Robust and efficient frontier pipelines for complex knowledge intensive tasks in the era of LLMs

CNI Seminar series, IISC

Venktesh Viswanathan

Postdoctoral Researcher, TU Delft

TUDelft

Knowledge Intensive Language Tasks

Efficient ⟷ **Effective**

Conversational AI

Web Search

Live fact checking

Knowledge-base Construction

Fact Checking Articles

Financial Auditing

Factual Article Generation

Indexing and Precompute

Test time Reasoning

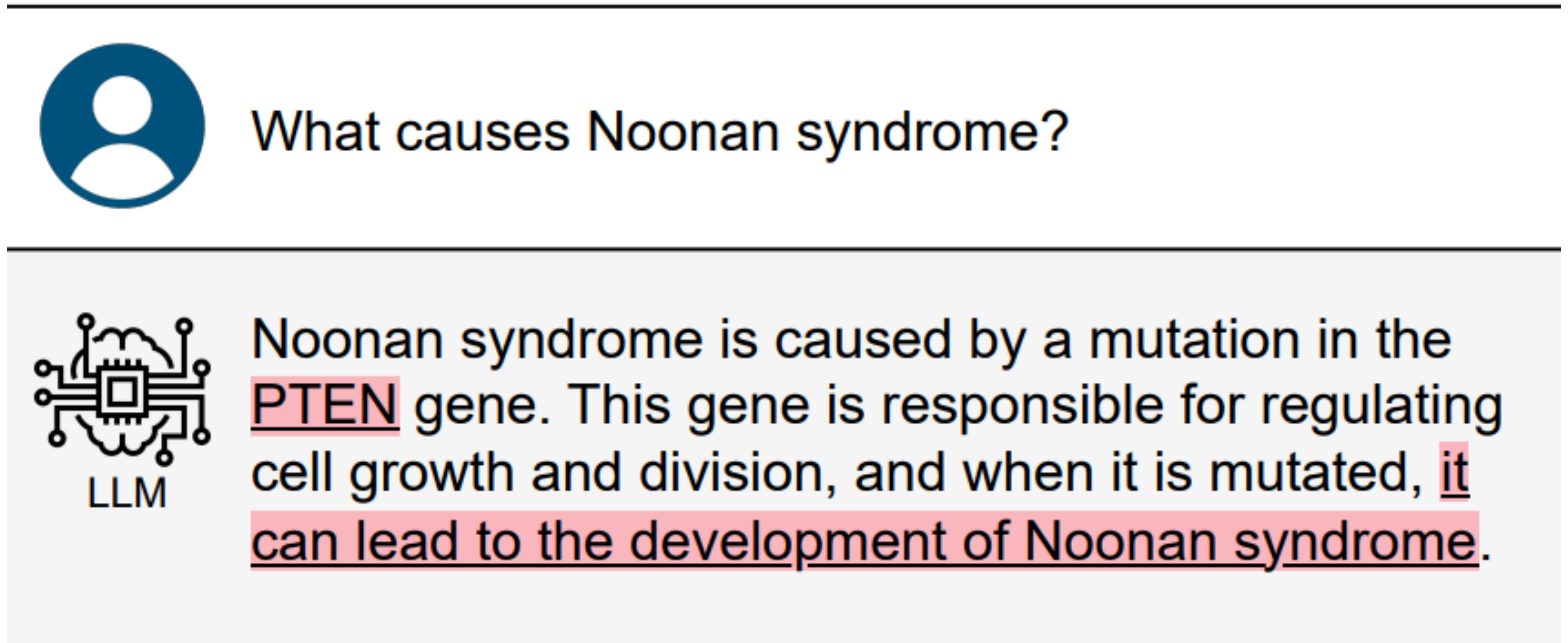
[Hoffart et al, WWW' 16], [Fetahu et al. '16, '17]

LLM Agents as general purpose solvers

TU Delft

Credit: Lil'Log

Hallucination in closed book setting



Now imagine a LLM citing fake cases when a resident is preparing his report
Or a lawyer preparing his arguments

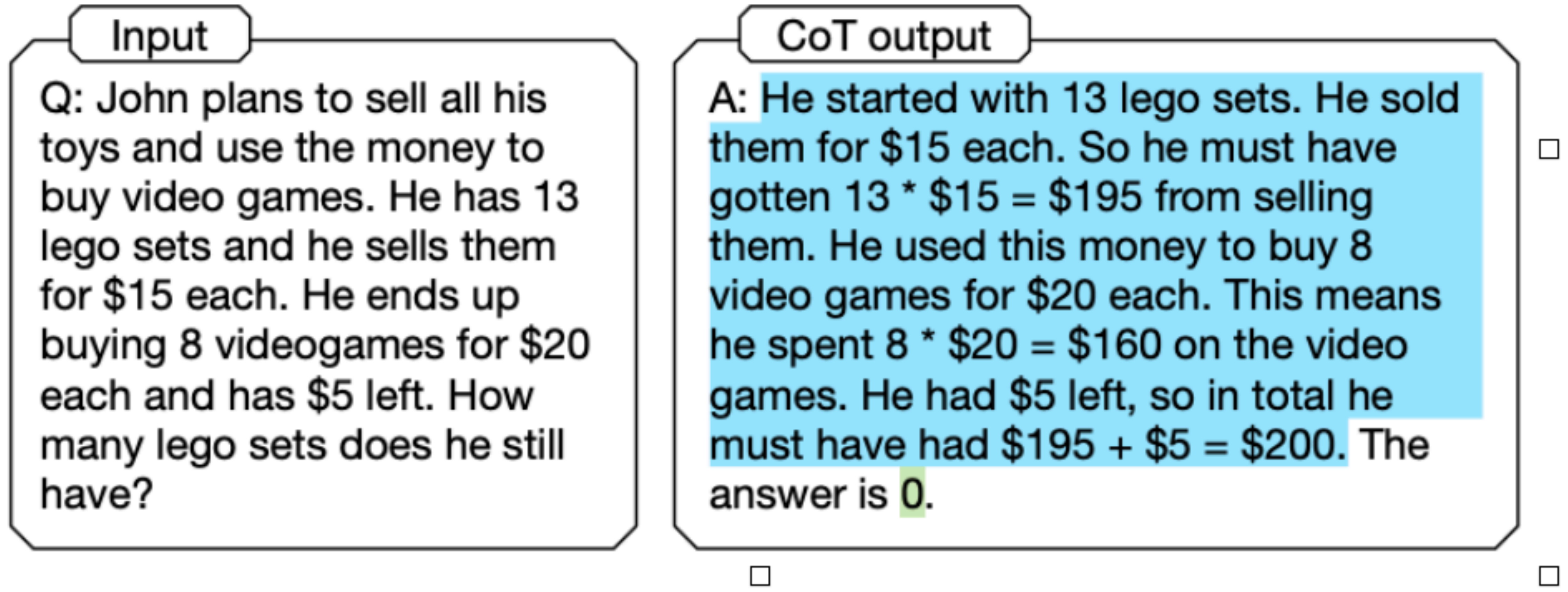
There’s no provenance even if answer is correct.

Hallucinations - catastrophic effects

How to mitigate hallucination and establish provenance?

Ask LLM to explain itself ?

