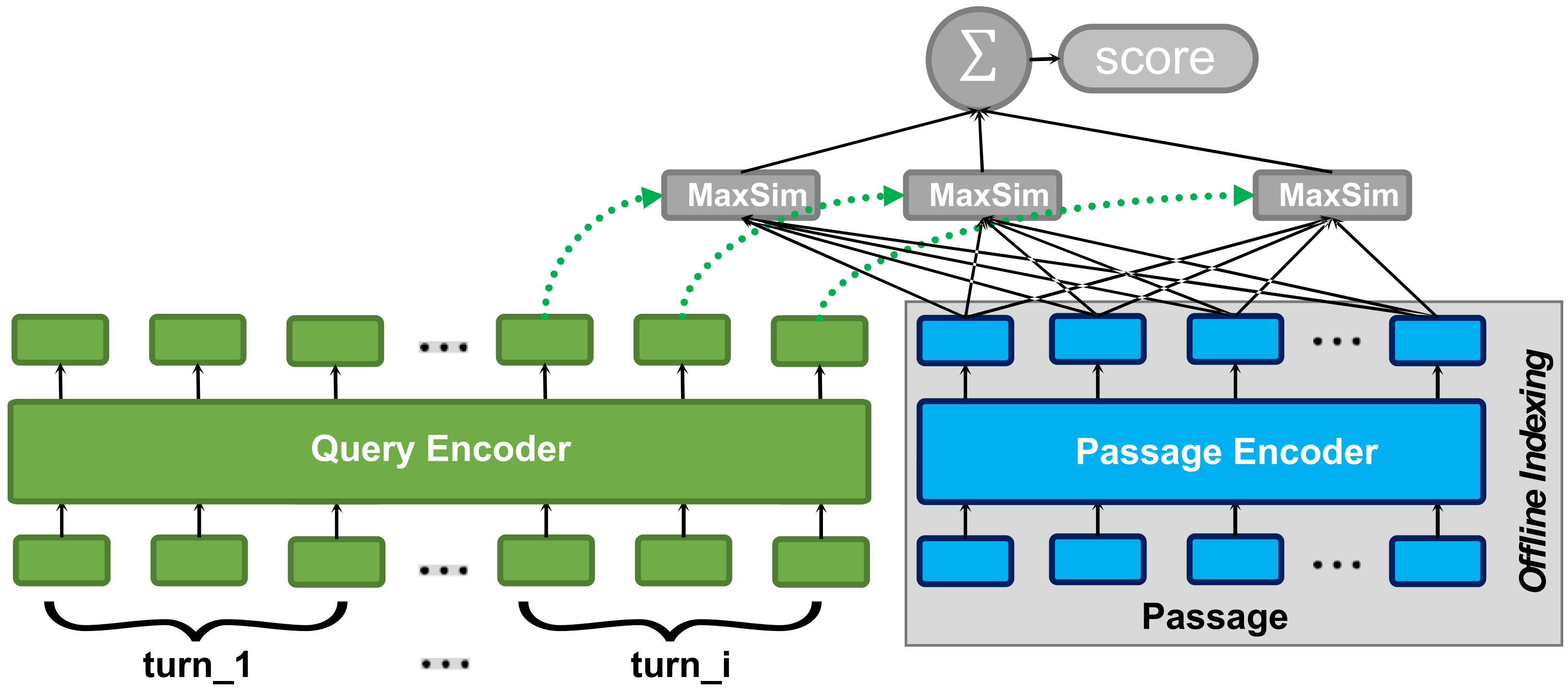
Turn	Query / Answer
1	What are the alternatives to cow milk?
	Commercially available **almond milk** is often enriched with vitamins A and D, as well as calcium, to make it seem more like regular milk. (Almonds provide some calcium, but not as much as cow's milk.) Almond milk is low in calories compared to other milk substitutes and contains some heart-healthy monounsaturated fat. You might also try **oat**, **hemp** or **rice** milk. Oat milk provides more fiber than other milk substitutes and was shown in a Swedish study to help lower LDL ("bad") cholesterol and total cholesterol. Hemp milk, made by pulverizing the seeds, blending them with water and straining out the solid residue, provides both omega-6 and omega-3 essential fatty acids in a healthy three-to-one ratio.
2	Which is healthiest?

LATE INTERACTION



Khattab, Omar, and Matei Zaharia. "Colbert: Efficient and effective passage search via contextualized late interaction over bert." In SIGIR 2020.

