



JOHNS HOPKINS  
WHITING SCHOOL  
*of* ENGINEERING

# 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.

# 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?



# Google

A screenshot of a Google search interface. The search bar at the top contains the text "google". Below the search bar is a list of search suggestions, each preceded by a magnifying glass icon:

- google maps
- google classroom
- google docs
- google drive
- google news
- google flights
- google earth
- google photos
- google calendar
- google slides

At the bottom of the search interface are two buttons: "Google Search" and "I'm Feeling Lucky". A small text link "Report inappropriate predictions" is located at the very bottom. To the right of the search interface, there is a large hand cursor icon pointing towards the bottom right corner.

# 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

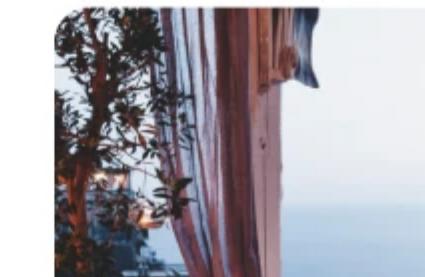
Nestled in the hills above Positano, La Tagliata provides panoramic views of the Amalfi Coast. This family-run restaurant is renowned for its authentic Italian dishes, particularly handmade pastas and grilled meats. Due to its popularity, especially on weekends, it's advisable to book in advance.

LA TAGLIATA



## Il Ritrovo

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



# Search really hasn't changed

# 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/](http://www.nist.gov/standards/)

### Wolves Outside Your Data

[www.janerodgers.blog/wolves](http://www.janerodgers.blog/wolves)

### Digital Protection

[www.clearlaw.net/digital](http://www.clearlaw.net/digital)

# Keyword Search

## Query

Find websites  
explaining data privacy

## Documents

### Data Encryption Standards

[www.nist.gov/standards/](http://www.nist.gov/standards/)



### Wolves Outside Your Data

[www.janerodgers.blog/wolves](http://www.janerodgers.blog/wolves)



### Digital Protection

[www.clearlaw.net/digital](http://www.clearlaw.net/digital)



# Semantic Search

## Query

Find websites  
explaining data privacy

## Documents

**Data Encryption Standards**  
[www.nist.gov/standards/](http://www.nist.gov/standards/)

**Wolves Outside Your Data**  
[www.janerodgers.blog/wolves](http://www.janerodgers.blog/wolves)

**Digital Protection**  
[www.clearlaw.net/digital](http://www.clearlaw.net/digital)

# Semantic Search

## Query

Find websites  
explaining data privacy

## Documents

### Data Encryption Standards

[www.nist.gov/standards/](http://www.nist.gov/standards/)



### Wolves Outside Your Data

[www.janerodgers.blog/wolves](http://www.janerodgers.blog/wolves)



### Digital Protection

[www.clearlaw.net/digital](http://www.clearlaw.net/digital)



# Instruction-based Search

## Query

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

## Documents

### Data Encryption Standards

[www.nist.gov/standards/](http://www.nist.gov/standards/)

### Wolves Outside Your Data

[www.janerodgers.blog/wolves](http://www.janerodgers.blog/wolves)

### Digital Protection

[www.clearlaw.net/digital](http://www.clearlaw.net/digital)

# Instruction-based Search

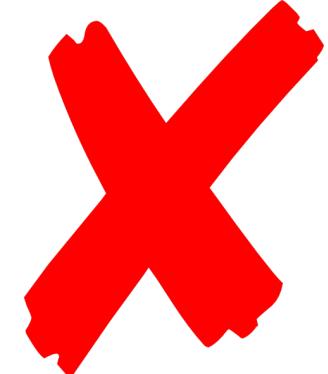
## Query

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

## Documents

### Data Encryption Standards

[www.nist.gov/standards/](http://www.nist.gov/standards/)



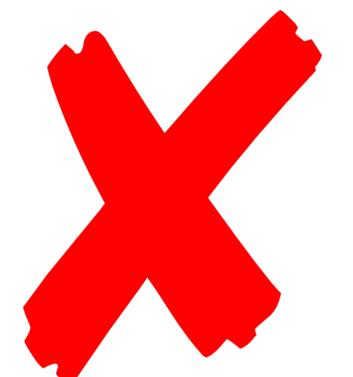
### Wolves Outside Your Data

[www.janerodgers.blog/wolves](http://www.janerodgers.blog/wolves)



### Digital Protection

[www.clearlaw.net/digital](http://www.clearlaw.net/digital)



# Instruction-based Search

## Query

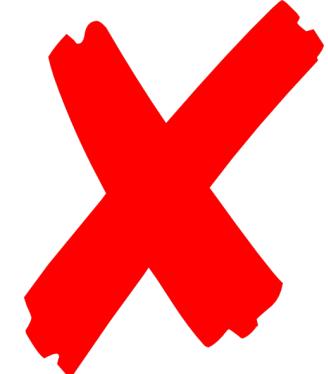
Find websites  
explaining data privacy  
**and uses extended  
metaphors**

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

## Documents

**Data Encryption Standards**

[www.nist.gov/standards/](http://www.nist.gov/standards/)



**Wolves Outside Your Data**

[www.janerodgers.blog/wolves](http://www.janerodgers.blog/wolves)



**Digital Protection**

[www.clearlaw.net/digital](http://www.clearlaw.net/digital)



# Instruction-based Search

## Query

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

## Documents

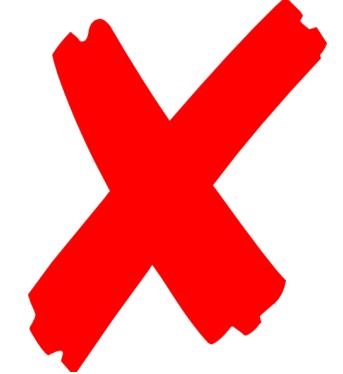
### **Wolves Outside Your Data**

[www.janerodgers.blog/wolves](http://www.janerodgers.blog/wolves)



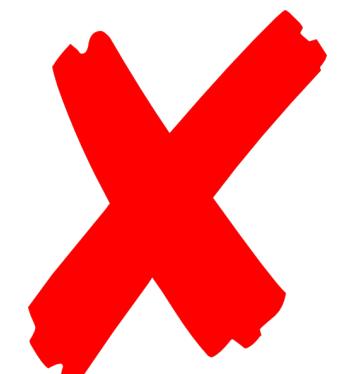
### **Data Encryption Standards**

[www.nist.gov/standards/](http://www.nist.gov/standards/)



### **Digital Protection**

[www.clearlaw.net/digital](http://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](http://www.janerodgers.blog/wolves)



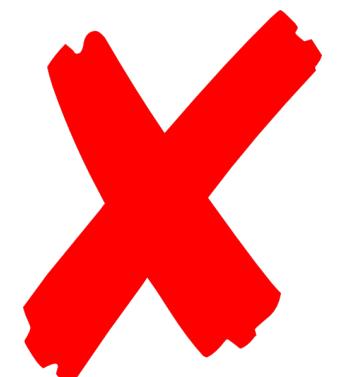
### **Data Encryption Standards**

[www.nist.gov/standards/](http://www.nist.gov/standards/)



### **Digital Protection**

[www.clearlaw.net/digital](http://www.clearlaw.net/digital)