Unfaithful Reasoning

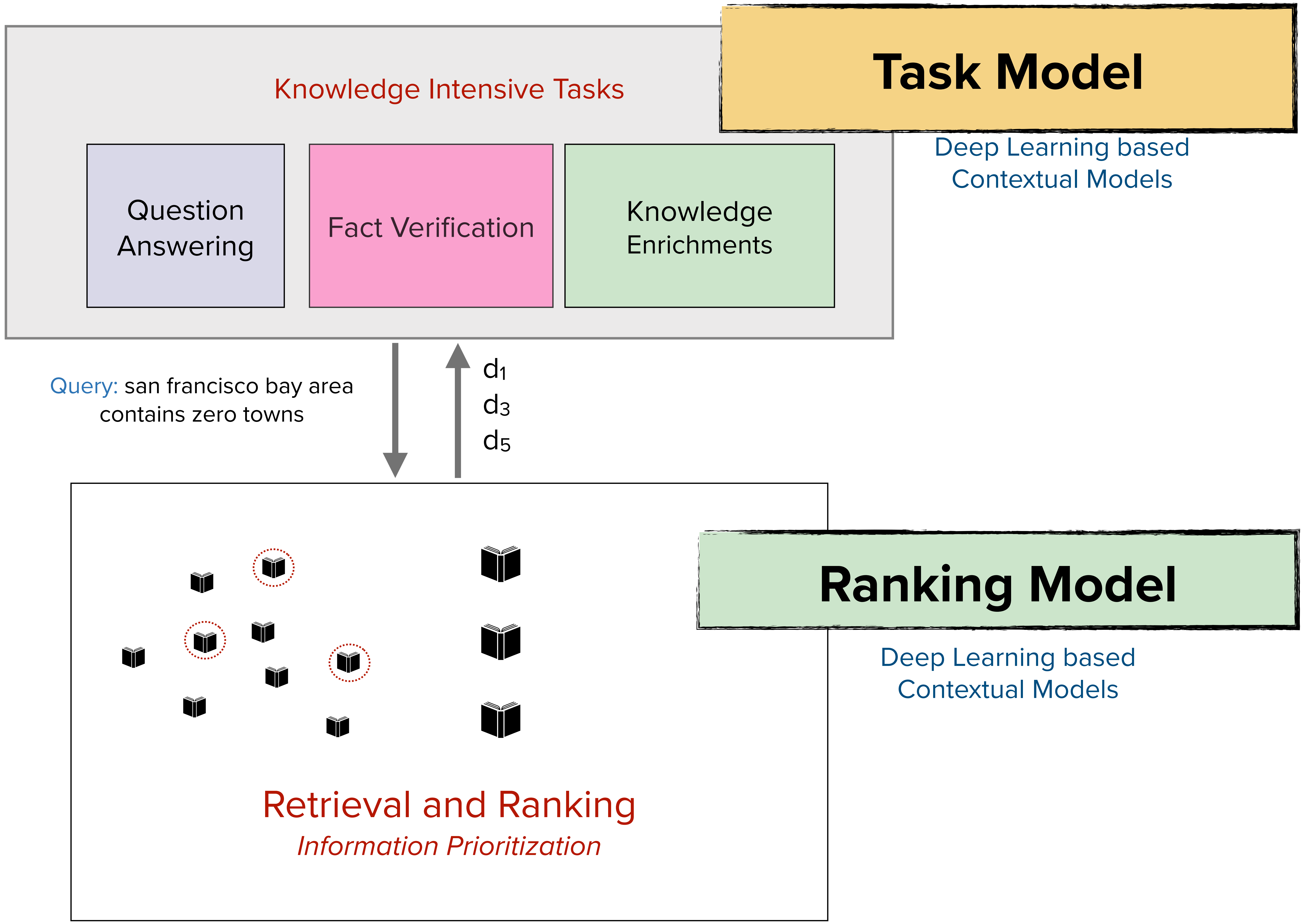


Knowledge Gaps

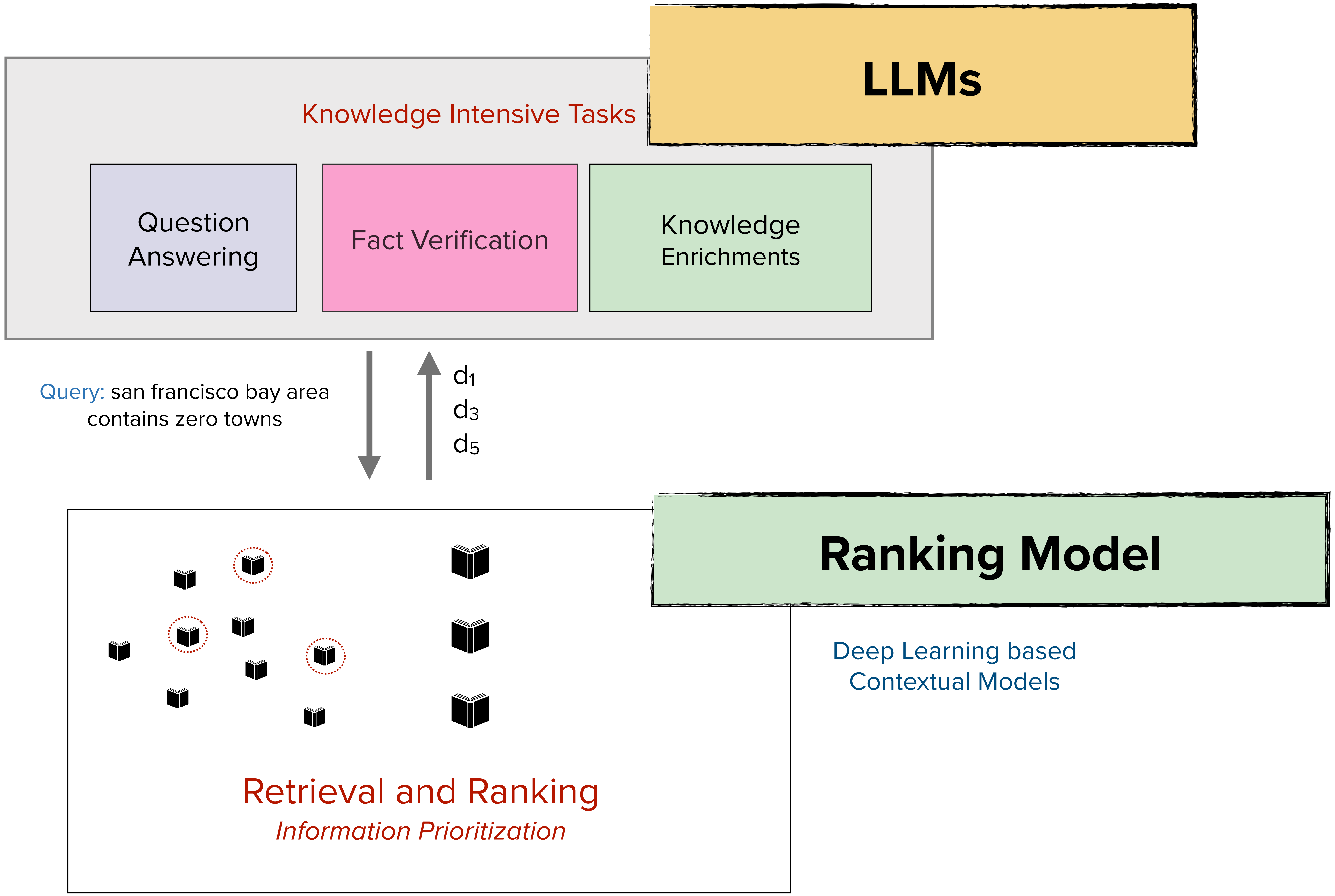
Question: What is the mouth of the river which serves as the mouth of the Bumping River?

FEW-SHOT-COT. : [Answer]: There is no river named Bumping River.

