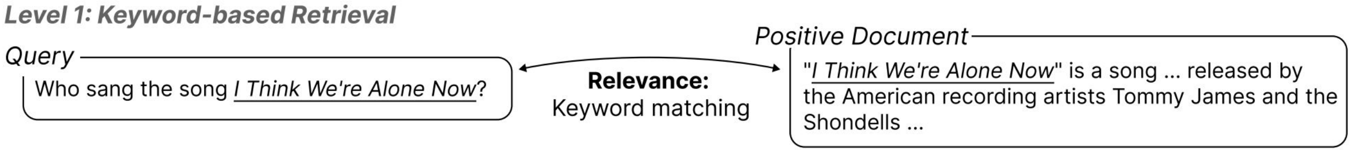
BRIGHT: A Realistic and Challenging Benchmark for Reasoning-Intensive Retrieval

Hongjin Su*[h] Howard Yen*[p] Mengzhou Xia*[p] Weijia Shi [w] Niklas Muennighoff [s]
Han-yu Wang [h] Haisu Liu [h] Quan Shi [p] Zachary S. Siegel [p] Michael Tang [p]
Ruoxi Sun [g] Jinsung Yoon [g] Sercan Ö. Arık [g] Danqi Chen [p] Tao Yu [h]

[h] The University of Hong Kong [p] Princeton University [s] Stanford University
[w] University of Washington [g] Google Cloud AI Research

Prior work: keyword-based retrieval



Level 1: Keyword-based Retrieval

Query

Who sang the song *I Think We're Alone Now*?

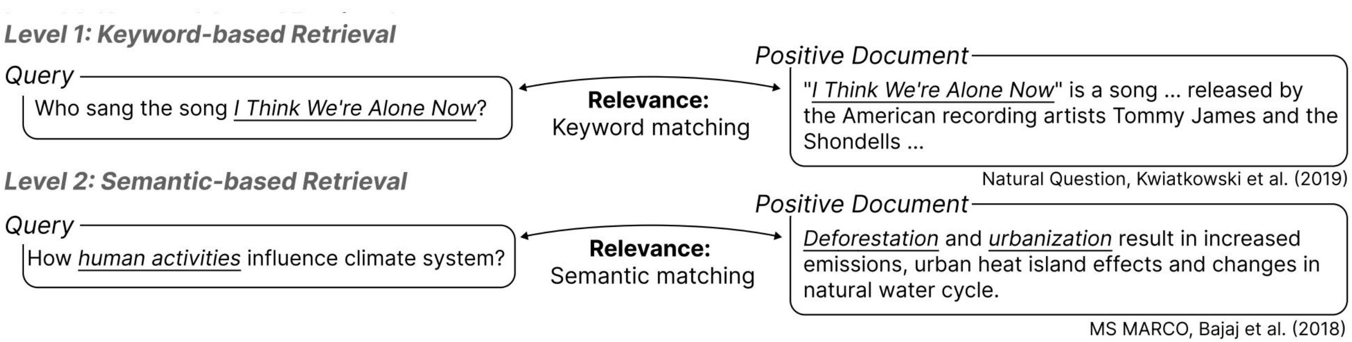
Relevance:
Keyword matching

Positive Document

"*I Think We're Alone Now*" is a song ... released by the American recording artists Tommy James and the Shondells ...

Natural Question, Kwiatkowski et al. (2019)

Prior work: keyword-based retrieval



Level 1: Keyword-based Retrieval

Query

Who sang the song *I Think We're Alone Now*?

Relevance:
Keyword matching

Positive Document

"*I Think We're Alone Now*" is a song ... released by the American recording artists Tommy James and the Shondells ...

Natural Question, Kwiatkowski et al. (2019)

Level 2: Semantic-based Retrieval

Query

How *human activities* influence climate system?