ZERO-SHOT QUERY CONTEXTUALIZATION



Krasakis, Antonios Minas, Andrew Yates, and Evangelos Kanoulas. "Zero-shot query contextualization for conversational search." In SIGIR 2022

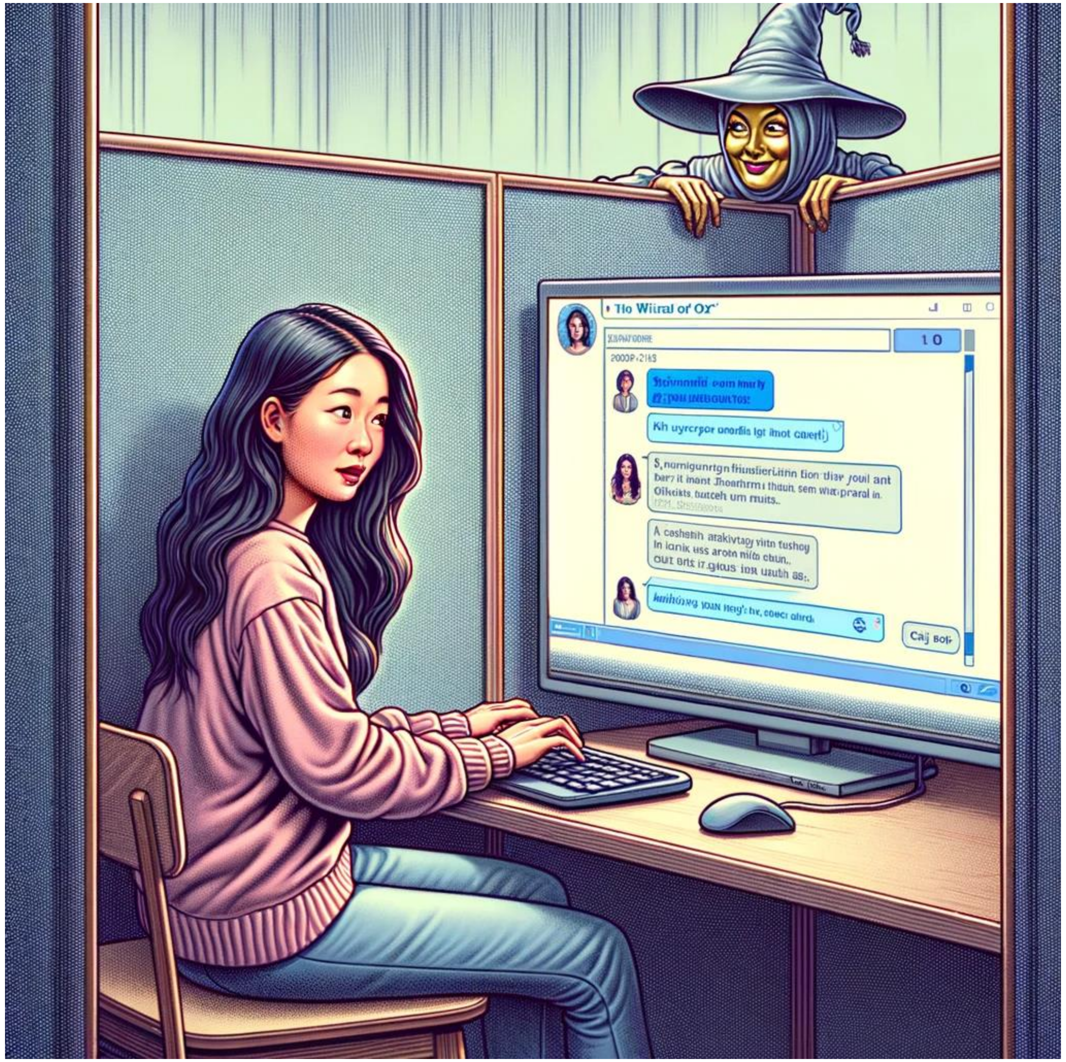
ZERO-SHOT QUERY CONTEXTUALIZATION

$turn_3$	tell me about $\overbrace{\text{lung cancer}}^{\text{Resolution (R)}}$.
$turn_4$	what are $\underbrace{\text{its}}_{\text{Anaphora (A)}}$ symptoms?

