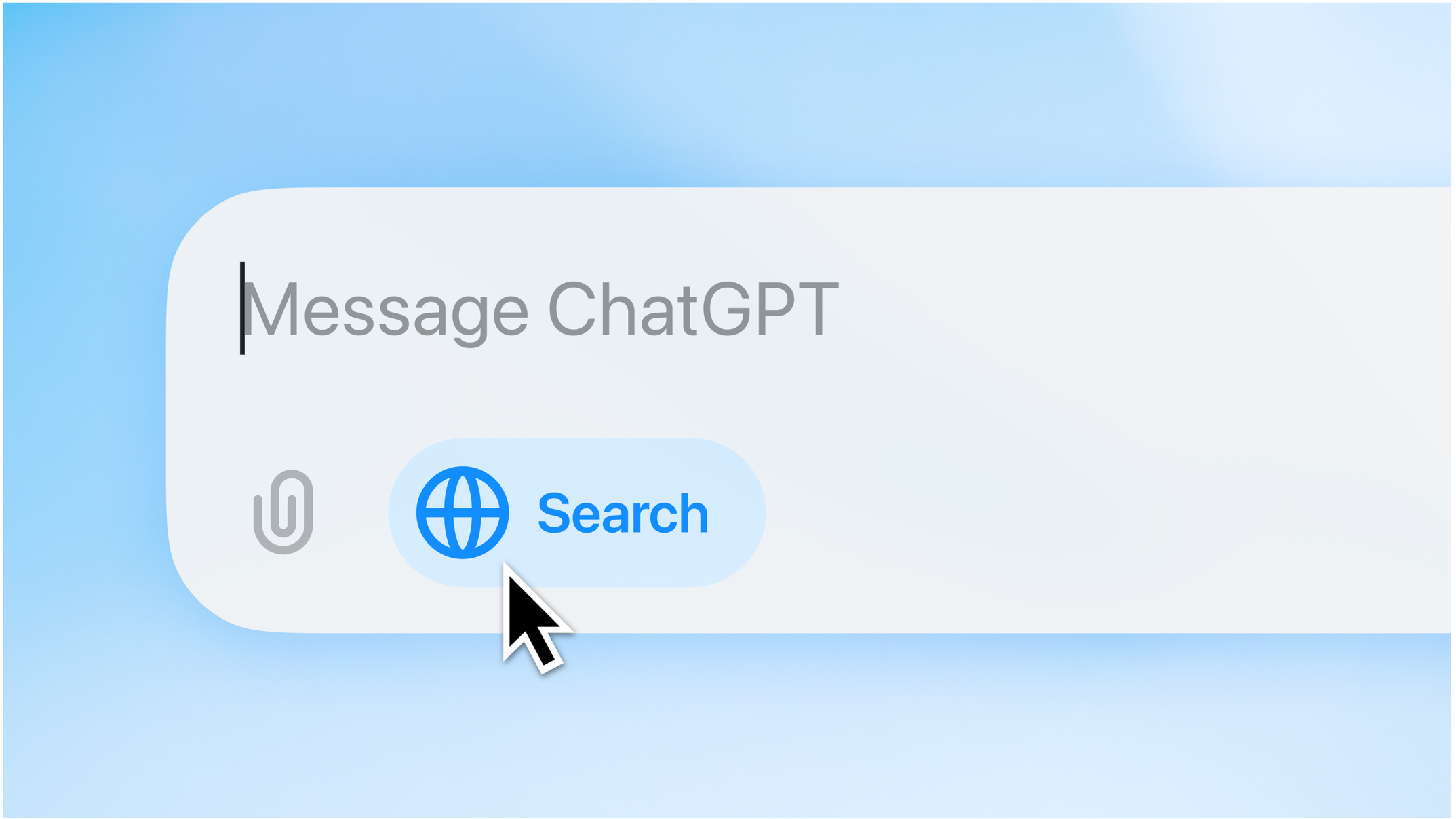

# New Frontiers in IR: Instruction Following and Reasoning

Orion Weller


Message ChatGPT
Search

# Instruction Following

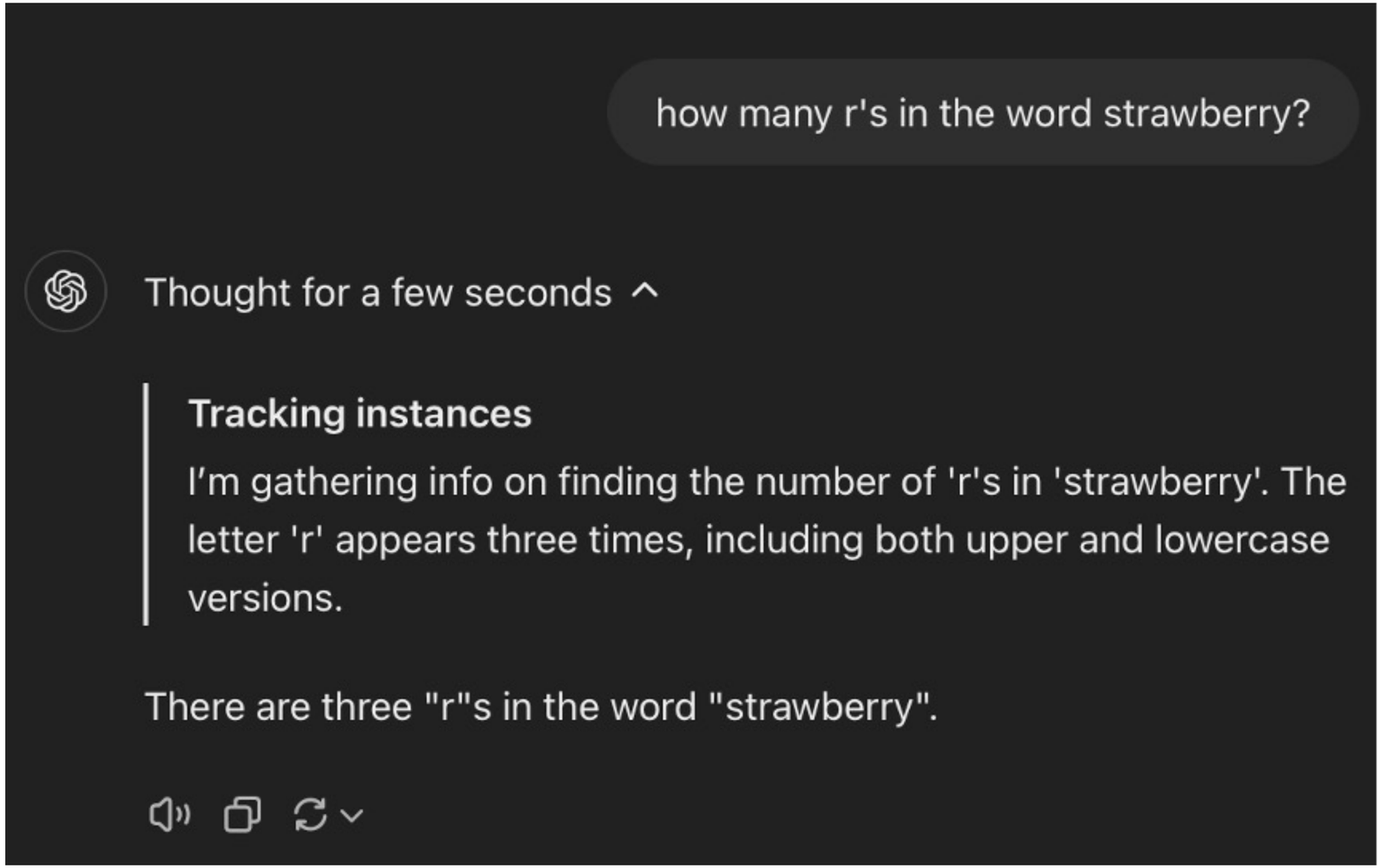
Generate a haiku about information retrieval in the style of a pirate and mention RAG.

Scourin′ the sea wide,

RAG finds treasure in the fog—

Answers hoist with pride!

# Reasoning


how many r's in the word strawberry?
Thought for a few seconds
Tracking instances
I'm gathering info on finding the number of 'r's in 'strawberry'. The letter 'r' appears three times, including both upper and lowercase versions.
There are three "r"s in the word "strawberry".

# What about in retrieval?


Search
Google!
BETA
Search the web using Google
Google Search
I'm feeling lucky
More Google!
Copyright ©1999 Google Inc.


google

google **maps**

google **classroom**

google **docs**

google **drive**

google **news**

google **flights**

google **earth**

google **photos**

google **calendar**

google **slides**

Google Search I'm Feeling Lucky

*Report inappropriate predictions*

# SearchGPT

where can i grab dinner in positano on friday night?

Search

Planning a dinner in Positano on a Friday night offers a delightful combination of exquisite cuisine and stunning coastal views. Here are some top recommendations to consider:

### La Tagliata

Nestled in the hills above Positano, La Tagliata provides panoramic views of the Amalfi Coast. This family-run restaurant is renowned for its authentic Italian dishes, particularly handmade pastas and grilled meats. Due to its popularity, especially on weekends, it's advisable to book in advance. LA TAGLIATA

### Il Ritrovo

Located in the heart of Positano, Il Ritrovo offers a cozy ambiance with a menu

# Search really hasn’t changed

Despite using LLMs, we’re just adding a wrapper around the results

# Keyword Search

## Query

Find websites
explaining data privacy

# Keyword Search

## Query

Find websites
explaining data privacy

## Documents

**Data Encryption Standards**
www.nist.gov/standards/

**Wolves Outside Your Data**
www.janerodgers.blog/wolves

**Digital Protection**
www.clearlaw.net/digital

# Semantic Search

## Query

Find websites
explaining data privacy

## Documents

**Data Encryption Standards**
www.nist.gov/standards/

**Wolves Outside Your Data**
www.janerodgers.blog/wolves

**Digital Protection**
www.clearlaw.net/digital

# Instruction-based Search

## Query

Find websites explaining data privacy **and uses extended metaphors**

**Can’t solve this by using LLMs to rerank!**

## Documents

**Data Encryption Standards**
www.nist.gov/standards/

**Wolves Outside Your Data**
www.janerodgers.blog/wolves

**Digital Protection**
www.clearlaw.net/digital

# Instruction-based Search

## Query

Find websites explaining data privacy **and uses extended metaphors**

## Documents

**Wolves Outside Your Data**
www.janerodgers.blog/wolves

**Data Encryption Standards**
www.nist.gov/standards/

**Digital Protection**
www.clearlaw.net/digital

# Prompt and Reasoning-based Search

## Query

Find websites explaining data privacy **and uses extended metaphors. Have really high recall or I will lose my job**

## Documents

**Wolves Outside Your Data**
www.janerodgers.blog/wolves

**Data Encryption Standards**
www.nist.gov/standards/

**Digital Protection**
www.clearlaw.net/digital

# What is an instruction in IR?

Doc attributes:
date, length, source

NLU aspects:
sentiment, style

And more!