RAG to the rescue

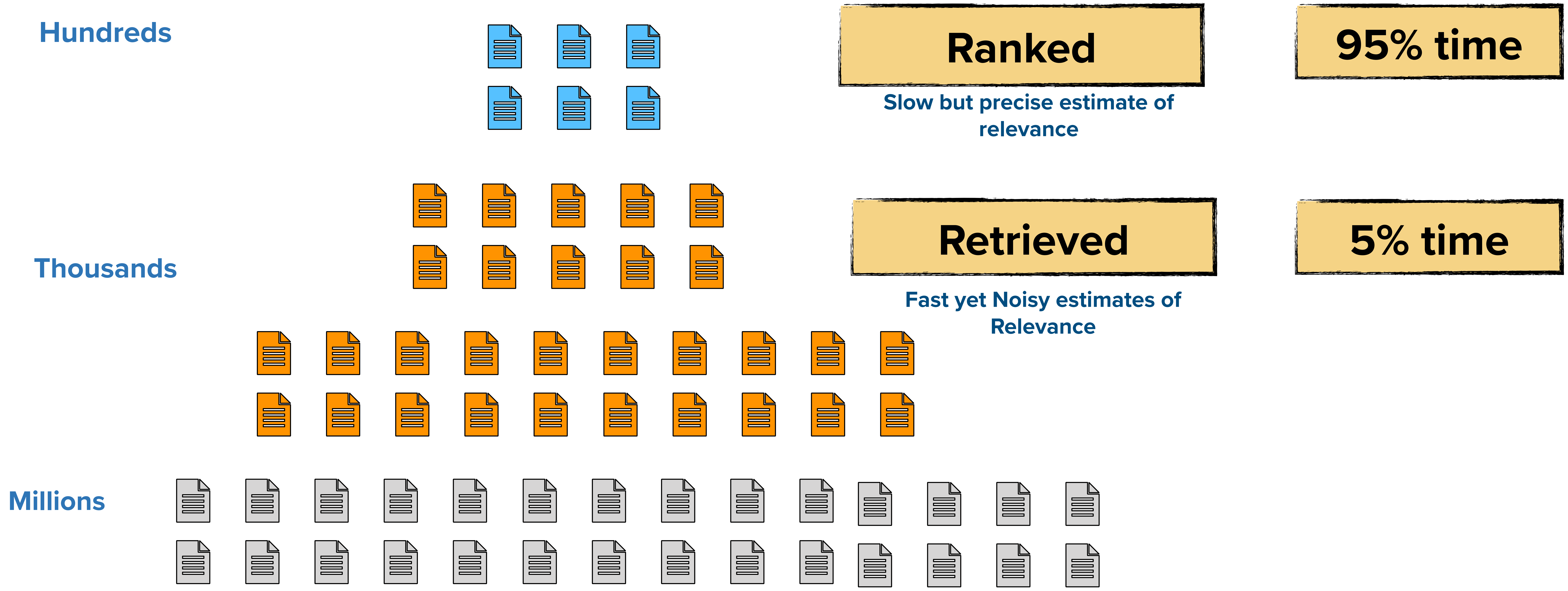


RAG to the Rescue



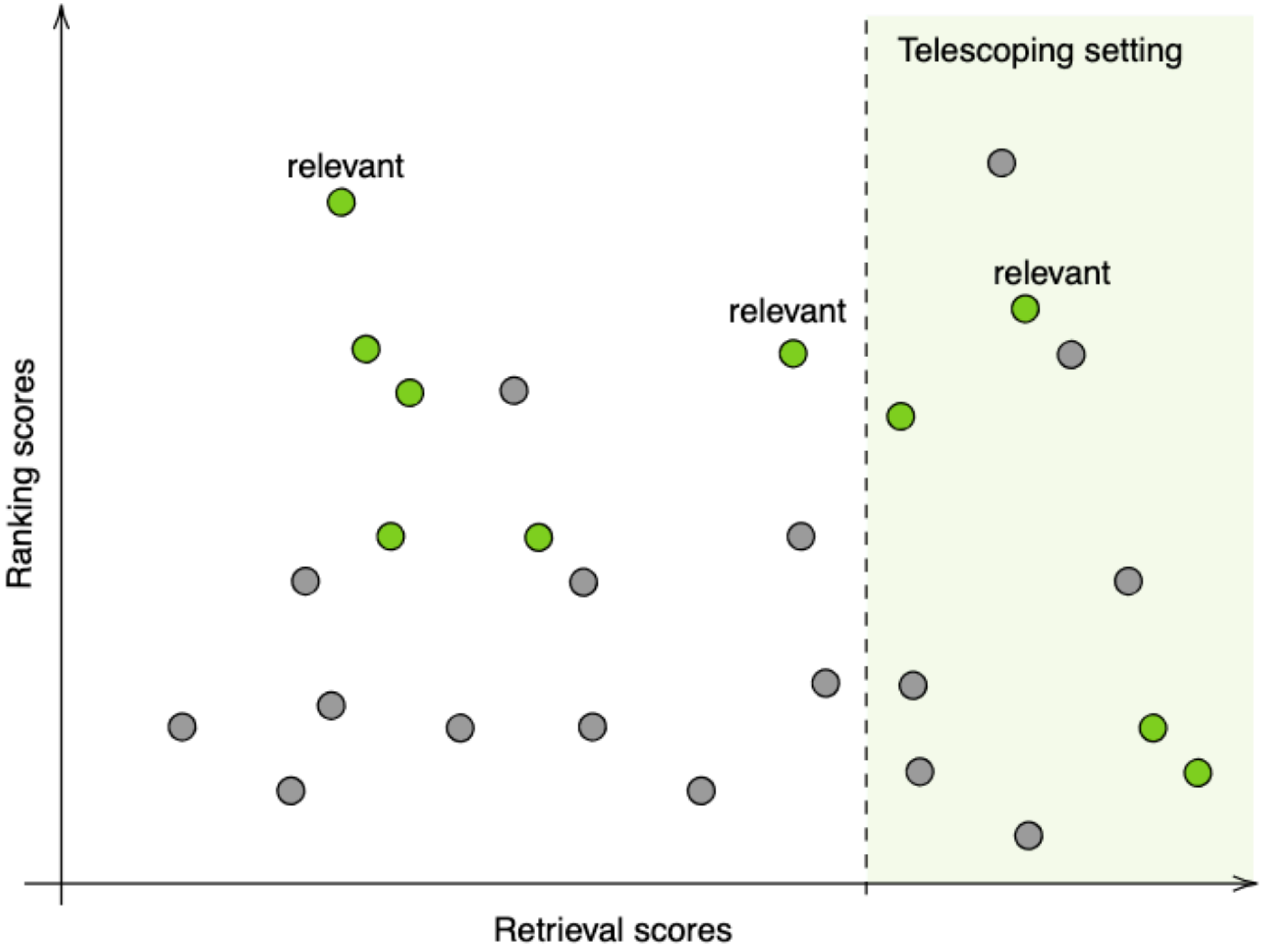
9

Telescoping view of retrieve-rerank pipelines



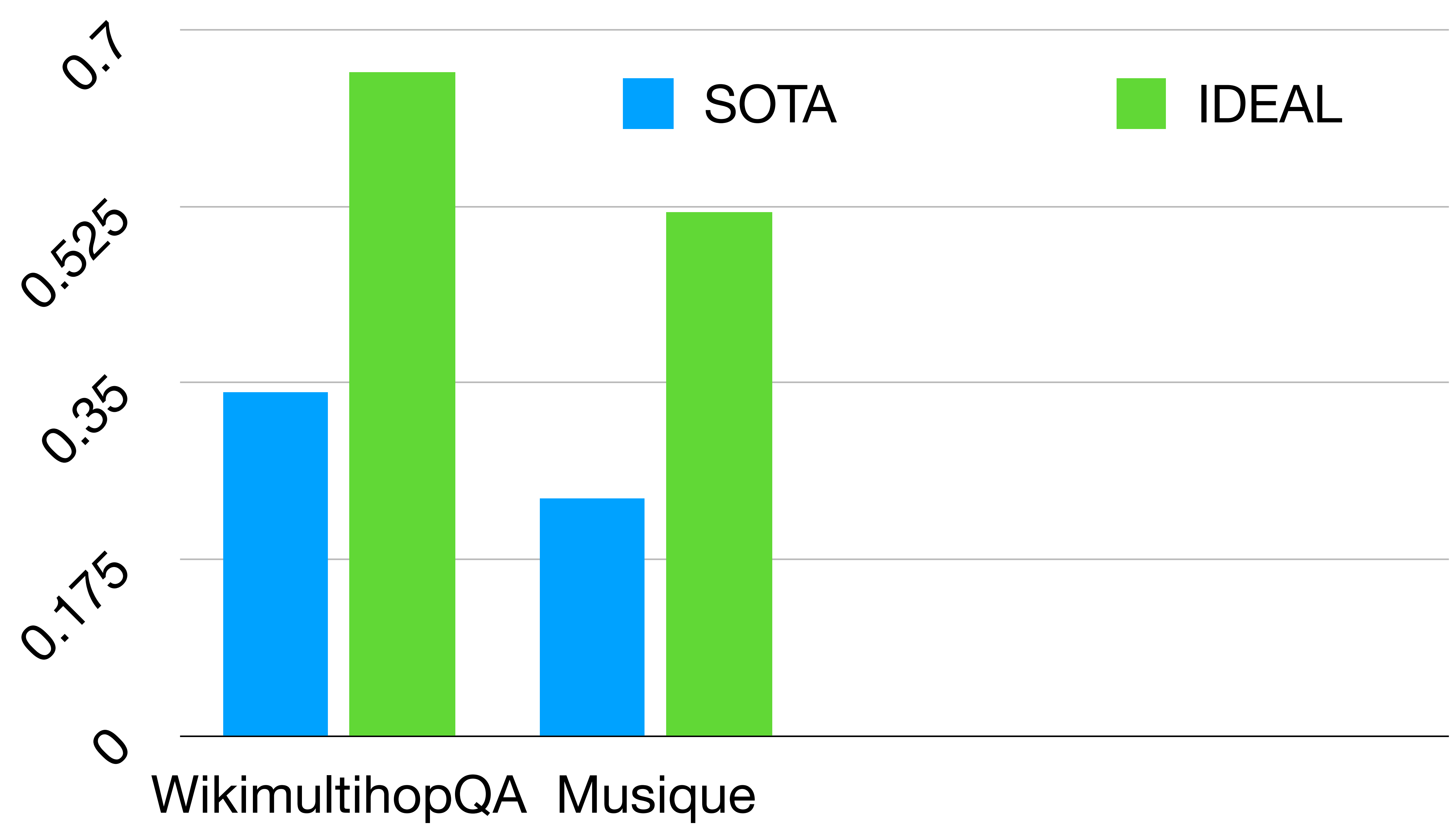
Bounded recall problem

- RAG pipelines require the most relevant document to appear within top-5 or top-10 to fit i context of most affordable LLMs.
- Classical re-ranking approaches are limited by recall of first-stage retrieval.
- How do we capture more relevant documents ?
- How do we ensure the relevant documents are ranked higher and answer the question ?

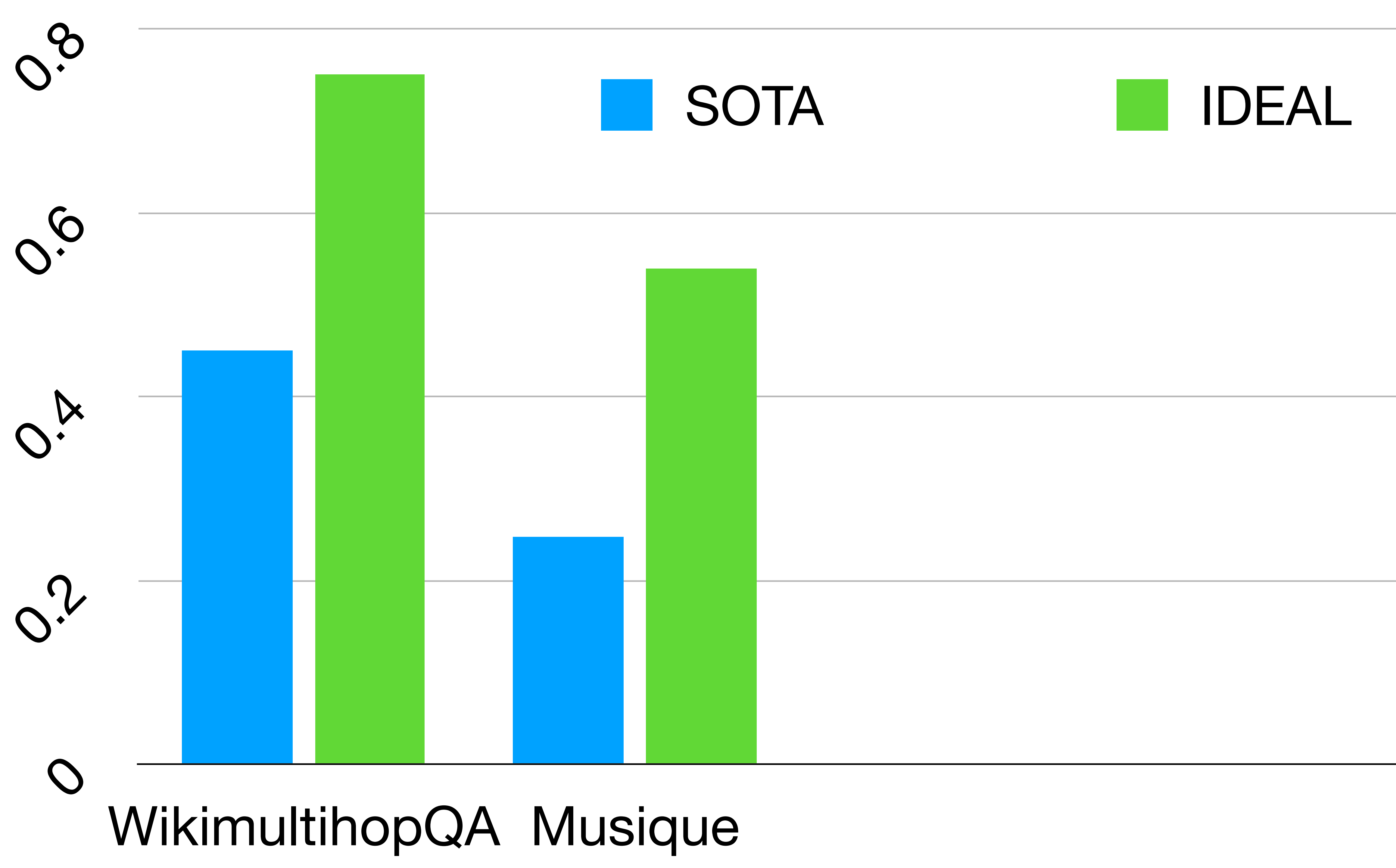


Retrieval and Reasoning Gap in complex QA

Retrieval Gap

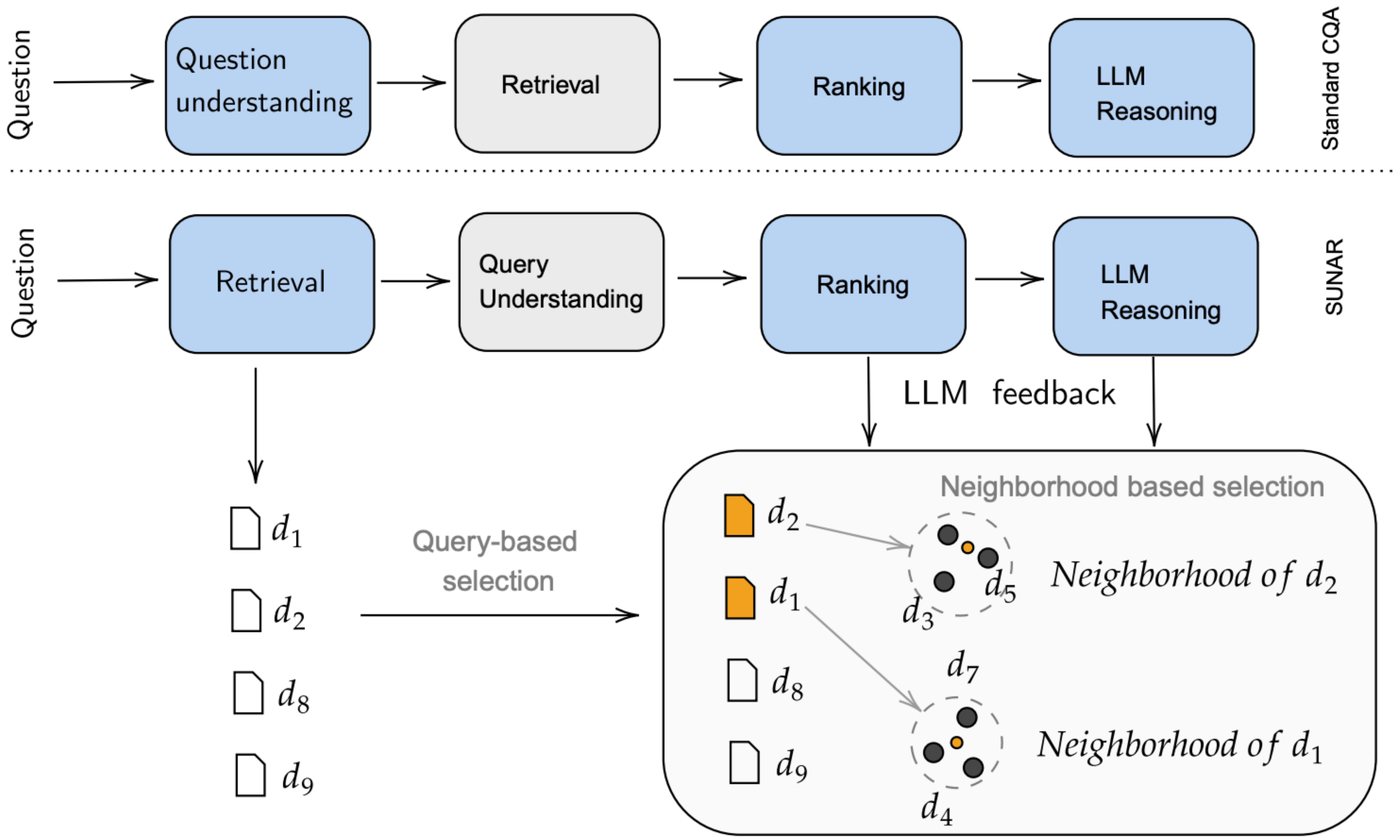


Reasoning gap



TUDelft

Semantic-Uncertainty based Neighborhood Aware Retrieval



Solving the Retrieval gap through LLM uncertainty based feedback

TU Delft

SUNAR- Deep Dive

Algorithm 1 The SUNAR Algorithm

Input: Initial retrieved list R, batch size b, re-ranking budget c, document graph G