# What is an instruction in IR?

# What is an instruction in IR?

Doc attributes:  
date, length, source

# What is an instruction in IR?

Doc attributes:  
date, length, source

NLU aspects:  
sentiment, style

# What is an instruction in IR?

Doc attributes:  
date, length, source

NLU aspects:  
sentiment, style

Logical conditions  
AND/OR/NOT

# 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

We're so used to prompting LMs

Let's use retrievers the **same way**





# Promptriever fast embedder



**Promptriever**  
**fast embedder**



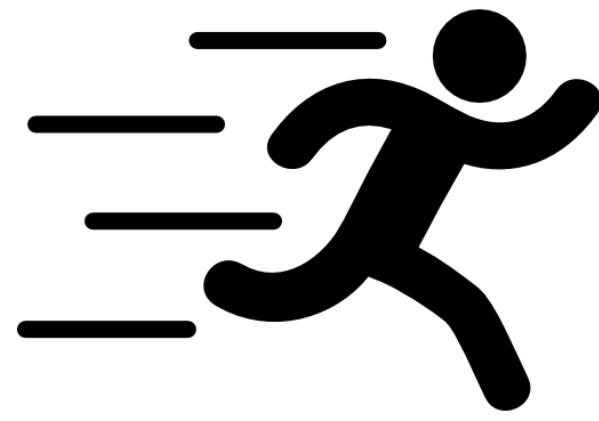
**Rank1**  
**strong but slow**



**Promptriever**  
**fast embedder**



**Rank1**  
**strong but slow**



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



Orion Weller



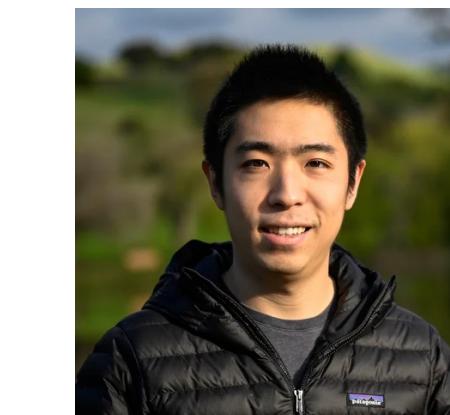
Benjamin  
Van Durme



Dawn Lawrie



Ashwin Paranjape



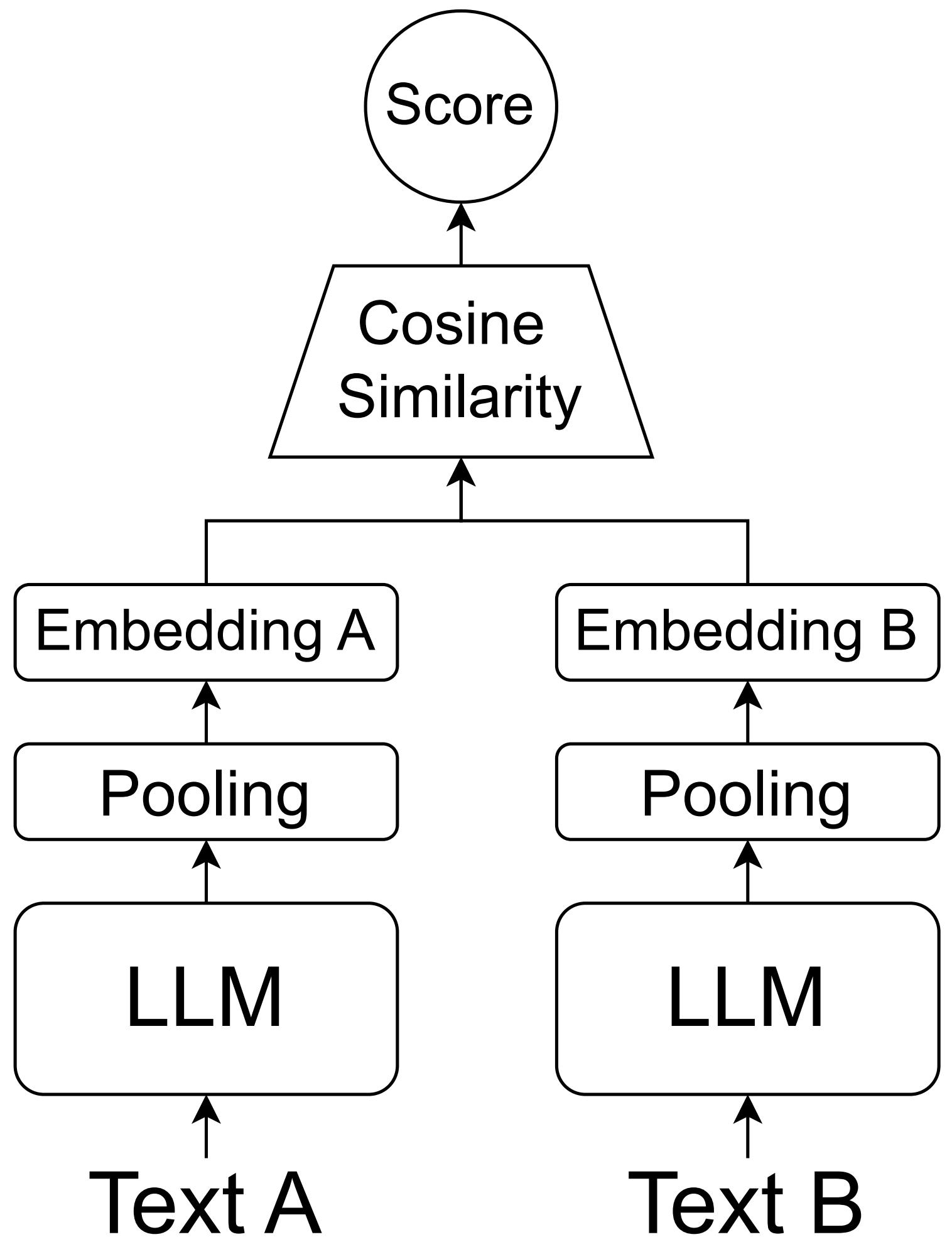
Yuhao Zhang



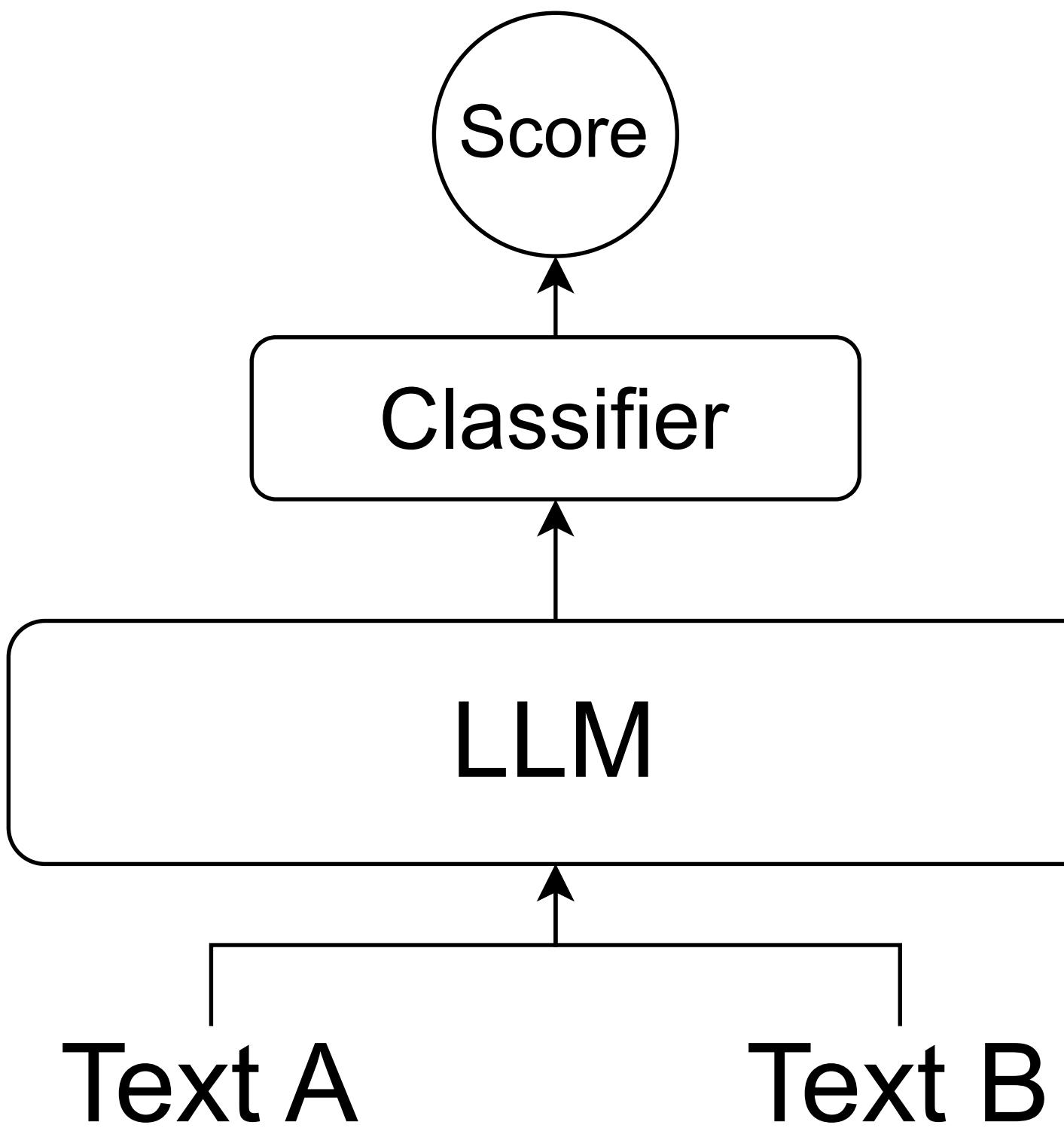
Jack Hessel



# Bi-Encoder



# Cross-Encoder

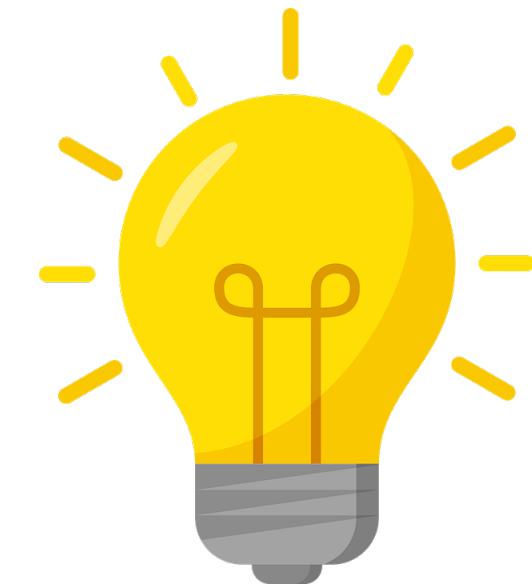


# Can we make promptable retrievers?

Can bi-encoders take instructions?

# 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

## 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 ...



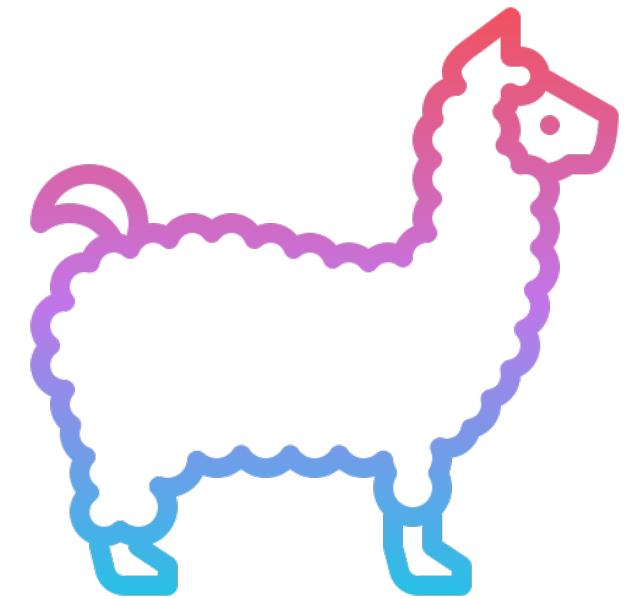
## Instruction

Volcanoes are classified into different types based on their shape and eruption style. A document is relevant if it describes a **specific type of volcano that has not been directly observed erupting, and provides information about its formation or characteristics.**

