Open-World Evaluation for Retrieving Diverse Perspectives

Hung-Ting Chen, Eunsol Choi

NAACL 2025

NYU | COURANT

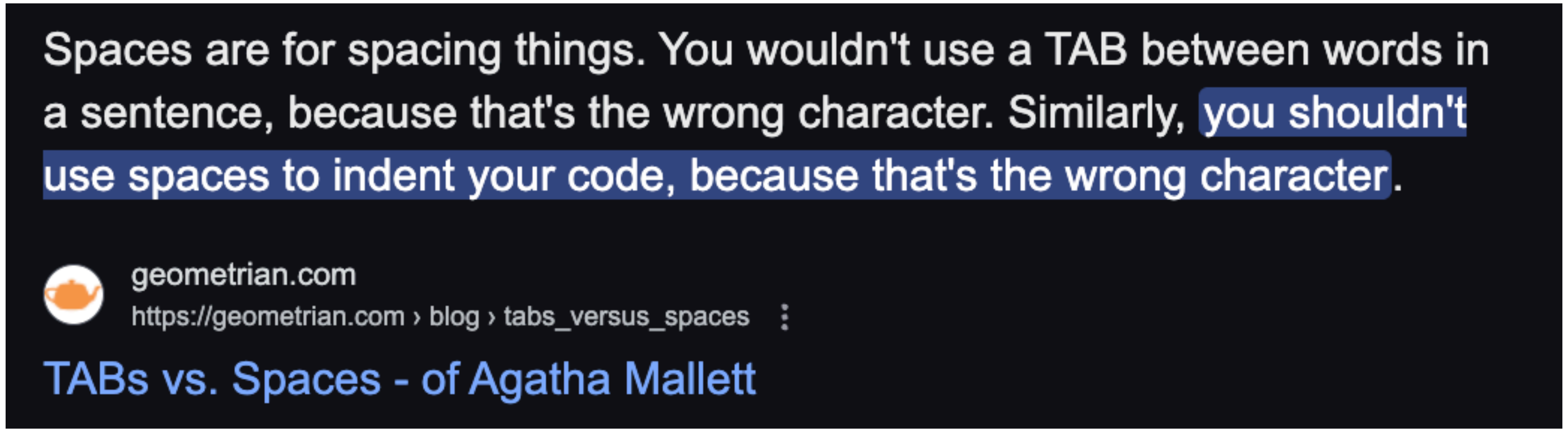
Seeking Information on Subjective Topics

Tabs VS Spaces. Which character should be used for indentation?

You’re Happy with the Results

Tabs VS Spaces. Which character should be used for indentation?

:



Spaces are for spacing things. You wouldn't use a TAB between words in a sentence, because that's the wrong character. Similarly, you shouldn't use spaces to indent your code, because that's the wrong character.

geometrian.com
https://geometrian.com › blog › tabs_versus_spaces

TABs vs. Spaces - of Agatha Mallett

Until…

👤 Tabs VS Spaces. Which character should be used for indentation?

PEP 8 ：

Spaces are the preferred indentation method.

Tabs should be used solely to remain consistent with code that is already indented with tabs.

Users need comprehensive information

Tabs VS Spaces. Which character should be used for indentation?

Pro Tabs

Tabs avoid the malformatting that happens when multiple contributors to a codebase have different space indentation level settings.

Tabs allow for an easier editing experience.

Tab characters use fewer bytes of storage than for any indentation greater than one character.

Pro Spaces

Tabs and spaces are redundant ways of indenting source code. Having only one way is better than having two ways.

The tab character is hard to insert into input fields, for example in search-and-replace forms, since it is also used to switch to the

Users need comprehensive information

We study whether retrieval systems can present diverse information given subjective questions.

Retrieval Systems

Traditional

Who was the actor that played Ben stone on Law & Order?

y* = Michael Moriaty

Retriever

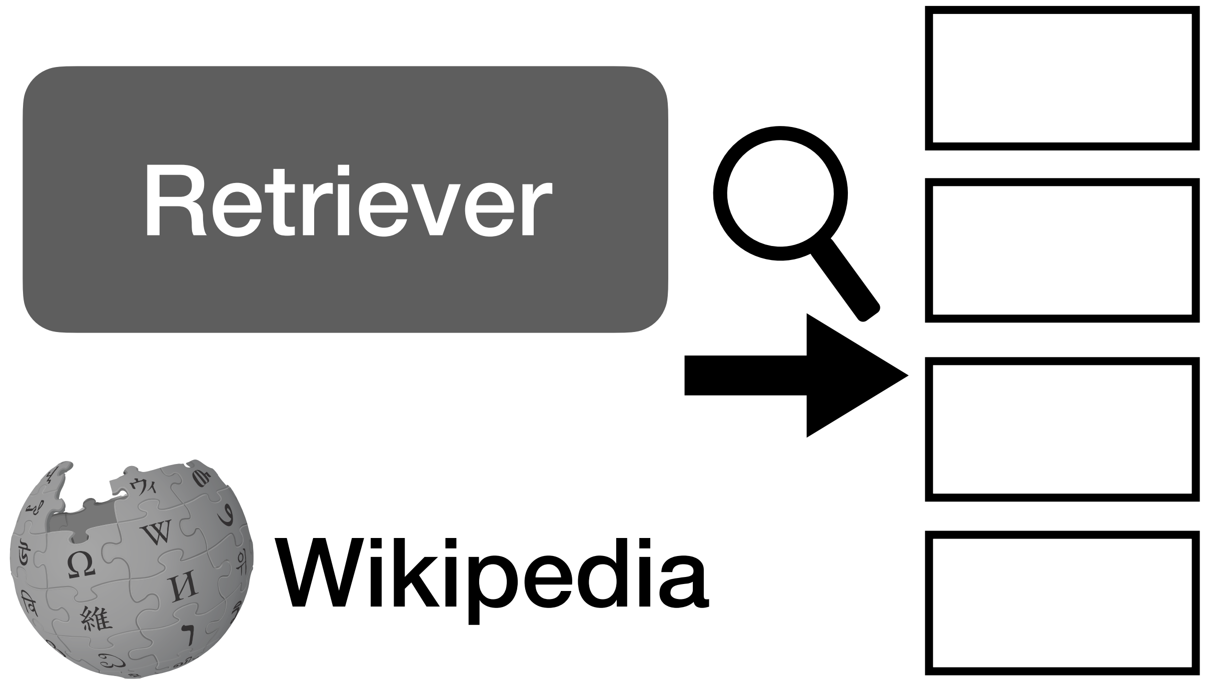
Wikipedia

Retrieval Systems

Traditional

Who was the actor that played Ben stone on Law & Order?