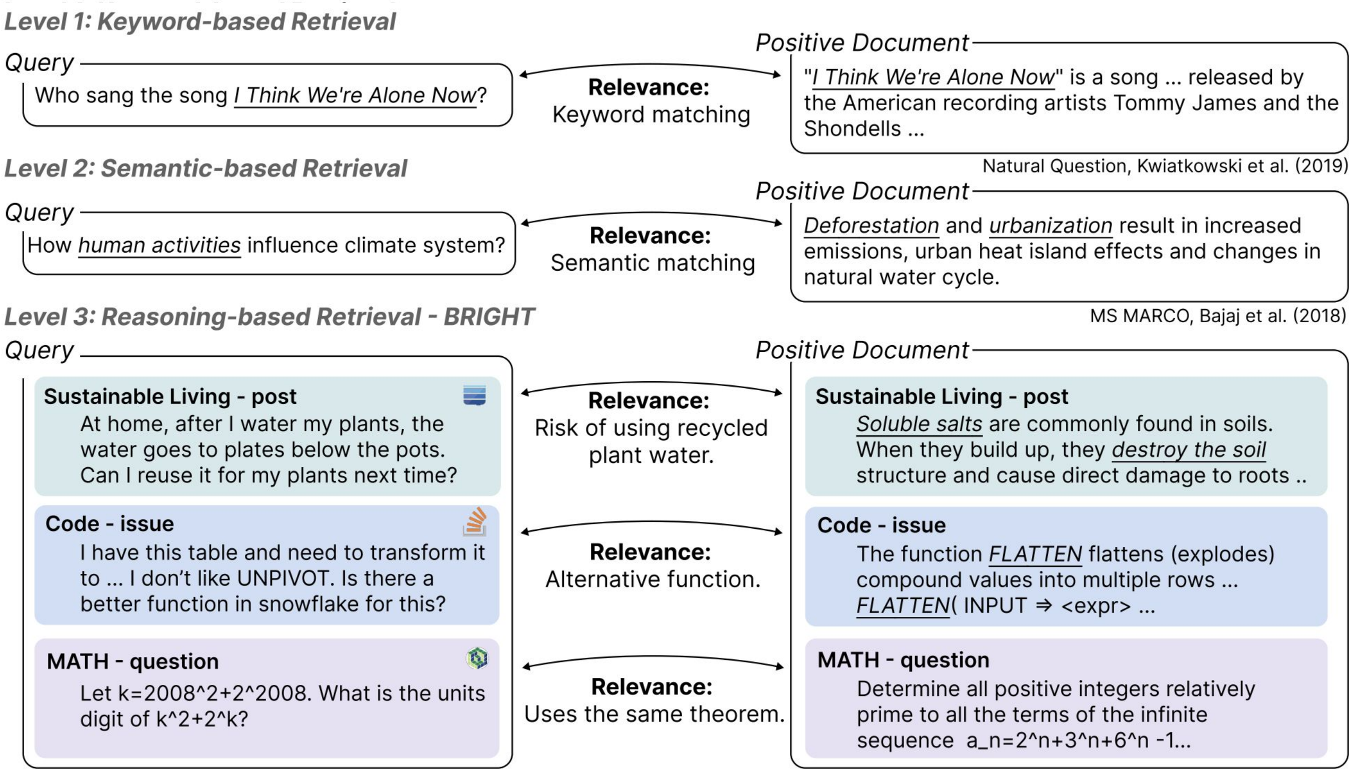
Relevance:
Semantic matching

Positive Document

Deforestation and *urbanization* result in increased emissions, urban heat island effects and changes in natural water cycle.

MS MARCO, Bajaj et al. (2018)

Our focus: reasoning-based retrieval



Level 1: Keyword-based Retrieval

Query

Who sang the song *I Think We're Alone Now*?

Relevance:
Keyword matching

Positive Document

"*I Think We're Alone Now*" is a song ... released by the American recording artists Tommy James and the Shondells ...

Natural Question, Kwiatkowski et al. (2019)

Level 2: Semantic-based Retrieval

Query

How *human activities* influence climate system?

Relevance:
Semantic matching

Positive Document

Deforestation and *urbanization* result in increased emissions, urban heat island effects and changes in natural water cycle.

MS MARCO, Bajaj et al. (2018)

Level 3: Reasoning-based Retrieval - BRIGHT

Query

Sustainable Living - post
At home, after I water my plants, the water goes to plates below the pots. Can I reuse it for my plants next time?

Code - issue
I have this table and need to transform it to ... I don't like UNPIVOT. Is there a better function in snowflake for this?

MATH - question
Let k=2008^2+2^2008. What is the units digit of k^2+2^k?

Relevance:
Risk of using recycled plant water.

Relevance:
Alternative function.

Relevance:
Uses the same theorem.

Positive Document

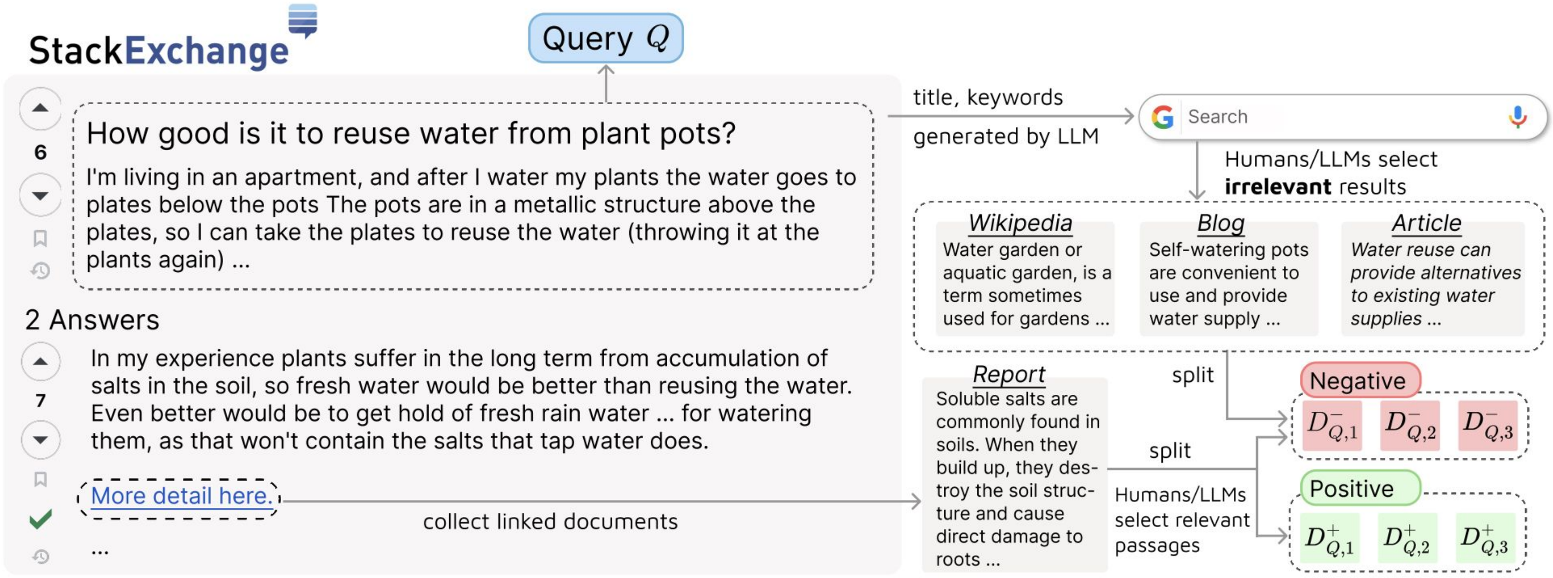
Sustainable Living - post
Soluble salts are commonly found in soils. When they build up, they *destroy the soil* structure and cause direct damage to roots ..