Logical conditions
AND/OR/NOT

# We’re so used to prompting LMs

## Let’s use retrievers the **same way**

# Promptriever

**fast embedder**

# Rank1

**strong but slow**

**Promptriever**

**fast embedder**

**Rank1**

**strong but slow**

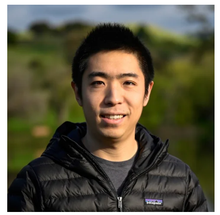
# Promptriever: Instruction-Trained Retrievers Can Be Prompted Like Language Models


Orion Weller

Benjamin Van Durme

Dawn Lawrie

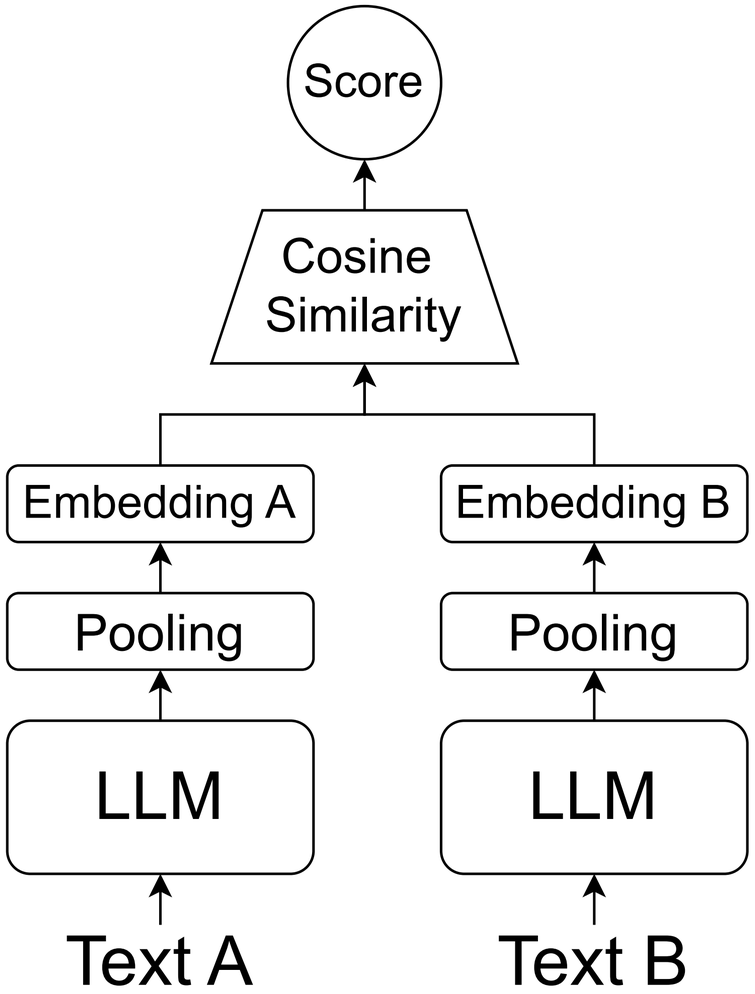
Ashwin Paranjape



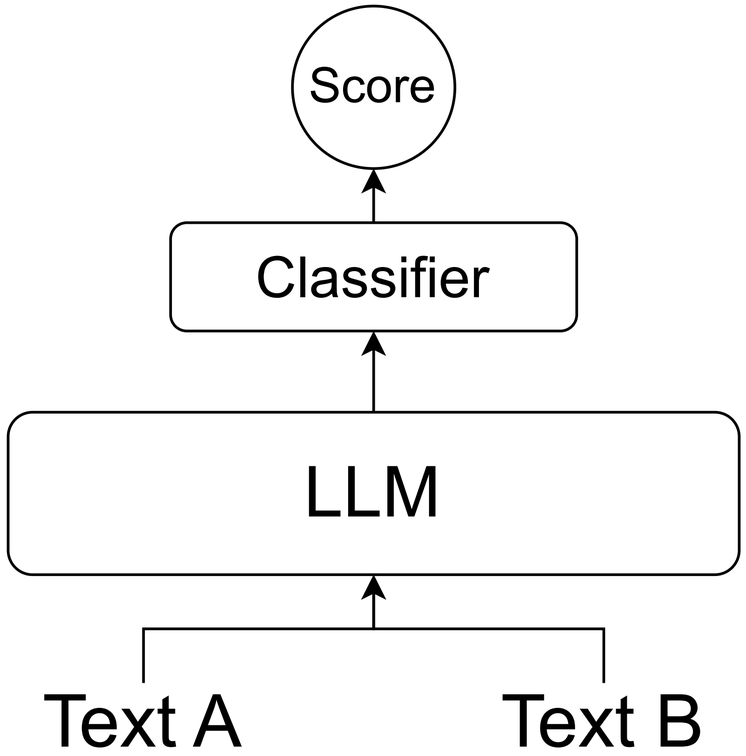
Yuhao Zhang

Jack Hessel

# Bi-Encoder




# Cross-Encoder



# Can we make promptable retrievers?

## Can bi-encoders take instructions?

Training data for instruction-following

# Generating Training Data

**Query**

Which type of volcano eruption has not been seen?

# Generating Training Data

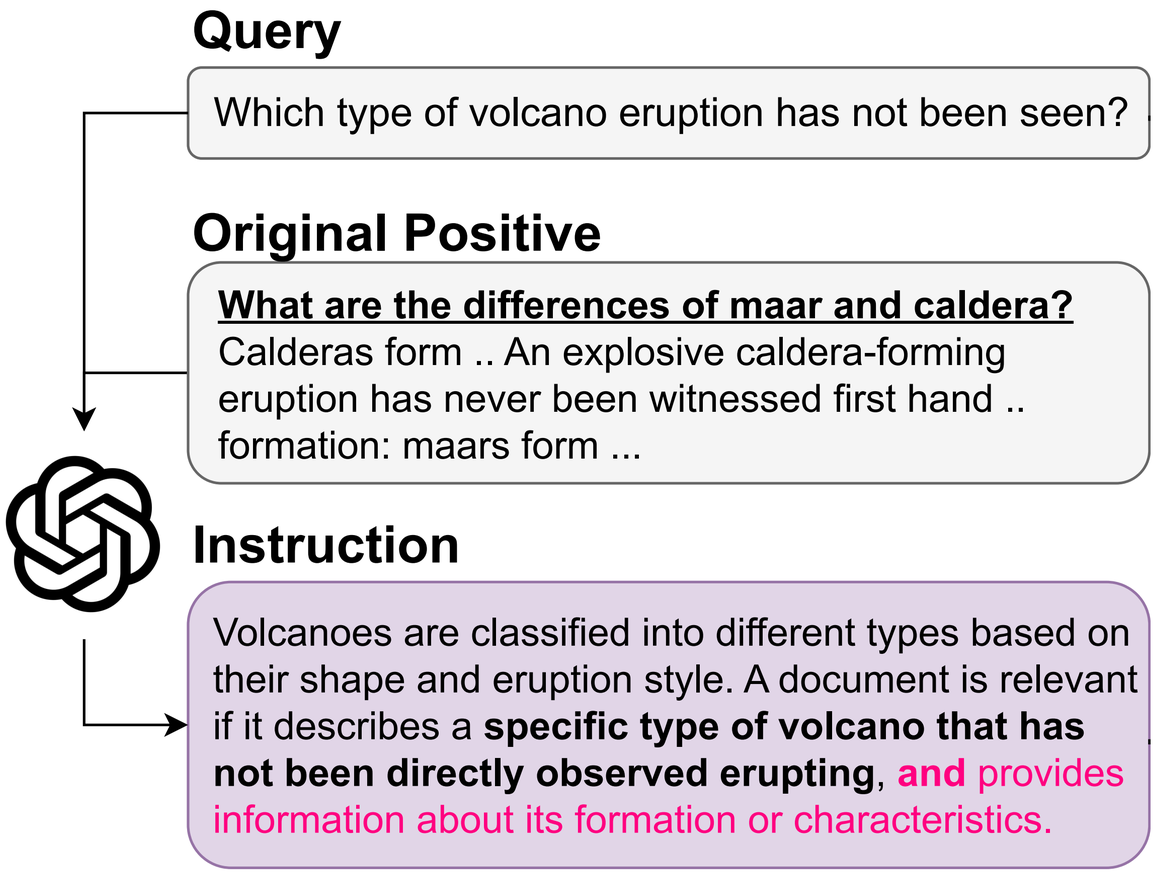
## Query

Which type of volcano eruption has not been seen?

## Original Positive

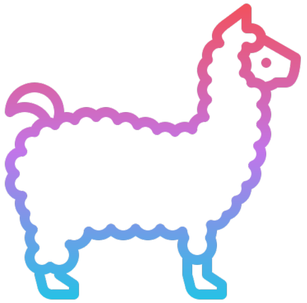
**What are the differences of maar and caldera?**
Calderas form .. An explosive caldera-forming eruption has never been witnessed first hand .. formation: maars form ...

# Generating Training Data




**We have to generate synthetic instructions**

# Experimental Settings



Started from RepLLaMA's (using LLaMA-2) training recipe for a direct comparison


Evaluation is done on in-domain data (MSMarco), instruction data, and out of domain data

# Evaluation Data

**Query:** I am looking for articles of Teflon use that may pose potential health risks.

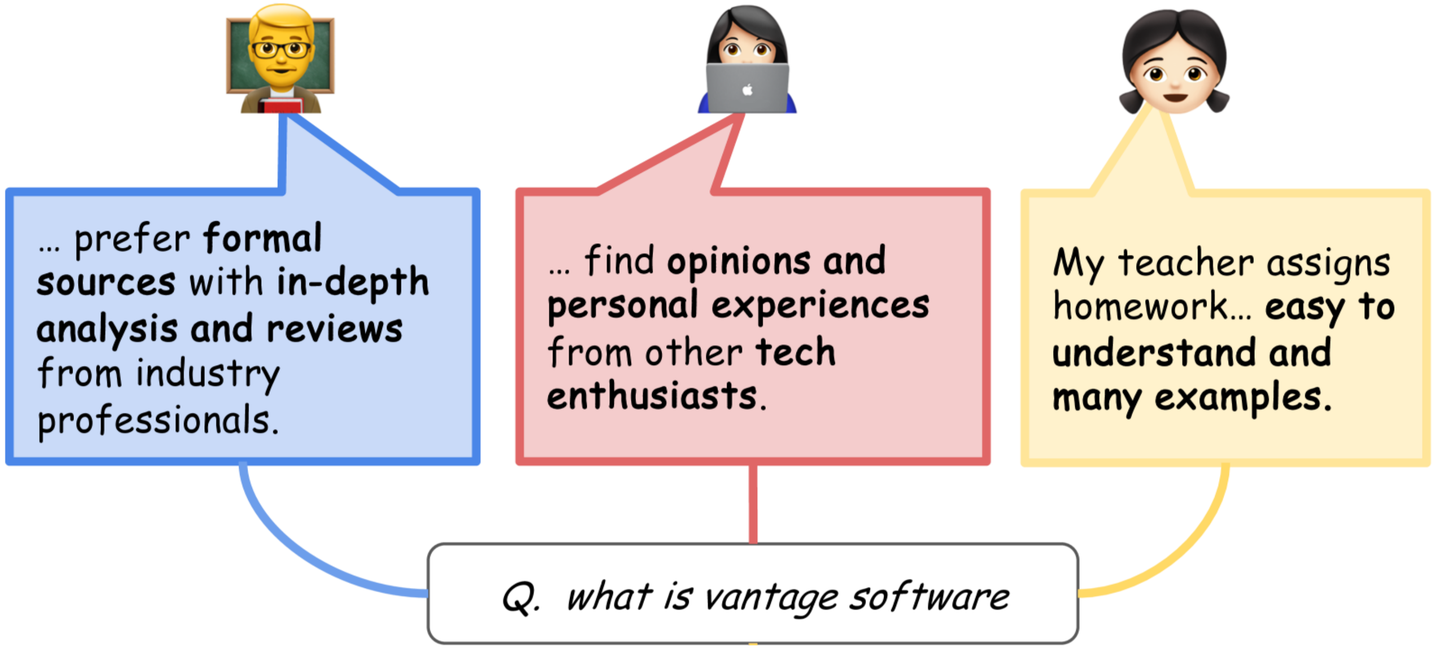
**Instruction:** Find articles referencing Teflon as potential health risk. Relevant documents must give a reason or explain why the use of Teflon is a concern. The application of Teflon ~~can be of any means~~, **has to be related to chemicals**, **specifically** as long as the health risk is affecting human.