y^* = Michael Moriaty

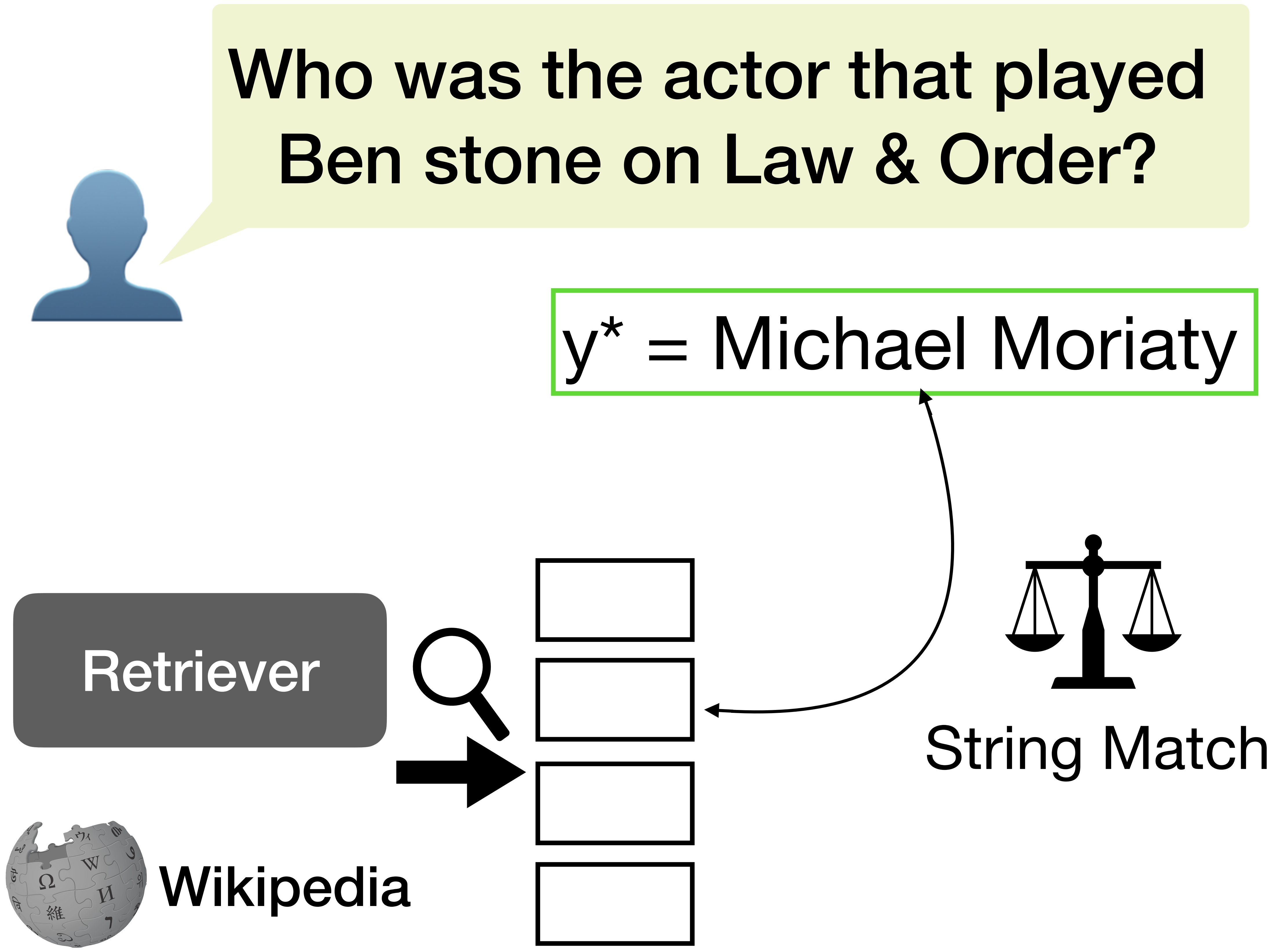


Retriever

Wikipedia

Retrieval Systems

Traditional

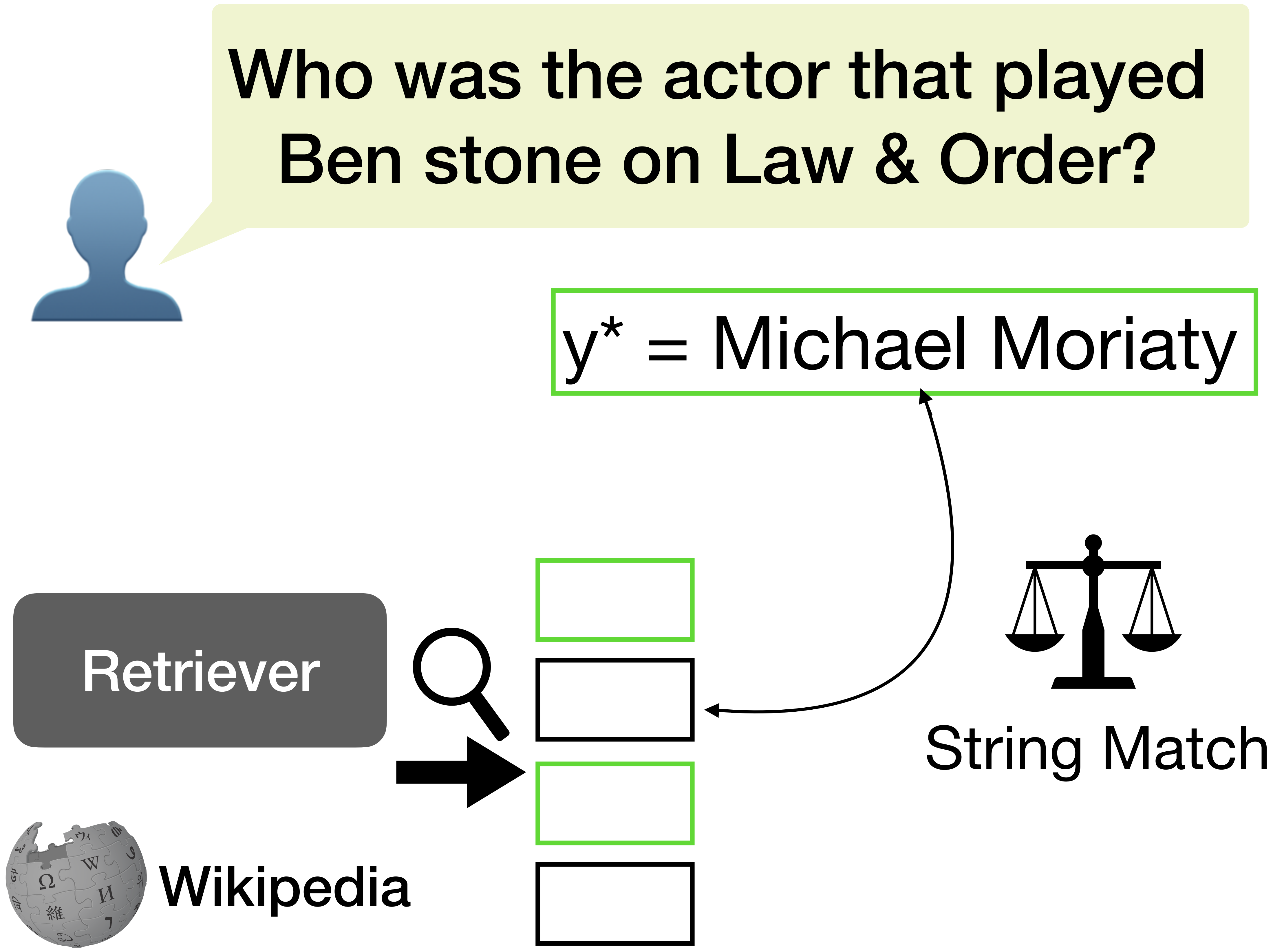


Who was the actor that played Ben stone on Law & Order?

y^* = Michael Moriaty

Retrieval Systems

Traditional



Who was the actor that played Ben stone on Law & Order?

y* = Michael Moriaty

Retriever

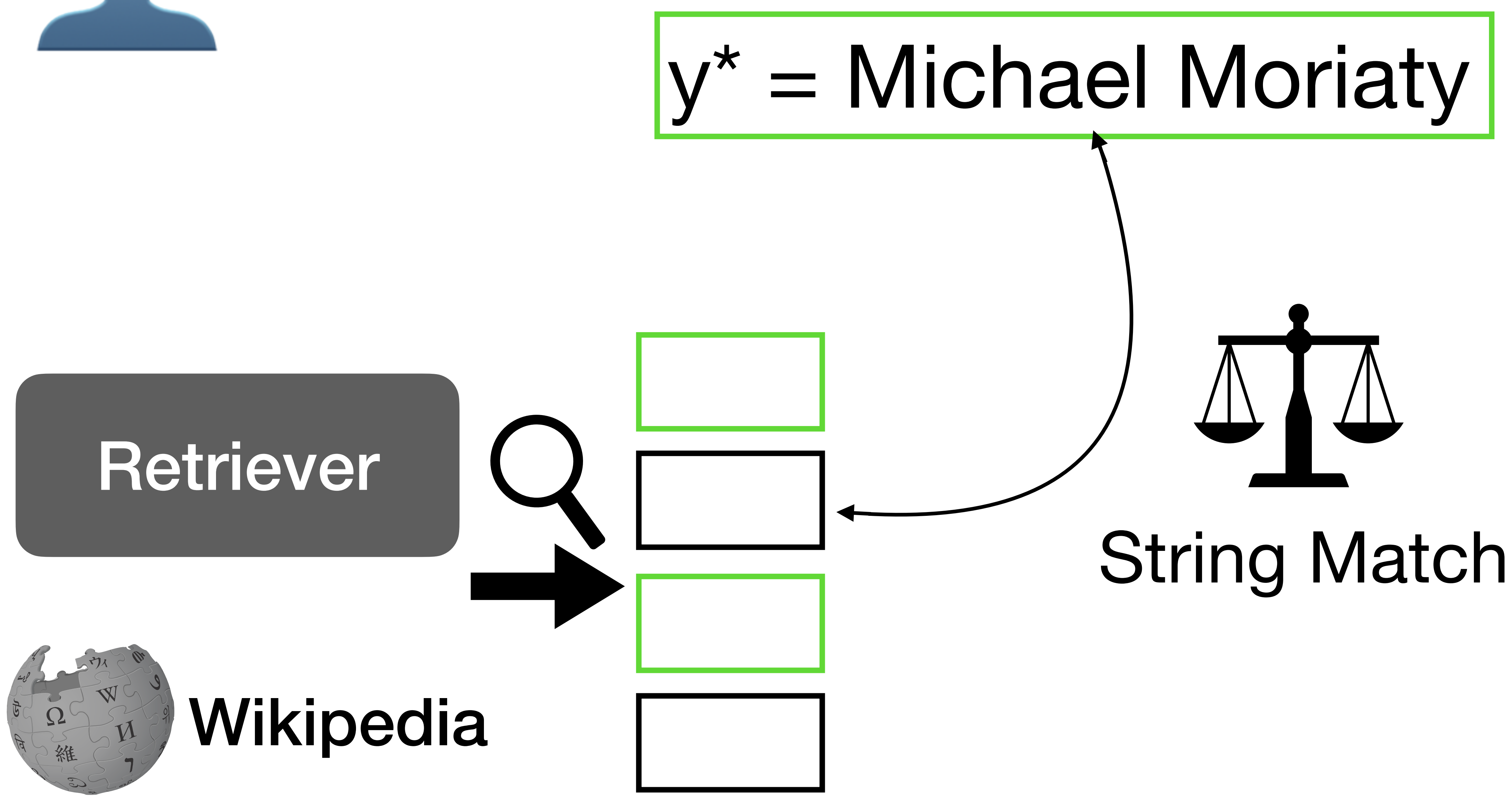
Wikipedia

String Match

Retrieval Systems

Traditional

Who was the actor that played Ben stone on Law & Order?



y* = Michael Moriaty

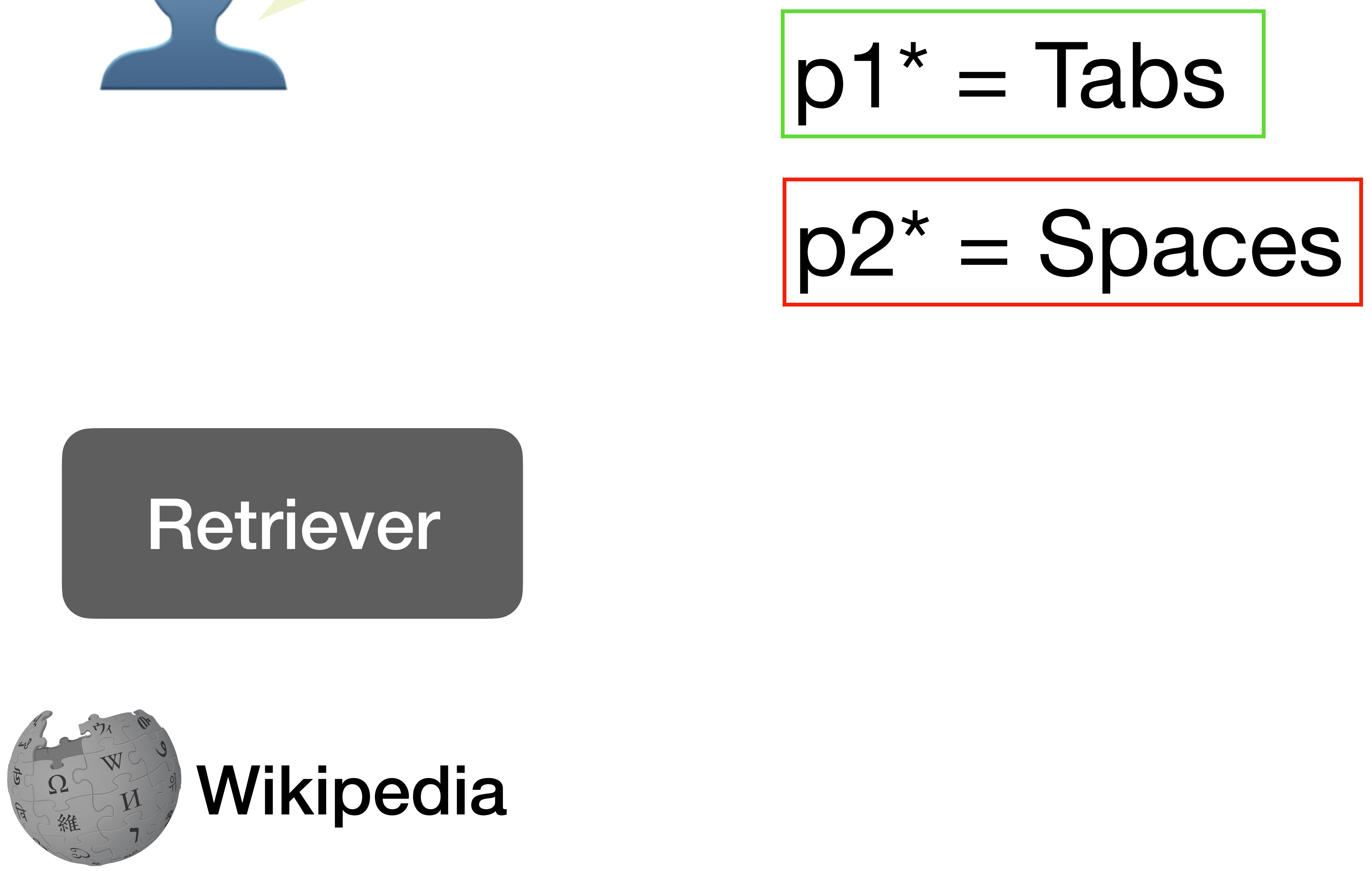
Retriever

Wikipedia

String Match

Our Setting

Tabs VS Spaces. Which character should be used for indentation?