$sim(\vec{A}_{last\_turn}, \vec{R})$	0.13
$sim(\vec{A}_{ZeCo^2}, \vec{R})$	0.47
$\Delta sim(\vec{A} \rightarrow \vec{R})$	0.34
$\Delta Recall$	$+0.24$

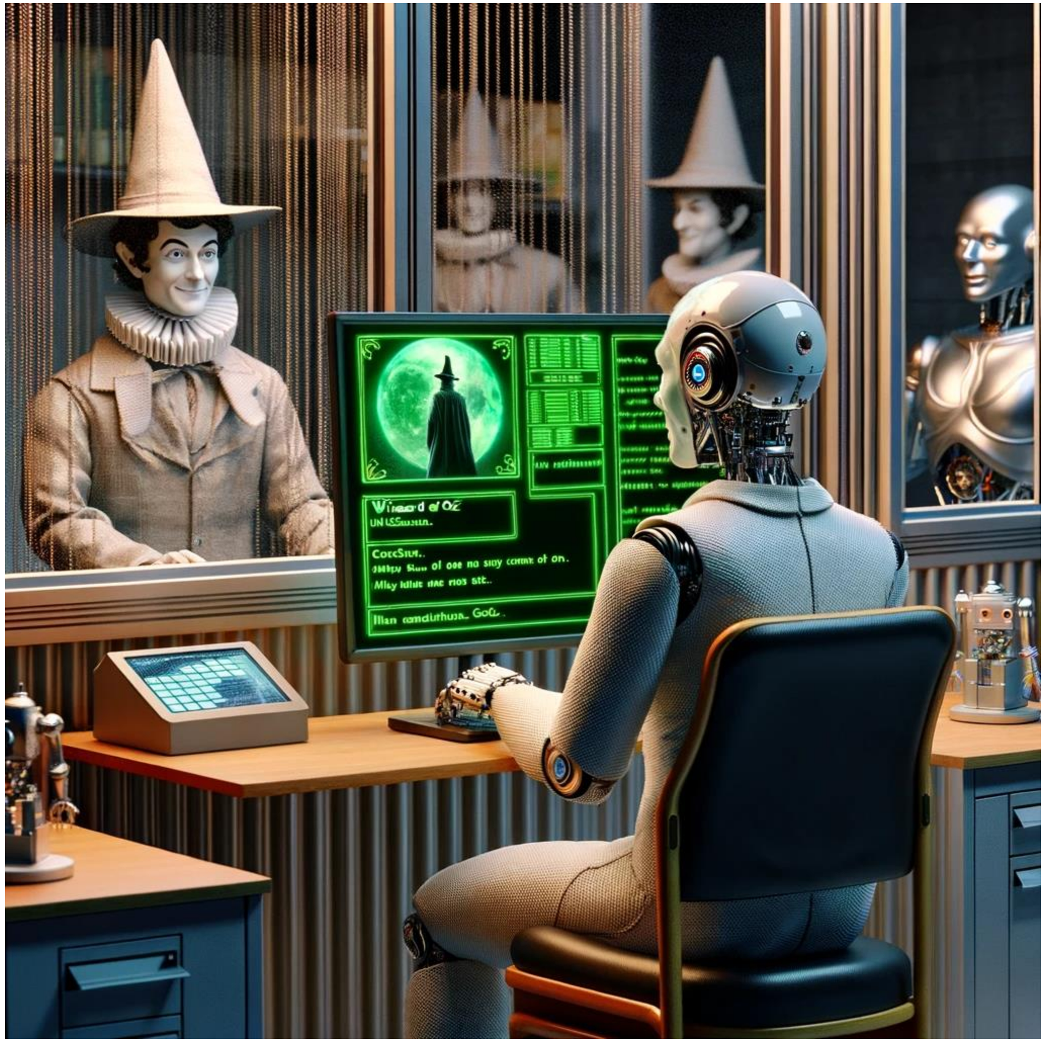
Krasakis, Antonios Minas, Andrew Yates, and Evangelos Kanoulas. "Zero-shot query contextualization for conversational search." In SIGIR 2022

WIZARD OF OZ



DALL*E

WIZARD OF OZ



DALL*E

Abbasiantaeb, Zahra, et al. "Let the LLMs Talk: Simulating Human-to-Human Conversational QA via Zero-Shot LLM-to-LLM Interactions." To appear in WSDM 2024.