Output: Re-Ranked pool R^+

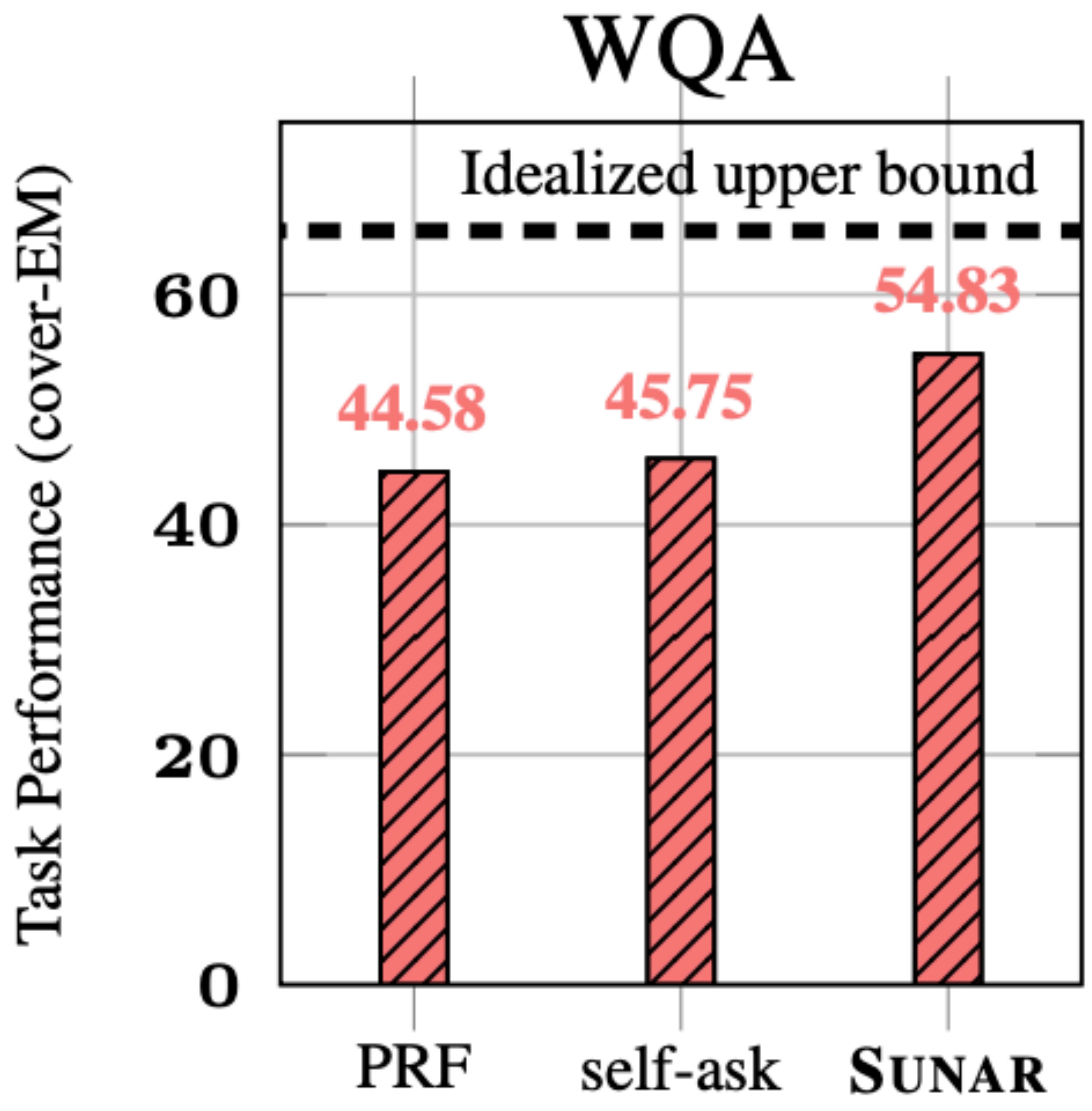
$R^+ \leftarrow \emptyset$ ▷ Re-Ranking results
$C \leftarrow R$ ▷ Re-ranking pool
$N \leftarrow \emptyset$ ▷ Neighbor pool
do
  $B \leftarrow \text{SCORE}(\text{top } b \text{ from } P\text{, subject to } c)$
  $\{sa_1 \ldots sa_m\} \leftarrow \phi(\mathbb{P}_{LLM}(sq_1, B))$
  $\{ac_1 .. ac_s\} \leftarrow \sigma(sa_1 .. sa_m)$ ▷ Clustering

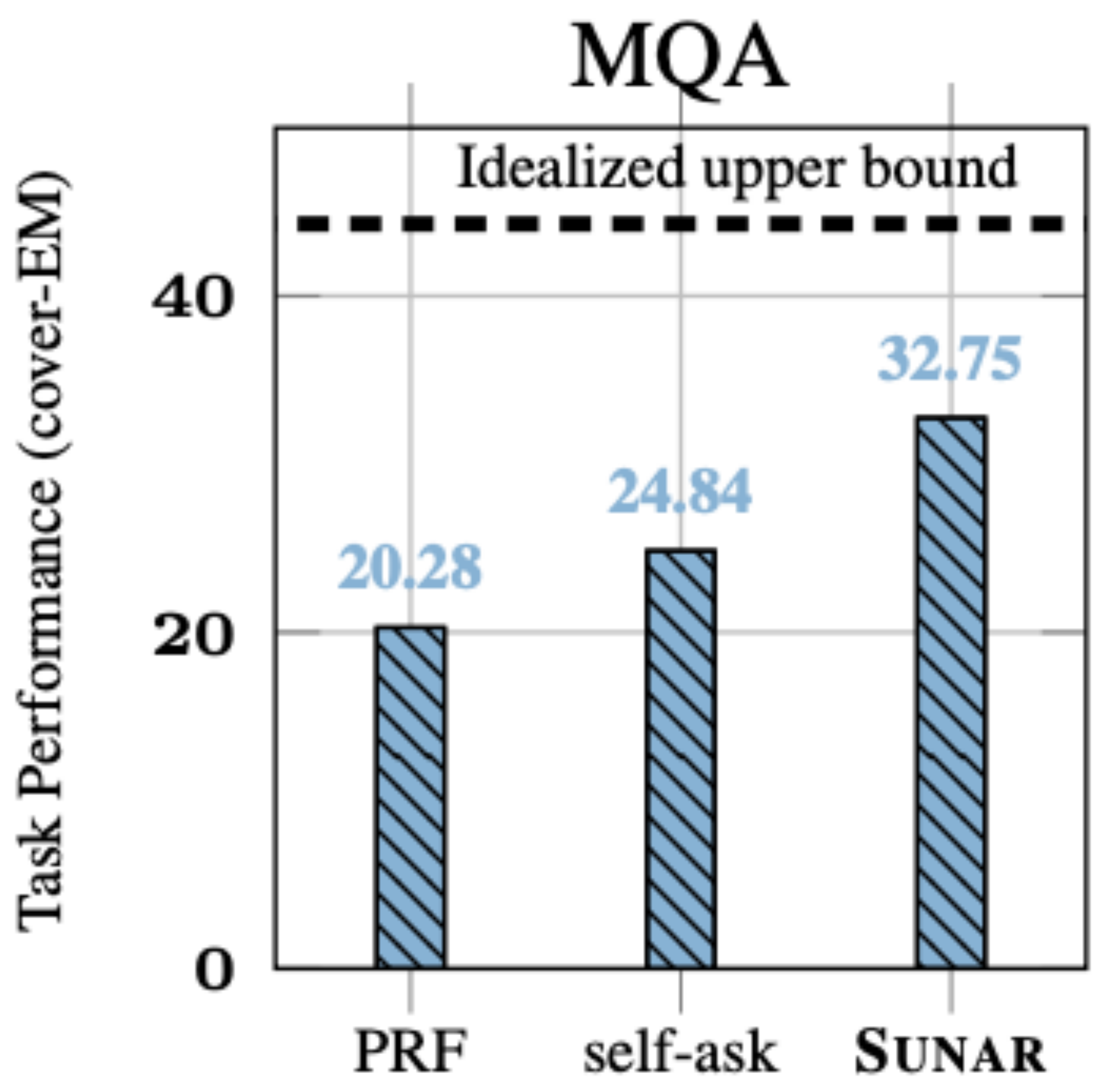
  $B \leftarrow \text{RESCORE}(B, 1/s)$ ▷ Rescore batch
  $R^+ \leftarrow R^+ \cup B$ ▷ Add batch to results

  // Discard Batches
  $R \leftarrow R \setminus B$
  $N \leftarrow N \setminus B$
  $N \leftarrow N \cup (\text{NEIGHBOURS}(B, G) \setminus R^+)$

  //Alternate R and N
  $C \leftarrow \begin{cases} R & \text{if } C = F \\ N & \text{if } C = N \end{cases}$
while $|R^+| < c$

Bridging retrieval gap and downstream reasoning enhancement





Outperforms existing state-of-the-art approaches & LLM agnostic

Method	MQA	WQA
Methods (w/o query understanding)		
ZERO-SHOT-COT (Kojima et al., 2023)	8.62	30.42
FEW-SHOT-COT (Wei et al., 2023)	15.02	32.83
FEW-SHOT-COT +PRF (Li et al., 2022)	16.69	35.55
SUNAR$_R$ (ours)	21.32	40.96
Methods (w/ query understanding)		
Self-RAG (Asai et al., 2024)	17.80	35.25
ReAct (Yao et al., 2023)	21.41	43.25
DecomP (Khot et al., 2023)	21.01	44.08
SearChain (Xu et al., 2024)	21.72	44.42
SELF-ASK +PRF (Li et al., 2022)	20.28	44.58
SELF-ASK (Press et al., 2023)	24.84	45.75
NAR (w/ query understanding) (ours)		
SUNAR$_R$	28.11	47.67
SUNAR	**32.75** †	**54.83**†
Golden Evidence (Ideal Upper Bound)		
FEW-SHOT-COT	44.28	65.55

Method	MQA	WQA
gpt-40-mini		
SELF-ASK	26.76	37.33
SUNAR	**32.19**	**48.16**
Llama 3.1 (8B)		
SELF-ASK	5.43	25.83
SUNAR	**13.82**	**39.52**
Mistral v0.2 (7B)		
SELF-ASK	7.84	27.72
SUNAR	**26.12**	**40.23**

TU Delft

SUNAR helps tackle hallucination and knowledge gaps

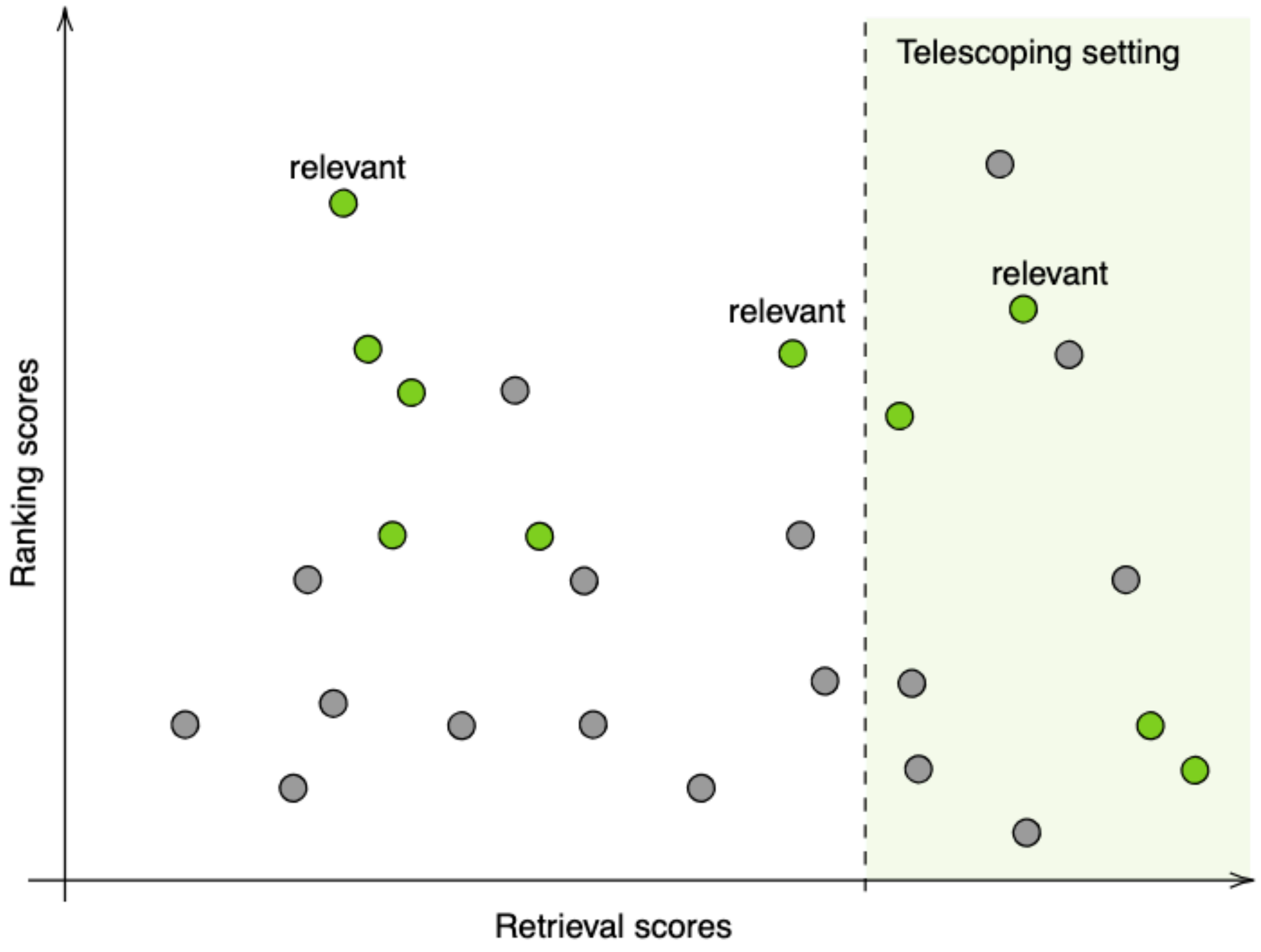
Method	Evidences
Question	**Where was the director of film Ronnie Rocket born?** [Dataset: **WQA**]
SELF-ASK	[Evidence 1]: This is a list of film series by director. [Evidence 2]: This is a list of notable directors in motion picture and television arts. [Final Answer]: Unknown
SUNAR (ours)	[Evidence 1]: Ronnie Rocket is an unfinished film project written by David Lynch, who also intended [. . .]. [Evidence 2]: David Keith Lynch was born in Missoula, Montana, on January 20, 1946. His father [. . .] . [Final Answer]: Missoula, Montana
Question	**Who did the screenwriter for Good Will Hunting play in Dazed and Confused?** [Dataset: **MQA**]
SELF-ASK	[Evidence 1]: Damon begins working alongside his younger brother, Stefan Salvatore, to resist greater[. . .]. [Evidence 2]: Damon Salvatore is a fictional character in The Vampire Diaries. He is portrayed by Ian Somerhalder in the television. [Final Answer]: Damon Salvatore
SUNAR (ours)	[Evidence 1]: Damon and Ben Affleck wrote G̈ood Will Hunting(̈1997), a screenplay[. . .]. [Evidence 2]: Benjamin Affleck- Boldt(born August 15, 1972) is an American actor . He later appeared in the independent coming- of- age comedyD̈azed and Confused as Fred O'Bannion [. . .]" [Final Answer]: Fred O'Bannion