p1* = Tabs

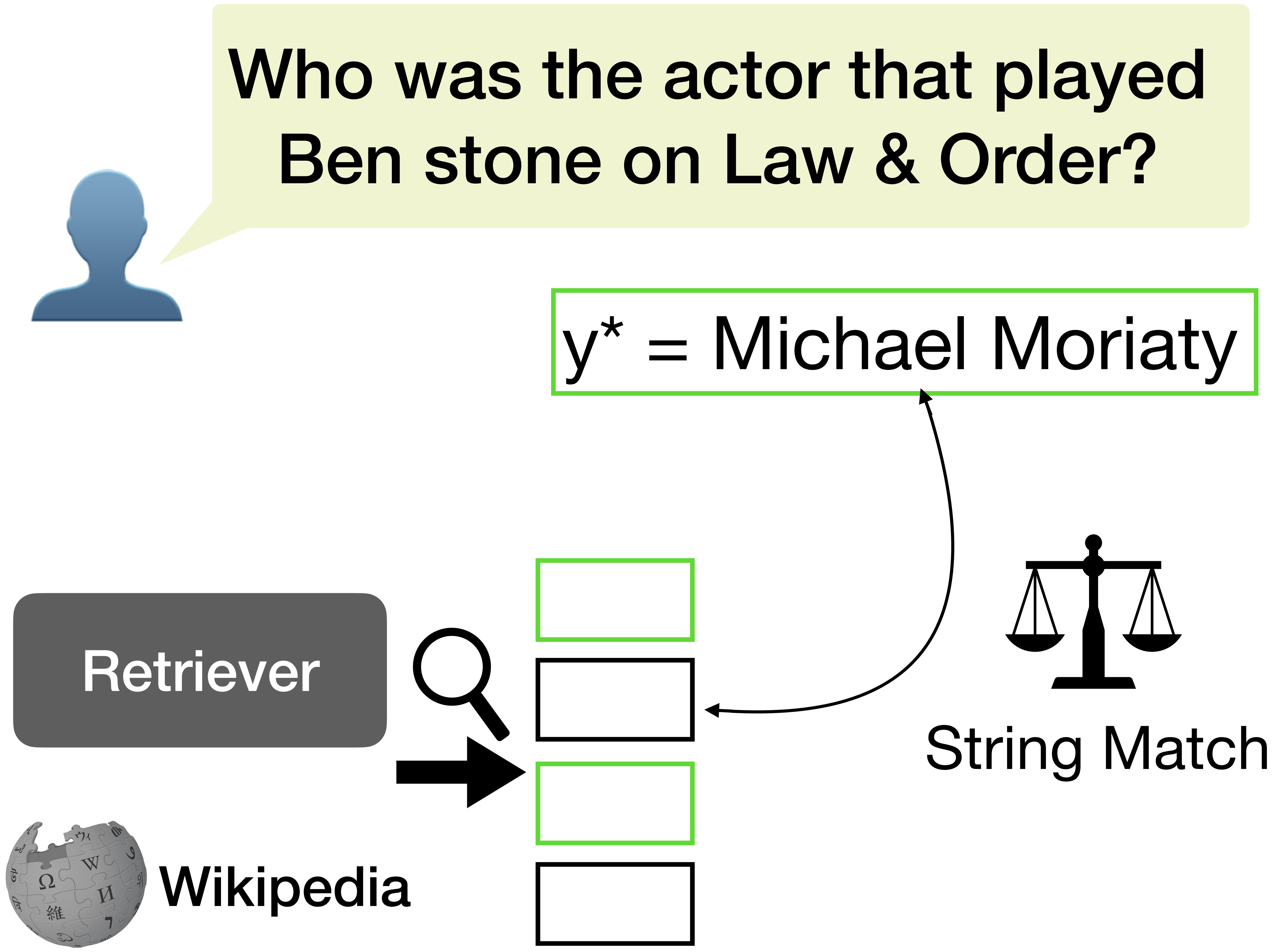
p2* = Spaces

Retriever

Wikipedia

Retrieval Systems

Traditional



Who was the actor that played Ben stone on Law & Order?

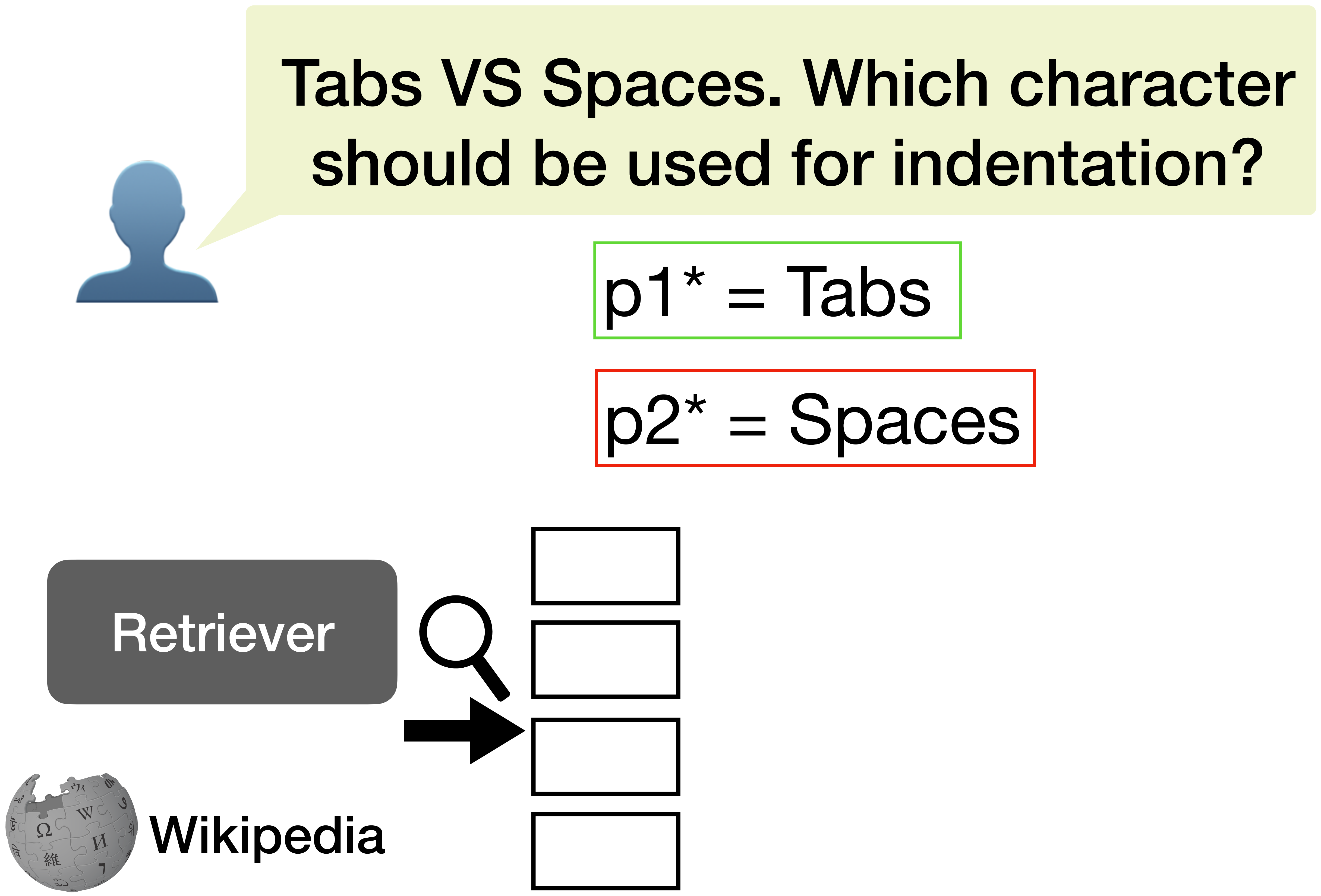
y^* = Michael Moriaty

Retriever

Wikipedia

String Match

Our Setting



Tabs VS Spaces. Which character should be used for indentation?

$p1^*$ = Tabs

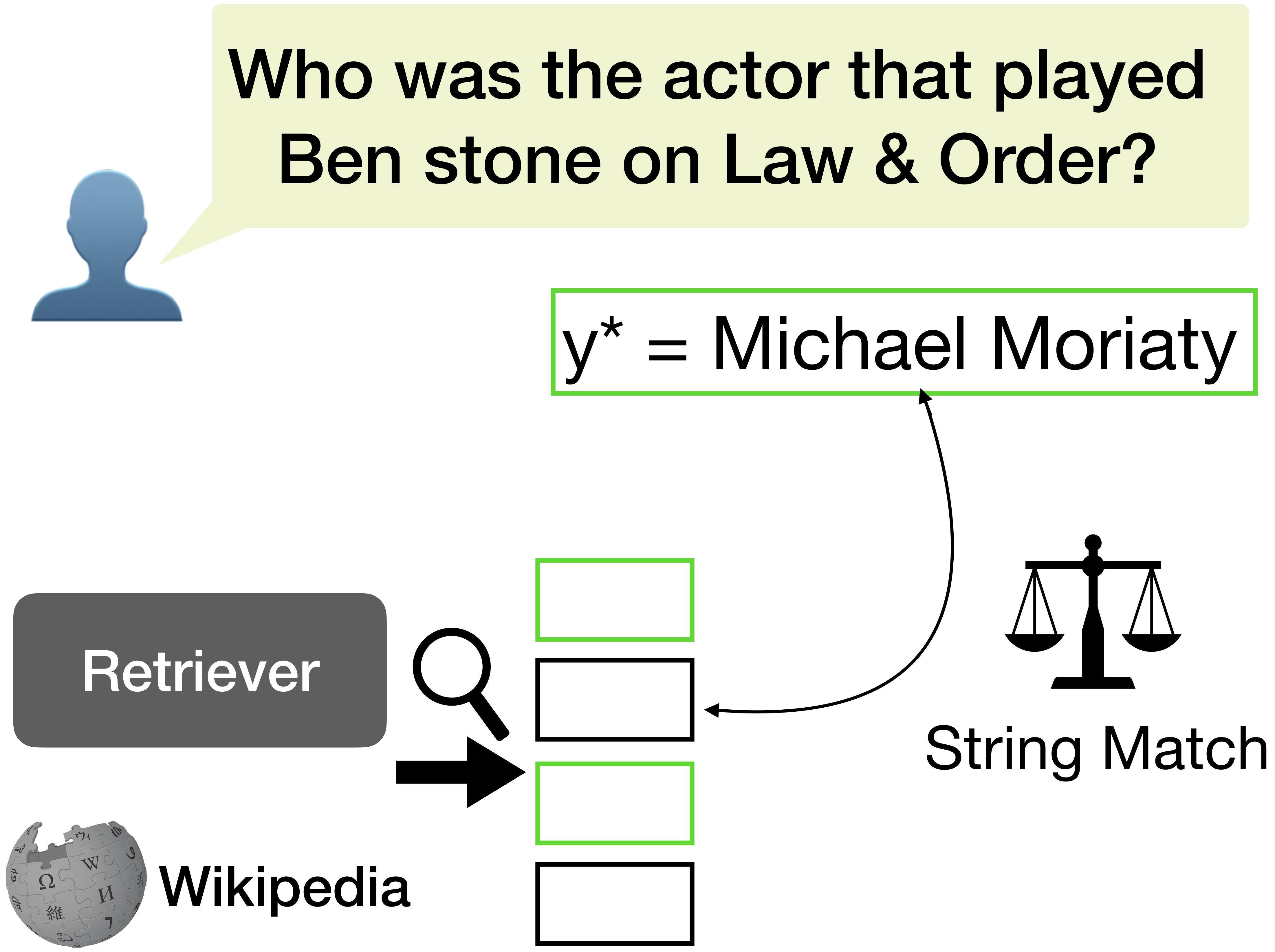
$p2^*$ = Spaces

Retriever

Wikipedia

Retrieval Systems

Traditional



Who was the actor that played Ben stone on Law & Order?