We have to generate synthetic instructions

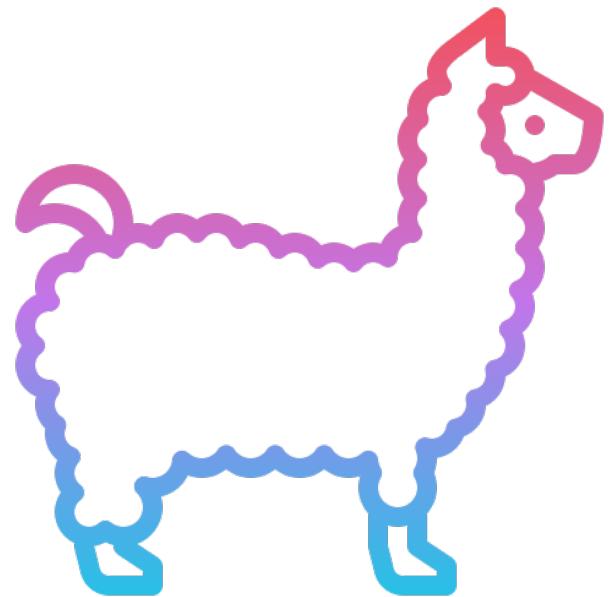
# Experimental Settings

# Experimental Settings



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

# 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.

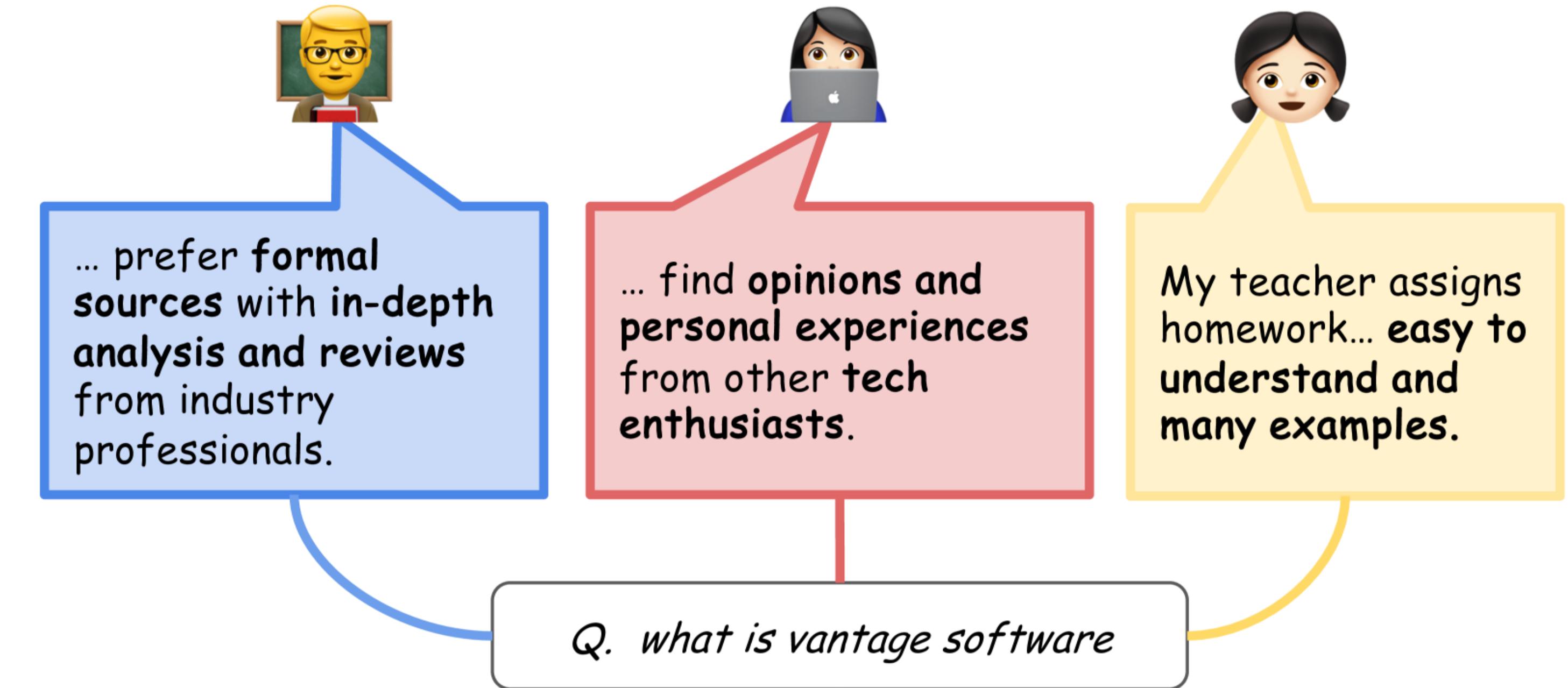
---

# FollowIR

p-MRR ranges  
from -100 to 100

# InstructIR

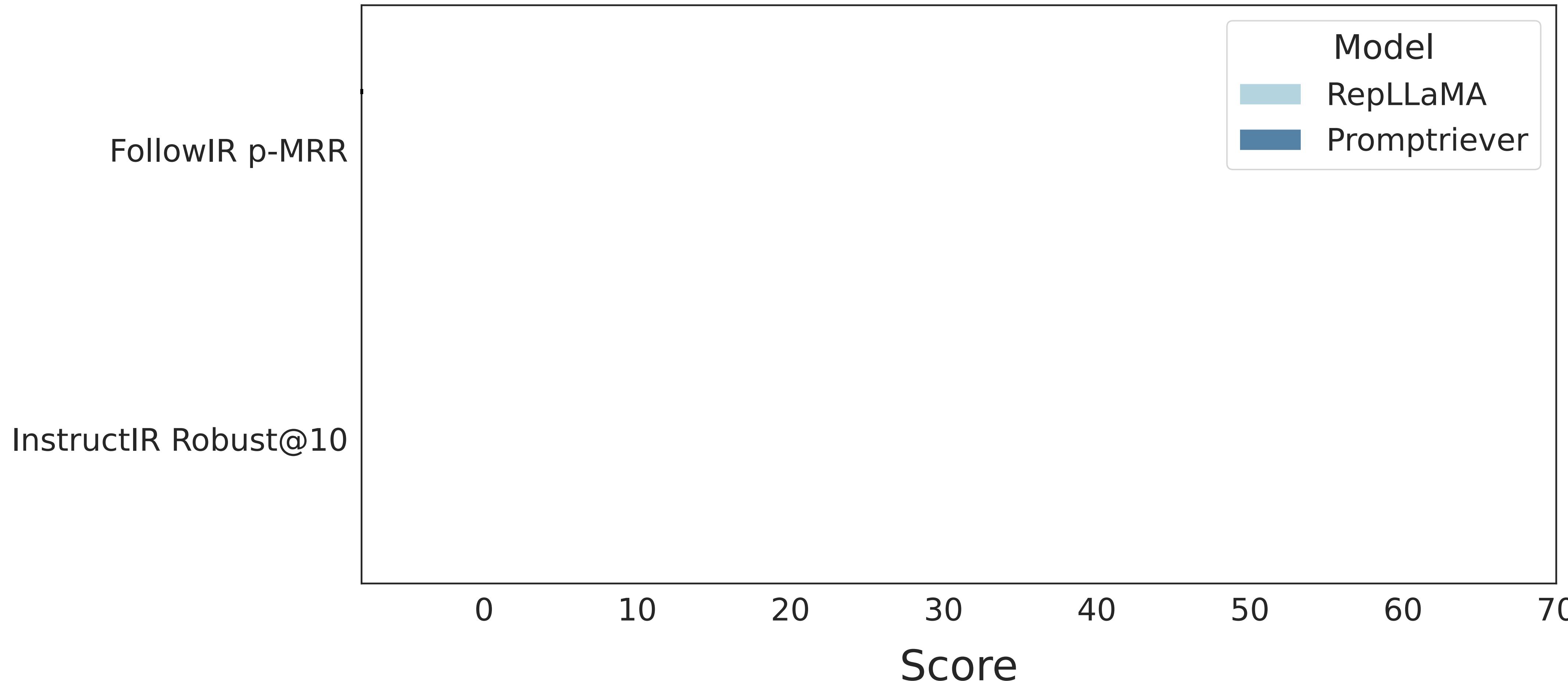
10 personas for each  
MS MARCO query



# Results

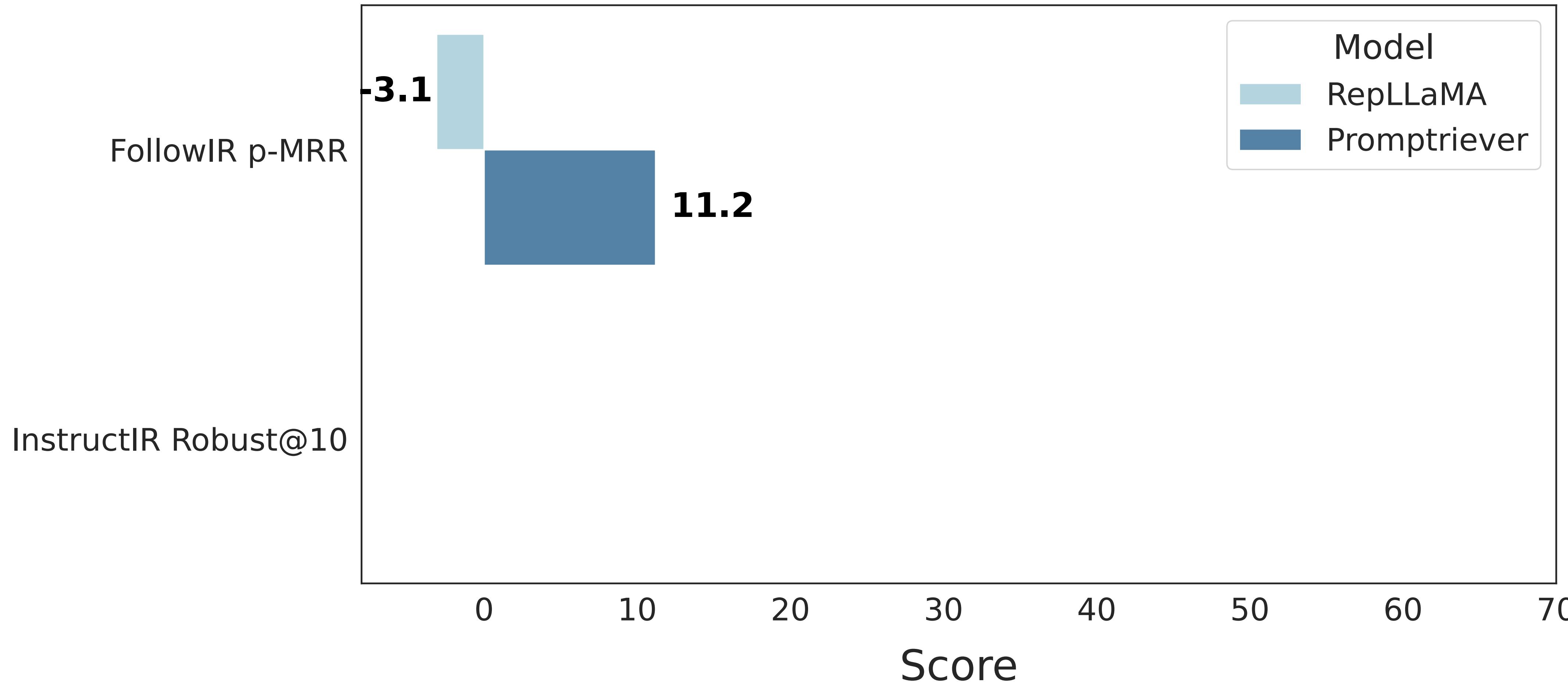
# Results

## Instruction Following



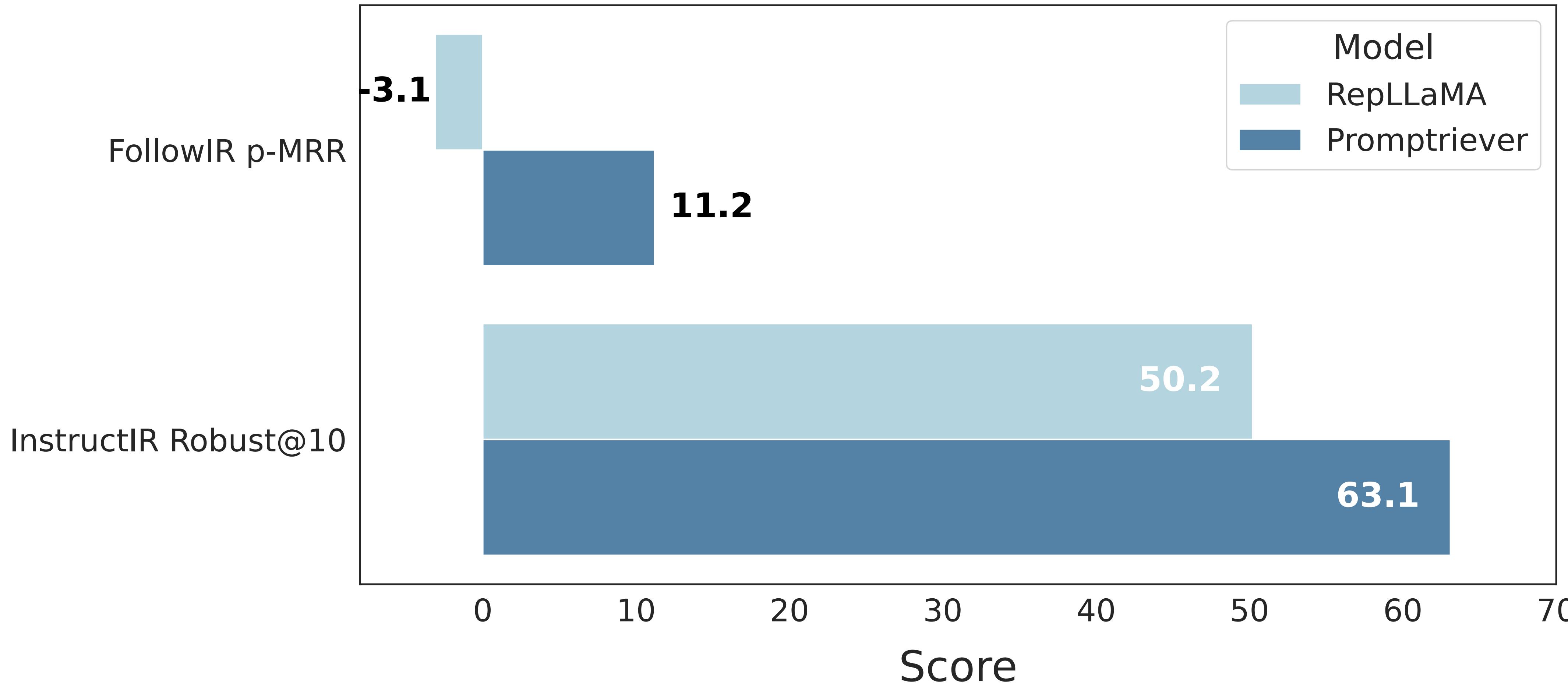
# Results

## Instruction Following



# Results

## Instruction Following



# Results

# Results

# Results

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

# Results

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

- We show with no prompt (standard)

# 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

# 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

# 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

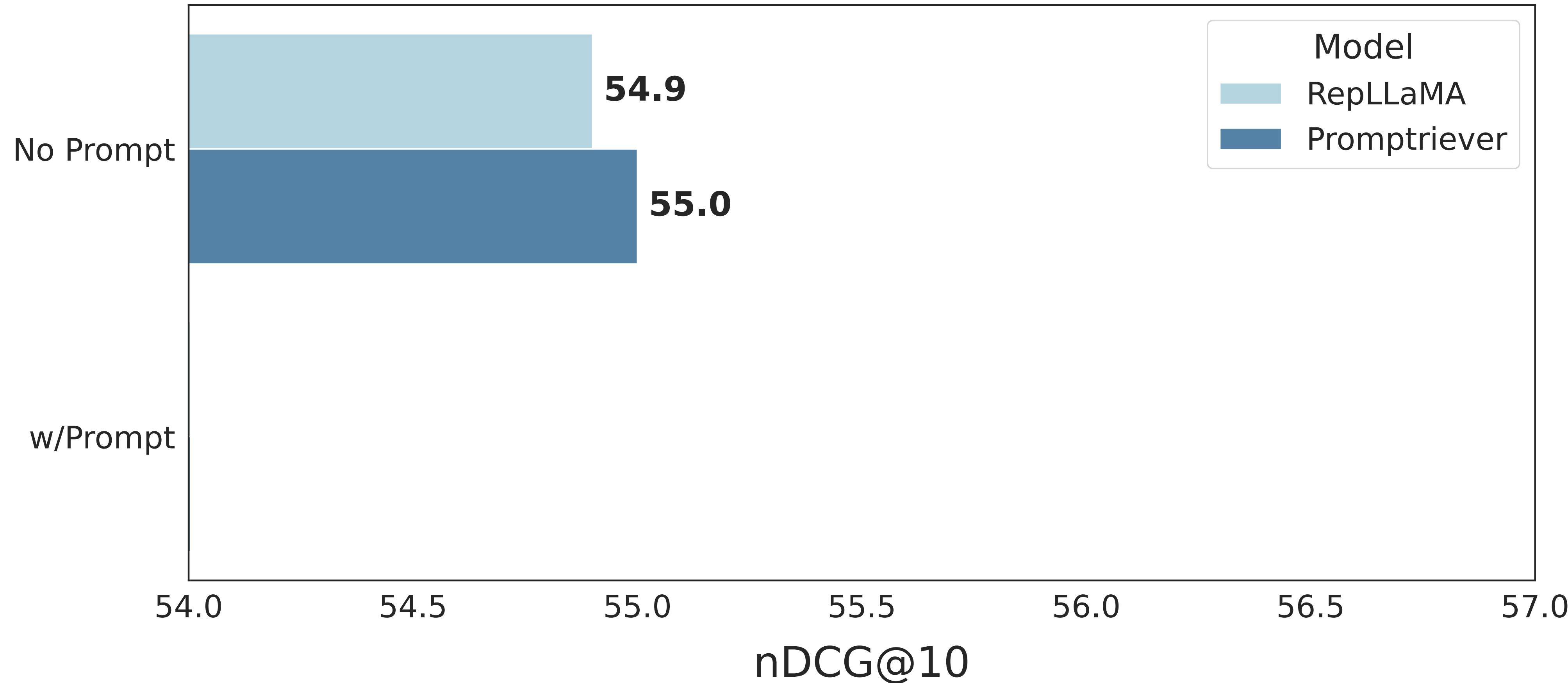
---

- 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.

...