Code - issue
The function *FLATTEN* flattens (explodes) compound values into multiple rows ...
FLATTEN(INPUT ⇒ <expr> ...

MATH - question
Determine all positive integers relatively prime to all the terms of the infinite sequence a_n=2^n+3^n+6^n -1...

Data collection - StackExchange

Relevance: A document is considered relevant to a query only if it is cited in an accepted or highly voted answer and unanimously confirmed by annotators and domain experts that it helps reason through the query with critical concepts or theories.



How good is it to reuse water from plant pots?

I'm living in an apartment, and after I water my plants the water goes to plates below the pots The pots are in a metallic structure above the plates, so I can take the plates to reuse the water (throwing it at the plants again) ...

2 Answers

In my experience plants suffer in the long term from accumulation of salts in the soil, so fresh water would be better than reusing the water. Even better would be to get hold of fresh rain water ... for watering them, as that won't contain the salts that tap water does.

More detail here.

...

collect linked documents

title, keywords generated by LLM

Humans/LLMs select **irrelevant** results

Wikipedia
Water garden or aquatic garden, is a term sometimes used for gardens ...

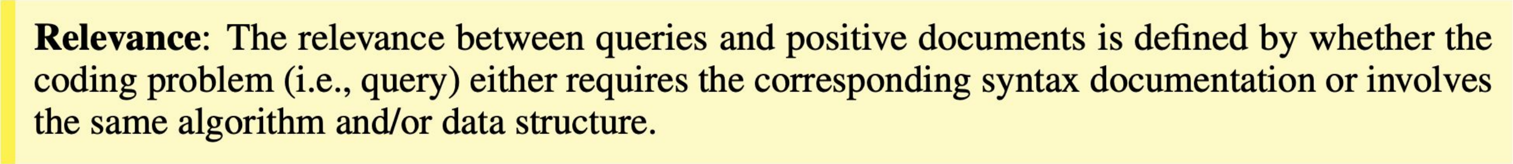
Blog
Self-watering pots are convenient to use and provide water supply ...

Article
Water reuse can provide alternatives to existing water supplies ...

Report
Soluble salts are commonly found in soils. When they build up, they destroy the soil structure and cause direct damage to roots ...

Humans/LLMs select relevant passages

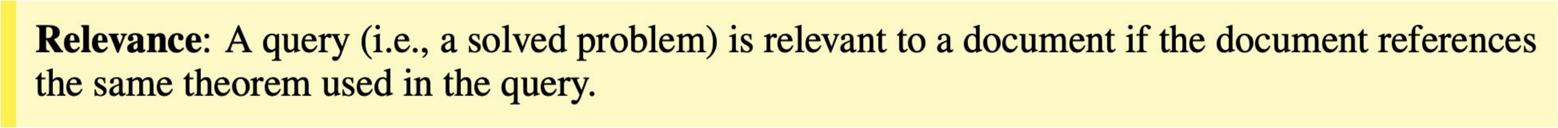
Data collection - Coding



Relevance: The relevance between queries and positive documents is defined by whether the coding problem (i.e., query) either requires the corresponding syntax documentation or involves the same algorithm and/or data structure.

- Pony: Coding problems as queries, required syntax documentation as positive documents
- Leetcode: Coding problems as queries, solved problems using the same algorithm as positive documents