# FollowIR

p-MRR ranges
from -100 to 100

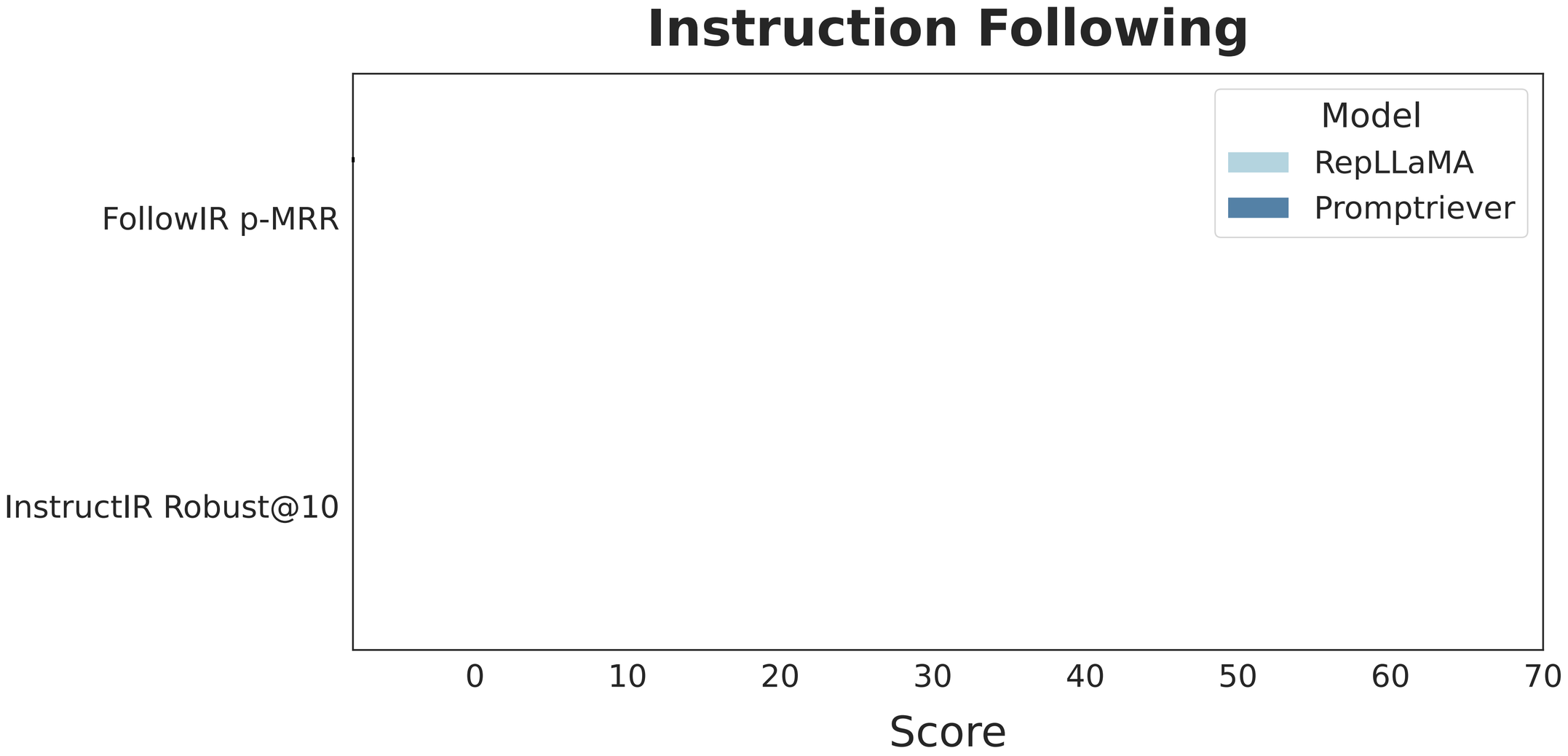
# InstructIR

**10 personas for each MS MARCO query**

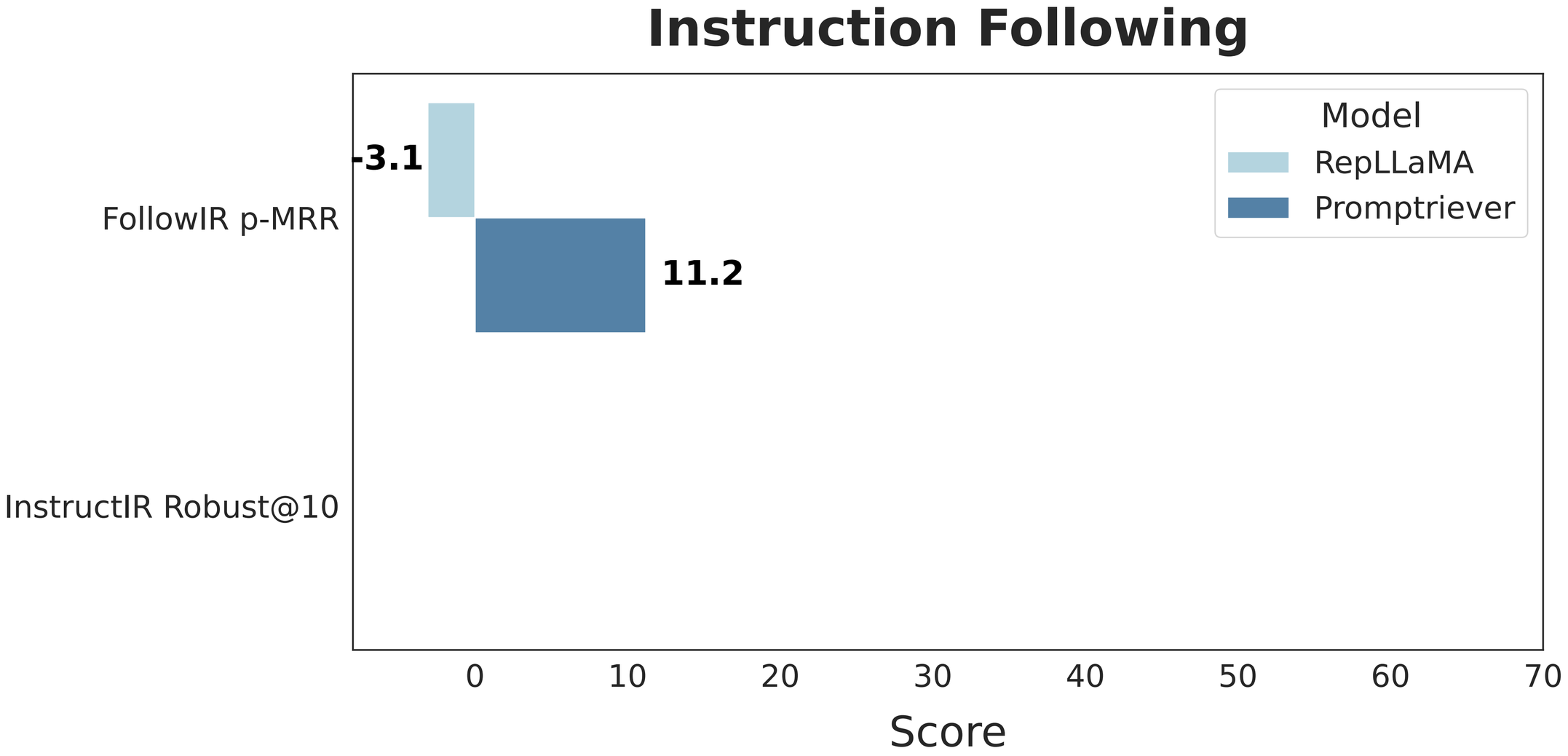


# Results

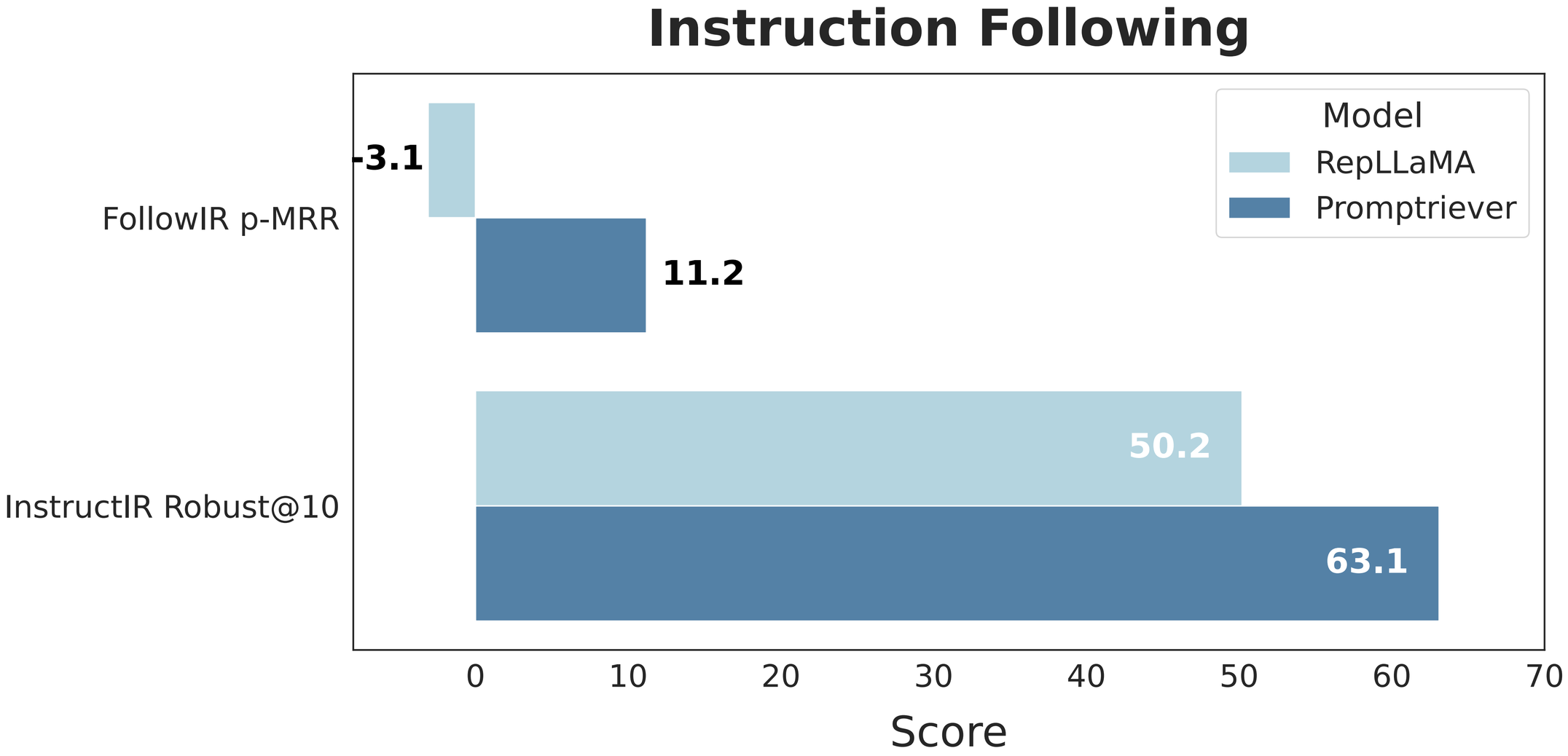
# Results


Instruction Following
Model
RepLLaMA
Promptriever
FollowIR p-MRR
InstructIR Robust@10
0
10
20
30
40
50
60
70
Score

# Results


Instruction Following
Model
RepLLaMA
Promptriever
-3.1
11.2
FollowIR p-MRR
InstructIR Robust@10
0
10
20
30
40
50
60
70
Score

# Results


Instruction Following
Model
RepLLaMA
Promptriever
FollowIR p-MRR
InstructIR Robust@10
-3.1
11.2
50.2
63.1
0
10
20
30
40
50
60
70
Score

# Results

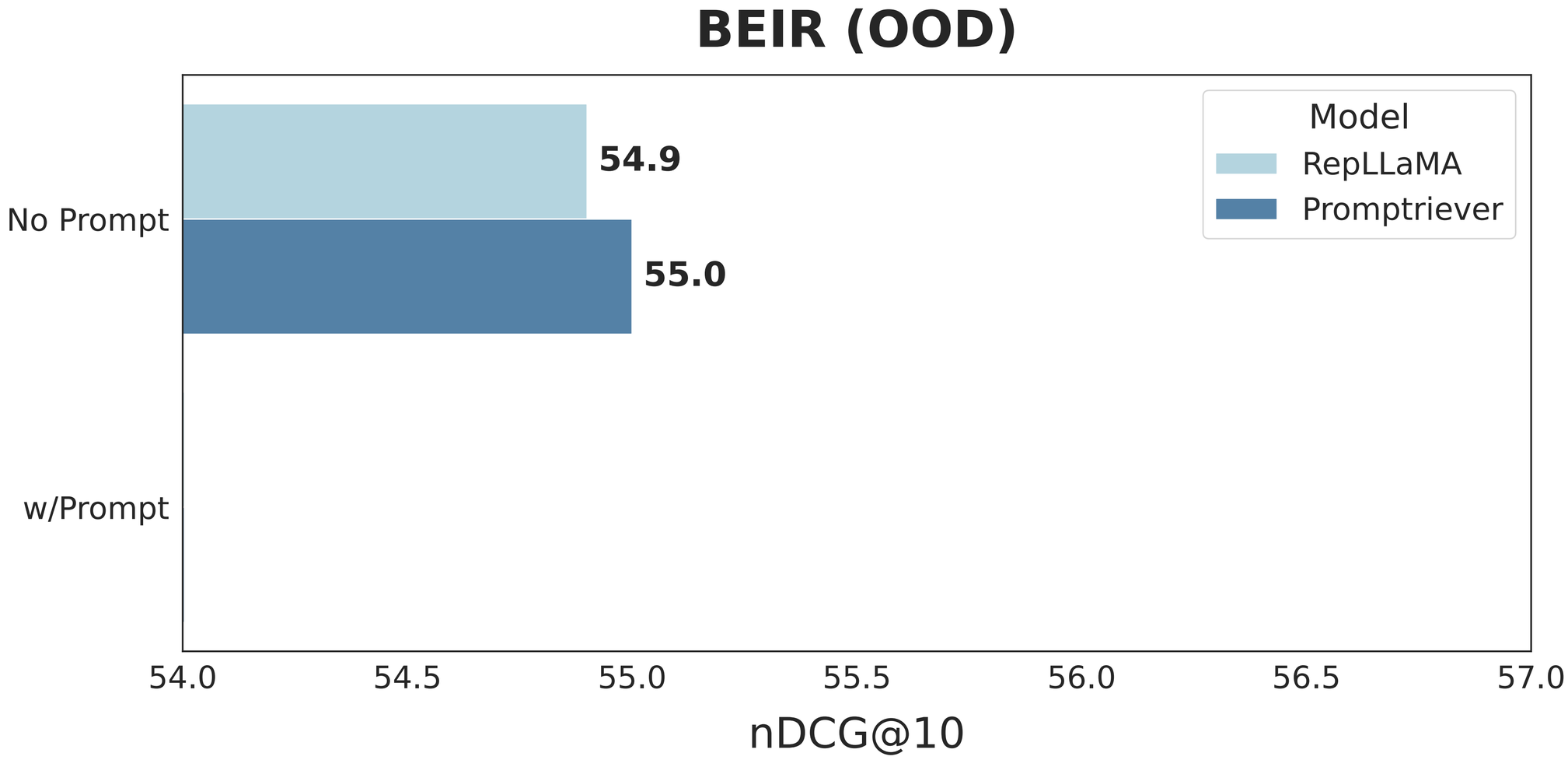
# Results

When evaluating on data with no set instruction, what prompt to use?