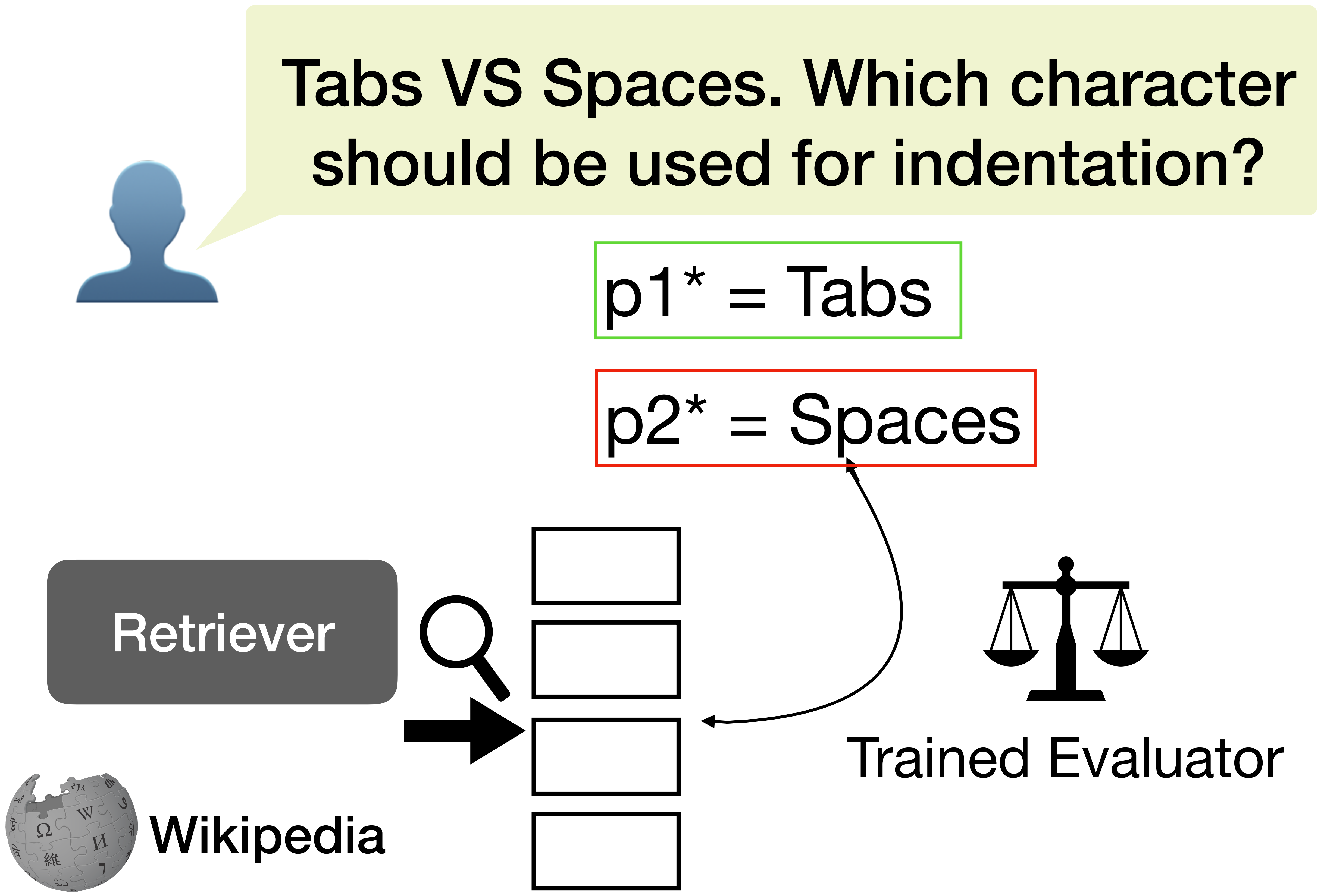
y* = Michael Moriaty

Retriever

Wikipedia

String Match

Our Setting



Tabs VS Spaces. Which character should be used for indentation?

p1* = Tabs

p2* = Spaces

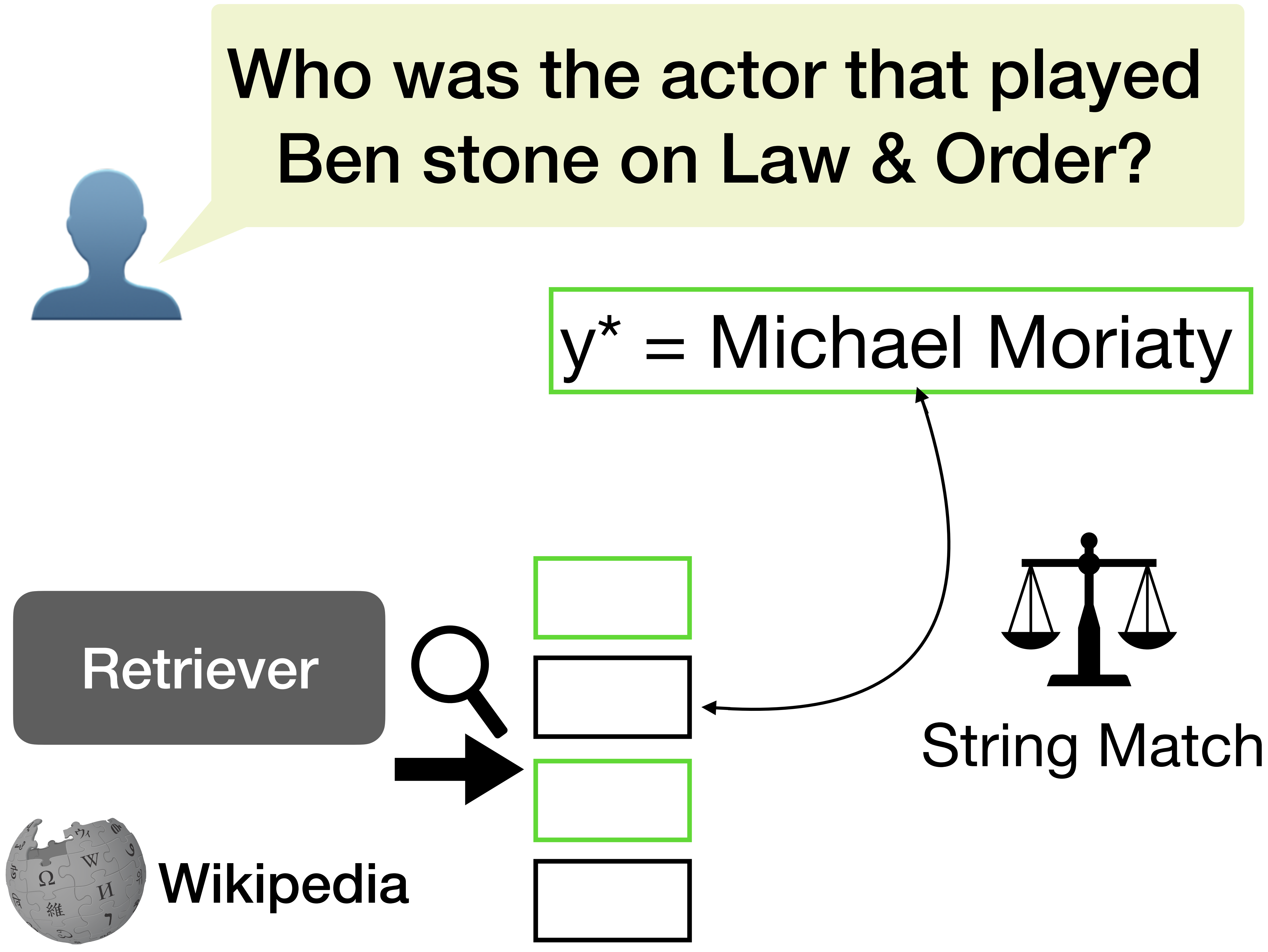
Retriever

Wikipedia

Trained Evaluator

Retrieval Systems

Traditional



Who was the actor that played Ben stone on Law & Order?