TU Delft

Online Relevance Estimation

TUDelft

Telescoping systems and drawbacks

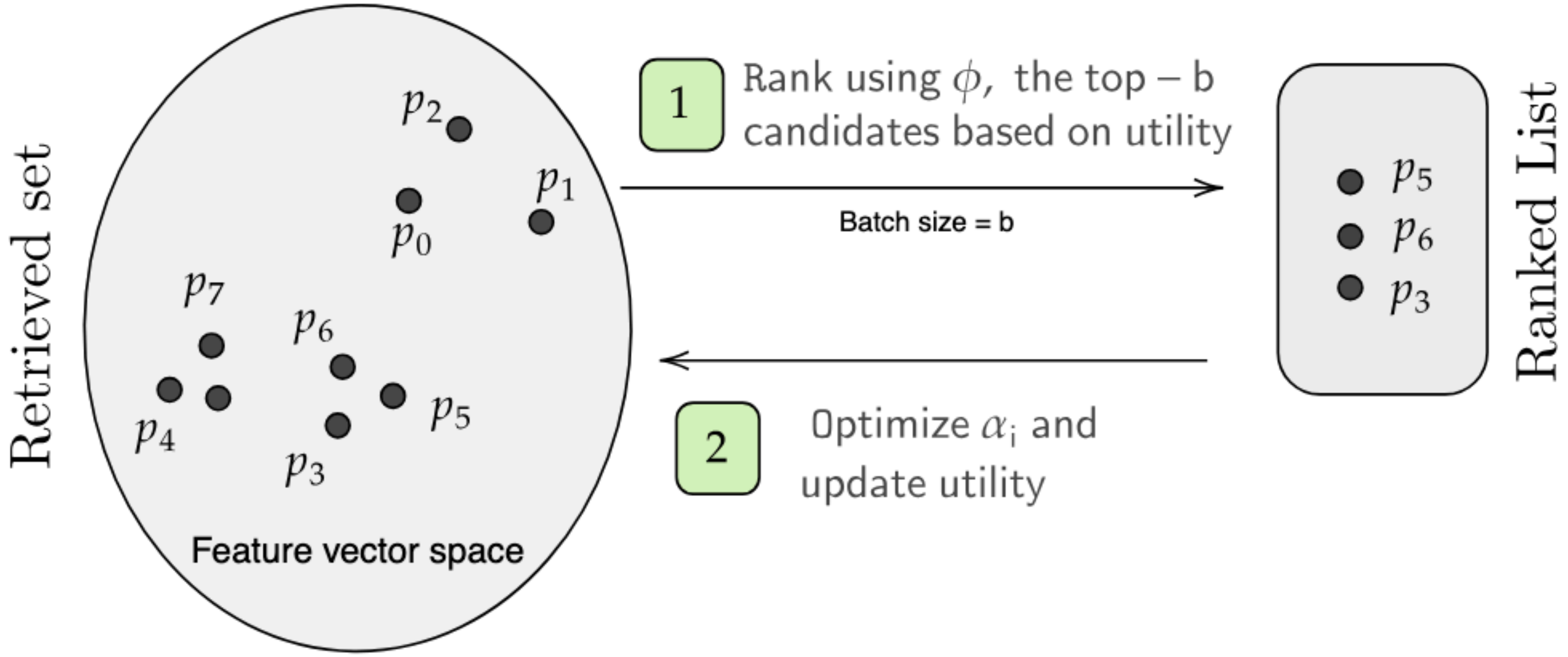


- Telescoping approaches involve progressive filtering of documents through less-precise retrieval methods

- Key is capturing relevant documents with low retrieval scores that current approaches ignore.