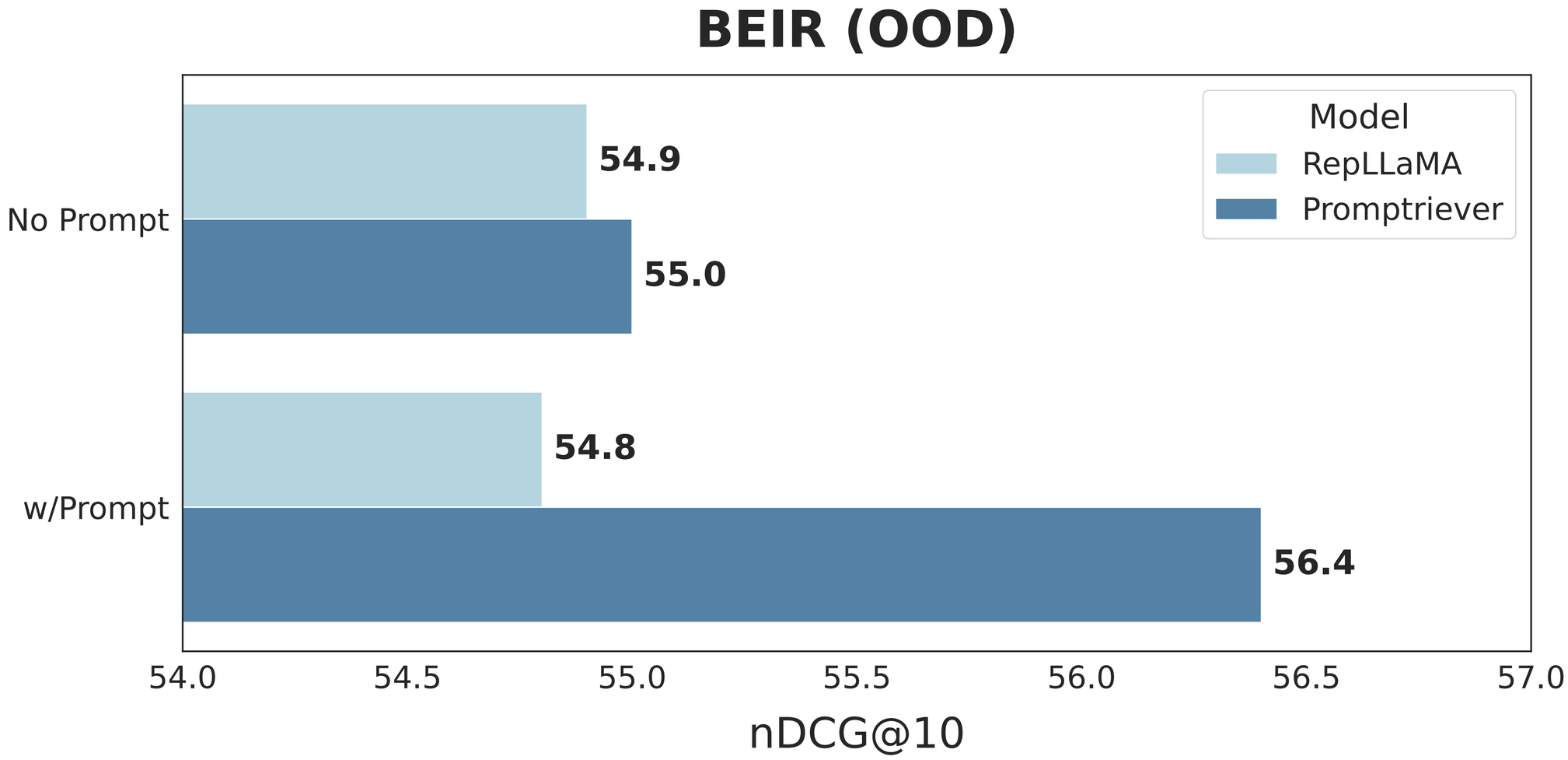
- We show with no prompt (standard)
- We come up with 10 generic prompts and use the best on the dev set for test

| **Prompts** |
| --- |
| • Be careful when assigning relevance as your job is on the line and I will give you a 1000 dollar tip. |
| • Think carefully about these conditions when determining relevance. |
| … |

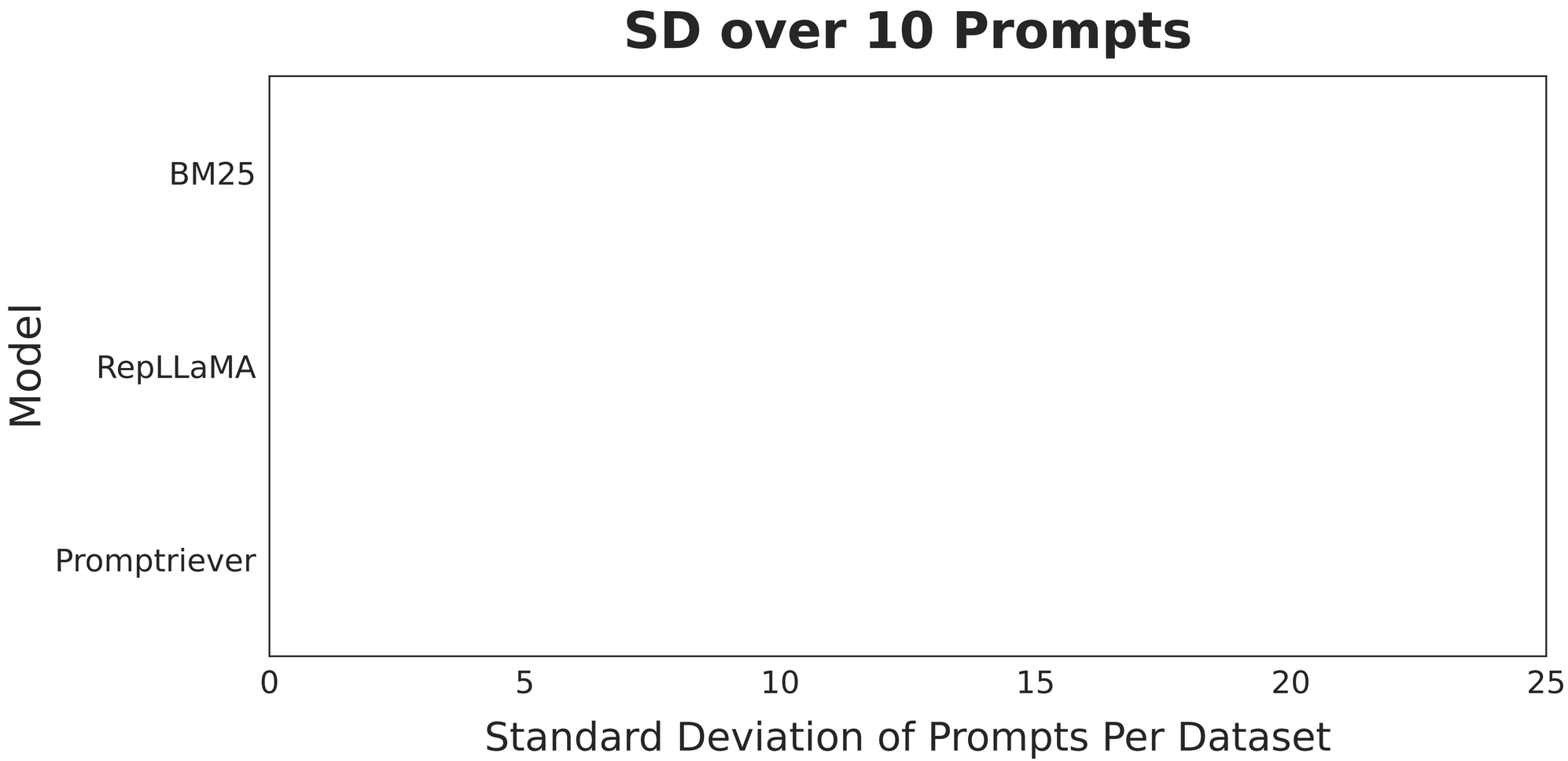
# Results


BEIR (OOD)
Model
RepLLaMA
Promptriever
54.9
55.0
No Prompt
w/Prompt
54.0
54.5
55.0
55.5
56.0
56.5
57.0
nDCG@10

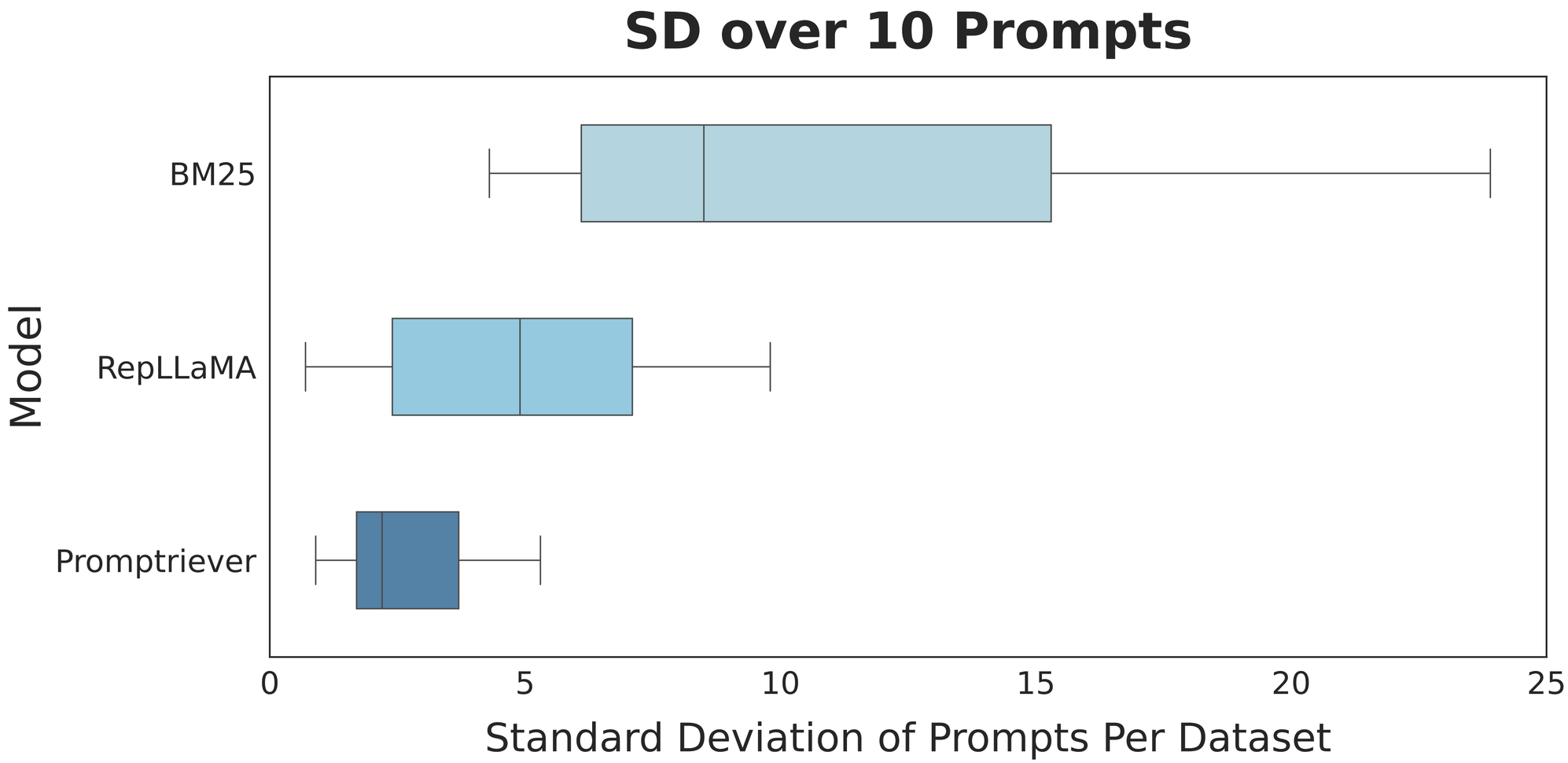
# Results


BEIR (OOD)
Model
RepLLaMA
Promptriever
No Prompt
w/Prompt
54.9
55.0
54.8
56.4
54.0
54.5
55.0
55.5
56.0
56.5
57.0
nDCG@10