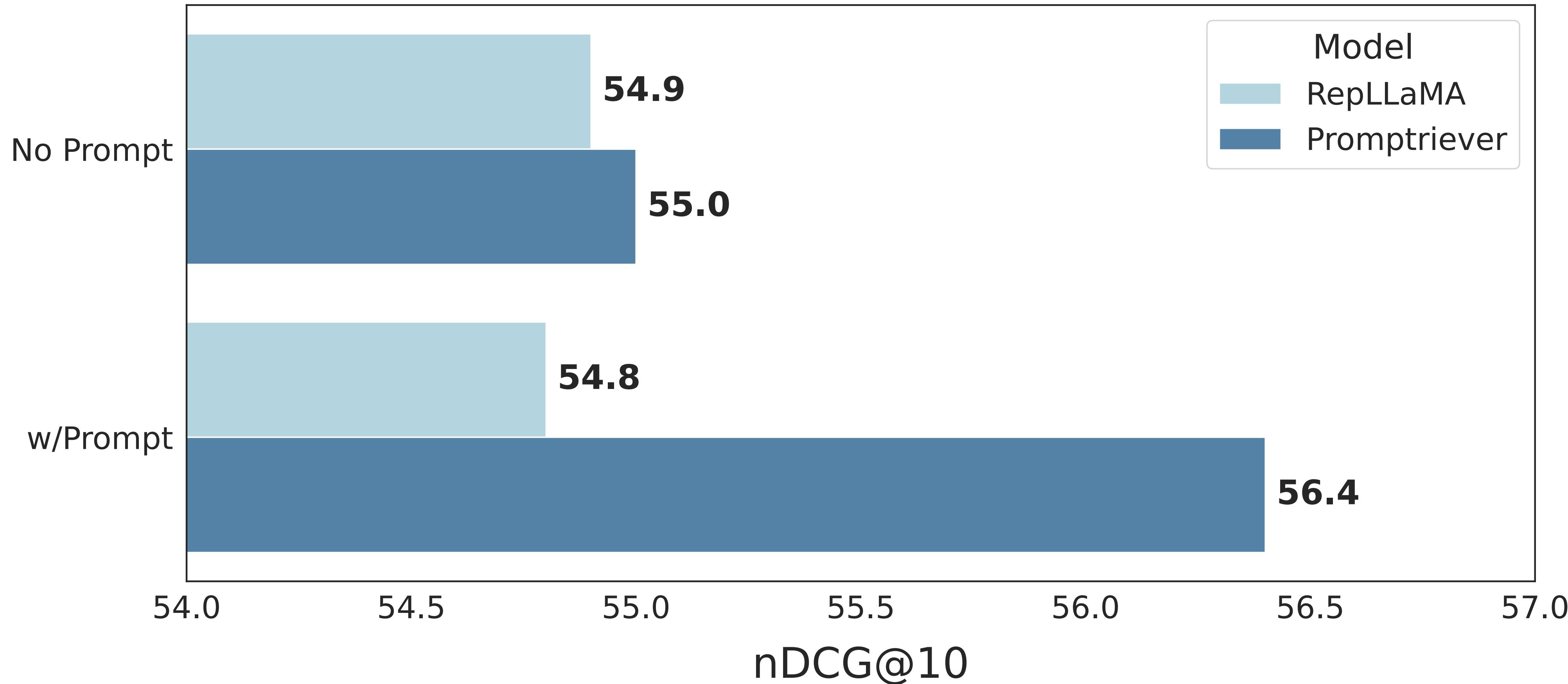
# Results

## BEIR (OOD)

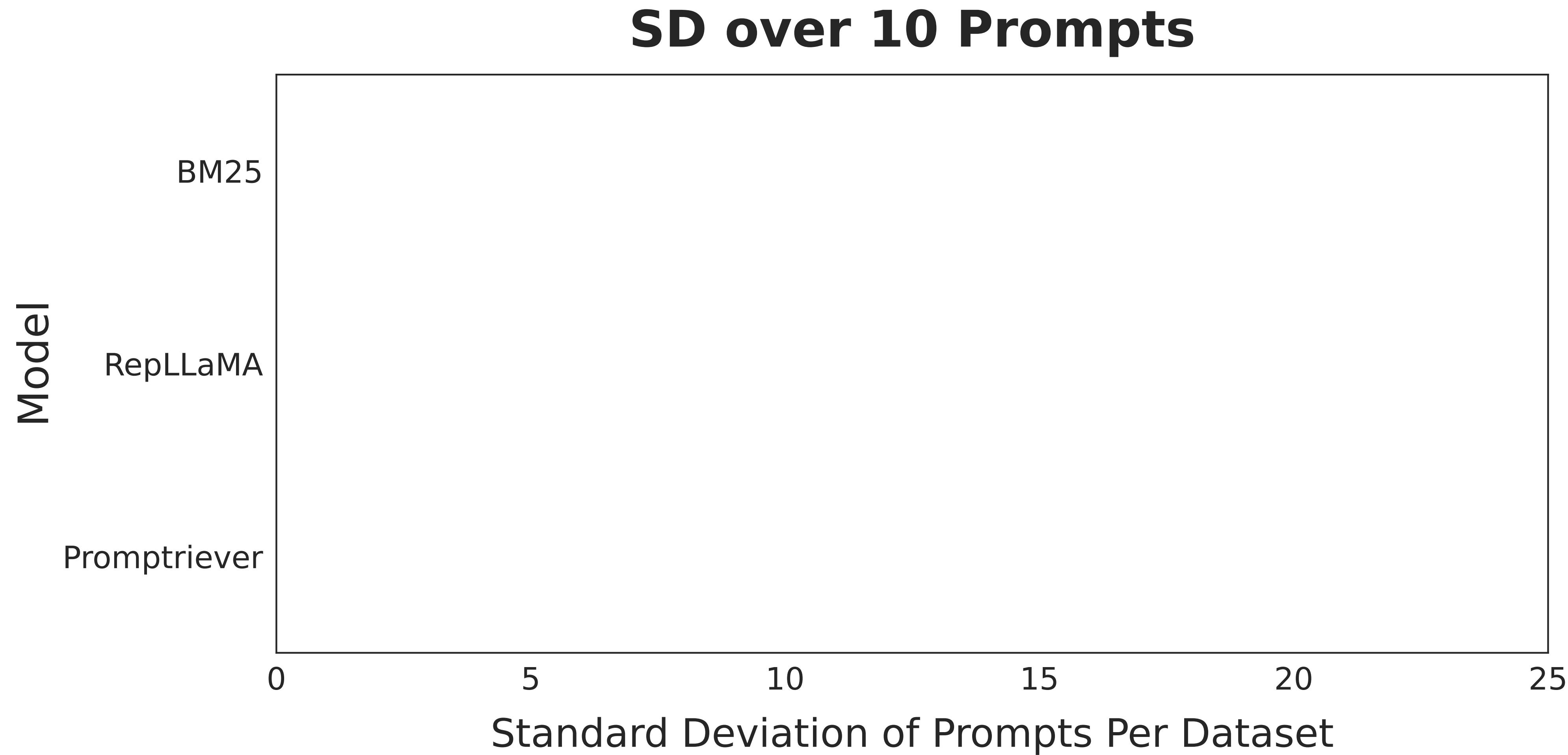


# Results

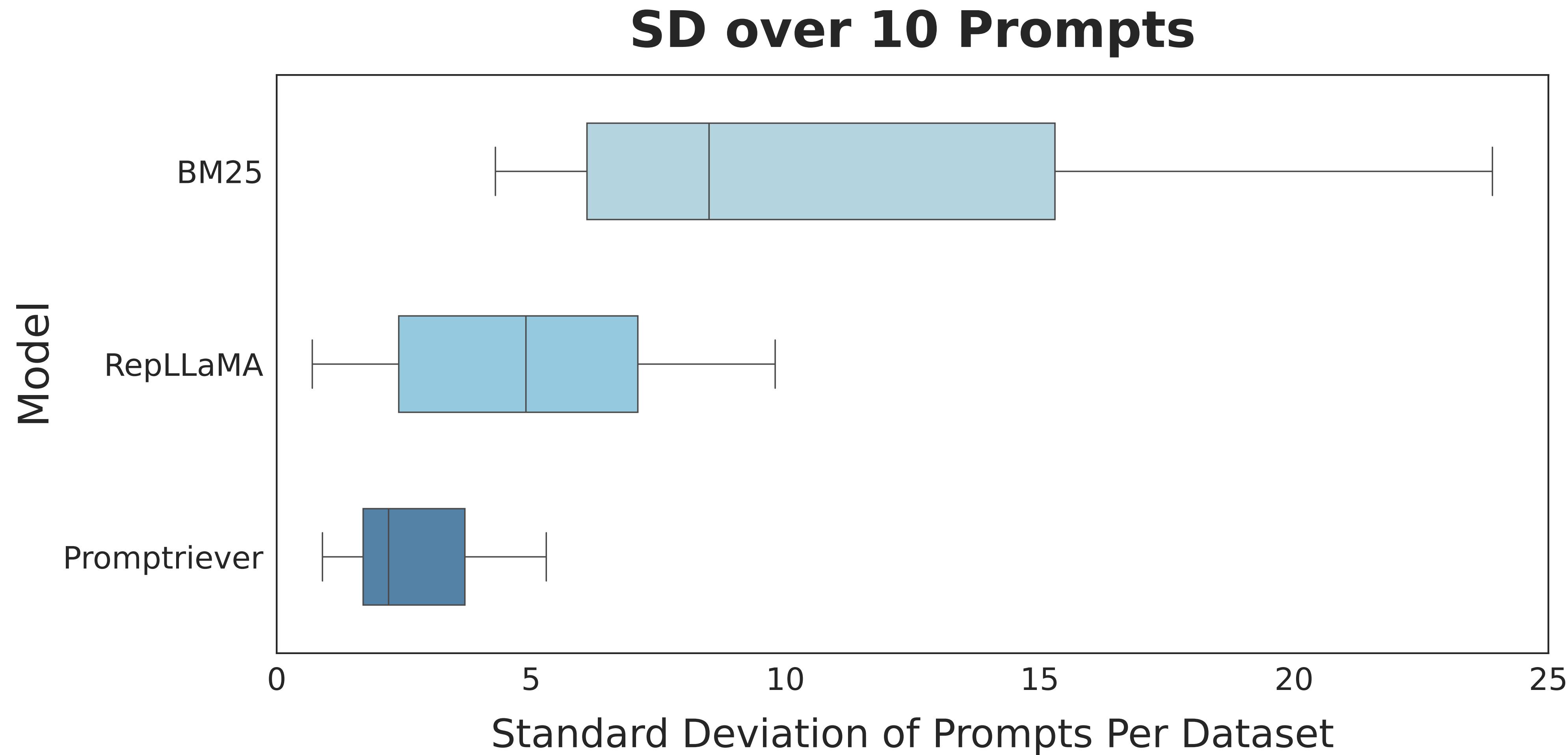
## BEIR (OOD)



# Results



# Results



# Summary

# Summary

- **With the right data**, you can have retrievers that are promptable just like LM

# Summary

- **With the right data**, you can have retrievers that are promptable just like LM
- **Unlocks new types of queries!**

# 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



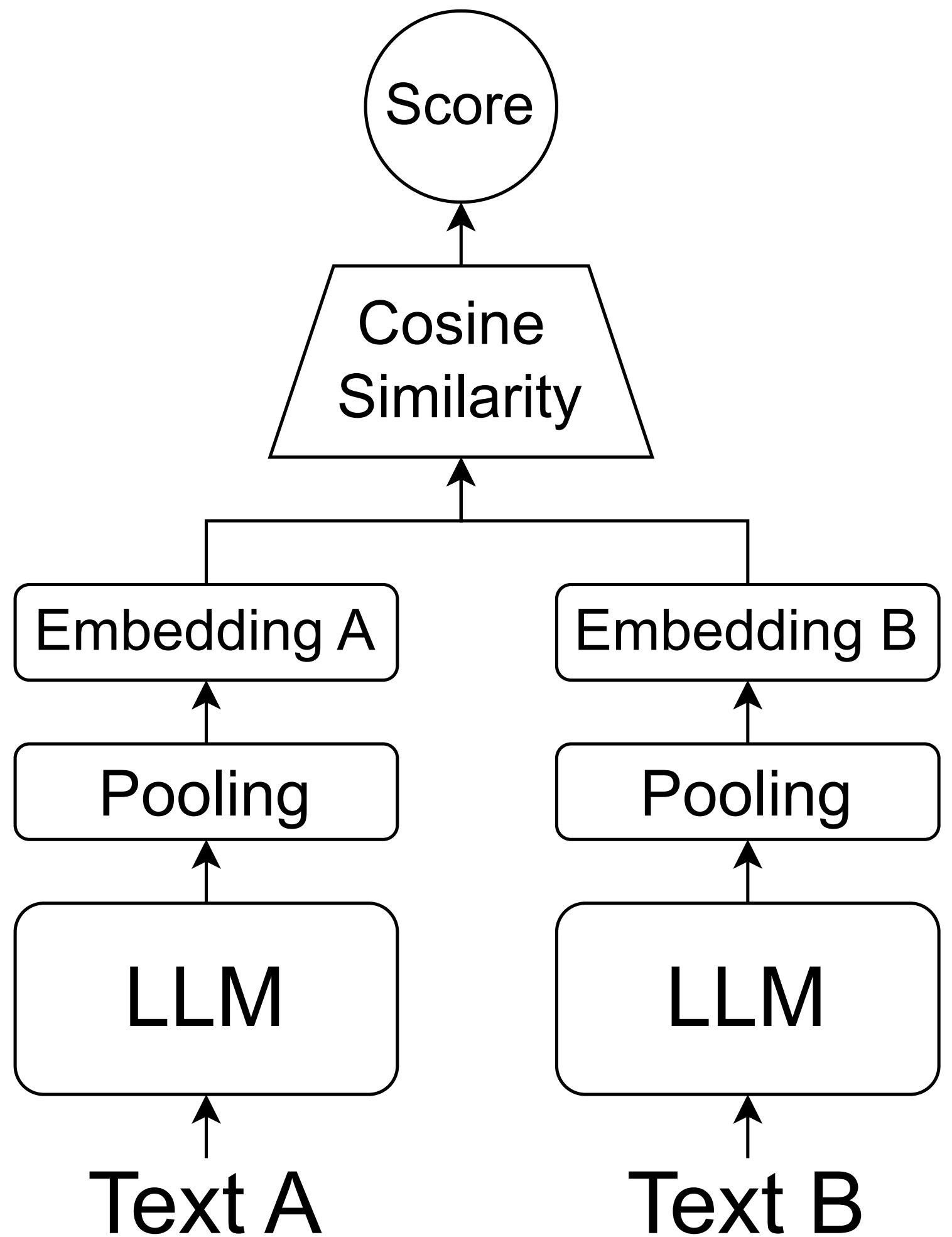
Dawn Lawrie



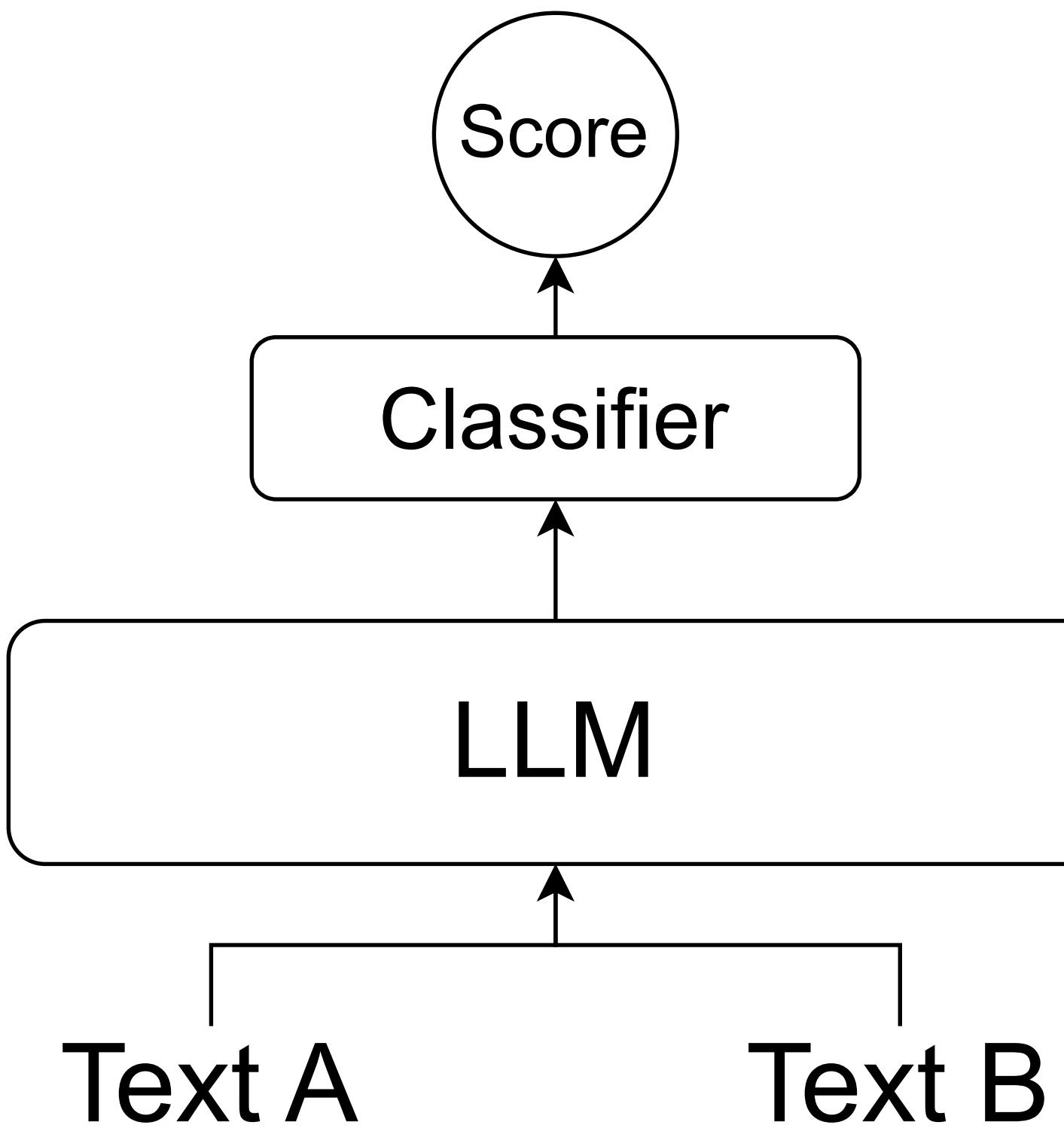
Benjamin  
Van Durme



# Bi-Encoder



# Cross-Encoder



# Test-Time Compute

how many r's in the word strawberry?