WIZARD OF OZ

VS

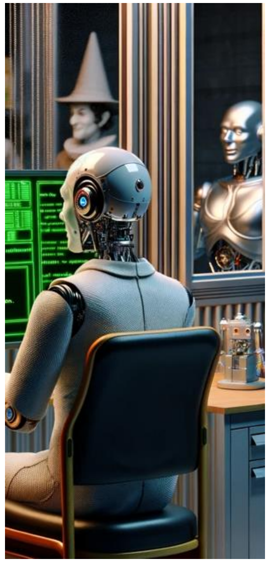
Annot. level	Metric	$answer_{QuAC}$	$answer_{Sim}$	Tie
Question	Correctness	11.31%	**38.6%**	50.0%
	Naturalness	7.1%	**42.1%**	50.7%
	Completeness	5.26%	**53.8%**	40.8%
Conversation	Preference	6.12%	**87.7%**	6.18%

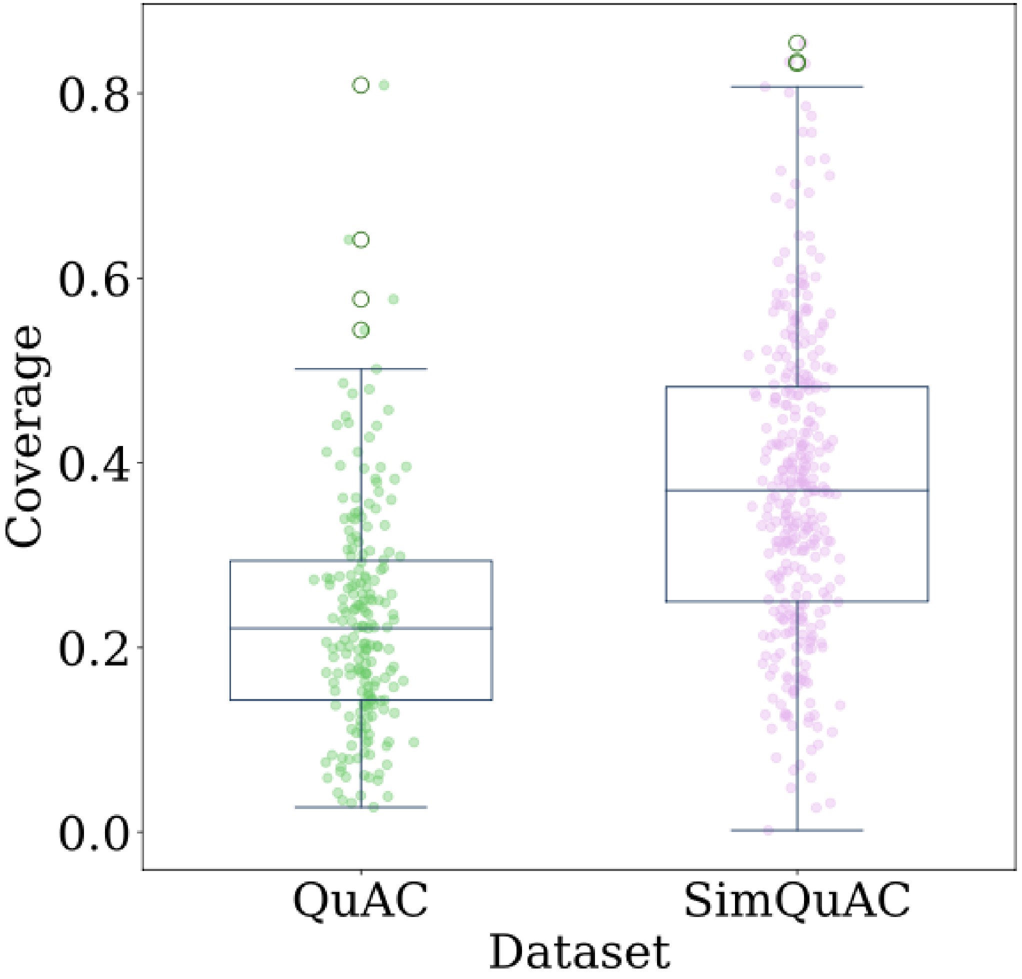
DALL*E

Abbasiantaeb, Zahra, et al. "Let the LLMs Talk: Simulating Human-to-Human Conversational QA via Zero-Shot LLM-to-LLM Interactions." To appear in WSDM 2024.