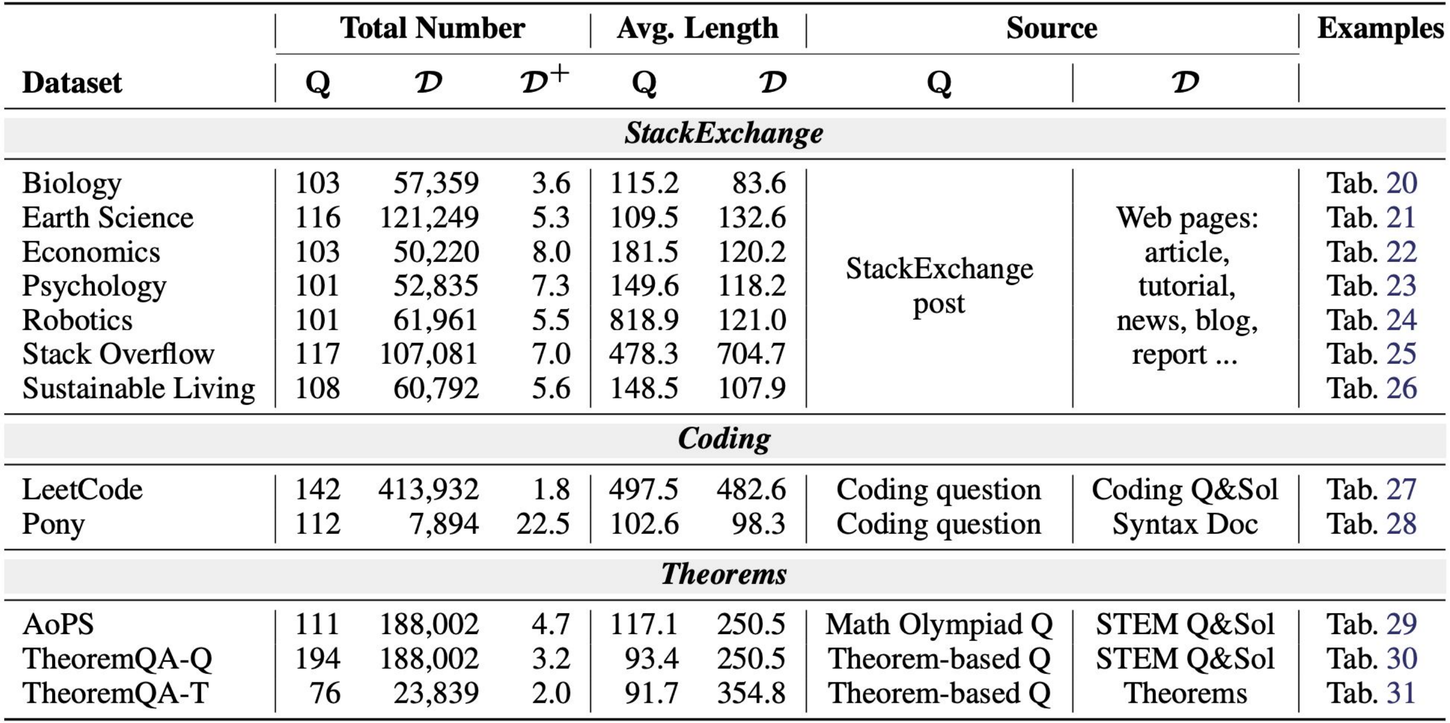
Data collection - Science



Relevance: A query (i.e., a solved problem) is relevant to a document if the document references the same theorem used in the query.

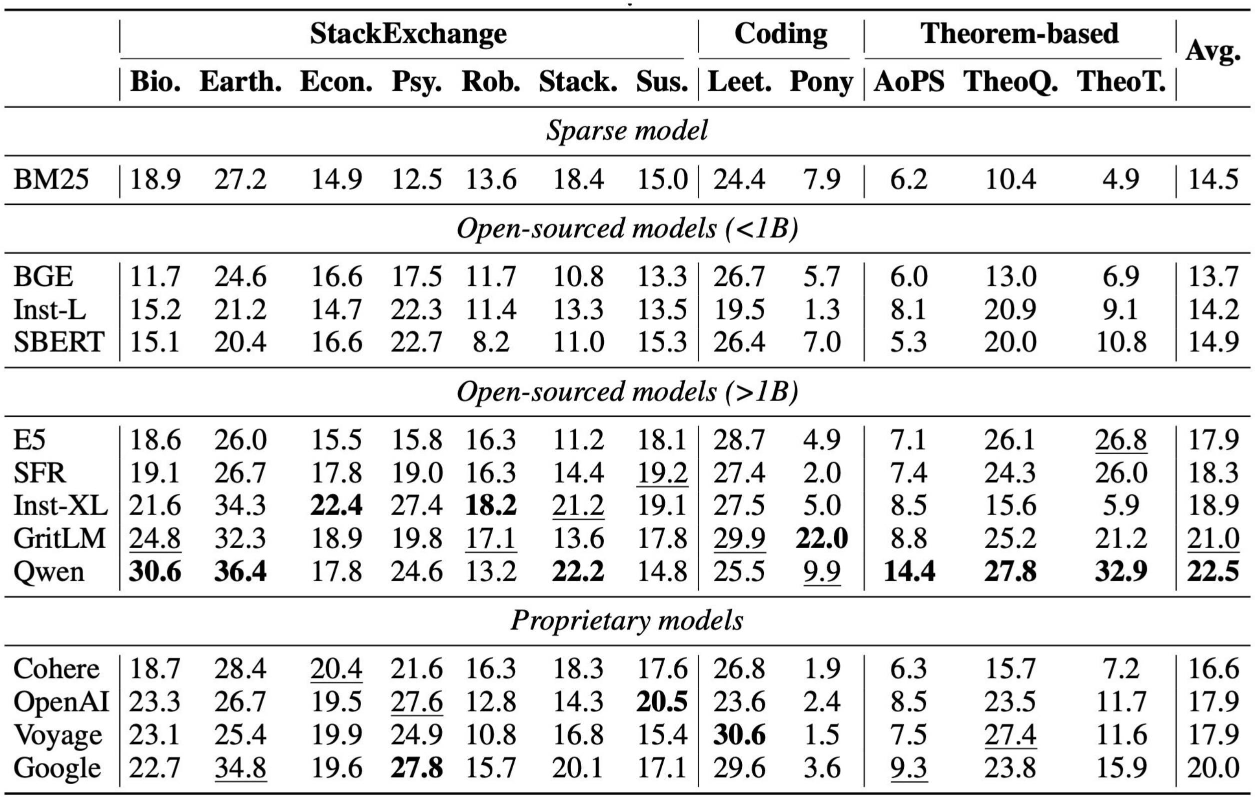
- TheoremQA: Scientific questions as queries, required theorems or solved problems using the same theorem as positive documents
- AoPS: Olympic math problems as queries, solved problems using the same technique as positive documents