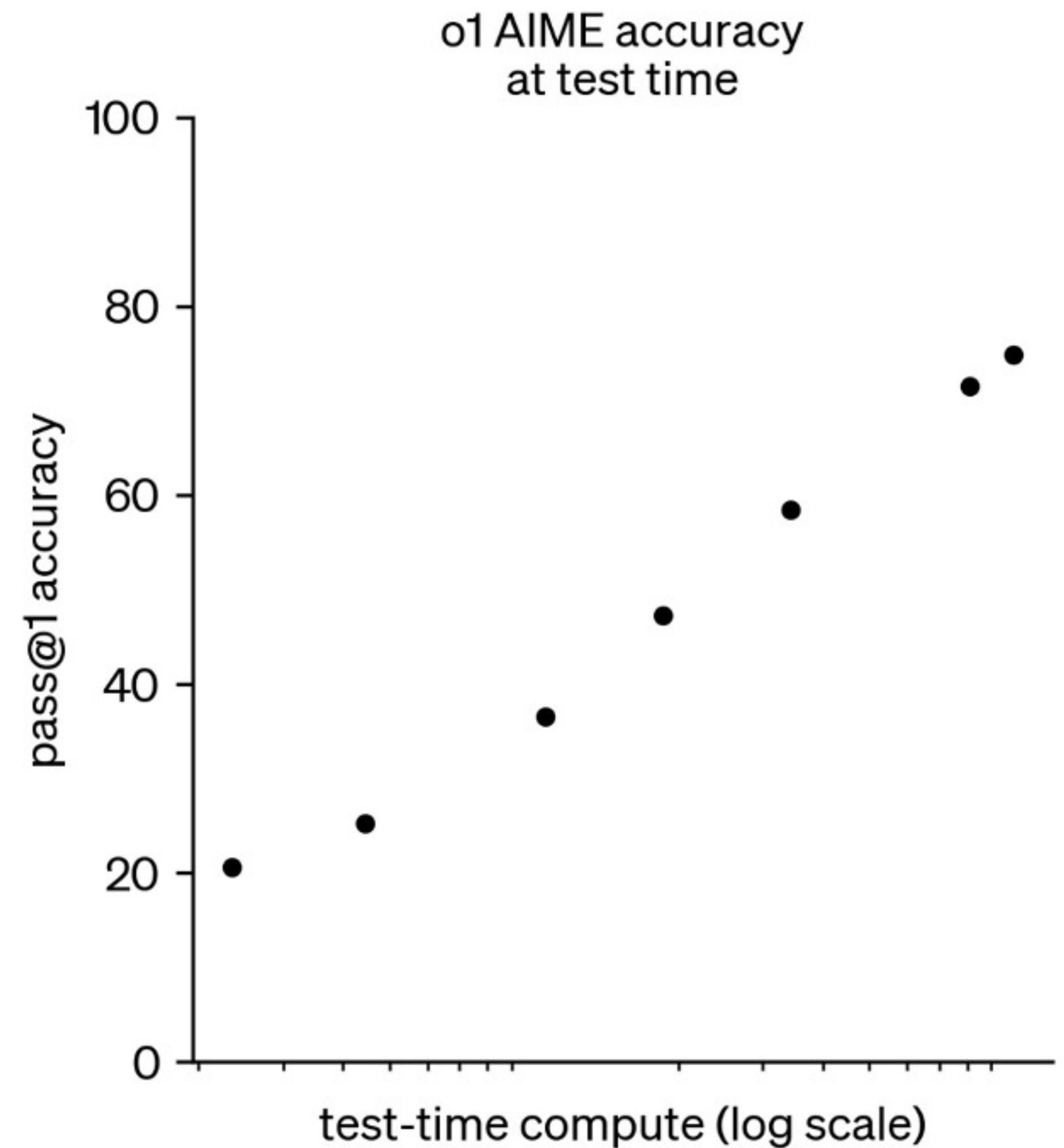
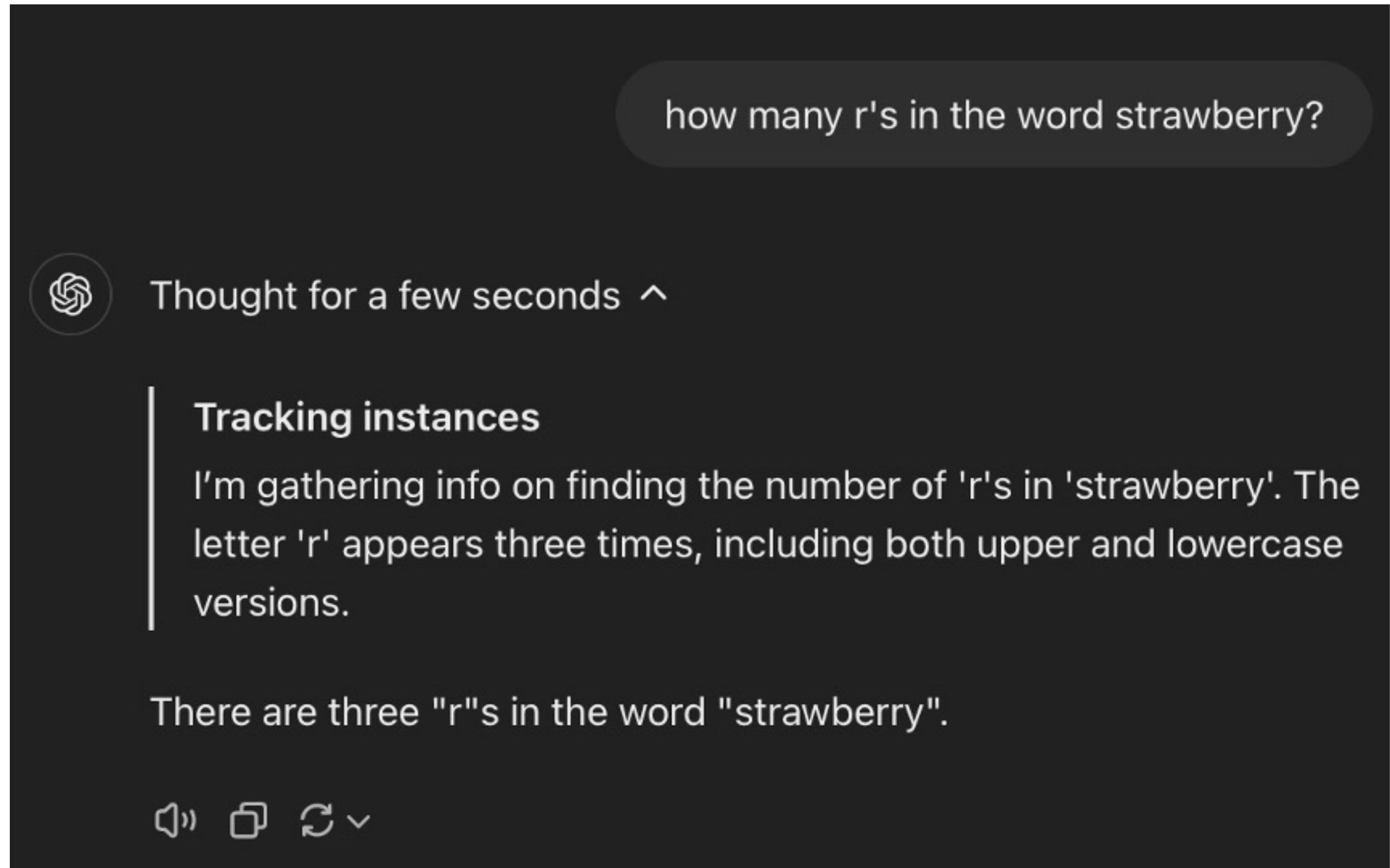
Thought for a few seconds ^

**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".

⟳ ⟲ ⌂ ↴

# 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.

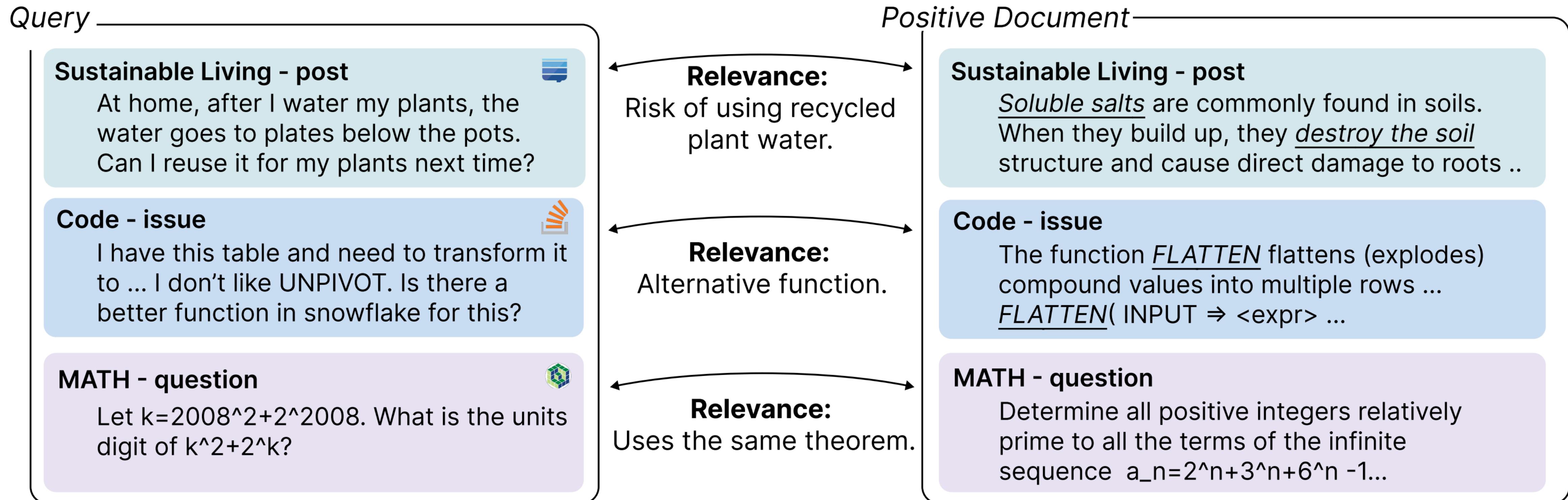
[But wait](#), "varies" might just mean that different individuals have different base colors, not that a single leopard changes its color. The passage doesn't mention anything about changing color; it's just describing their appearance. ....

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

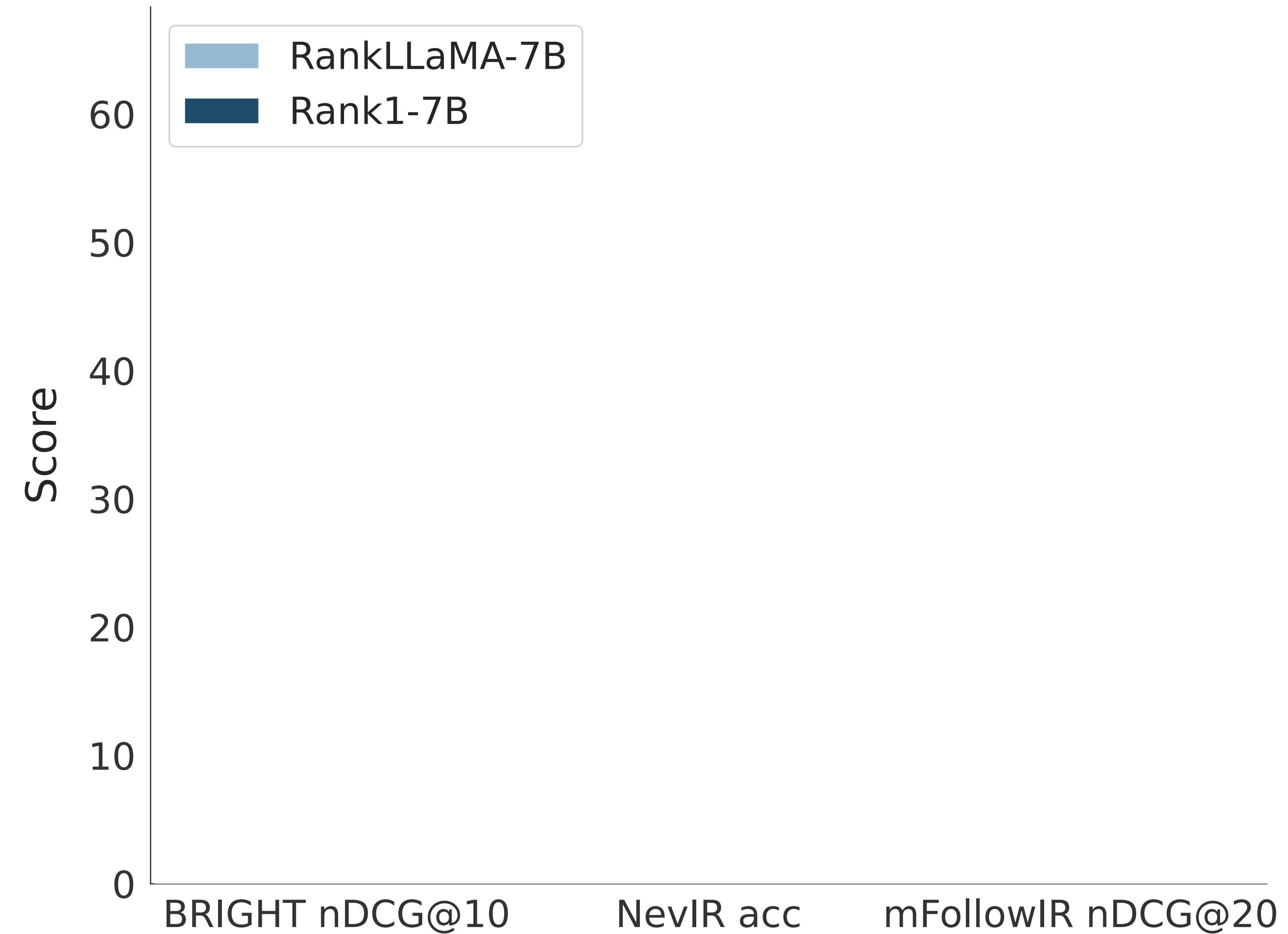
</think> true

# Results

# Results

We evaluate on a broad range of tasks:

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

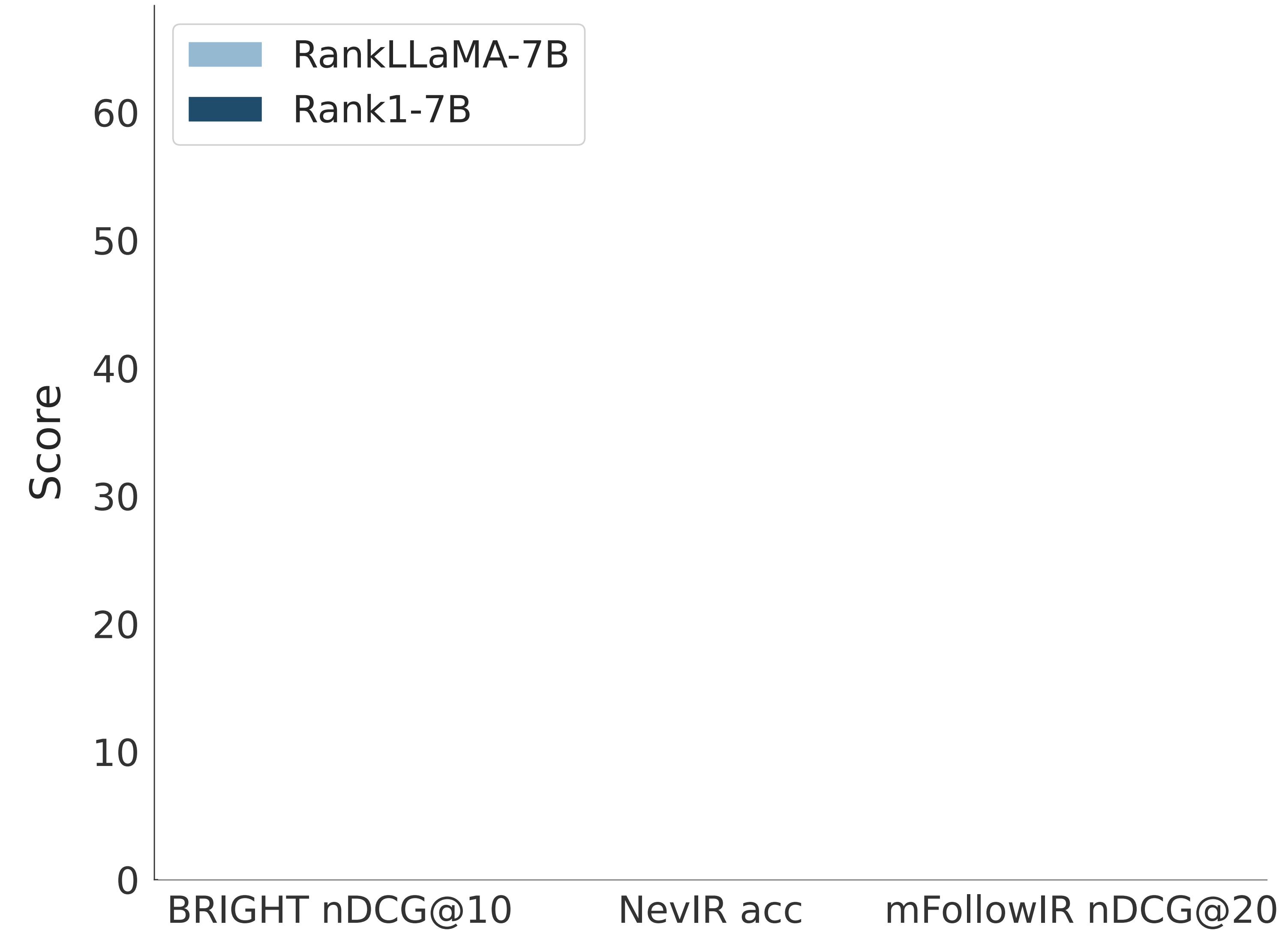


# Results

We evaluate on a broad range of tasks:

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

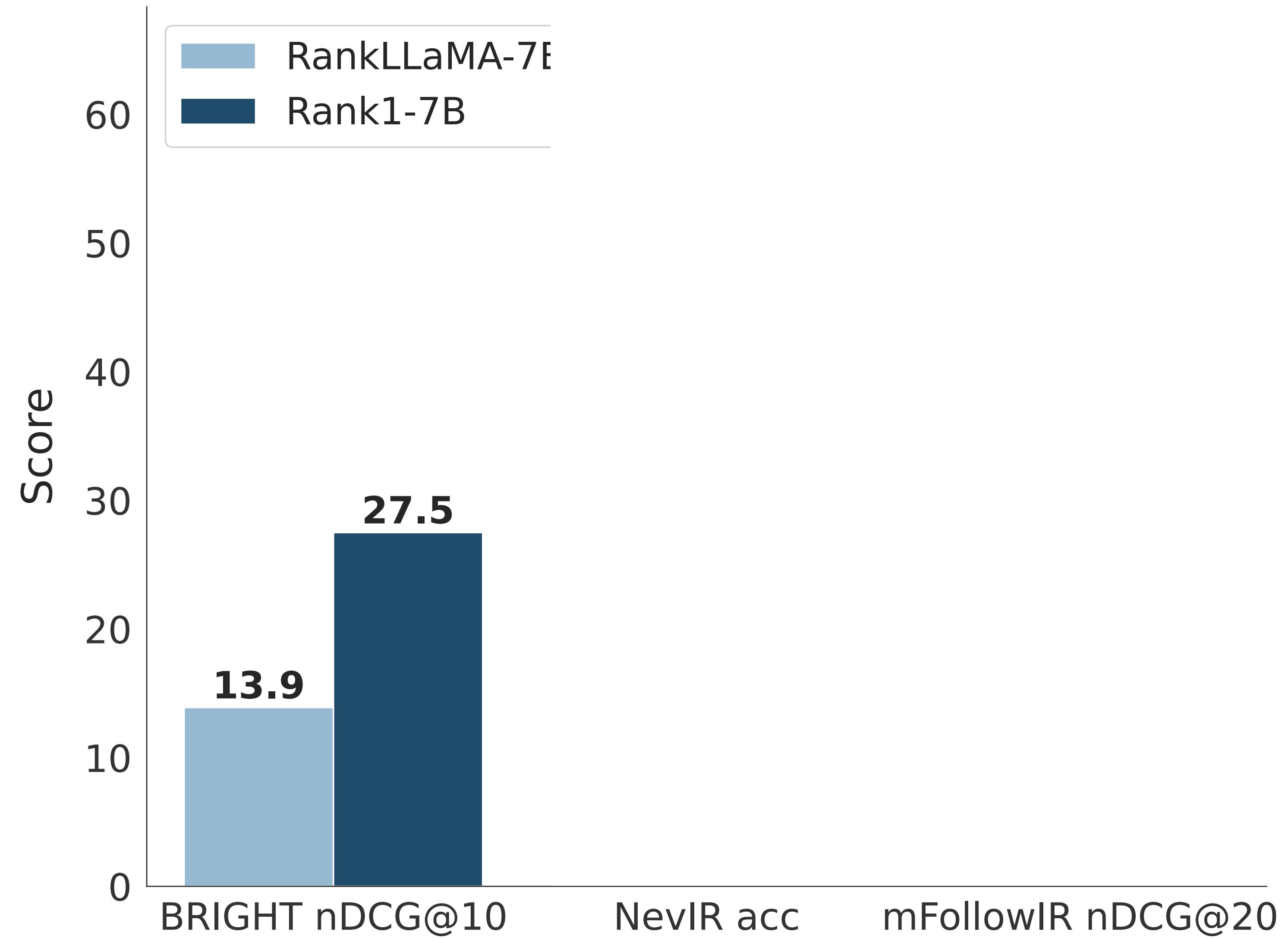
RankLLaMA was trained on **10x more data!**



# Results

We evaluate on a broad range of tasks:

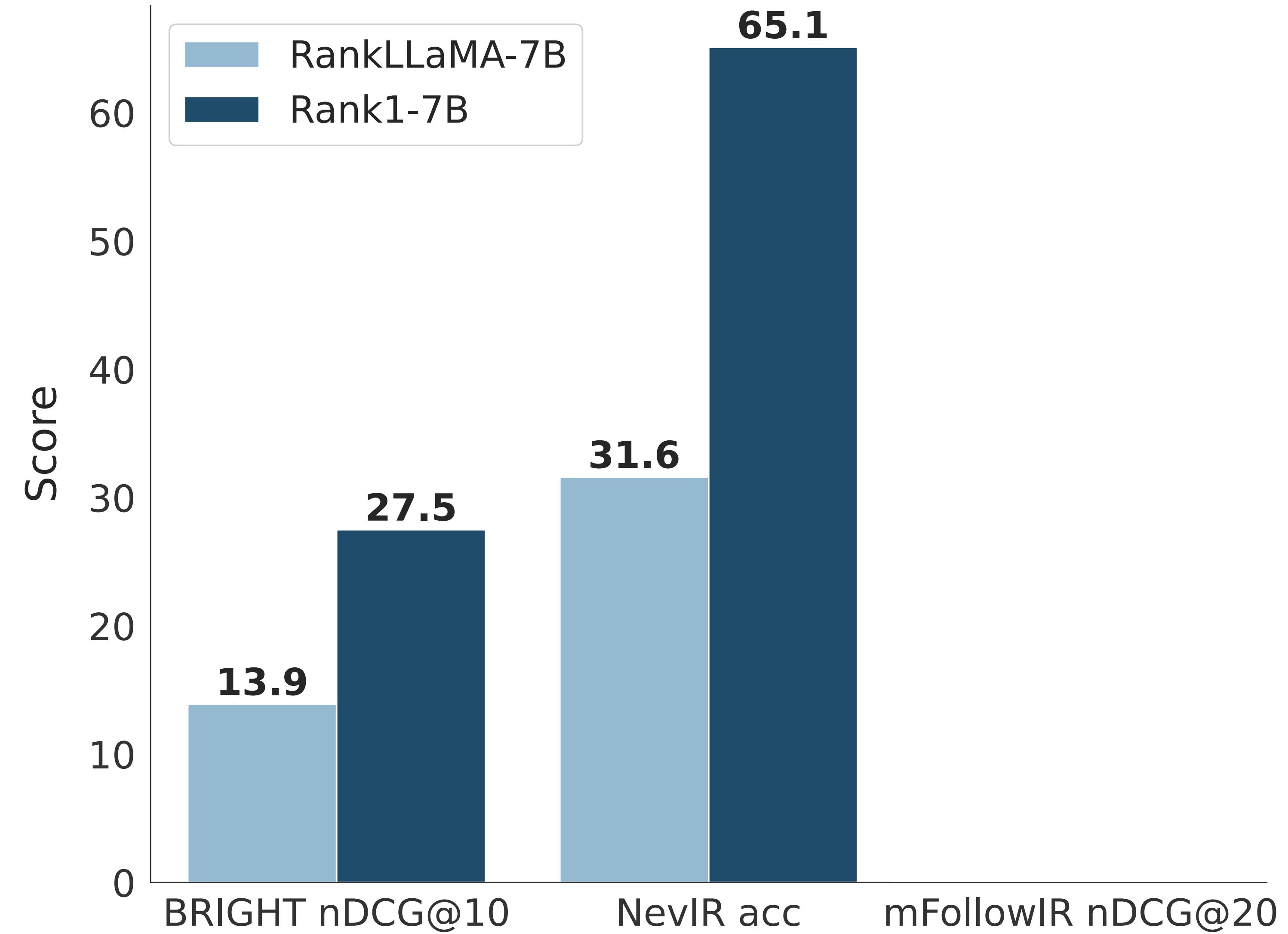
- BRIGHT  
*(reasoning)*
- NevIR *(negation)*
- mFollowIR  
*(instructions)*



# Results

We evaluate on a broad range of tasks:

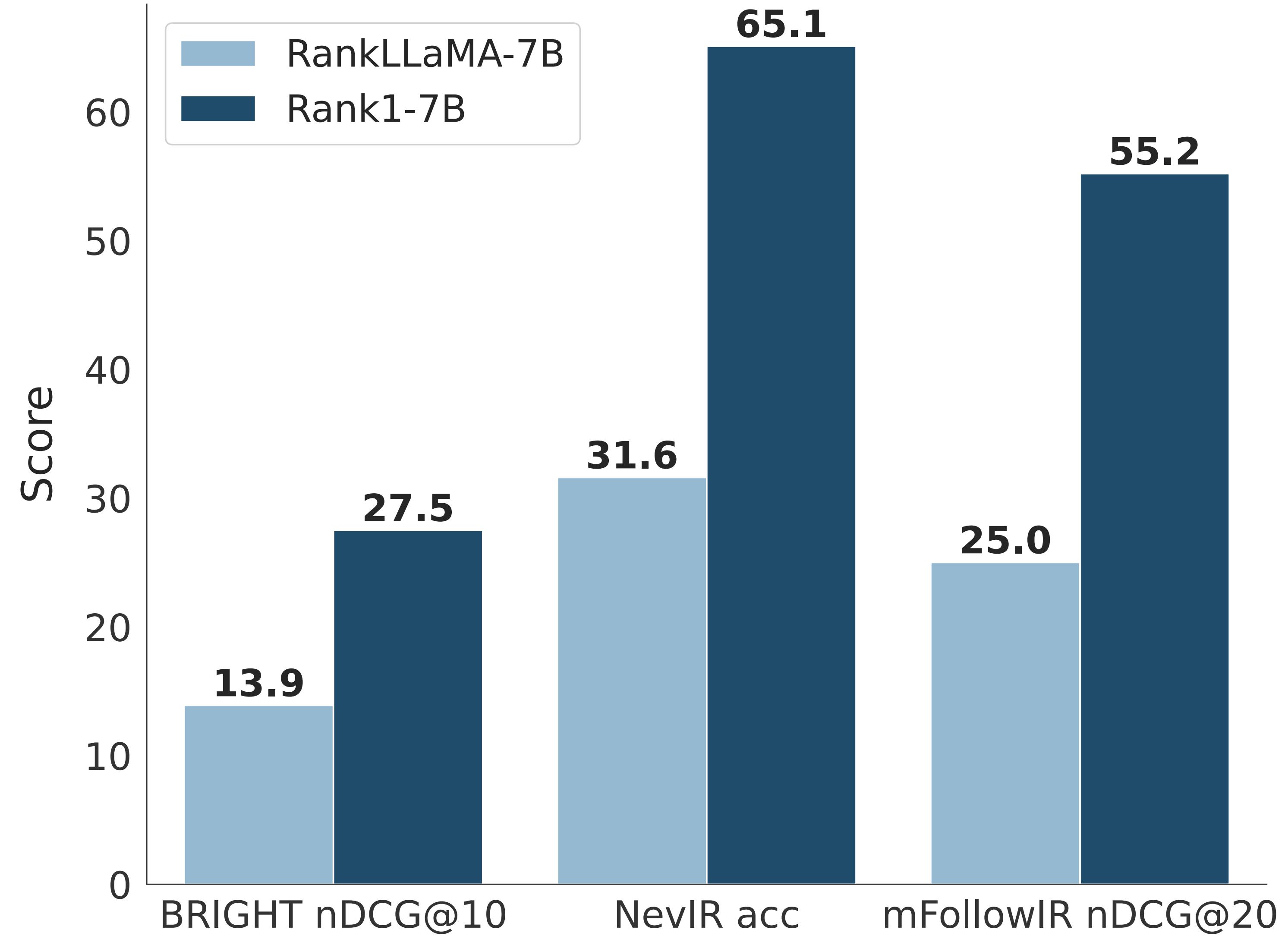
- BRIGHT (*reasoning*)
- NevIR (*negation*)
- mFollowIR (*instructions*)