# Results


SD over 10 Prompts
Model
BM25
RepLLaMA
Promptriever
0
5
10
15
20
25
Standard Deviation of Prompts Per Dataset

# Results


SD over 10 Prompts
BM25
RepLLaMA
Promptriever
Model
0
5
10
15
20
25
Standard Deviation of Prompts Per Dataset

# Summary

- **With the right data**, you can have retrievers that are promptable just like LM
- Unlocks new types of queries!
- Don’t need to be picky about keywords, just tell the model what you want in natural language

# Promptriever

**fast embedder**

# Rank1

**strong but slow**

## Promptriever

fast embedder

## Rank1

**strong but slow**

# Rank1: Test-Time Compute for Information Retrieval

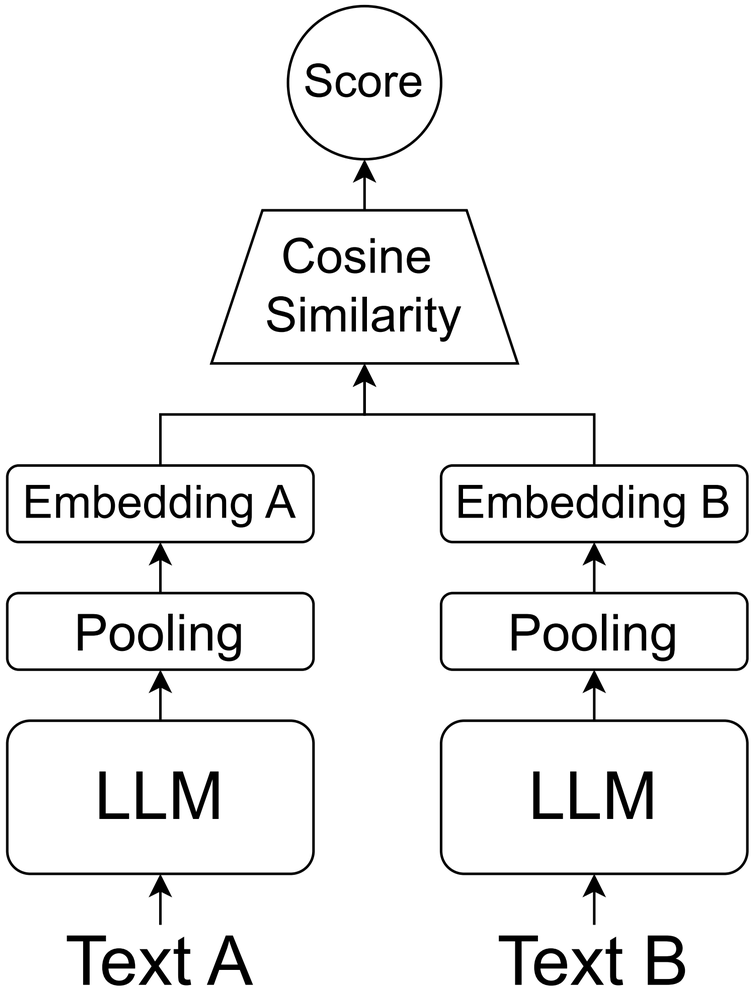




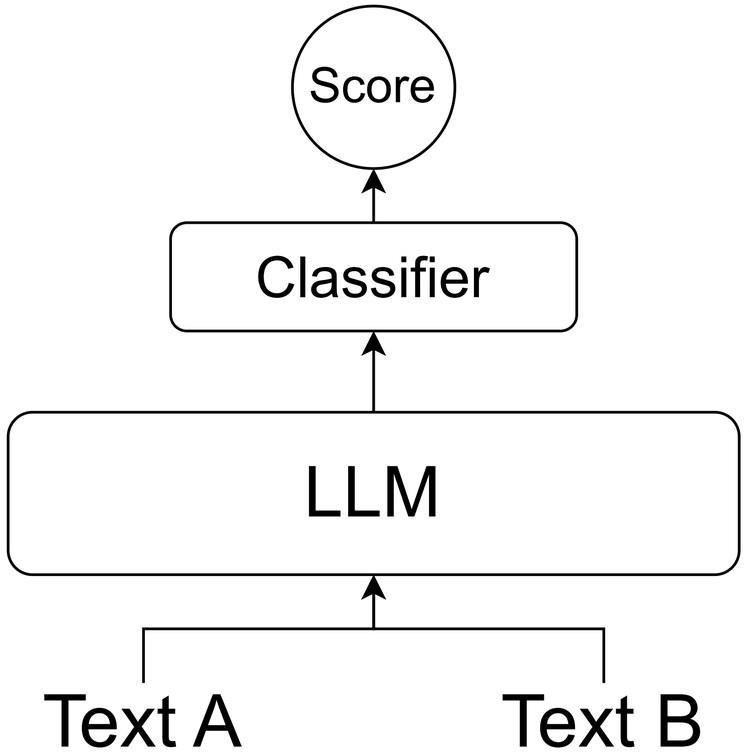


Orion Weller Kathryn Ricci Eugene Yang Andrew Yates Dawn Lawrie Benjamin Van Durme

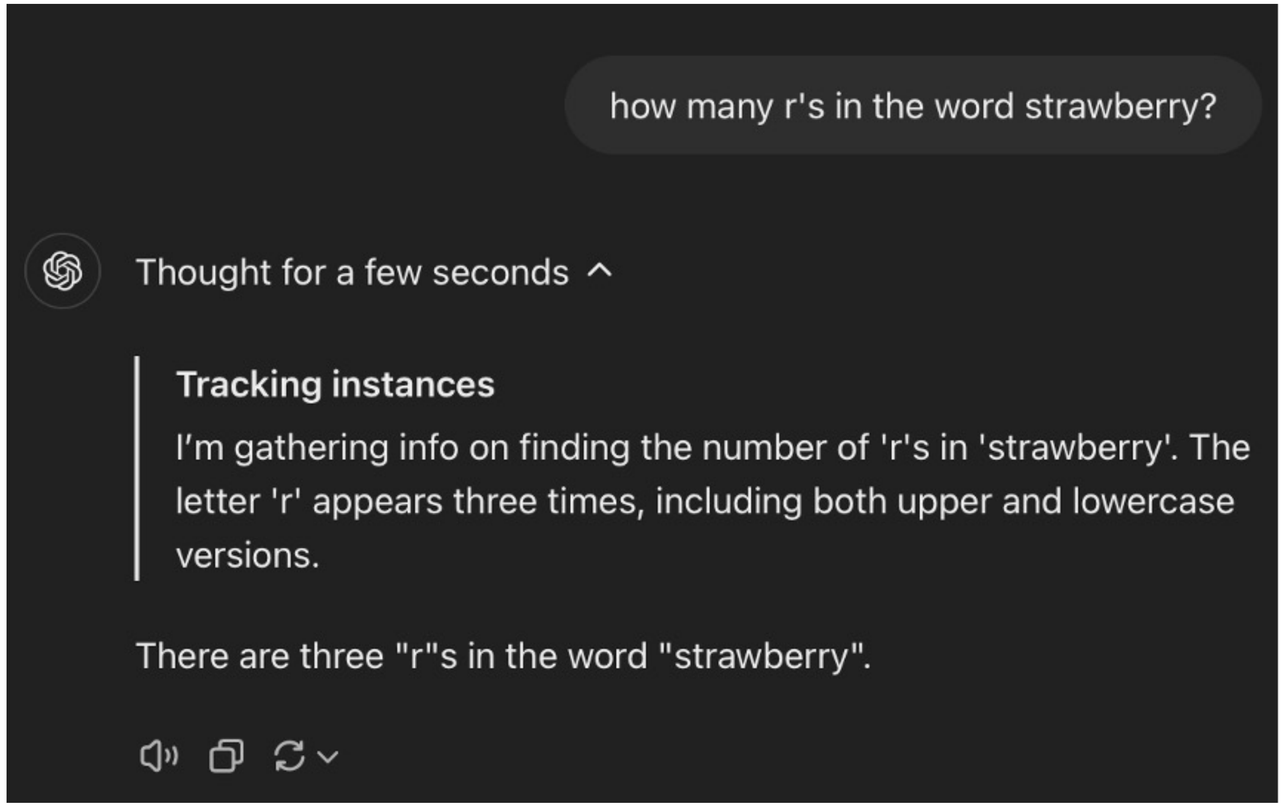
# Bi-Encoder

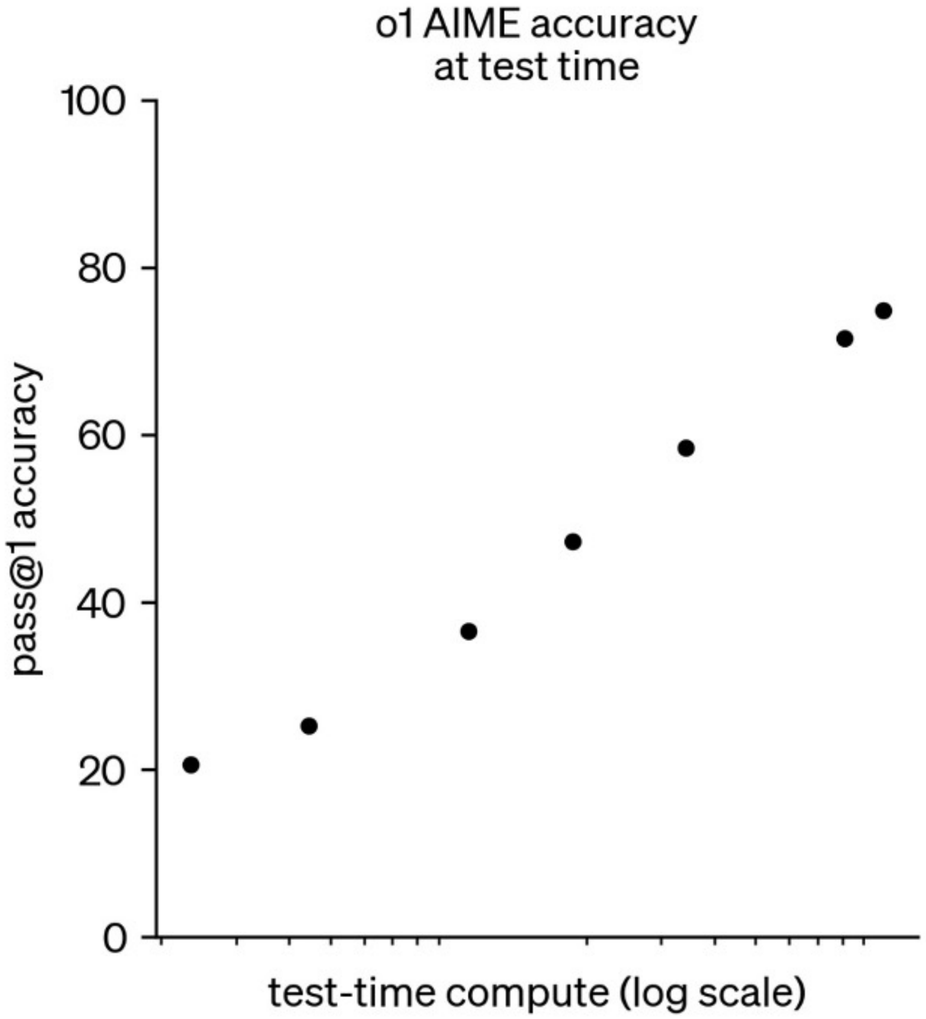



# Cross-Encoder



# Test-Time Compute



# What does test-time compute look like in IR?

# What does test-time compute look like in IR?

do snow leopards change color

Snow leopards have long, thick fur, and their base color varies from smoky gray to yellowish tan, with whitish underparts ...