TU Delft

Online Relevance Estimation

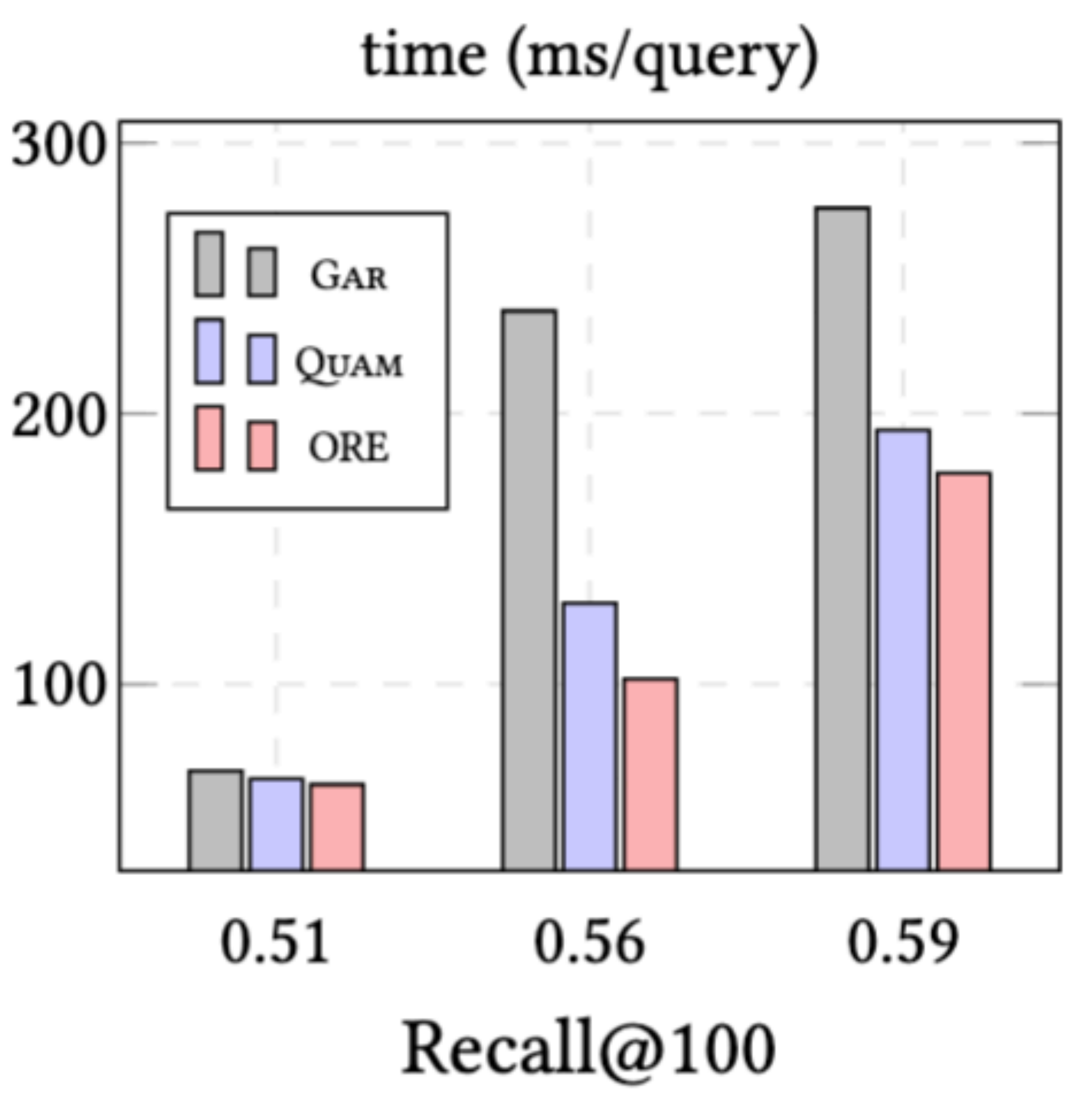


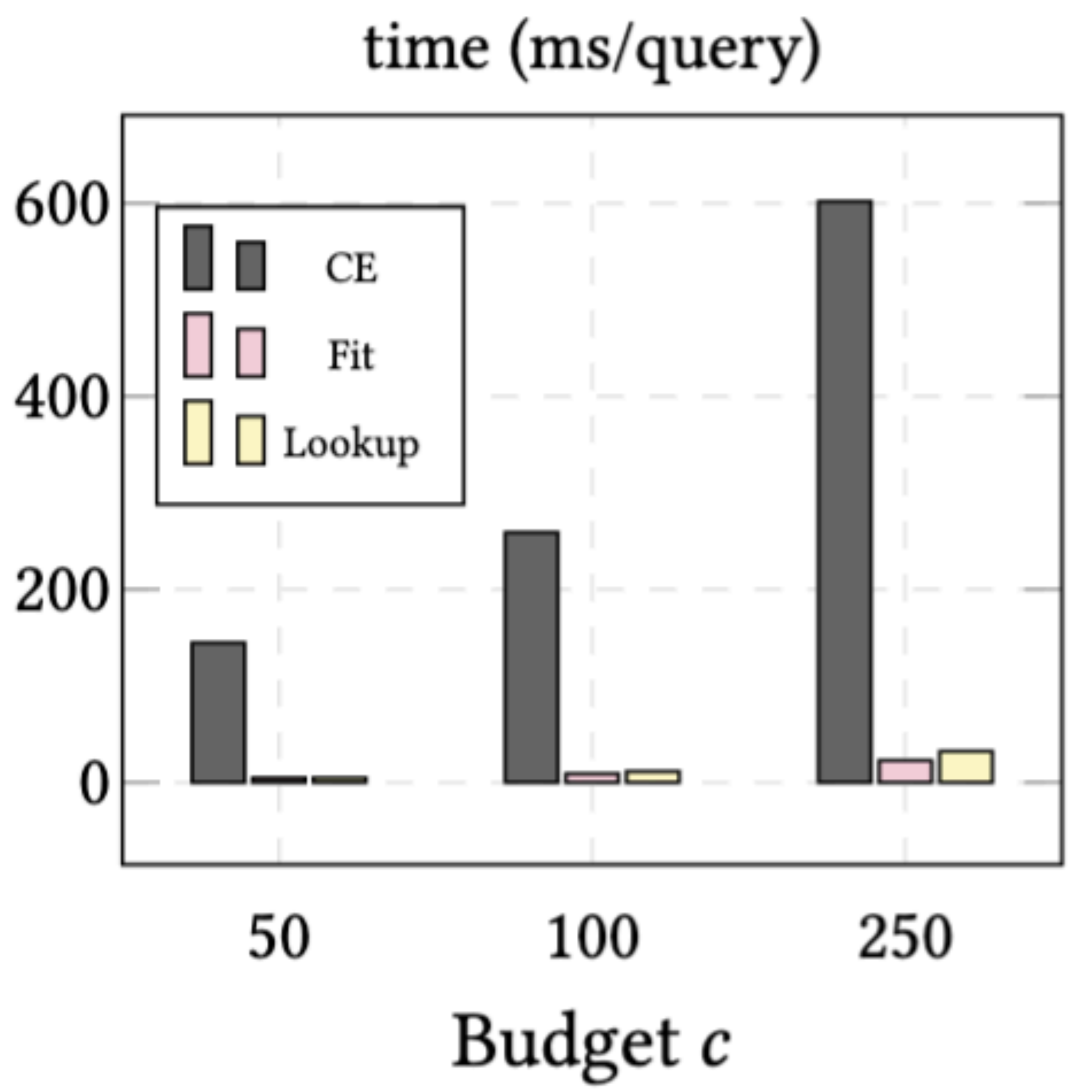
Features are flexible

Feature	Notation	Taxonomy	Source		Description
			Offline	Online	
x_1	$BM25(q, d)$	Q2DAff		✓	Lexical similarity between query and document.
x_2	$TCT(q, d)$	Q2DAff		✓	Semantic similarity between query and document.
x_3	$RM3(q', d)$	D2DAff		✓	Lexical similarity between expanded query using RM3 and document.
x_4	$BM25(d, d')$	D2DAff	✓		Lexical similarity between pair of documents.
x_5	$TCT(d, d')$	D2DAff	✓		Semantic similarity between pair of documents.
x_6	$\textsc{Laff}(d, d')$	D2DAff	✓		Learnt affinity or similarity between pair of documents [34].

Latency and Computational Efficiency

ORE offers 2x-7x speedup over SOTA based on ranker employed

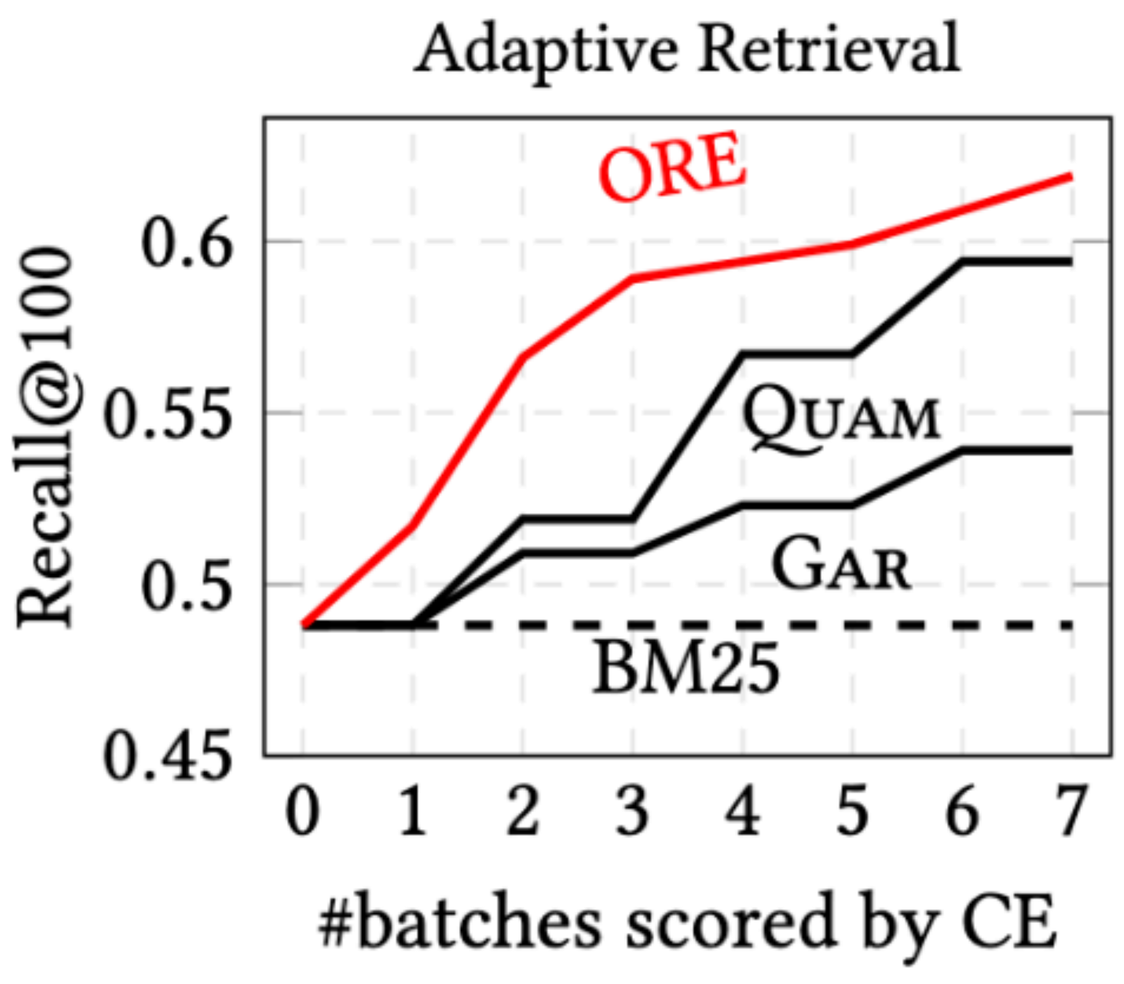




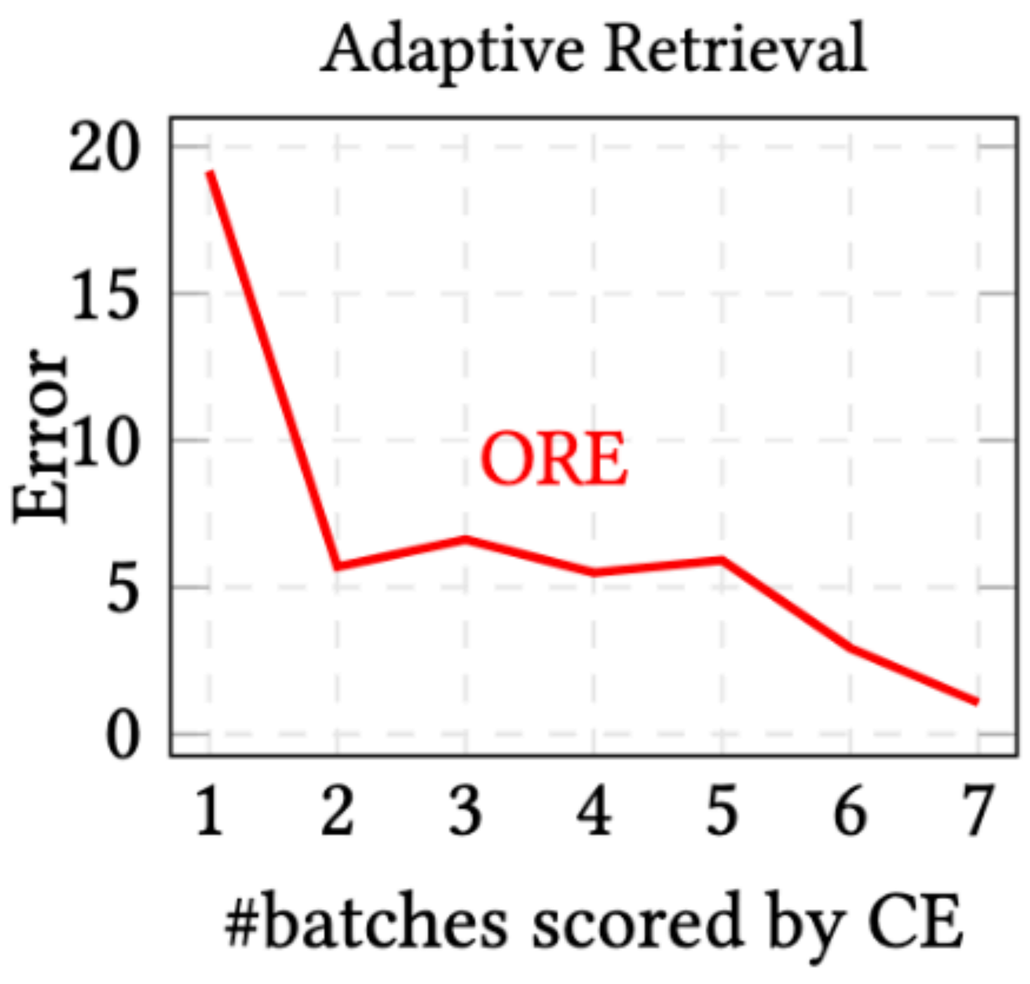
The online estimation component takes **10x** less time than ranker calls

TU Delft

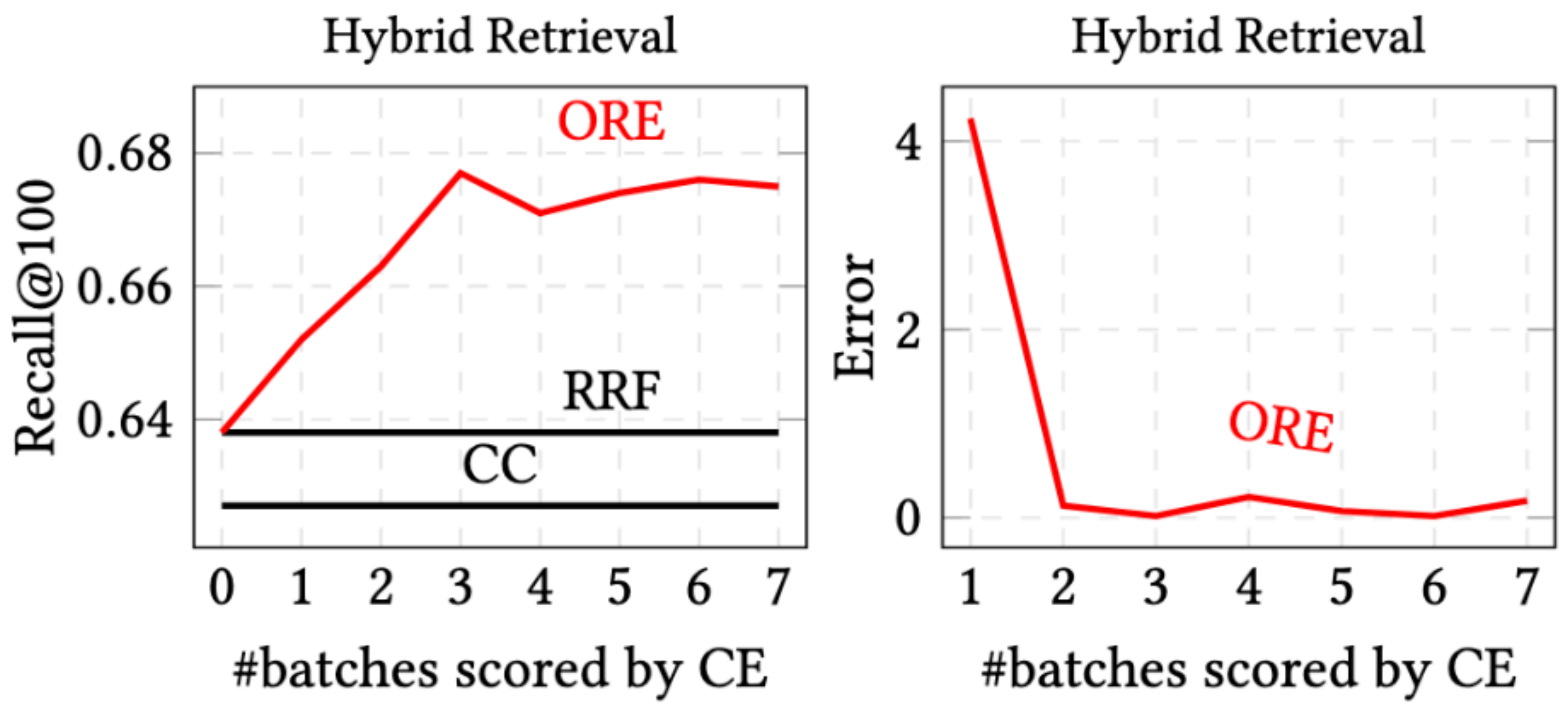
Sample Efficiency of ORE



(a)