Data statistics



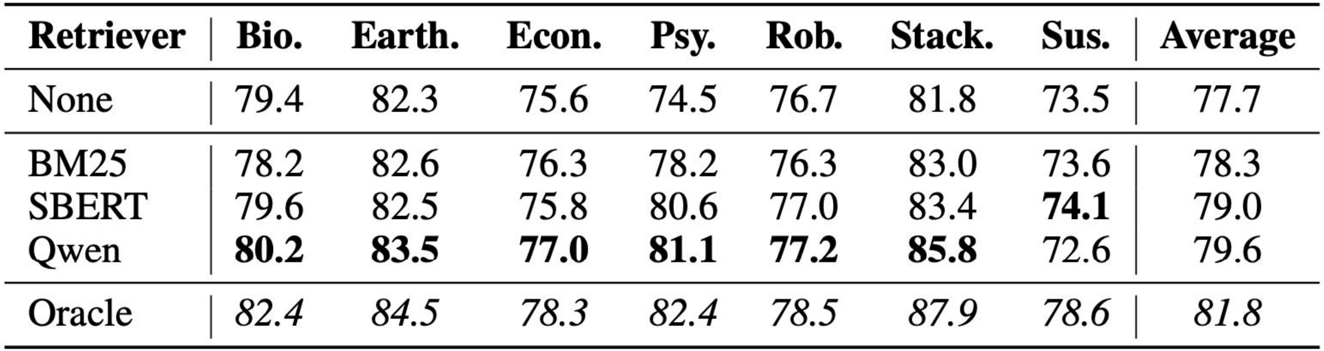
	Total Number			Avg. Length		Source		Examples
Dataset	**Q**	$\mathcal{D}$	$\mathcal{D}^+$	**Q**	$\mathcal{D}$	**Q**	$\mathcal{D}$	
StackExchange								
Biology	103	57,359	3.6	115.2	83.6	StackExchange post	Web pages: article, tutorial, news, blog, report ...	Tab. 20
Earth Science	116	121,249	5.3	109.5	132.6			Tab. 21
Economics	103	50,220	8.0	181.5	120.2			Tab. 22
Psychology	101	52,835	7.3	149.6	118.2			Tab. 23
Robotics	101	61,961	5.5	818.9	121.0			Tab. 24
Stack Overflow	117	107,081	7.0	478.3	704.7			Tab. 25
Sustainable Living	108	60,792	5.6	148.5	107.9			Tab. 26
Coding								
LeetCode	142	413,932	1.8	497.5	482.6	Coding question	Coding Q&Sol	Tab. 27
Pony	112	7,894	22.5	102.6	98.3	Coding question	Syntax Doc	Tab. 28
Theorems								
AoPS	111	188,002	4.7	117.1	250.5	Math Olympiad Q	STEM Q&Sol	Tab. 29
TheoremQA-Q	194	188,002	3.2	93.4	250.5	Theorem-based Q	STEM Q&Sol	Tab. 30
TheoremQA-T	76	23,839	2.0	91.7	354.8	Theorem-based Q	Theorems	Tab. 31

Main results



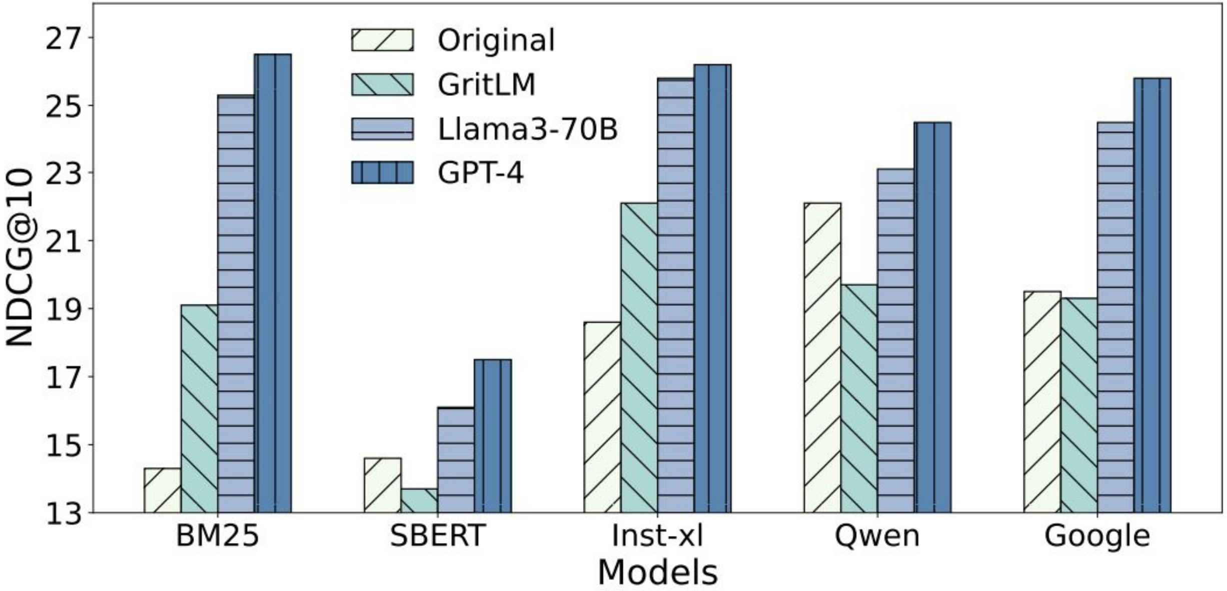
	StackExchange							Coding		Theorem-based			Avg.
	Bio.	Earth.	Econ.	Psy.	Rob.	Stack.	Sus.	Leet.	Pony	AoPS	TheoQ.	TheoT.	
Sparse model													
BM25	18.9	27.2	14.9	12.5	13.6	18.4	15.0	24.4	7.9	6.2	10.4	4.9	14.5
Open-sourced models (<1B)													
BGE	11.7	24.6	16.6	17.5	11.7	10.8	13.3	26.7	5.7	6.0	13.0	6.9	13.7
Inst-L	15.2	21.2	14.7	22.3	11.4	13.3	13.5	19.5	1.3	8.1	20.9	9.1	14.2
SBERT	15.1	20.4	16.6	22.7	8.2	11.0	15.3	26.4	7.0	5.3	20.0	10.8	14.9
Open-sourced models (>1B)													
E5	18.6	26.0	15.5	15.8	16.3	11.2	18.1	28.7	4.9	7.1	26.1	26.8	17.9
SFR	19.1	26.7	17.8	19.0	16.3	14.4	19.2	27.4	2.0	7.4	24.3	26.0	18.3
Inst-XL	21.6	34.3	**22.4**	27.4	**18.2**	21.2	19.1	27.5	5.0	8.5	15.6	5.9	18.9
GritLM	24.8	32.3	18.9	19.8	17.1	13.6	17.8	29.9	**22.0**	8.8	25.2	21.2	21.0
Qwen	**30.6**	**36.4**	17.8	24.6	13.2	**22.2**	14.8	25.5	9.9	**14.4**	**27.8**	**32.9**	**22.5**
Proprietary models													
Cohere	18.7	28.4	20.4	21.6	16.3	18.3	17.6	26.8	1.9	6.3	15.7	7.2	16.6
OpenAI	23.3	26.7	19.5	27.6	12.8	14.3	**20.5**	23.6	2.4	8.5	23.5	11.7	17.9
Voyage	23.1	25.4	19.9	24.9	10.8	16.8	15.4	**30.6**	1.5	7.5	27.4	11.6	17.9
Google	22.7	34.8	19.6	**27.8**	15.7	20.1	17.1	29.6	3.6	9.3	23.8	15.9	20.0