# What does test-time compute look like in IR?

❓ do snow leopards change color

Snow leopards have long, thick fur, and their base color varies from smoky gray to yellowish tan, with whitish underparts ...

🤖 <think> The user is asking if snow leopards change color. The passage provided describes their fur color as varying from smoky gray to yellowish tan, with some patterns. Hmm, the key here is whether "varies" implies that they change color over time, like seasonally or due to some factors.

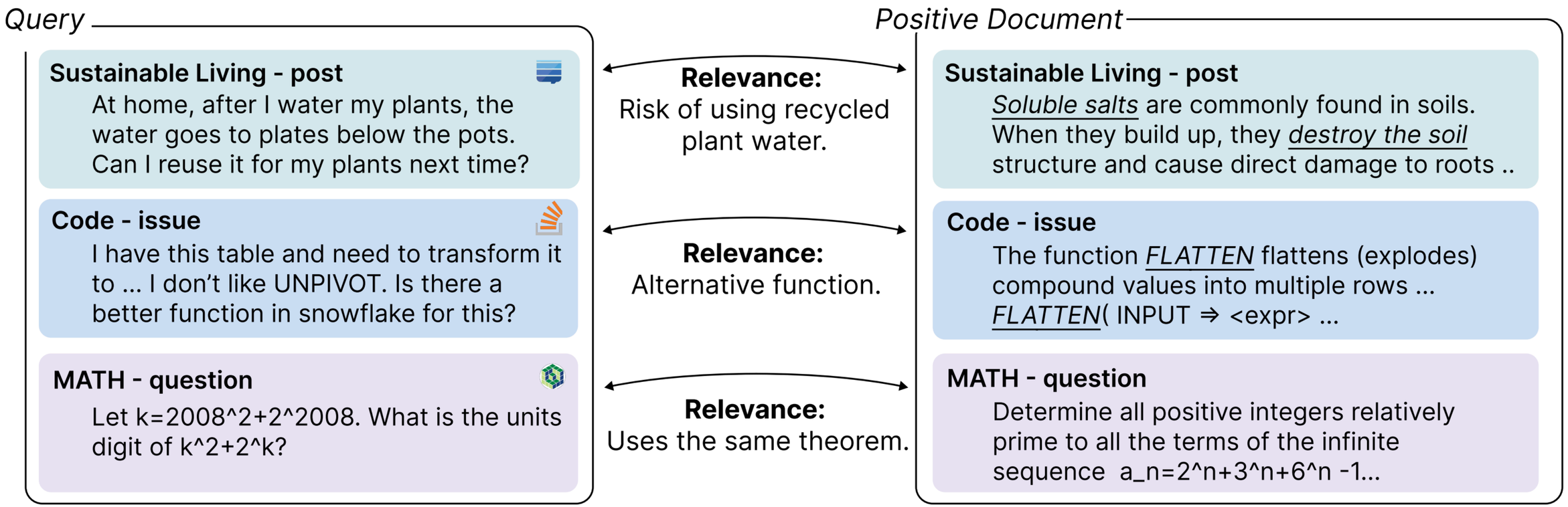
But wait, "varies" might just mean that different individuals have different base colors, not that a single leopard changes its color. The passage doesn't mention anything about changing color; it's just describing their appearance. ...

Therefore, the passage isn't directly relevant to the question of whether they change color. So the answer should be false.

</think> **false**

# Evaluation Data

# BRIGHT dataset



# What does this look like?

## Example: Can you find a similar leetcode problem?

**Model Reasoning**

<think>
Okay, let's tackle this problem. The user is asking whether the given passage is relevant to their query about finding problems with similar data structures or algorithms. The query specifically mentions problems that share similar data structures (like arrays, trees) or algorithms (DFS, DP, sorting, traversals). The example problem is about the rainwater trapping problem, which uses a two-pointer approach.

Now, looking at the passage provided, it's about the "Max Area" problem, which also uses a two-pointer approach. Both problems involve arrays and solving them with a two-pointer technique. The rainwater problem uses a two-pointer approach to find the maximum trapped water, while the max area problem does the same to find the maximum area between vertical lines.

The key here is that both problems share the same algorithmic approach (two pointers) and similar data structures (arrays). Even though the problems are different (trapping rain vs. max area), the underlying algorithm and data structure similarities make them relevant. The user is looking for problems that share these similarities, so the passage does fit that criteria. Therefore, the answer should be true
</think> true

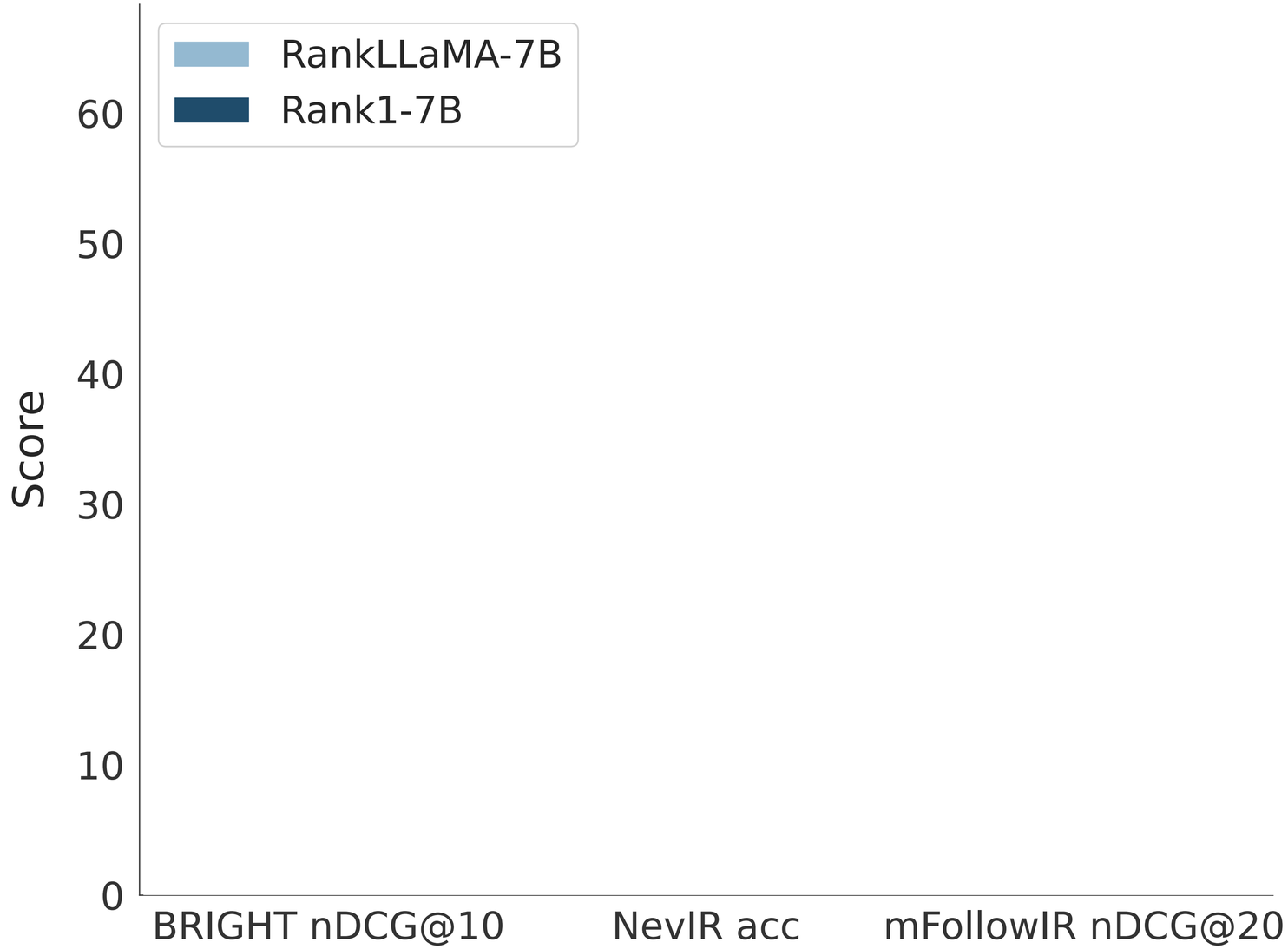
# Results

# Results

RankLLaMA was trained on **10x more data**!

We evaluate on a broad range of tasks:

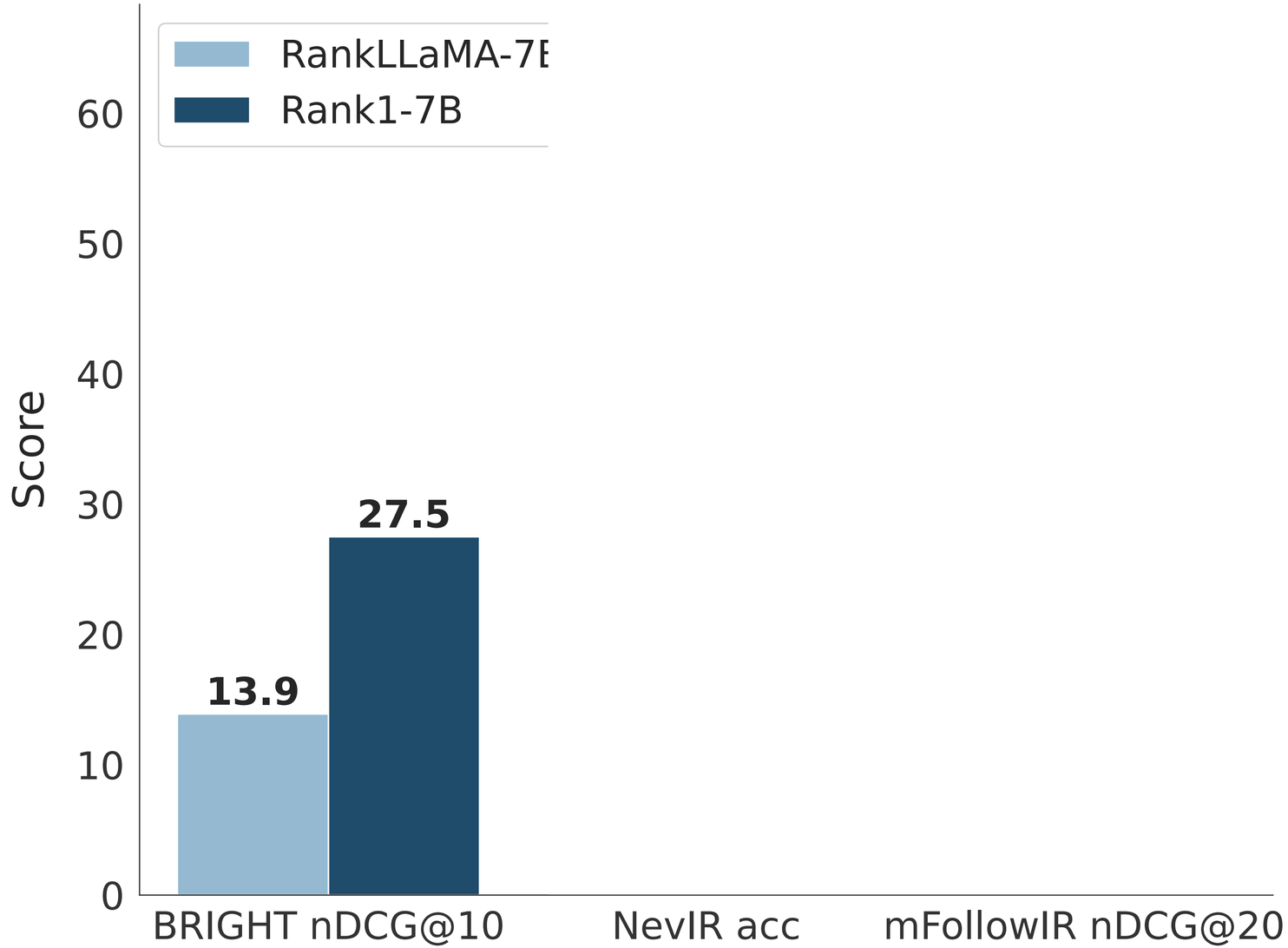
- BRIGHT (*reasoning*)
- NevIR (*negation*)
- mFollowIR (*instructions*)