(b)



Impressive Performance Gains

Dataset	Pipeline	$c = 50$		$c = 100$	
		nDCG@c	Recall@c	nDCG@c	Recall@c
DL21	**Hybrid**				
	RRF»MonoT5 [R]	0.576	0.401	0.558	0.520
	CC»MonoT5 [C]	0.584	0.419	0.569	0.545
	ORE	^{R}**0.604**	^{R}**0.444**	^{RC}**0.609**	^{RC}**0.609**
	Adaptive				
	BM25»MonoT5 [B]	0.436	0.242	0.433	0.331
	w/ GAR$_{BM25}$ [G]	0.457	0.290	0.465	0.414
	w/ QUAM$_{BM25}$ [Q]	0.478	0.310	**0.499**	0.454
	w/ ORE$_{BM25}$	$^{GQ}_{B}$**0.503**	$^{GQ}_{B}$**0.364**	$_{B}$0.481	$^{G}_{B}$**0.463**
	w/ GAR$_{TCT}$ [G]	0.502	0.331	**0.520**	0.489
	w/ QUAM$_{TCT}$ [Q]	0.491	0.311	0.518	0.477
	w/ ORE$_{TCT}$	$^{GQ}_{B}$**0.532**	$^{GQ}_{B}$**0.406**	$_{B}$0.512	$_{B}$**0.502**
DL22	**Hybrid**				
	RRF»MonoT5 [R]	0.452	0.260	0.430	0.341
	CC»MonoT5 [C]	0.459	0.278	0.433	0.362
	ORE	^{RC}**0.481**	^{R}**0.297**	^{RC}**0.459**	^{RC}**0.389**
	Adaptive				
	BM25»MonoT5 [B]	0.290	0.115	0.275	0.164
	w/ GAR$_{BM25}$ [G]	0.287	0.121	0.290	0.191
	w/ QUAM$_{BM25}$ [Q]	**0.308**	0.135	**0.303**	**0.196**
	w/ ORE$_{BM25}$	**0.292**	**0.137**	0.284	0.195
	w/ GAR$_{TCT}$ [G]	0.329	0.157	**0.348**	0.256
	w/ QUAM$_{TCT}$ [Q]	0.329	0.155	0.334	0.237
	w/ ORE$_{TCT}$	$^{GQ}_{B}$**0.364**	$^{GQ}_{B}$**0.206**	$_{B}$0.342	$_{B}$**0.260**

The Retrieval Gap

The Reasoning Gap

Numerical and Compositional Reasoning

Claim: Repealing the sales tax on boats in Rhode Island has spawned 2,000 companies, 7,000 jobs and close to $2 billion a year in sales activity

How many companies where there before the tax ?

Numerical and Compositional Reasoning

Claim: Repealing the sales tax on boats in Rhode Island has spawned 2,000 companies, 7,000 jobs and close to $2 billion a year in sales activity

How many jobs
where there
before the tax ?

Numerical and Compositional Reasoning

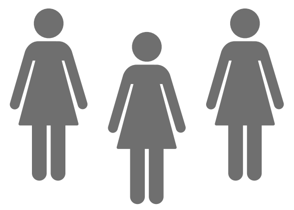
Claim: Repealing the sales tax on boats in Rhode Island has spawned 2,000 companies, 7,000 jobs and close to $2 billion a year in sales activity

What was the annual sales there before the tax ?

Numerical and Compositional Reasoning

Claim: Repealing the sales tax on boats in Rhode Island has spawned 2,000 companies, 7,000 jobs and close to $2 billion a year in sales activity

Need More than Prompting LLMs



Rationales or Explanations

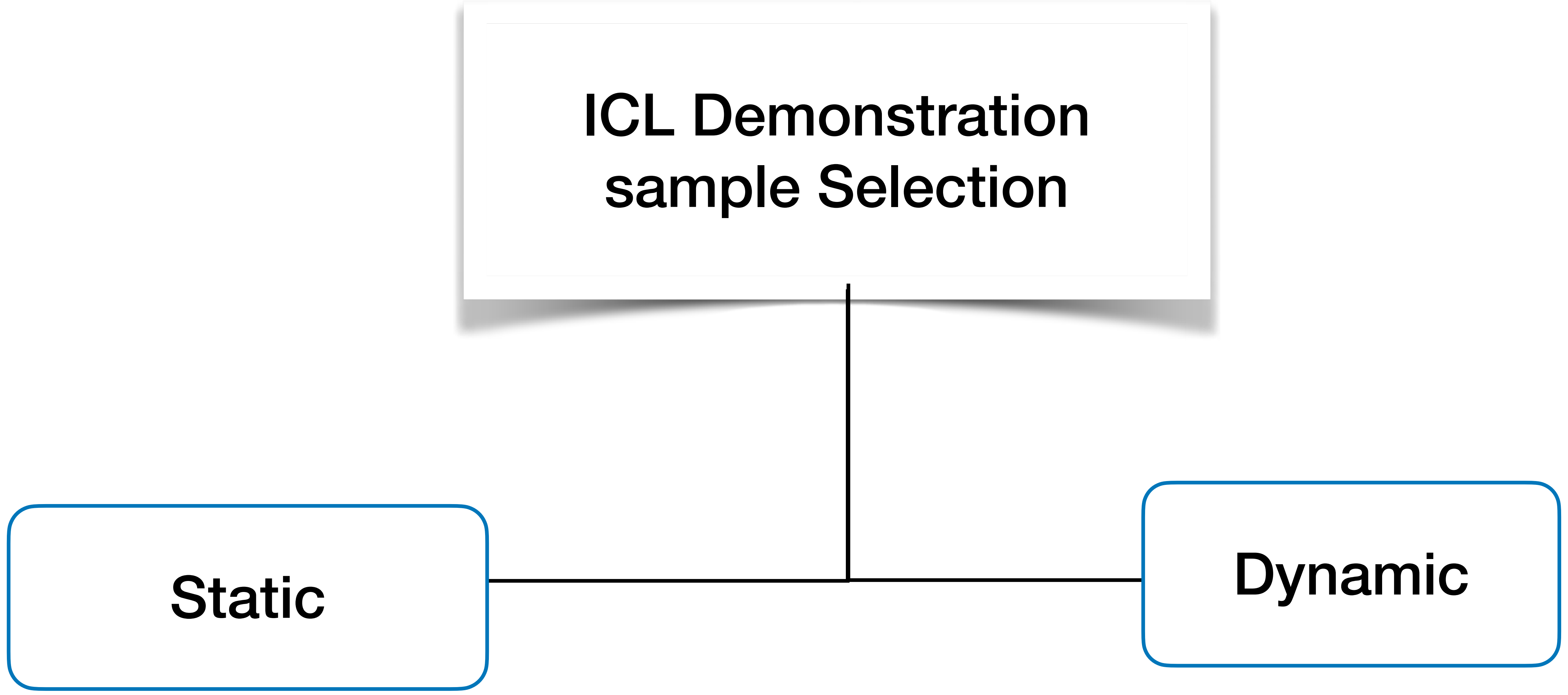
Need to compose abilities required to solve the task

Could be achieved through fine-tuning on required abilities.
Result: Smaller 1M param models outperform larger 1B param models

Or skill composition through In-Context Learning

Vishwanath, Setty & Anand, [SIGIR '24]

Demonstration Samples is all you need ?



FinQA Prompt

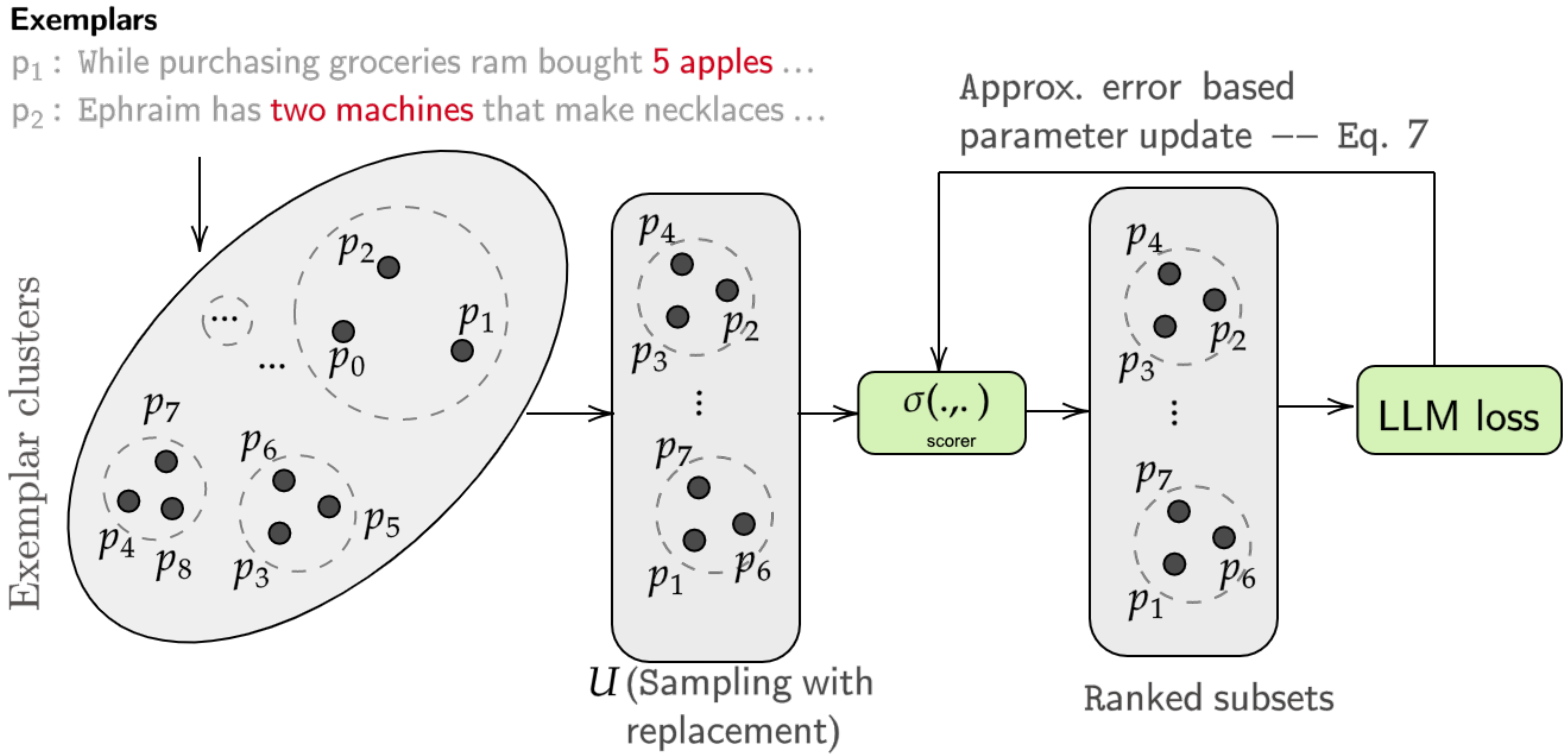
Instruction:`You are a helpful, respectful and honest assistant helping to solve math word problems or tasks requiring reasoning or math, using the information from given table and text.`

Exemplars :

Read the following table, and then answer the question:
[Table]: Year | 2016 | 2015 | 2014 |
share-based compensation expense | 30809 | 21056 | 29793 |
income tax benefit | 9879 | 6907 | 7126 |
[Question]: *how much percent did the income tax benefit increase from 2014 to 2016?*
[Explanation]: $x0 = (9879 - 7126)$,
ans=($x0/7126$)
[Answer]: The answer is increased 38.6%
...
...