WIZARD OF OZ

VS



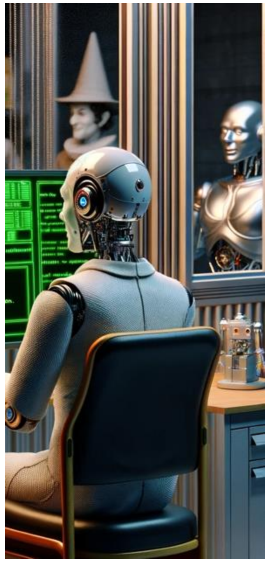


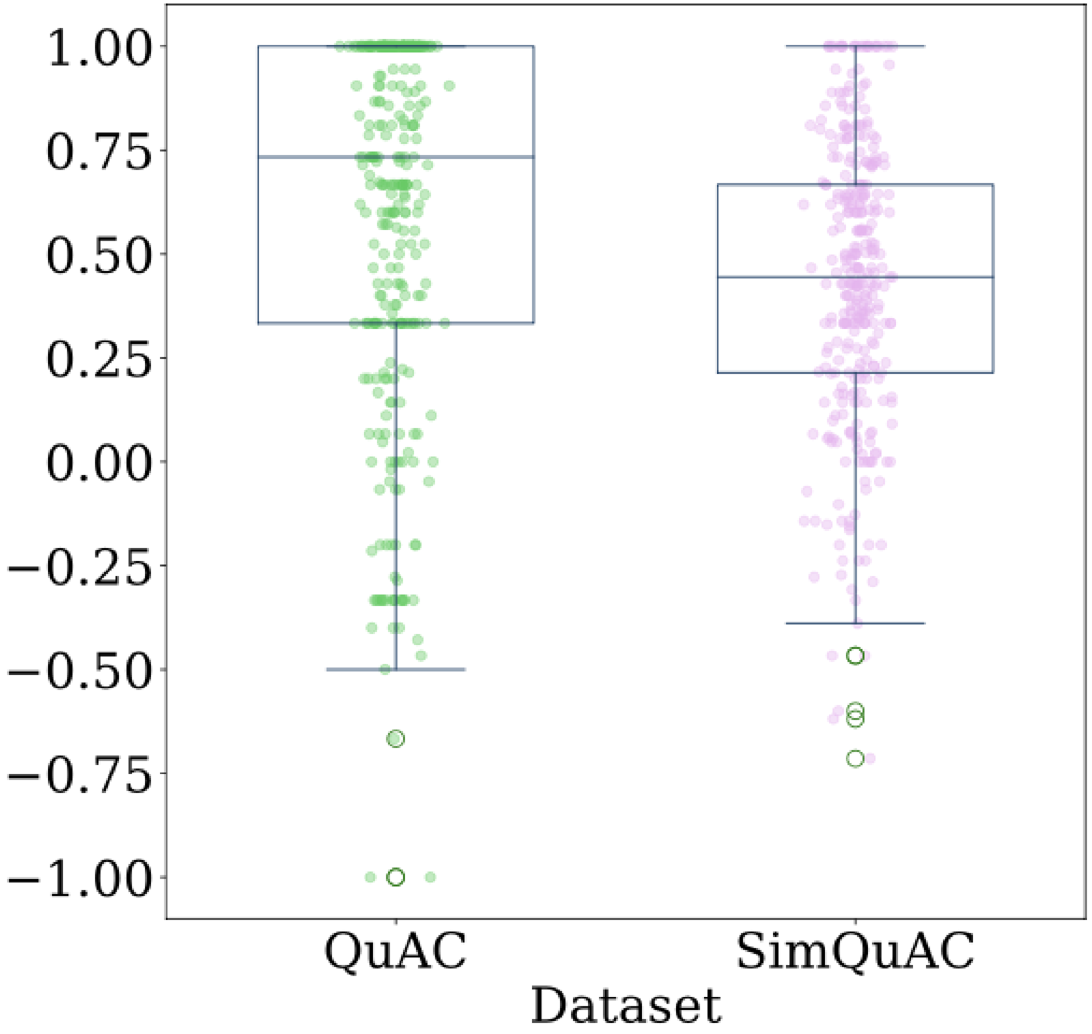
(a) Topic coverage

Abbasiantaeb, Zahra, et al. "Let the LLMs Talk: Simulating Human-to-Human Conversational QA via Zero-Shot LLM-to-LLM Interactions." To appear in WSDM 2024.

WIZARD OF OZ

vs





(b) Conversation flow

Abbasiantaeb, Zahra, et al. "Let the LLMs Talk: Simulating Human-to-Human Conversational QA via Zero-Shot LLM-to-LLM Interactions." To appear in WSDM 2024.

BROKEN TELEPHONE