# Results

We evaluate on a broad range of tasks:

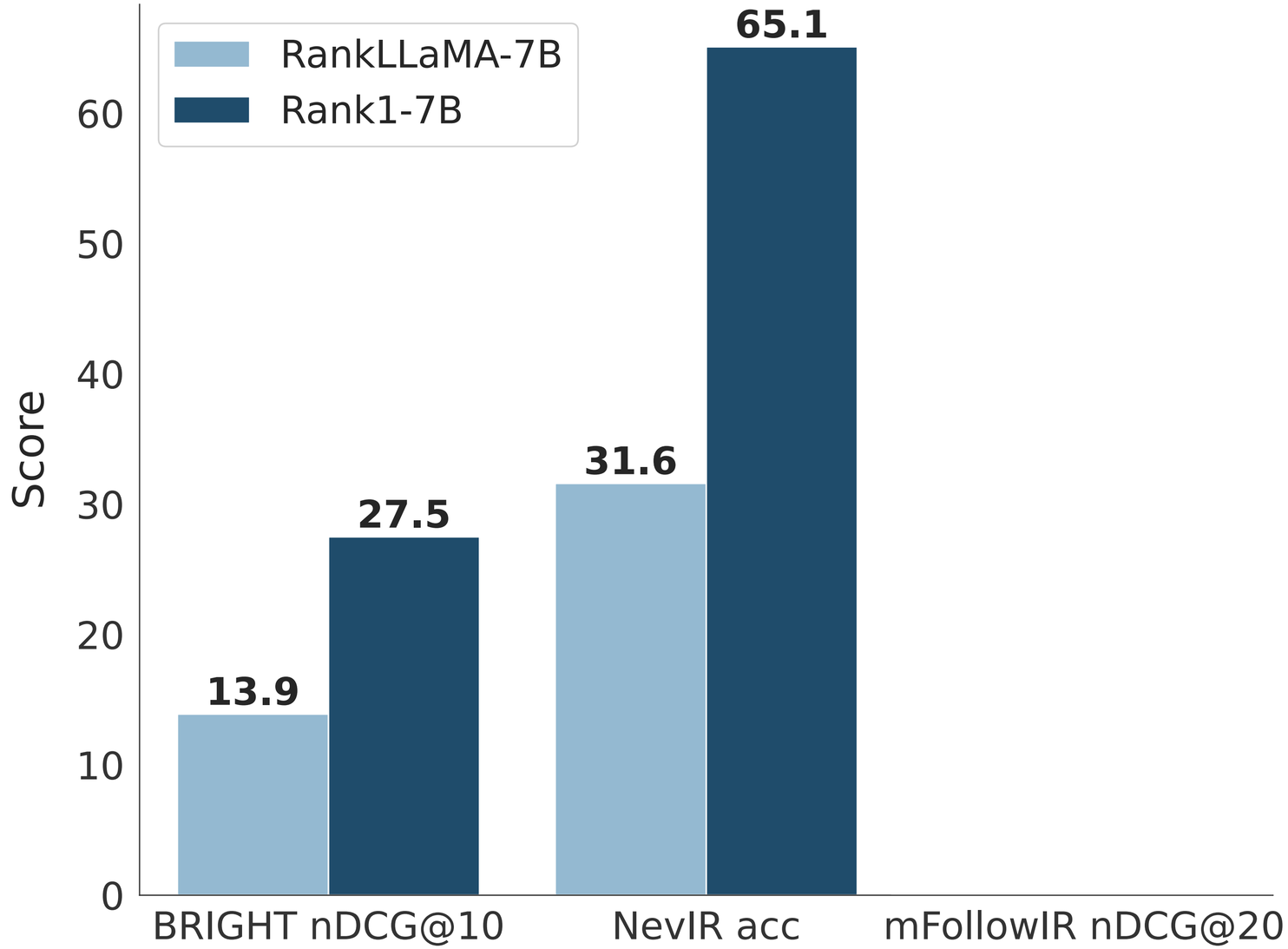
- BRIGHT (*reasoning*)
- NevIR (*negation*)
- mFollowIR (*instructions*)



# Results

We evaluate on a broad range of tasks:

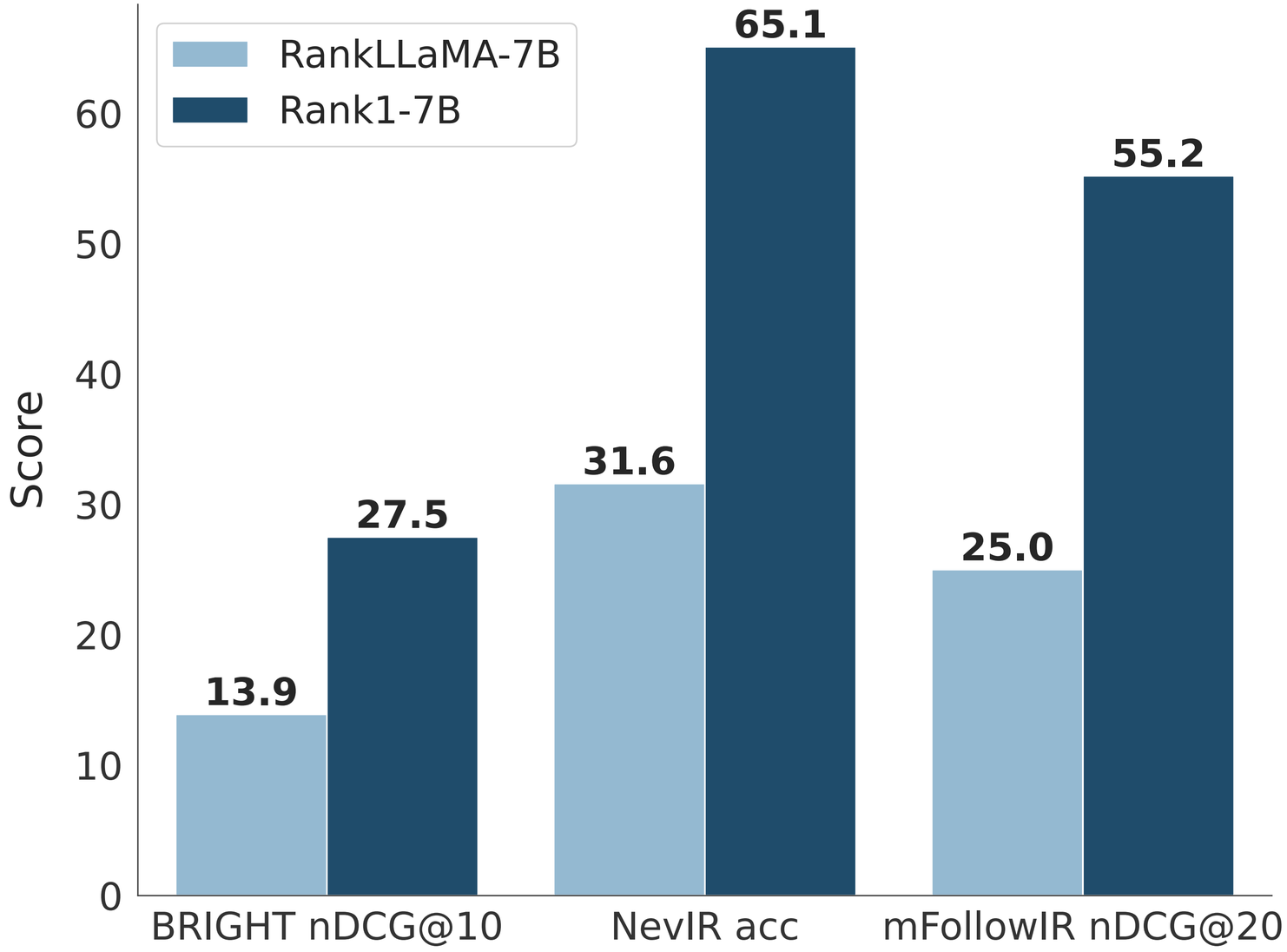
- BRIGHT (*reasoning*)
- NevIR (*negation*)
- mFollowIR (*instructions*)



# Results

We evaluate on a broad range of tasks:

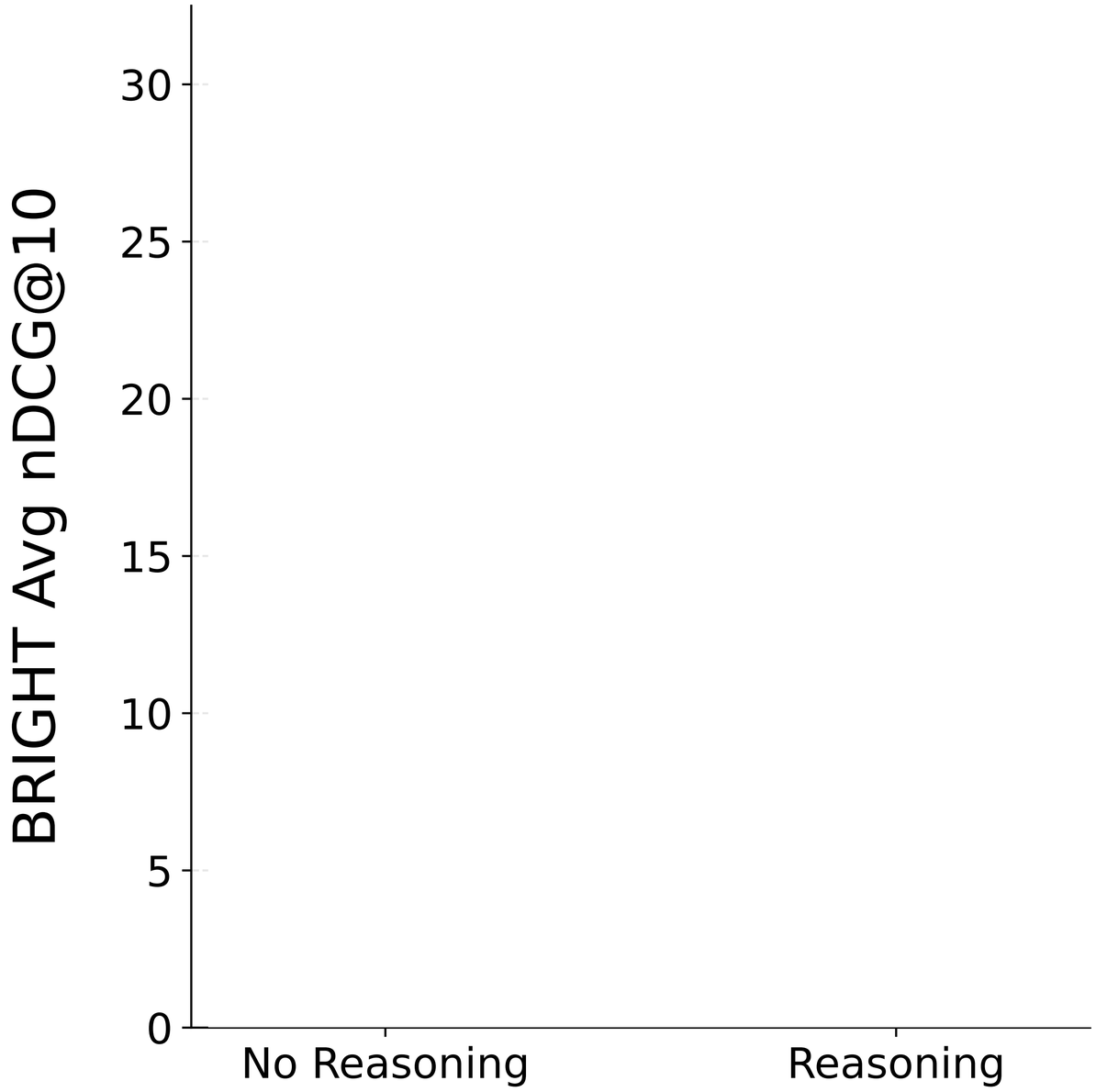
- BRIGHT (*reasoning*)
- NevIR (*negation*)
- mFollowIR (*instructions*)



# Results

What about a direct comparison of the reasoning chain?