Test Input : Read the following table, and then answer the question: Table: Question:
Explanation: [INS] Answer: [INS]

Smart Exploration and Exploitation for ICL Exemplars



Loss Modeling (Approximation) for efficient selection

Subset of k Exemplars ($S \subseteq \mathcal{S}$)

Loss modelling function;
Approximating $L(S, \mathcal{V})$

$$\overbrace{\sigma(\vec{\alpha}, S) = \frac{1}{m} \sum_{j=1}^{m} \sum_{i=1}^{n} \alpha_i (x_i \in S) E_{ij}} \tag{1}$$

ith exemplars contribution, low if important exemplar

Any transformer based encoder

$$E_{ij} = \frac{\phi(x_i)^T \phi(u_j)}{\|\phi(x_i)\| \|\phi(u_j)\|}$$

ith exemplar, $x_i \in S$

ith validation sample, $u_i \in \mathcal{V}$

Efficient Estimation of parameters

Update parameters to reduce approximation error

set of l subsets at timestep t with lowest validation loss

Validation Set

$$\mathcal{L}(\vec{\alpha}; U_t, V_t) = \sum_{S \in U_t} (L(S, \mathcal{V}) - \sigma(\vec{\alpha}, S))^2 + \sum_{S' \sim V_t} (L(S', \mathcal{V}) - \sigma(\vec{\alpha}, S'))^2$$

Remaining subsets; $V_t \leftarrow \mathcal{U} \setminus U_t$, where $\mathcal{U} \subset \mathcal{S}$

Negative samples from high-loss set V_t

Estimating loss here involves LLM calls and equivalent to arm pulling

TU Delft

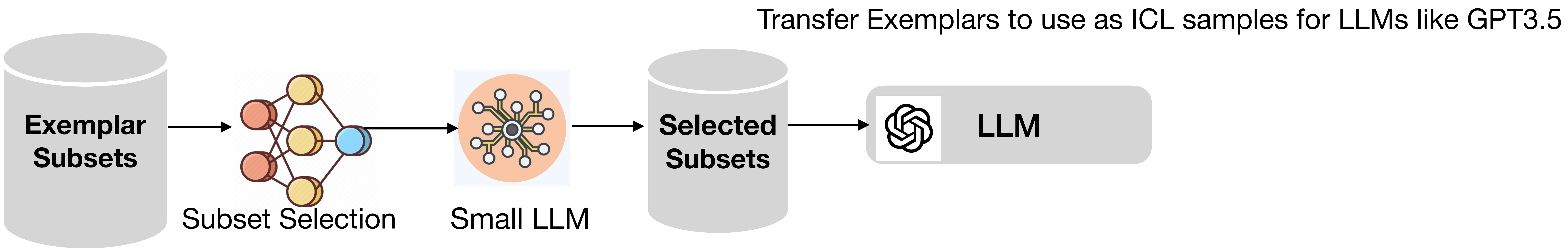
Summary

Algorithm 1: EXPLORA

Input: $\mathcal{U} \subseteq \mathcal{S}$: $\triangleright$ Initial exemplar subsets
Initialize: $U_0 \leftarrow$ set of random l subsets from $\mathcal{U}$
  $t \leftarrow 0$
  $\vec{\alpha} \leftarrow \mathcal{N}(0, 1)$ $\triangleright$ Sampling from a gaussian
while $t < T$ **do**
  Let $V_t \leftarrow \mathcal{U} \setminus U_t$
  $\vec{\alpha_t} \leftarrow \min_{\vec{\alpha}} \mathcal{L}(\vec{\alpha}, U_t, V_t)$ $\triangleright$ Eq. in previous slide
  $S_t^* = \arg\min_{S \in V_t} \sigma(\vec{\alpha_t}, S)$ $\triangleright$ Lowest loss subset
  $\tilde{S}_t = \arg\max_{S \in U_t} \sigma(\vec{\alpha_t}, S)$ $\triangleright$ Highest loss subset
  if $\sigma(\vec{\alpha_t}, S_t^*) < \sigma(\vec{\alpha_t}, \tilde{S}_t)$ **then**
    $U_t \leftarrow U_t \setminus \{\tilde{S}_t\}$ $\triangleright$ Remove $\tilde{S}_t$
    $U_{t+1} \leftarrow U_t \cup \{S_t^*\}$ $\triangleright$ add S_t^*
  end
  $t \leftarrow t + 1$
end
Output: U_T;Set of l subsets from $\mathcal{U}$ which have the lowest validation loss

TU Delft

Tune and Transfer

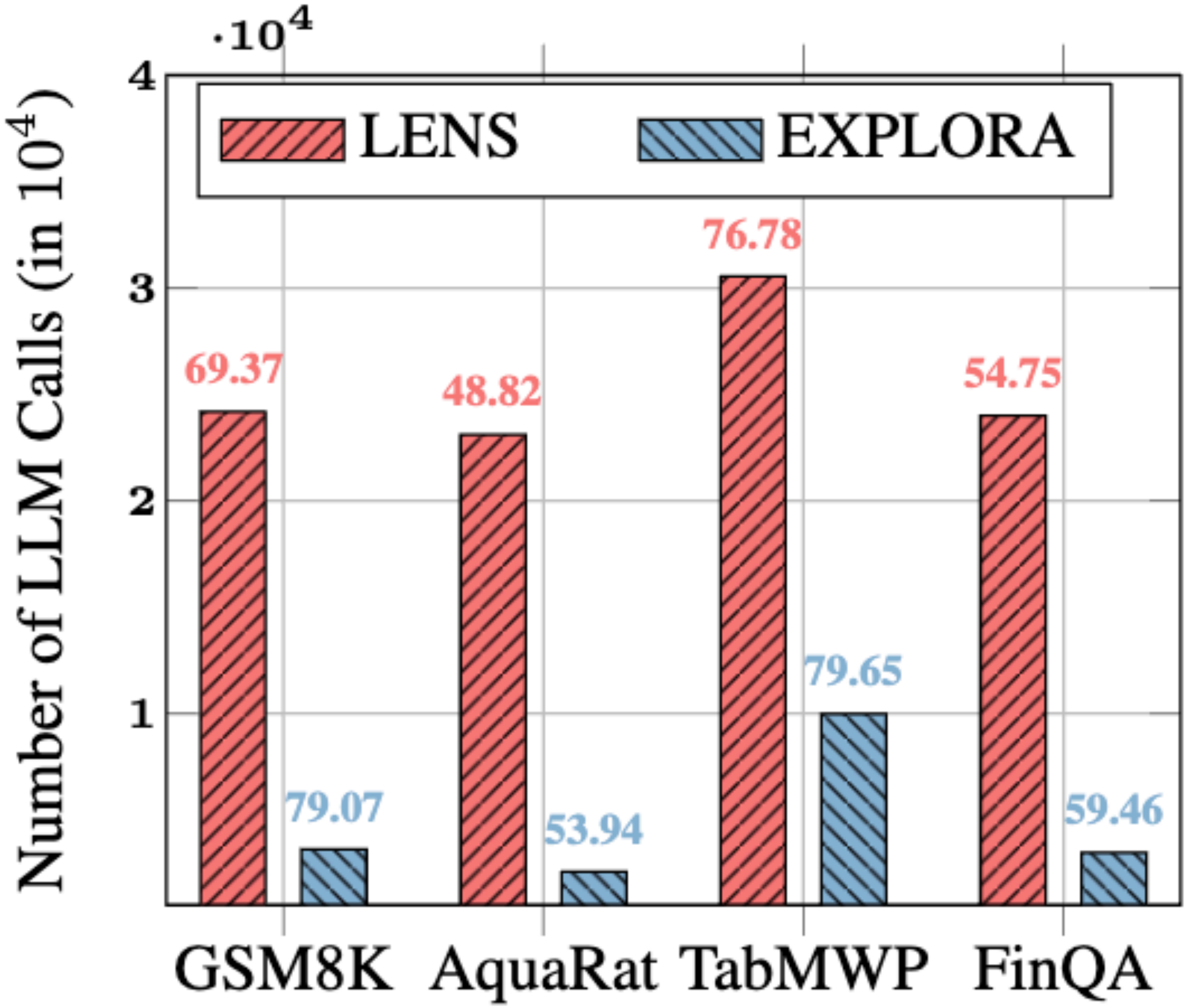


EXPLORA is Robust (Low Variance across test samples)

Datasets	**GSM**	**Aqua**	**Tab**	**Fin**
Zero-Shot COT	±5.18	±7.08	±1.84	±4.50
Few-Shot COT	±4.48	±12.03	±1.66	±4.76
KNN	±3.76	±5.49	±1.27	±4.17
MMR	±4.00	±10.53	±1.68	±6.10
Graph Cut	±6.38	±8.18	±2.03	±5.29
Facility Location	±4.23	±6.71	±1.74	±4.94
LENS	±5.04	±6.67	±1.72	±5.81
EXPLORA	±**3.39**	±**4.93**	±1.45	±**3.41**

TU Delft

EXPLORA is Resource Efficient



Results Transfers Well (L for Llama and M for Mistral)

Method	T	GSM	Aqua	Tab	Fin
EXP	L	79.07	53.94	79.65	54.66
	M	77.86	53.54	77.41	59.46
EXP+SC	L	85.82	63.78	86.76	61.16
	M	86.35	63.39	85.52	64.52
EXP+KNN+SC	L	85.89	64.17	85.74	63.64
	M	85.14	62.20	86.29	65.12
EXP+MMR+SC	L	86.20	62.99	87.81	64.60
	M	86.13	63.78	86.96	64.60

Prompts are transferred from Llama or Mistral to GPT3.5-turbo

TUDelft

A Recap

- Efficiency and Effectiveness are critical for practical robust RAG pipelines.
- Telescoping systems are limited in efficiency and suffer from Recall Boundedness.
- LLMs are still limited in reasoning.
- Test Time scaling for Retrieval is central to robust pipelines for complex knowledge intensive tasks.
- Careful selection of exemplars help in transferring abilities to LLMs through ICL.

TU Delft

Conclusion - Research Vision

- End-End Test Time Reasoning (TTR) has huge scope.
  - How do we incorporate Reasoning feedback (LLM) to improve retrieval
  - How can retrieval improve reasoning.
  - How to do this efficiently?

- How can we do this in a scalable manner ?

TUDelft