QA results

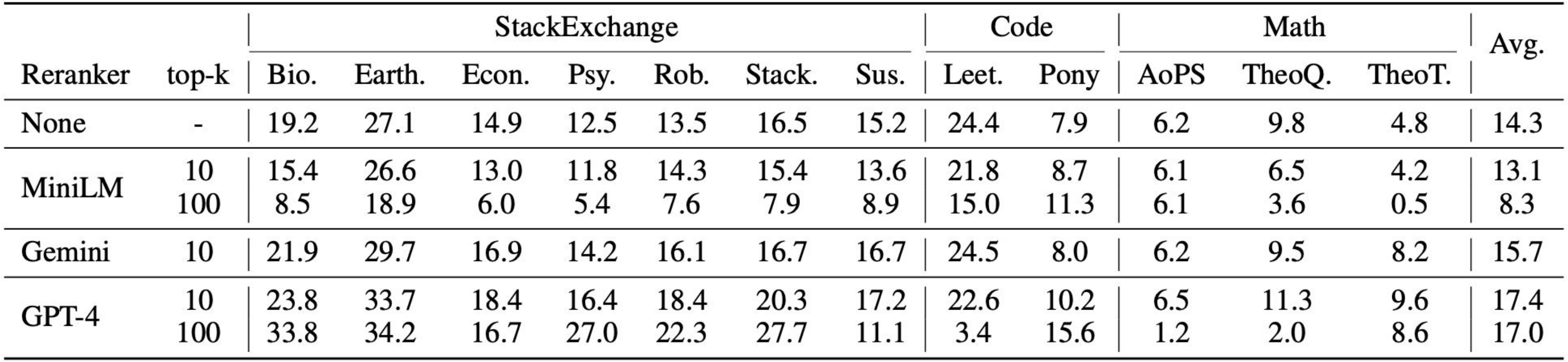


Retriever	**Bio.**	**Earth.**	**Econ.**	**Psy.**	**Rob.**	**Stack.**	**Sus.**	**Average**
None	79.4	82.3	75.6	74.5	76.7	81.8	73.5	77.7
BM25	78.2	82.6	76.3	78.2	76.3	83.0	73.6	78.3
SBERT	79.6	82.5	75.8	80.6	77.0	83.4	**74.1**	79.0
Qwen	**80.2**	**83.5**	**77.0**	**81.1**	**77.2**	**85.8**	72.6	79.6
Oracle	*82.4*	*84.5*	*78.3*	*82.4*	*78.5*	*87.9*	*78.6*	*81.8*