# Results

We evaluate on a broad range of tasks:

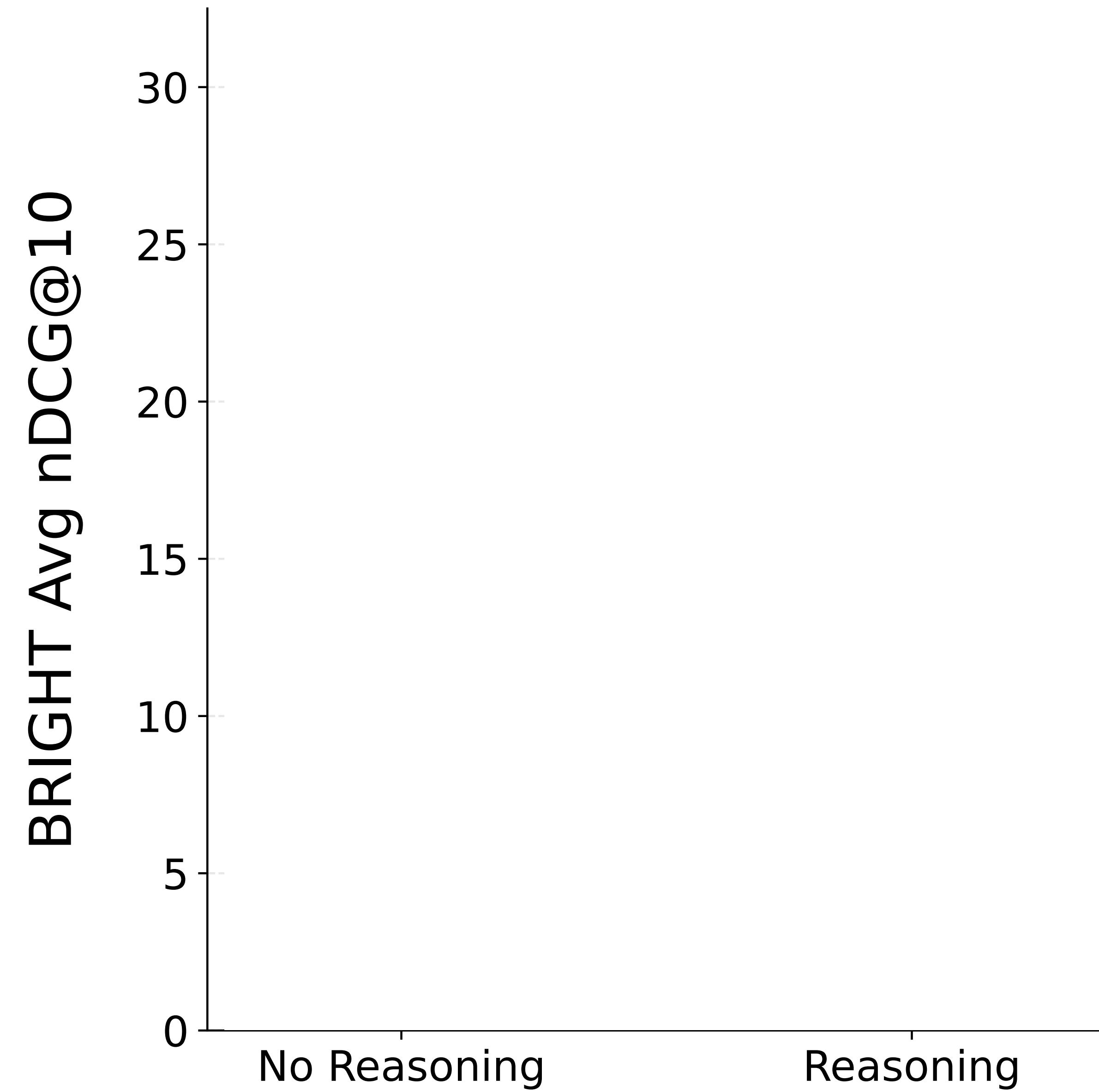
- BRIGHT (*reasoning*)
- NevIR (*negation*)
- mFollowIR (*instructions*)



# Results

What about a direct comparison of the reasoning chain?

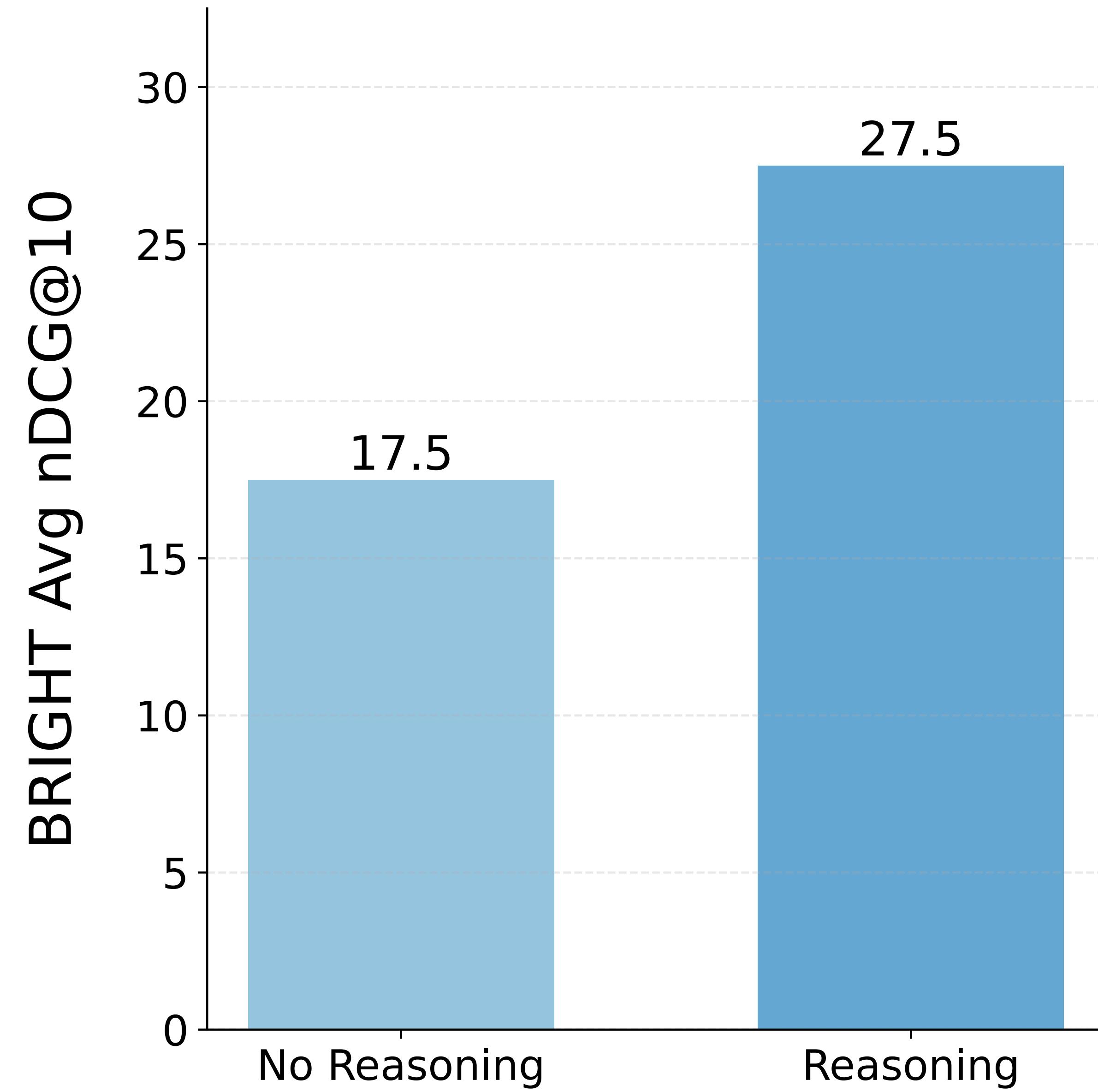
Same data, same model, no chain



# Results

What about a direct comparison of the reasoning chain?

Same data, same model, no chain



# Old Evaluation Data

# Results

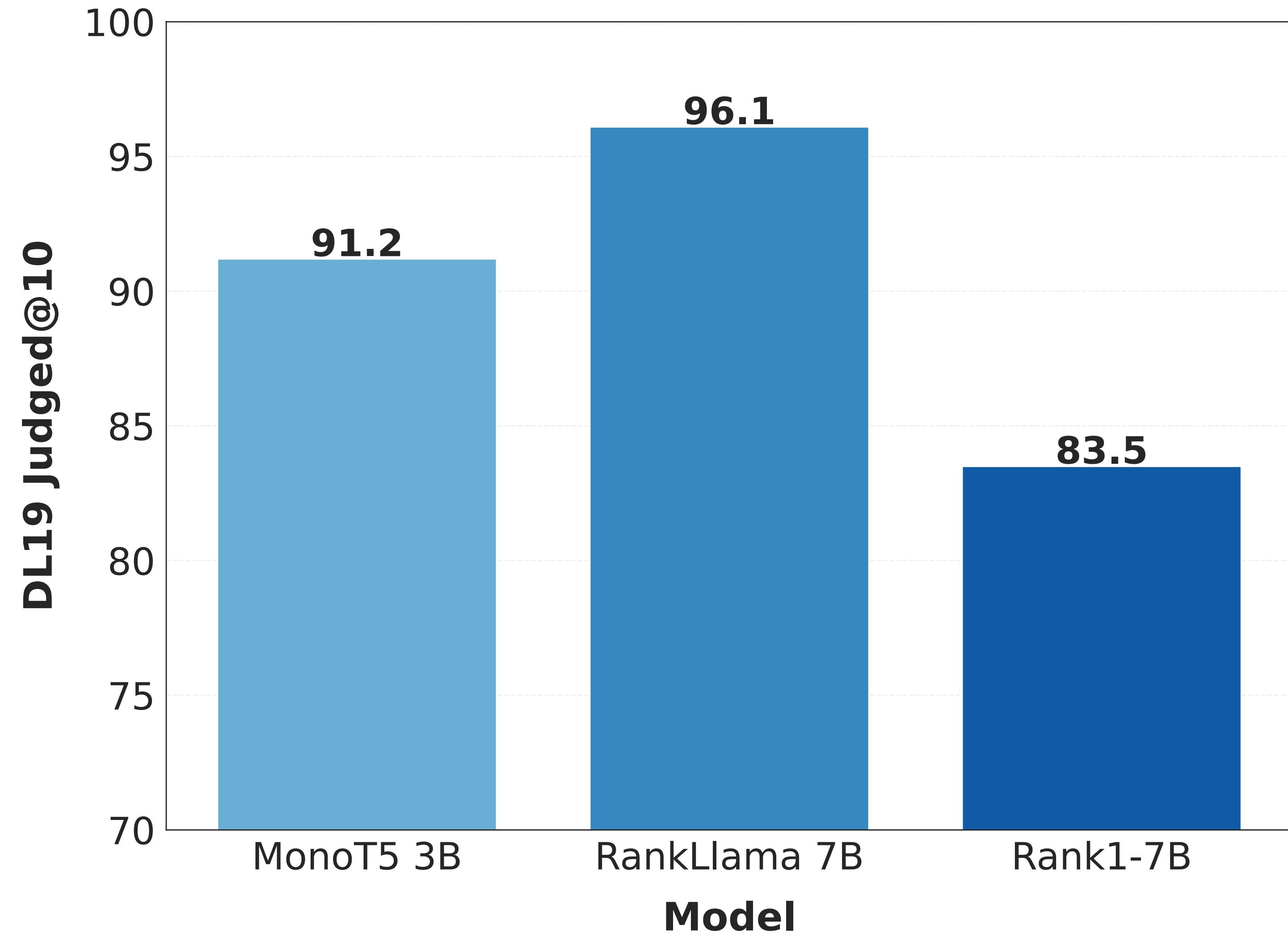
We were surprised by the low initial scores on DL19/DL20

Judged@10 is significant less!

# Results

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

Judged@10 is significant less!