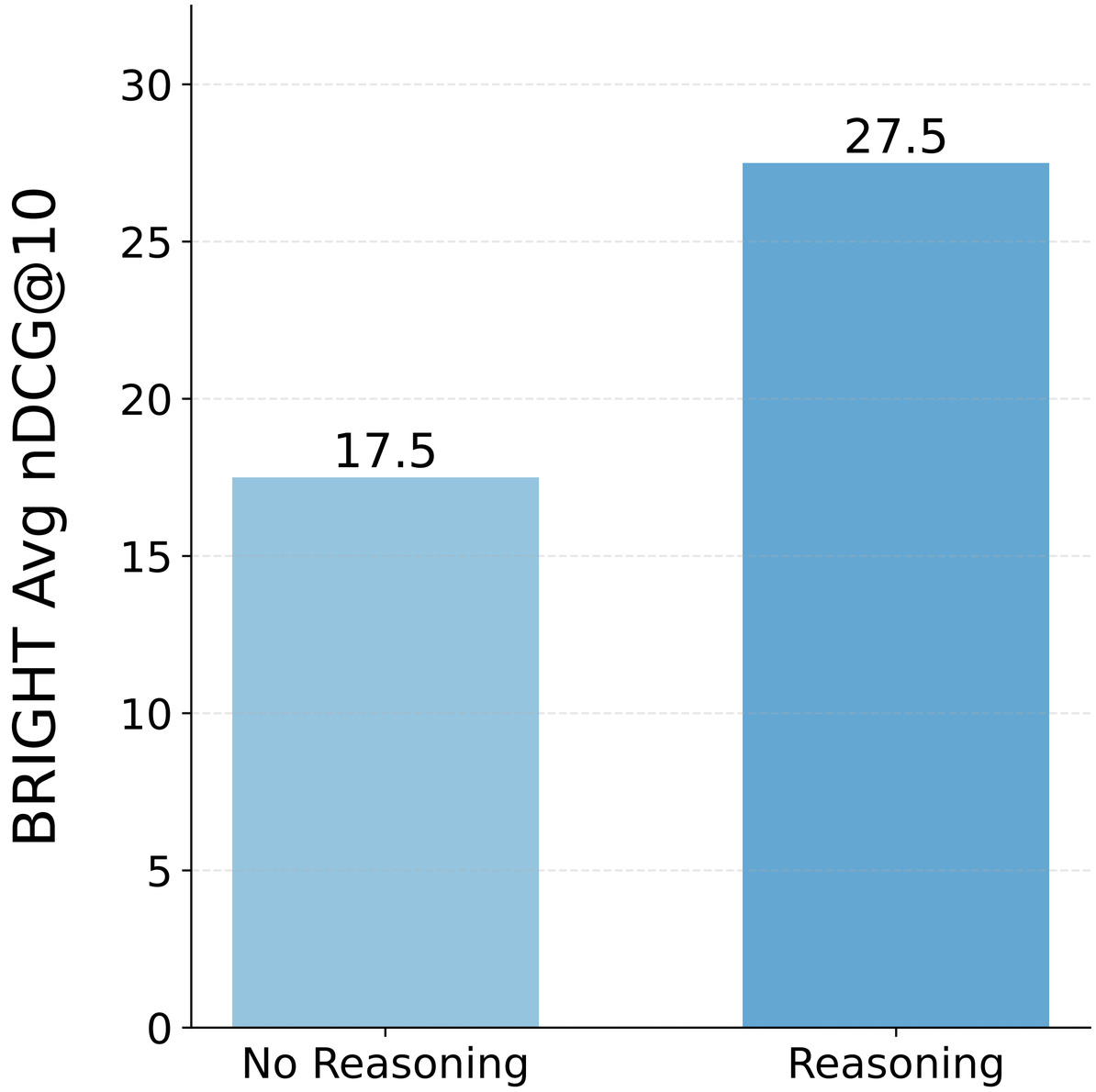
Same data, same model, no chain



# Results

What about a direct comparison of the reasoning chain?

Same data, same model, no chain

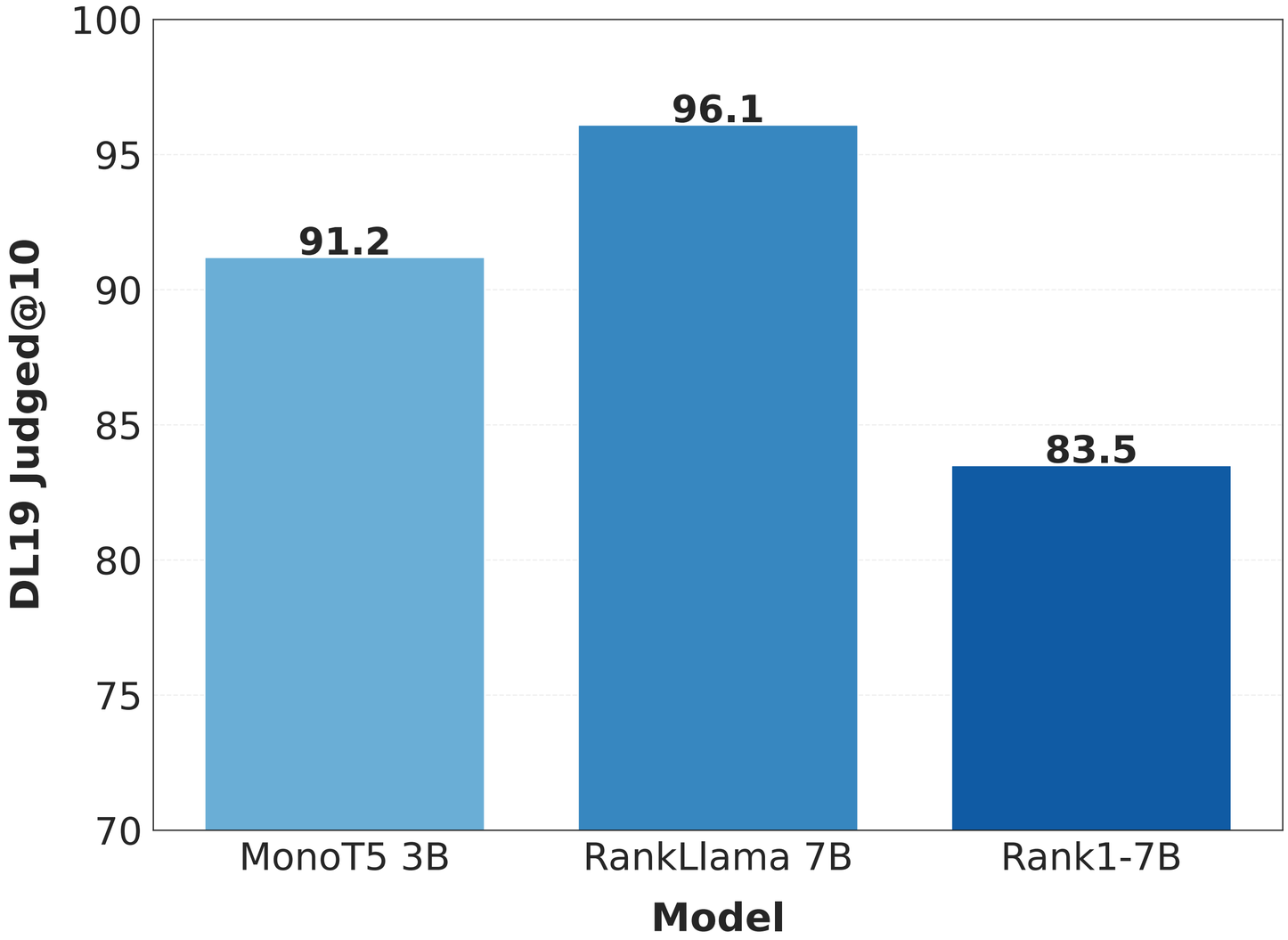


# Old Evaluation Data

# Results

We were surprised by the low initial scores on DL19/DL20

Judged@10 is significant less!

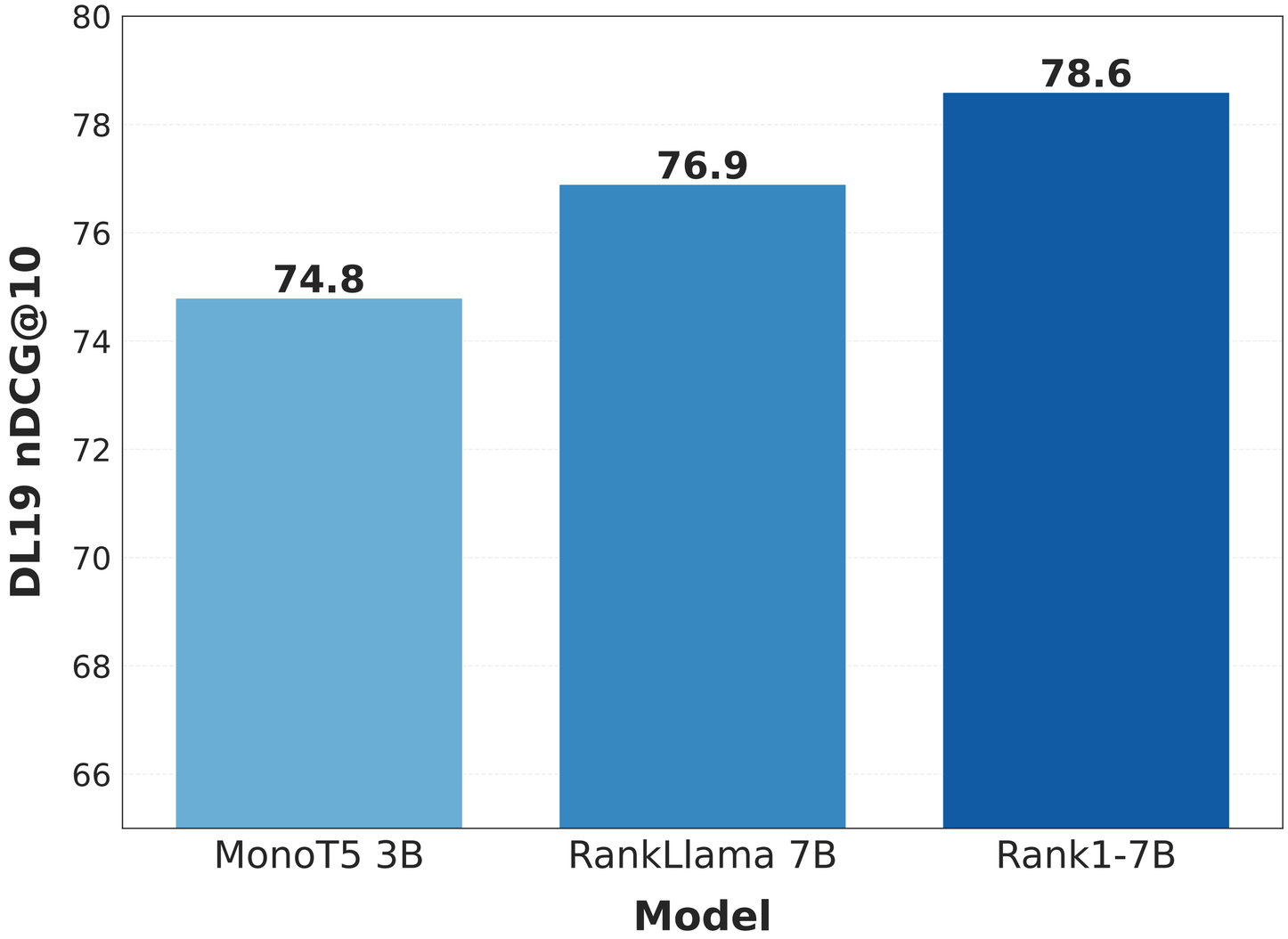


# Results

It's finding **new docs that previous systems didn't** — that are relevant!

We re-judged the unjudged + incorrect preds for all models

Perhaps the community should move on to newer eval datasets (**DL19 was before BERT**!)



# Summary

- Using test-time compute creates **promptable and reasoning rerankers -** no RL needed!
- It is slower, but much more powerful than previous approaches
- Only trained on general web data — training on in-domain data is likely to see huge gains

## Promptriever

fast embedder

## Rank1

**strong but slow**

# The goal: IR that works like LLMs

## Query

Find websites explaining data privacy **and uses extended metaphors. Have really high recall or I will lose my job**

## Documents

**Wolves Outside Your Data**
www.janerodgers.blog/wolves

**Data Encryption Standards**
www.nist.gov/standards/

**Digital Protection**
www.clearlaw.net/digital

# What does this mean?

New retrievers **benefit** from advances in LLMs

**Instruction-based search**: any query you can type - they can search

All models are open-data/open-source

oweller2@jh.edu