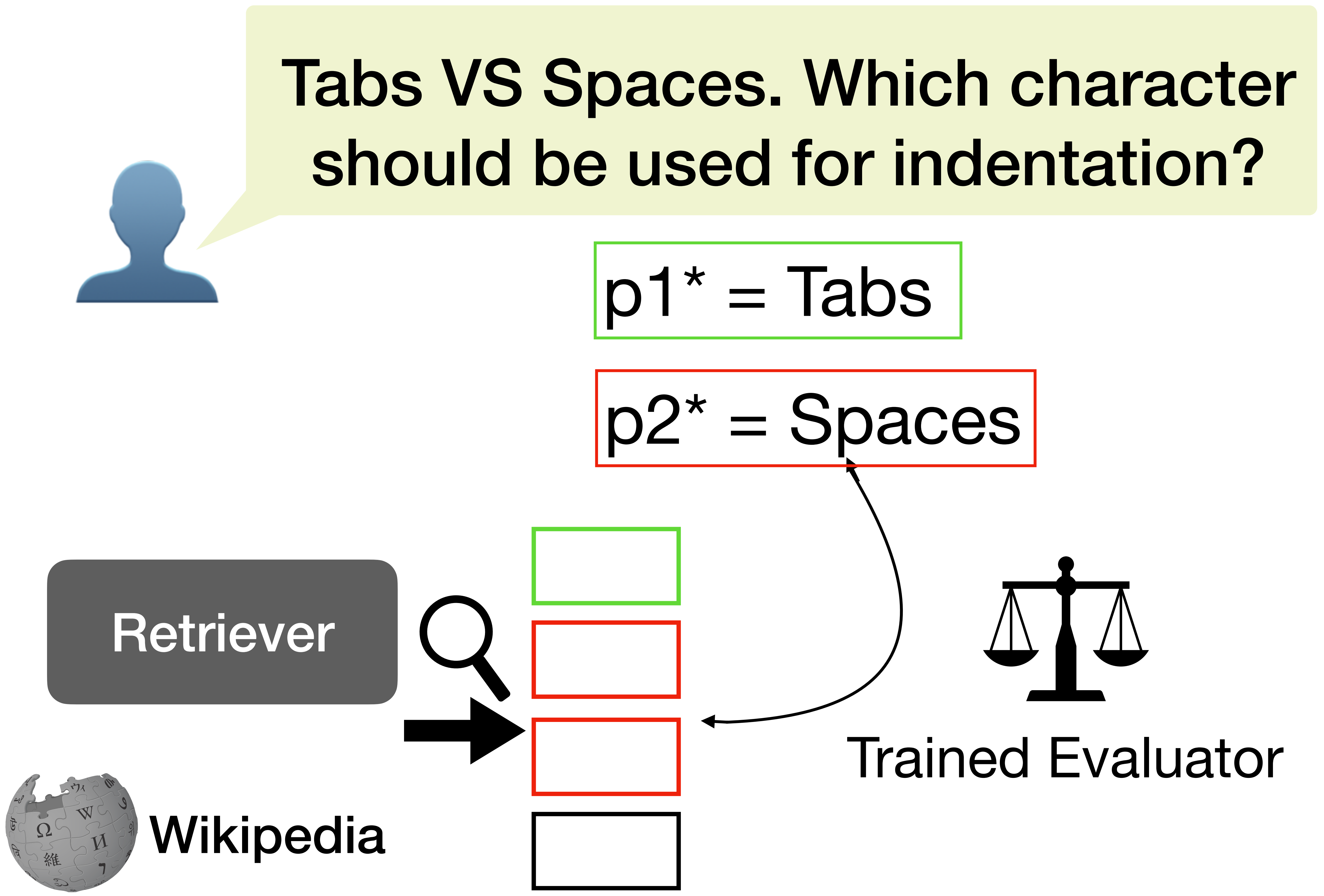
y* = Michael Moriaty

Retriever

Wikipedia

String Match

Our Setting



Tabs VS Spaces. Which character should be used for indentation?