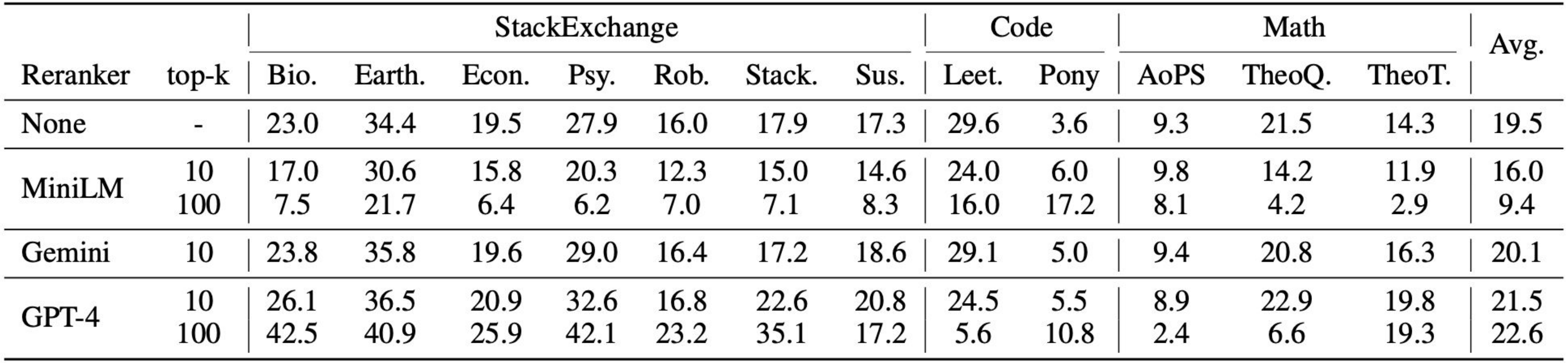
LLM reasoning



LLM reranking

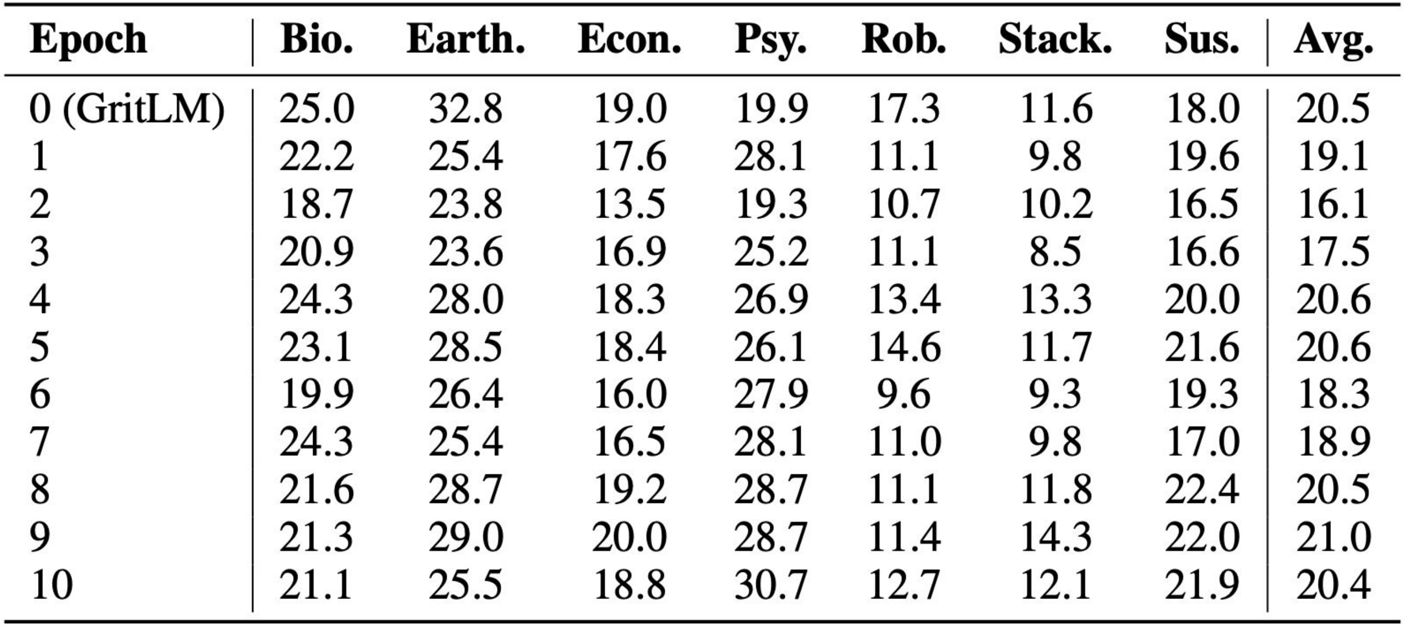


		StackExchange							Code		Math			Avg.
Reranker	top-k	Bio.	Earth.	Econ.	Psy.	Rob.	Stack.	Sus.	Leet.	Pony	AoPS	TheoQ.	TheoT.	
None	-	19.2	27.1	14.9	12.5	13.5	16.5	15.2	24.4	7.9	6.2	9.8	4.8	14.3
MiniLM	10	15.4	26.6	13.0	11.8	14.3	15.4	13.6	21.8	8.7	6.1	6.5	4.2	13.1
	100	8.5	18.9	6.0	5.4	7.6	7.9	8.9	15.0	11.3	6.1	3.6	0.5	8.3
Gemini	10	21.9	29.7	16.9	14.2	16.1	16.7	16.7	24.5	8.0	6.2	9.5	8.2	15.7
GPT-4	10	23.8	33.7	18.4	16.4	18.4	20.3	17.2	22.6	10.2	6.5	11.3	9.6	17.4
	100	33.8	34.2	16.7	27.0	22.3	27.7	11.1	3.4	15.6	1.2	2.0	8.6	17.0