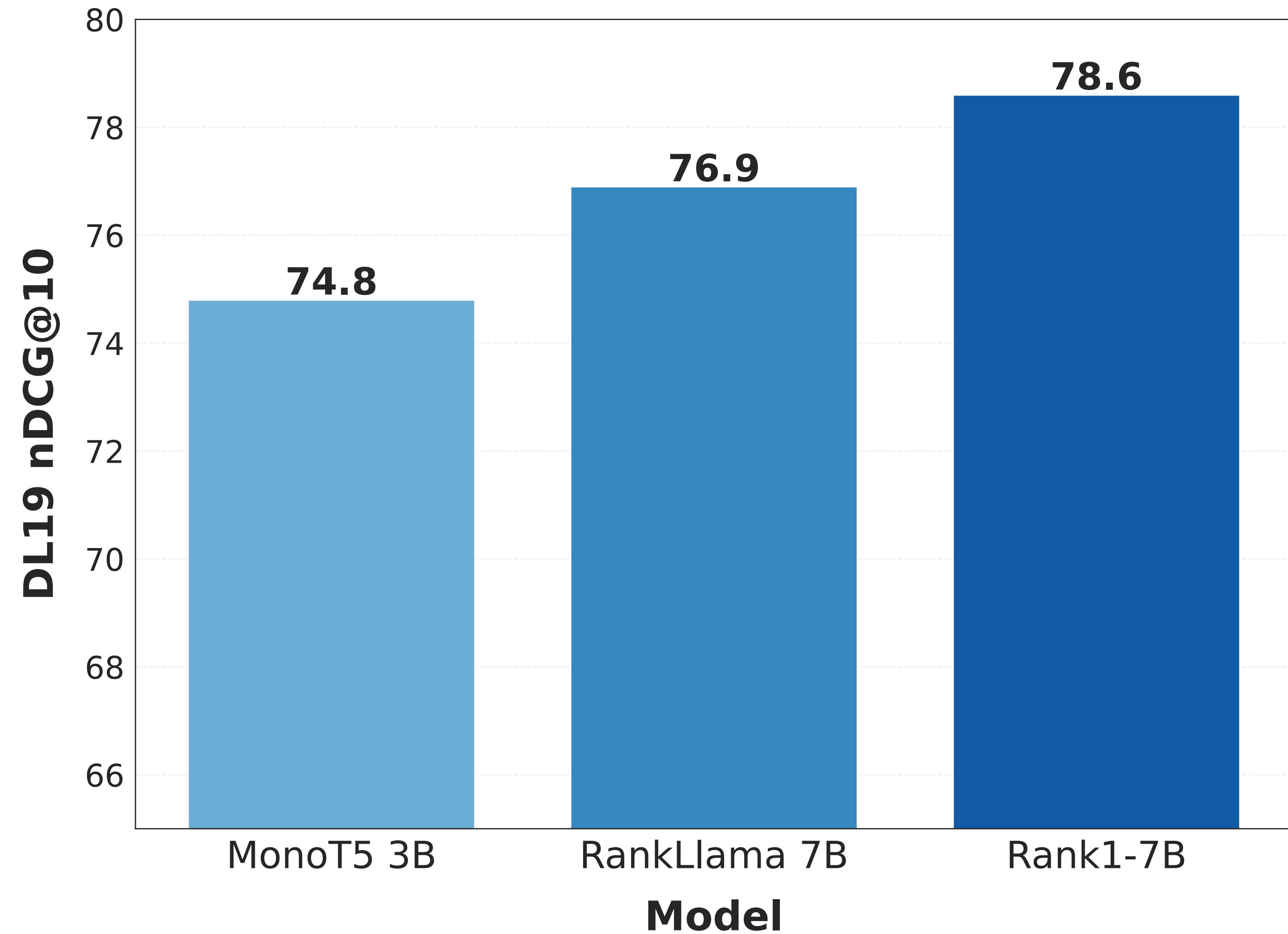


# Results

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

# Results

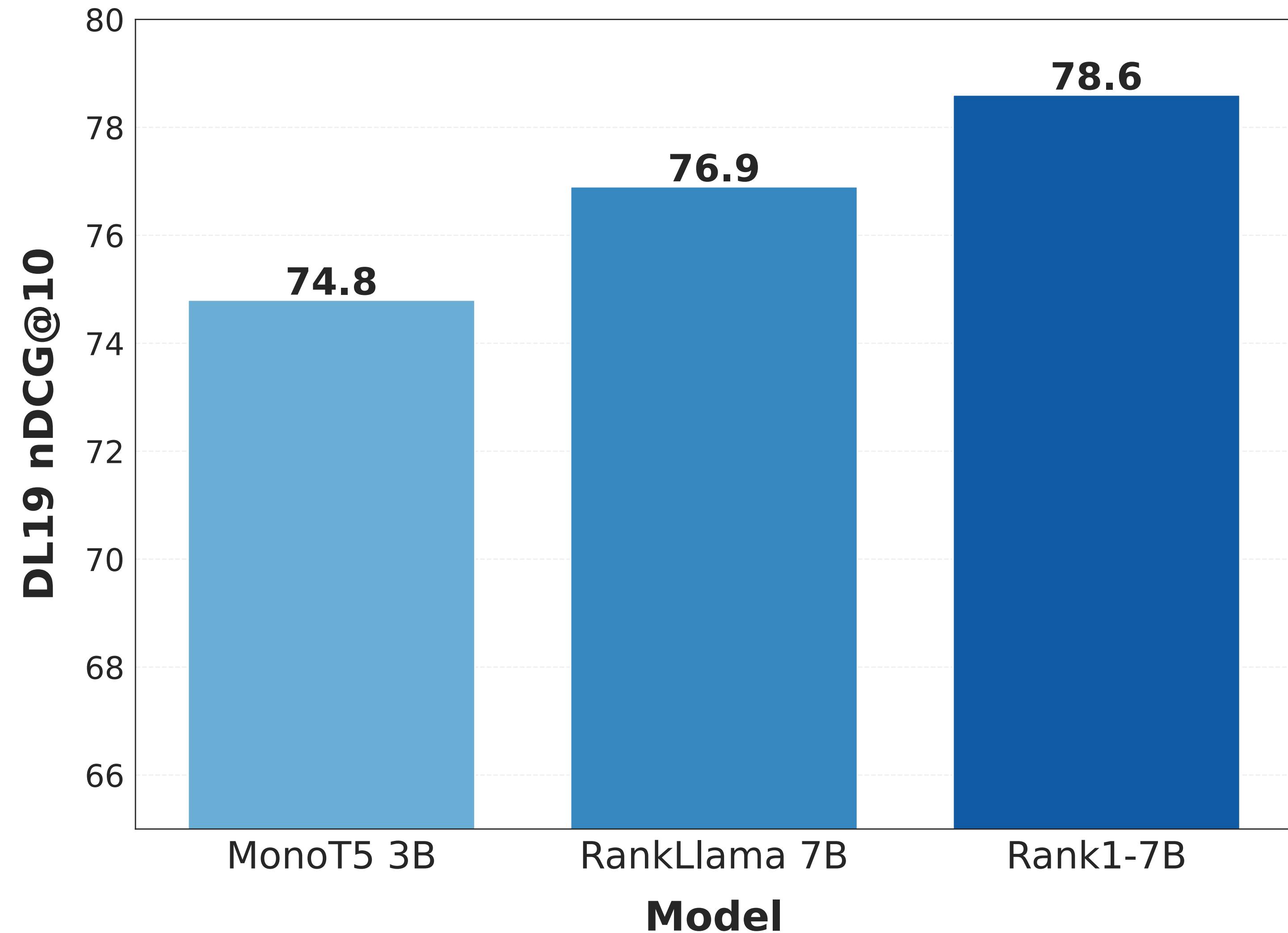
We re-judged the  
unjudged +  
incorrect preds  
for all models



# Results

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

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

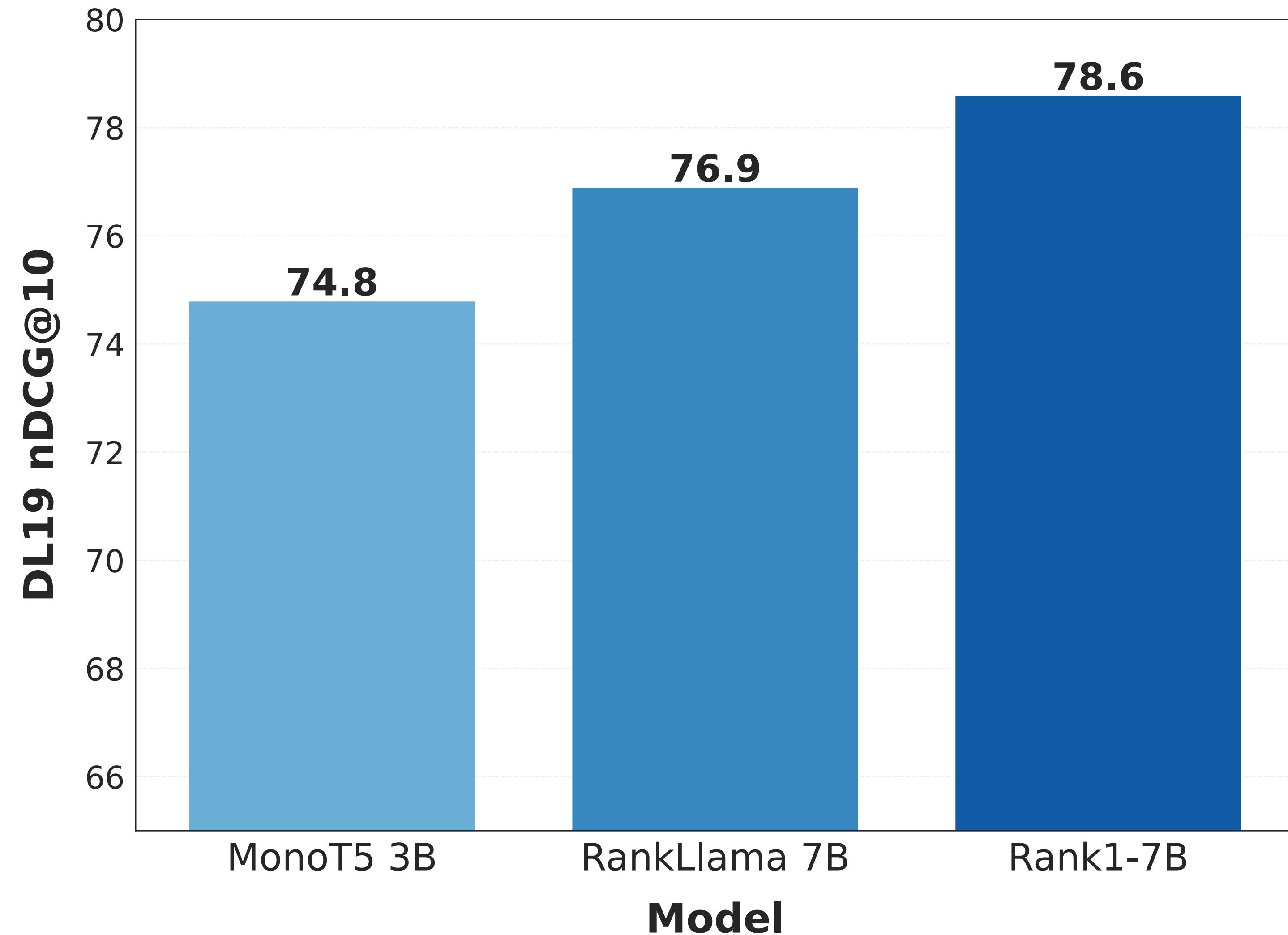


# Results

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

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

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



# Summary

# Summary

- Using test-time compute creates **promptable and reasoning rerankers** - no RL needed!

# Summary

- Using test-time compute creates **promptable and reasoning rerankers** - no RL needed!
- It is slower, but much more powerful than previous approaches

# 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](http://www.janerodgers.blog/wolves)



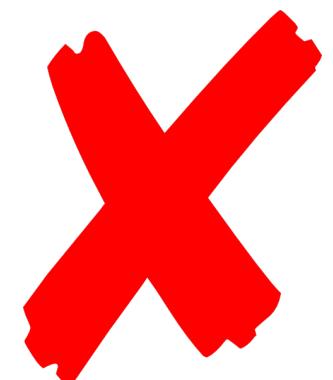
### Data Encryption Standards

[www.nist.gov/standards/](http://www.nist.gov/standards/)



### Digital Protection

[www.clearlaw.net/digital](http://www.clearlaw.net/digital)



# What does this mean?

# What does this mean?



New retrievers benefit  
from advances in LLMs

# What does this mean?

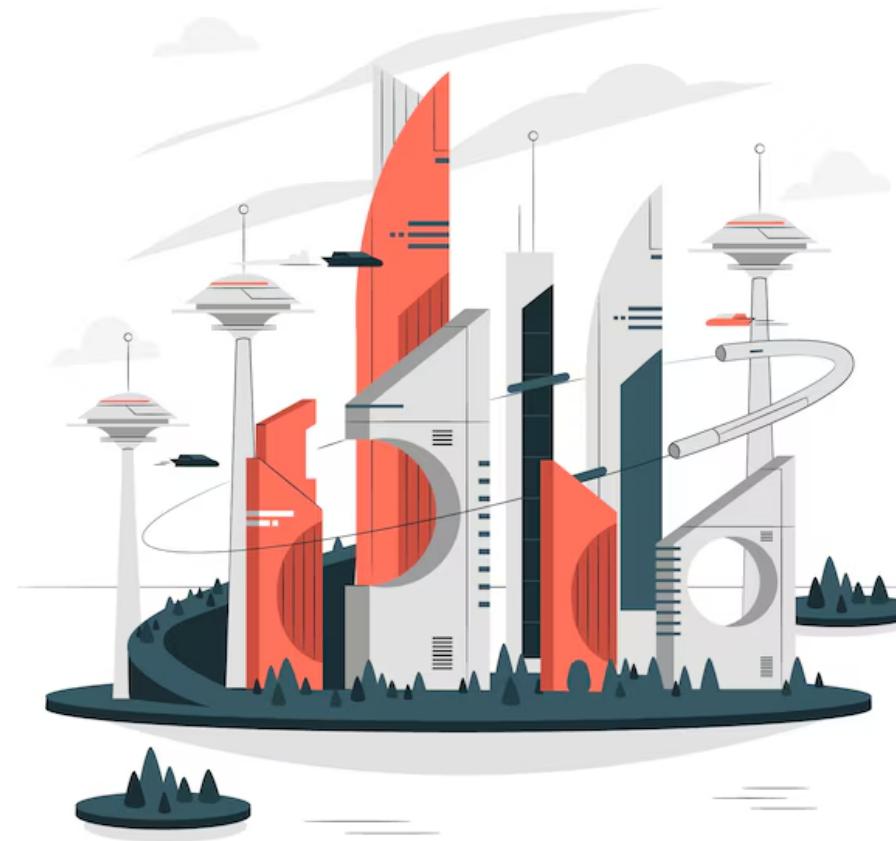


New retrievers benefit  
from advances in LLMs

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



# 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](mailto:oweller2@jh.edu)