p1* = Tabs

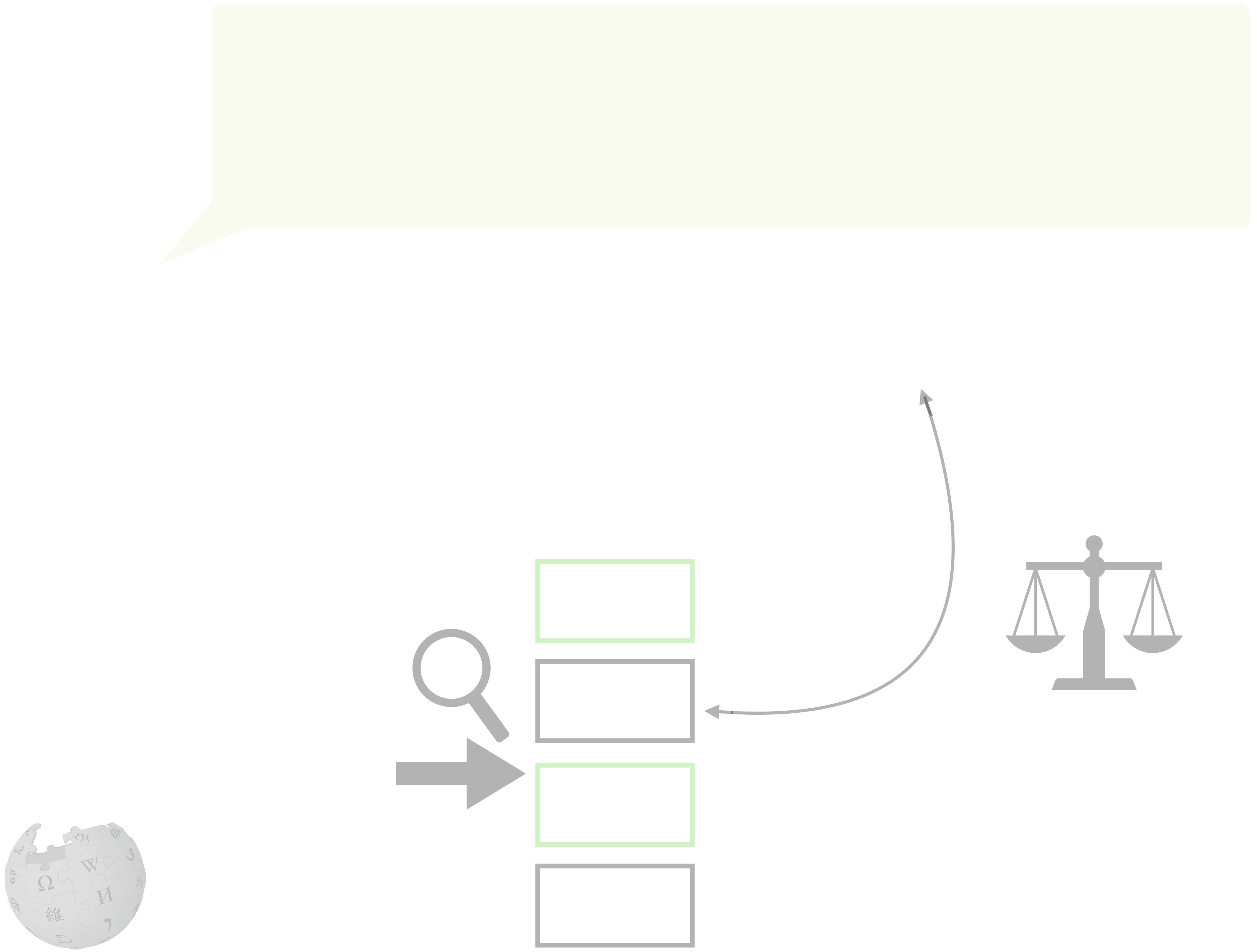
p2* = Spaces

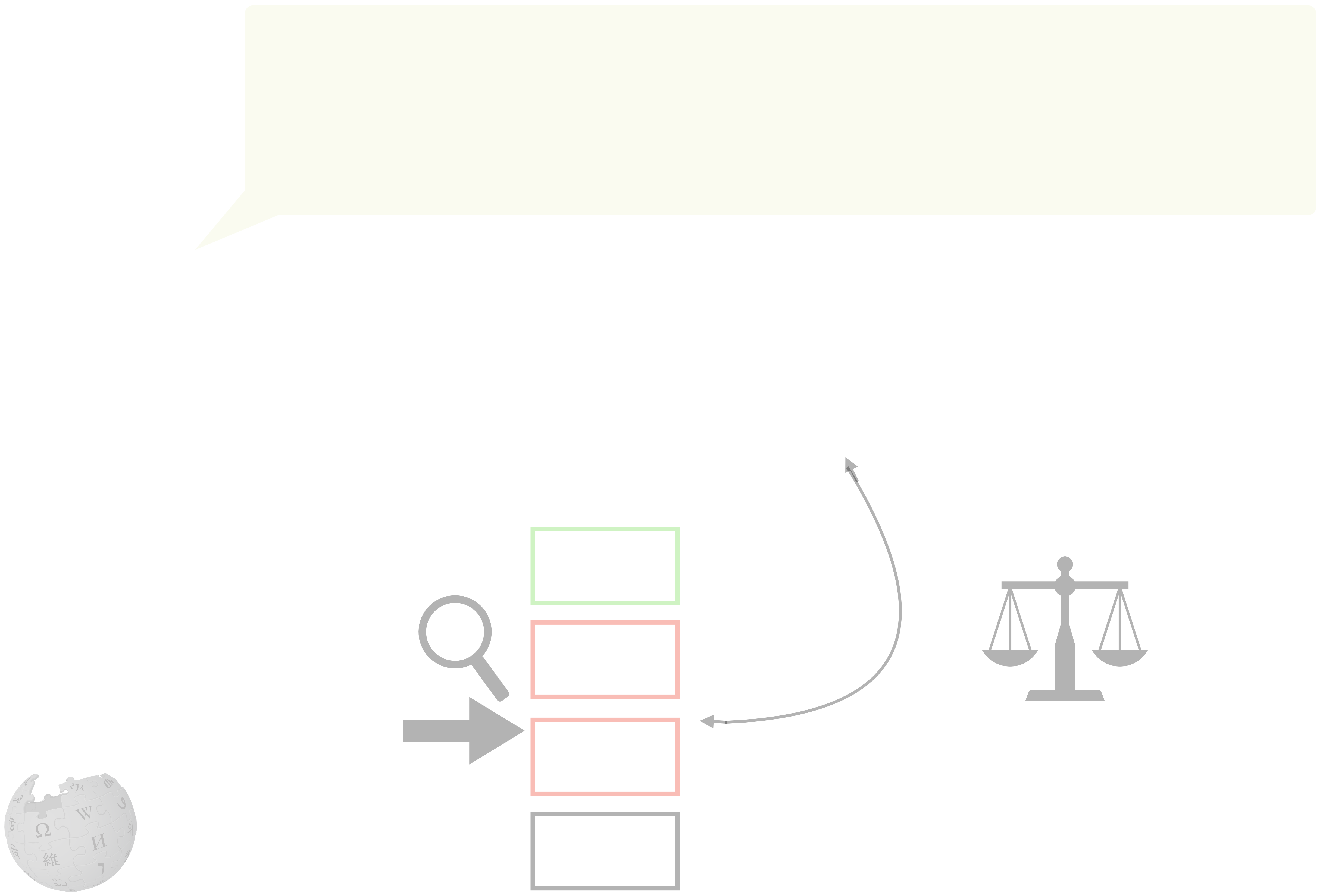
Retriever

Wikipedia

Trained Evaluator

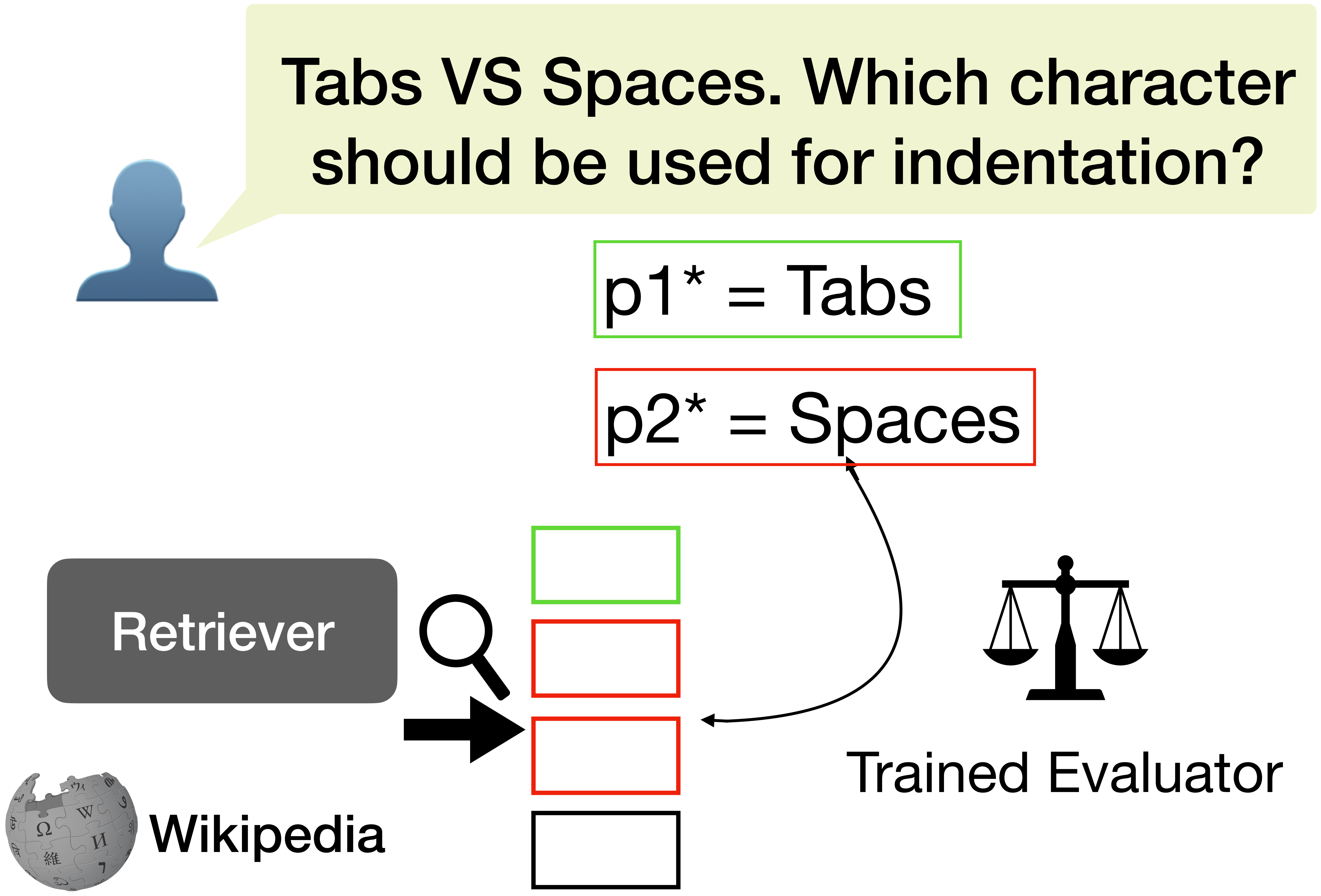
Retrieval Systems





Check if the retrieved document set contains all the ground truth perspectives

Overview



Tabs VS Spaces. Which character should be used for indentation?

p1* = Tabs

p2* = Spaces

Retriever

Wikipedia

Trained Evaluator

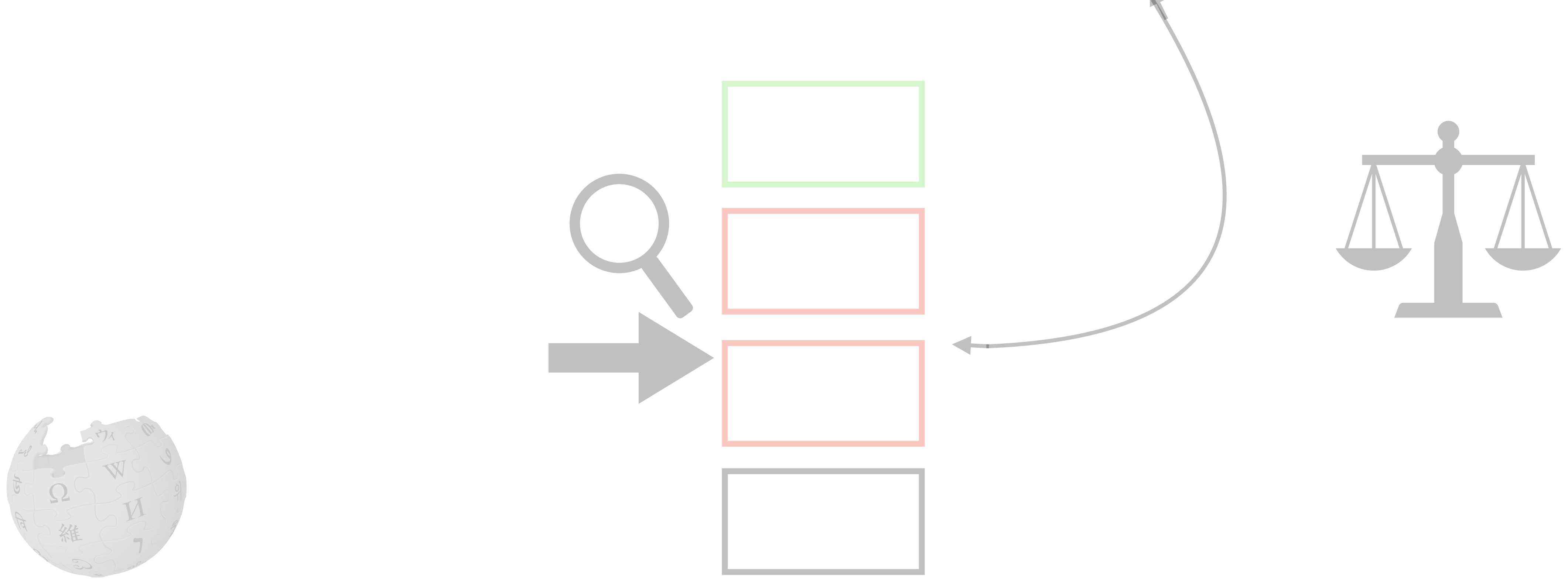
Overview

Tabs VS Spaces. Which character should be used for indentation?