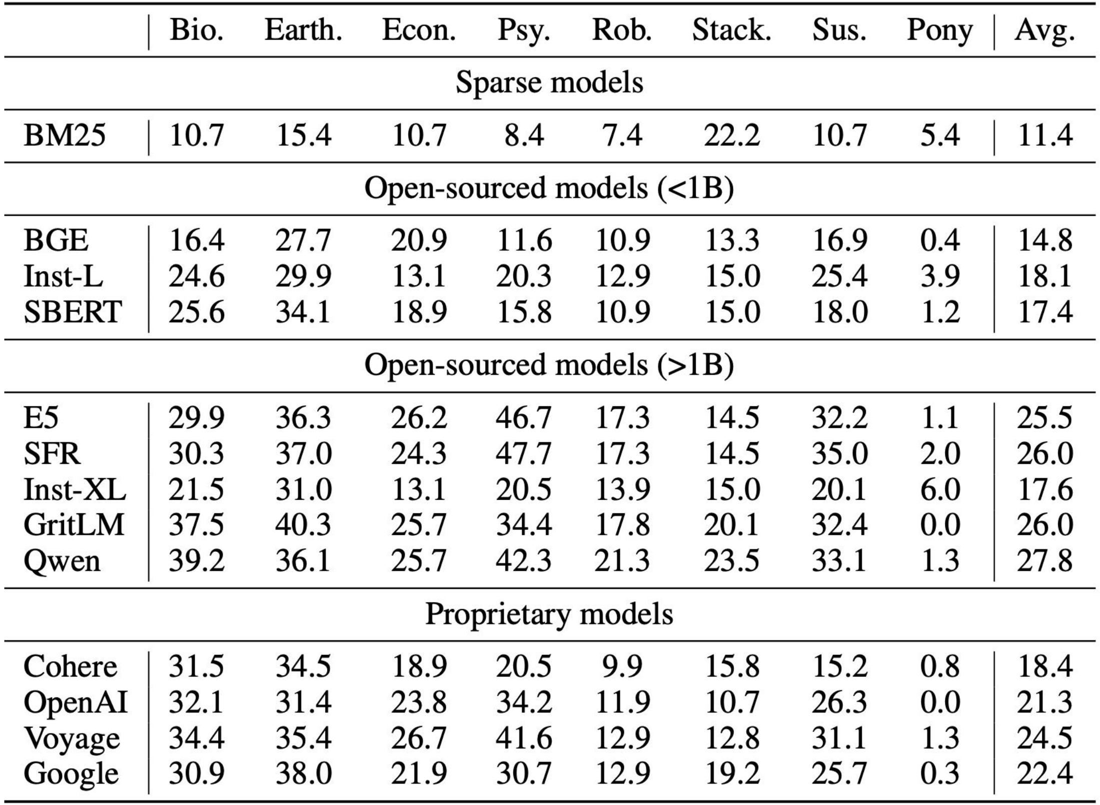
		StackExchange							Code		Math			Avg.
Reranker	top-k	Bio.	Earth.	Econ.	Psy.	Rob.	Stack.	Sus.	Leet.	Pony	AoPS	TheoQ.	TheoT.	
None	-	23.0	34.4	19.5	27.9	16.0	17.9	17.3	29.6	3.6	9.3	21.5	14.3	19.5
MiniLM	10	17.0	30.6	15.8	20.3	12.3	15.0	14.6	24.0	6.0	9.8	14.2	11.9	16.0
	100	7.5	21.7	6.4	6.2	7.0	7.1	8.3	16.0	17.2	8.1	4.2	2.9	9.4
Gemini	10	23.8	35.8	19.6	29.0	16.4	17.2	18.6	29.1	5.0	9.4	20.8	16.3	20.1
GPT-4	10	26.1	36.5	20.9	32.6	16.8	22.6	20.8	24.5	5.5	8.9	22.9	19.8	21.5
	100	42.5	40.9	25.9	42.1	23.2	35.1	17.2	5.6	10.8	2.4	6.6	19.3	22.6

Continue-training



Epoch	Bio.	Earth.	Econ.	Psy.	Rob.	Stack.	Sus.	Avg.
0 (GritLM)	25.0	32.8	19.0	19.9	17.3	11.6	18.0	20.5
1	22.2	25.4	17.6	28.1	11.1	9.8	19.6	19.1
2	18.7	23.8	13.5	19.3	10.7	10.2	16.5	16.1
3	20.9	23.6	16.9	25.2	11.1	8.5	16.6	17.5
4	24.3	28.0	18.3	26.9	13.4	13.3	20.0	20.6
5	23.1	28.5	18.4	26.1	14.6	11.7	21.6	20.6
6	19.9	26.4	16.0	27.9	9.6	9.3	19.3	18.3
7	24.3	25.4	16.5	28.1	11.0	9.8	17.0	18.9
8	21.6	28.7	19.2	28.7	11.1	11.8	22.4	20.5
9	21.3	29.0	20.0	28.7	11.4	14.3	22.0	21.0
10	21.1	25.5	18.8	30.7	12.7	12.1	21.9	20.4

Long-context retrieval



	Bio.	Earth.	Econ.	Psy.	Rob.	Stack.	Sus.	Pony	Avg.
Sparse models									
BM25	10.7	15.4	10.7	8.4	7.4	22.2	10.7	5.4	11.4
Open-sourced models (<1B)									
BGE	16.4	27.7	20.9	11.6	10.9	13.3	16.9	0.4	14.8
Inst-L	24.6	29.9	13.1	20.3	12.9	15.0	25.4	3.9	18.1
SBERT	25.6	34.1	18.9	15.8	10.9	15.0	18.0	1.2	17.4
Open-sourced models (>1B)									
E5	29.9	36.3	26.2	46.7	17.3	14.5	32.2	1.1	25.5
SFR	30.3	37.0	24.3	47.7	17.3	14.5	35.0	2.0	26.0
Inst-XL	21.5	31.0	13.1	20.5	13.9	15.0	20.1	6.0	17.6
GritLM	37.5	40.3	25.7	34.4	17.8	20.1	32.4	0.0	26.0
Qwen	39.2	36.1	25.7	42.3	21.3	23.5	33.1	1.3	27.8
Proprietary models									
Cohere	31.5	34.5	18.9	20.5	9.9	15.8	15.2	0.8	18.4
OpenAI	32.1	31.4	23.8	34.2	11.9	10.7	26.3	0.0	21.3
Voyage	34.4	35.4	26.7	41.6	12.9	12.8	31.1	1.3	24.5
Google	30.9	38.0	21.9	30.7	12.9	19.2	25.7	0.3	22.4

Thank you!