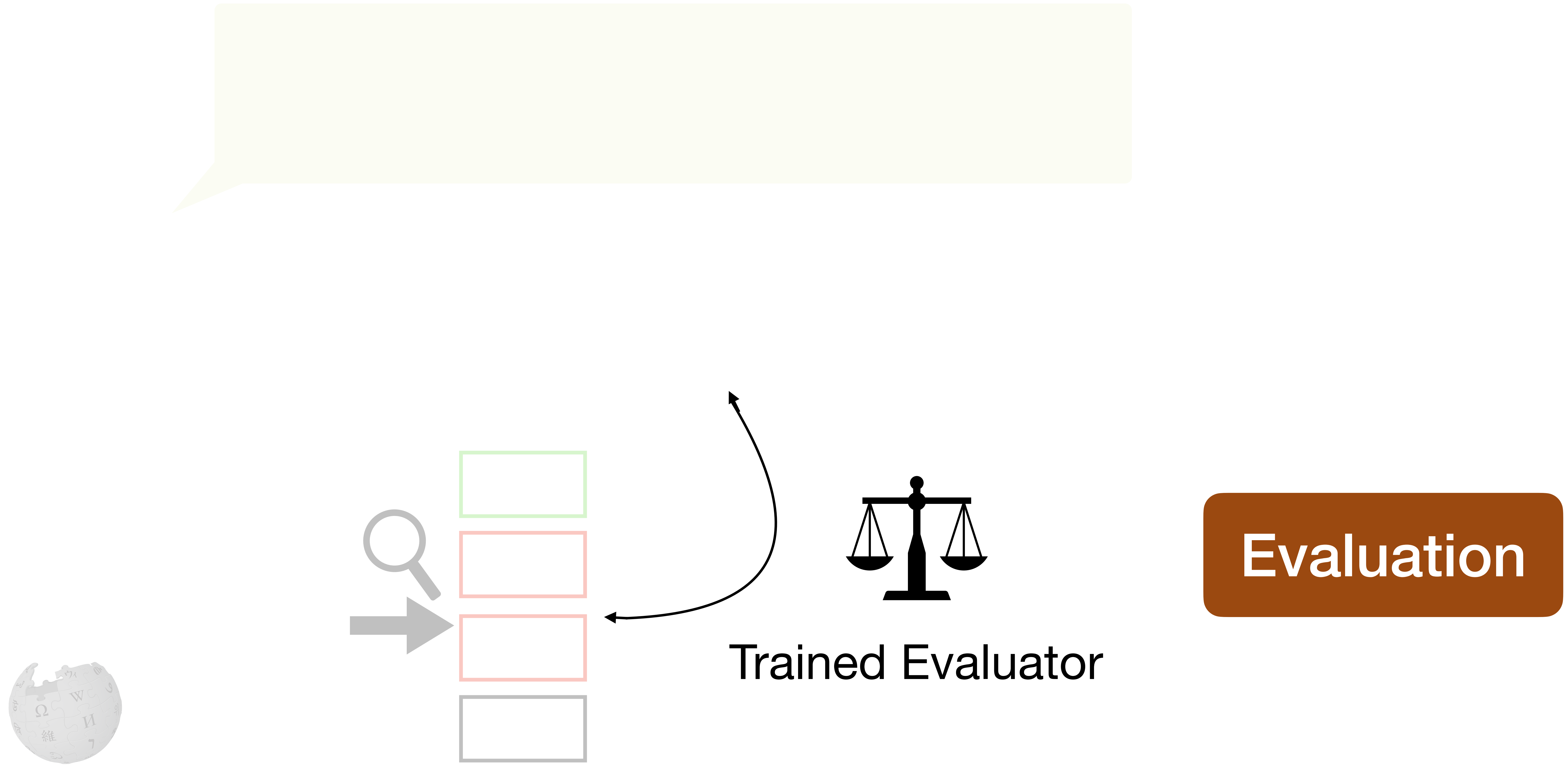
$p1^*$ = Tabs

$p2^*$ = Spaces

Data



Overview

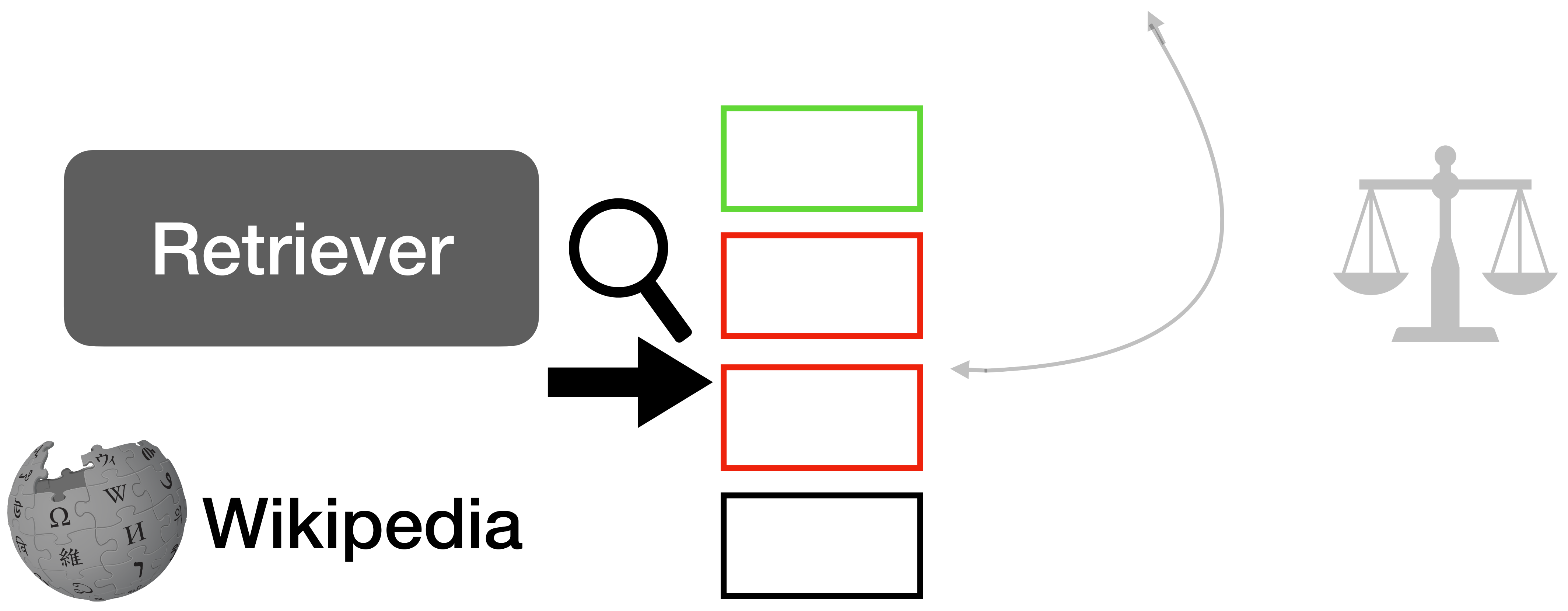


Trained Evaluator

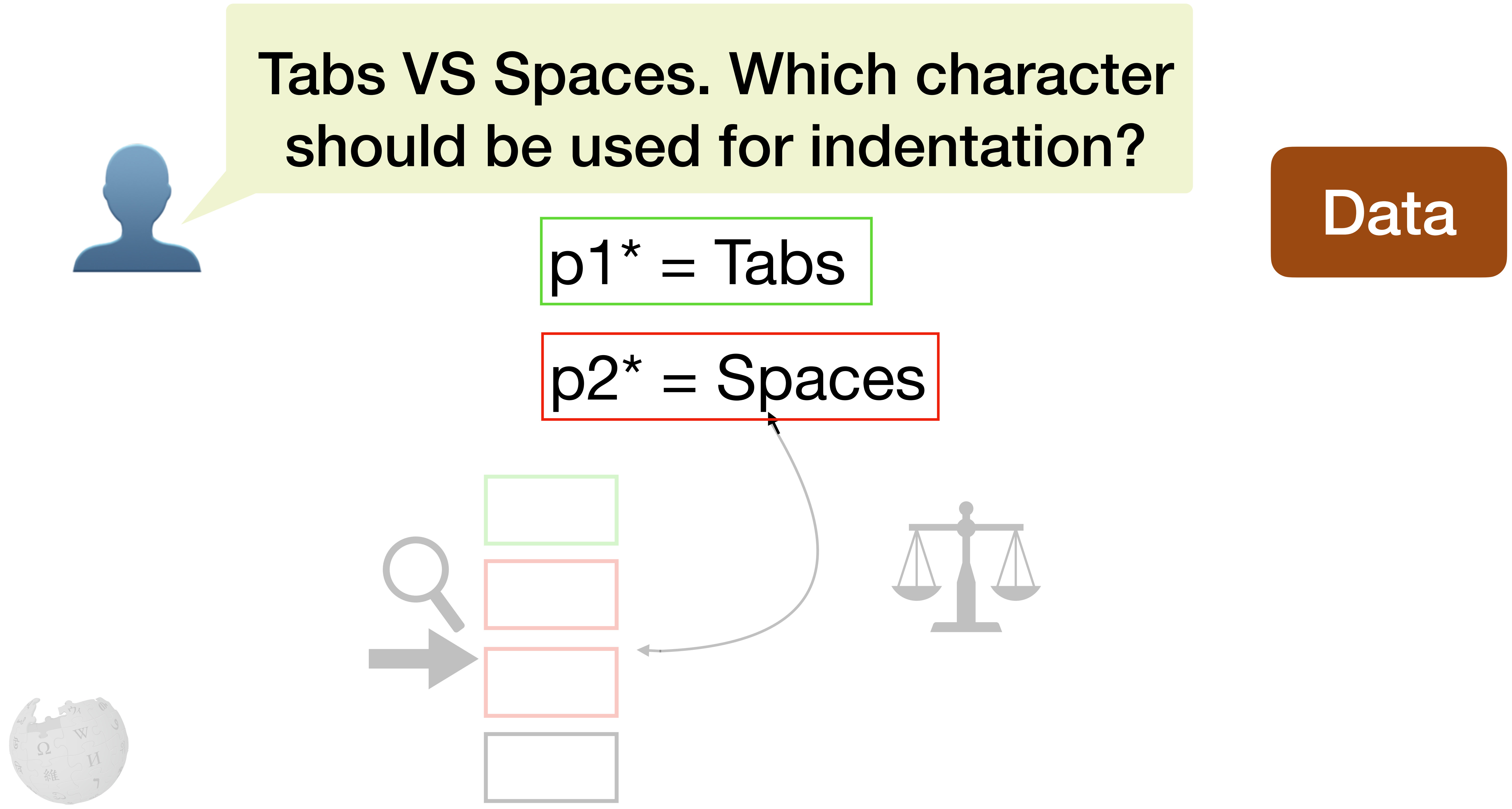
Evaluation

Overview

Testing Retrieval Systems



Benchmark Construction



Tabs VS Spaces. Which character should be used for indentation?

p1* = Tabs

p2* = Spaces

Data

	Debate Questions		Survey Questions
	Arguana	**Kialo**	**OpinionQA**
Question	Does an increase in choice lead to unhappiness and stress?	What’s the best Twitter/X alternative?	Do you feel safer with a gun in your household?

	Debate Questions		Survey Questions
	Arguana	**Kialo**	**OpinionQA**
Question	Does an increase in choice lead to unhappiness and stress?	What’s the best Twitter/X alternative?	Do you feel safer with a gun in your household?
Perspective Set	• An increase in choice leads to unhappiness and stress. • An increase in choice does not lead to unhappiness and stress.	• The best Twitter/X alternative is Facebook. • The best Twitter/X alternative is Kialo. • The best Twitter/X alternative is Instagram. • The best Twitter/X alternative is reddit.	• Having a gun in the household increases my sense of safety. • Having a gun in the household does not increase my sense of safety.

	Debate Questions		Survey Questions
	Arguana	**Kialo**	**OpinionQA**
Question	Does an increase in choice lead to unhappiness and stress?	What's the best Twitter/X alternative?	Do you feel safer with a gun in your household?
Perspective Set	• An increase in choice leads to unhappiness and stress. • An increase in choice does not lead to unhappiness and stress.	• The best Twitter/X alternative is Facebook. • The best Twitter/X alternative is Kialo. • The best Twitter/X alternative is Instagram. • The best Twitter/X alternative is reddit.	• Having a gun in the household increases my sense of safety. • Having a gun in the household does not increase my sense of safety.

Obtain data from these sources, and generate missing fields with GPT-4 🤖

BERDS: Benchmark for Retrieval Diversity for Subjective questions

	Debate Questions		Survey Questions
	Arguana	**Kialo**	**OpinionQA**
Question	Does an increase in choice lead to unhappiness and stress?	What's the best Twitter/X alternative?	Do you feel safer with a gun in your household?
Perspective Set	• An increase in choice leads to unhappiness and stress. • An increase in choice does not lead to unhappiness and stress.	• The best Twitter/X alternative is Facebook. • The best Twitter/X alternative is Kialo. • The best Twitter/X alternative is Instagram. • The best Twitter/X alternative is reddit.	• Having a gun in the household increases my sense of safety. • Having a gun in the household does not increase my sense of safety.

Obtain data from these sources, and generate missing fields with GPT-4 🤖

BERDS: Benchmark for Retrieval Diversity for Subjective questions

	Debate Questions		Survey Questions
	Arguana	**Kialo**	**OpinionQA**
Question	Does an increase in choice lead to unhappiness and stress?	What's the best Twitter/X alternative?	Do you feel safer with a gun in your household?
Perspective Set	• An increase in choice leads to unhappiness and stress. • An increase in choice does not lead to unhappiness and stress.	• The best Twitter/X alternative is Facebook. • The best Twitter/X alternative is Kialo. • The best Twitter/X alternative is Instagram. • The best Twitter/X alternative is reddit.	• Having a gun in the household increases my sense of safety. • Having a gun in the household does not increase my sense of safety.

Obtain data from these sources, and generate missing fields with GPT-4 🤖

1. ~1K examples for each data source.

BERDS: Benchmark for Retrieval Diversity for Subjective questions

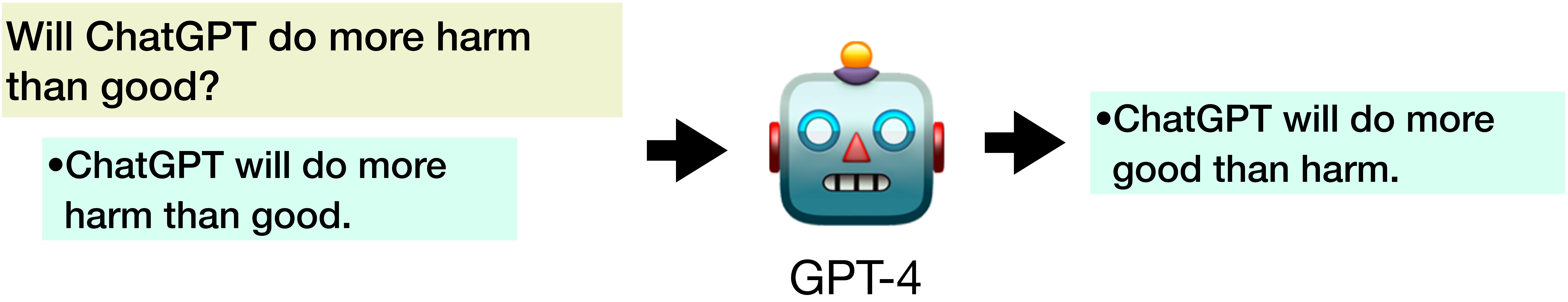
	Debate Questions		Survey Questions
	Arguana	**Kialo**	**OpinionQA**
Question	Does an increase in choice lead to unhappiness and stress?	What's the best Twitter/X alternative?	Do you feel safer with a gun in your household?
Perspective Set	• An increase in choice leads to unhappiness and stress. • An increase in choice does not lead to unhappiness and stress.	• The best Twitter/X alternative is Facebook. • The best Twitter/X alternative is Kialo. • The best Twitter/X alternative is Instagram. • The best Twitter/X alternative is reddit.	• Having a gun in the household increases my sense of safety. • Having a gun in the household does not increase my sense of safety.

Obtain data from these sources, and generate missing fields with GPT-4 🤖

1. ~1K examples for each data source.

2. Most examples contain two (binary) perspectives.

Benchmark Construction Example: Kialo



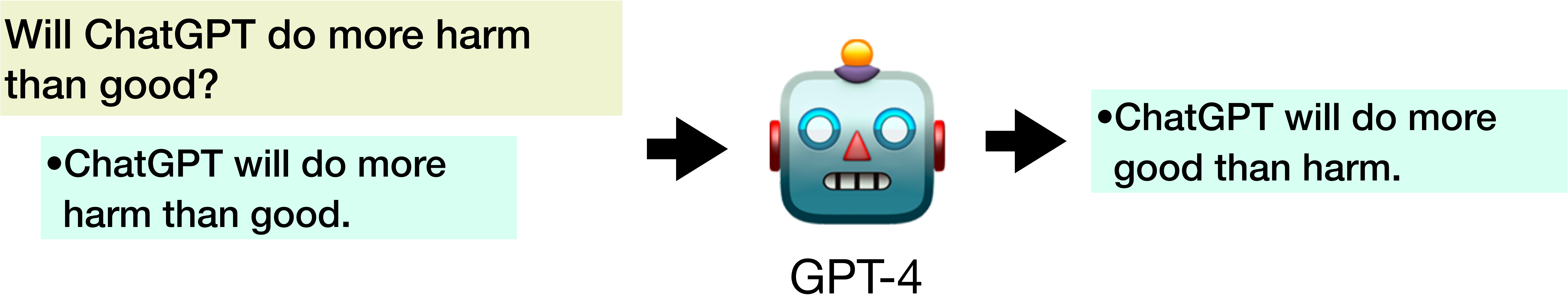
Will ChatGPT do more harm than good?

- ChatGPT will do more harm than good.

GPT-4

- ChatGPT will do more good than harm.

Benchmark Construction Example: Kialo



Will ChatGPT do more harm than good?

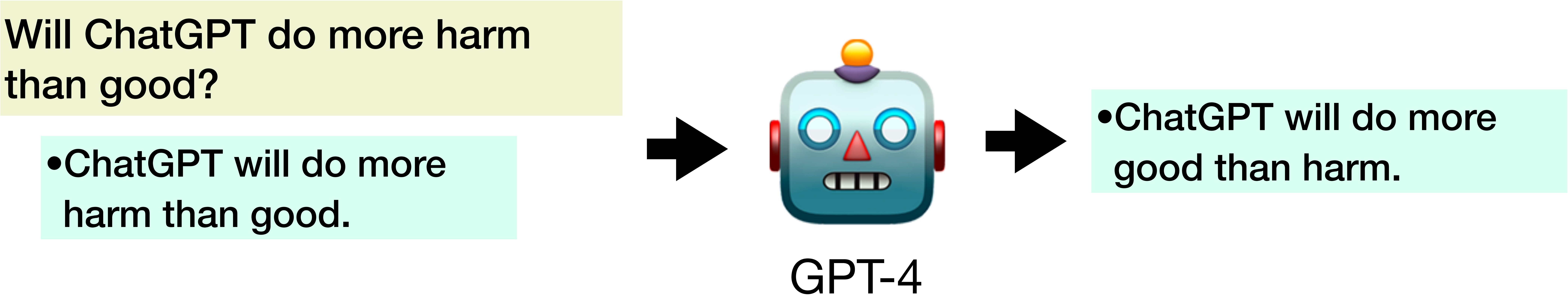
- ChatGPT will do more harm than good.

GPT-4

- ChatGPT will do more good than harm.

✓ The data generation task is easy for GPT-4.

Benchmark Construction Example: Kialo



Will ChatGPT do more harm than good?

- ChatGPT will do more harm than good.

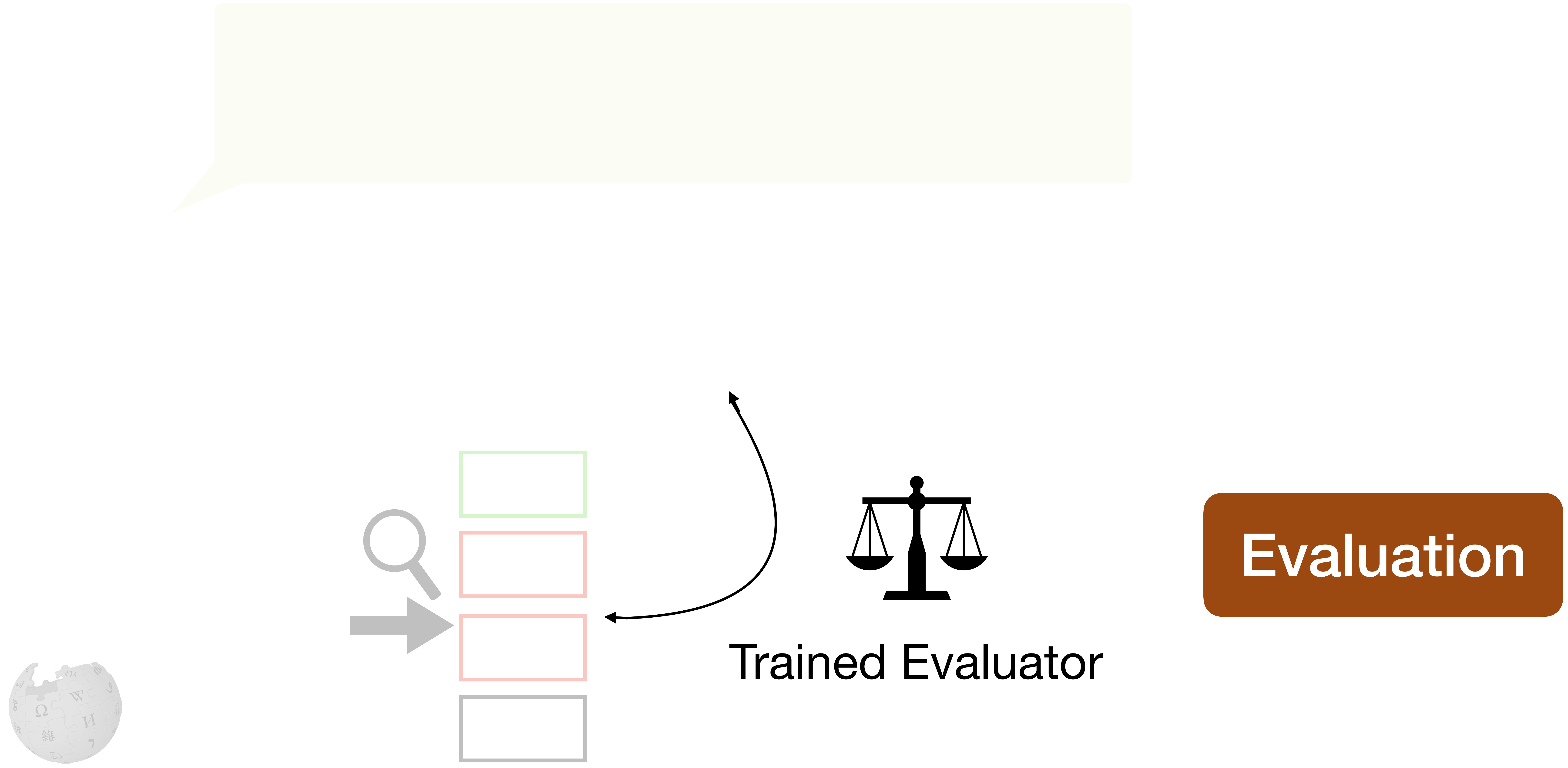
GPT-4

- ChatGPT will do more good than harm.

✓ The data generation task is easy for GPT-4.

✓ The generated data is of high quality.

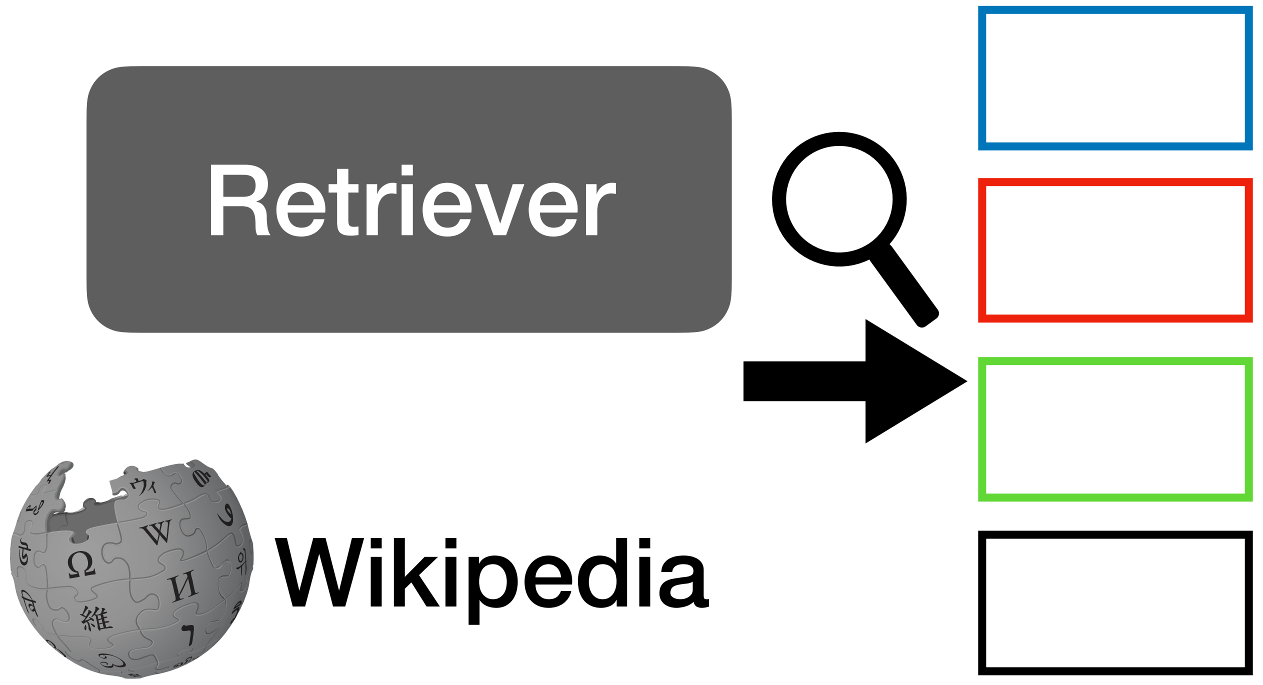
Evaluation



Background: Evaluation for Multi-Answer QA

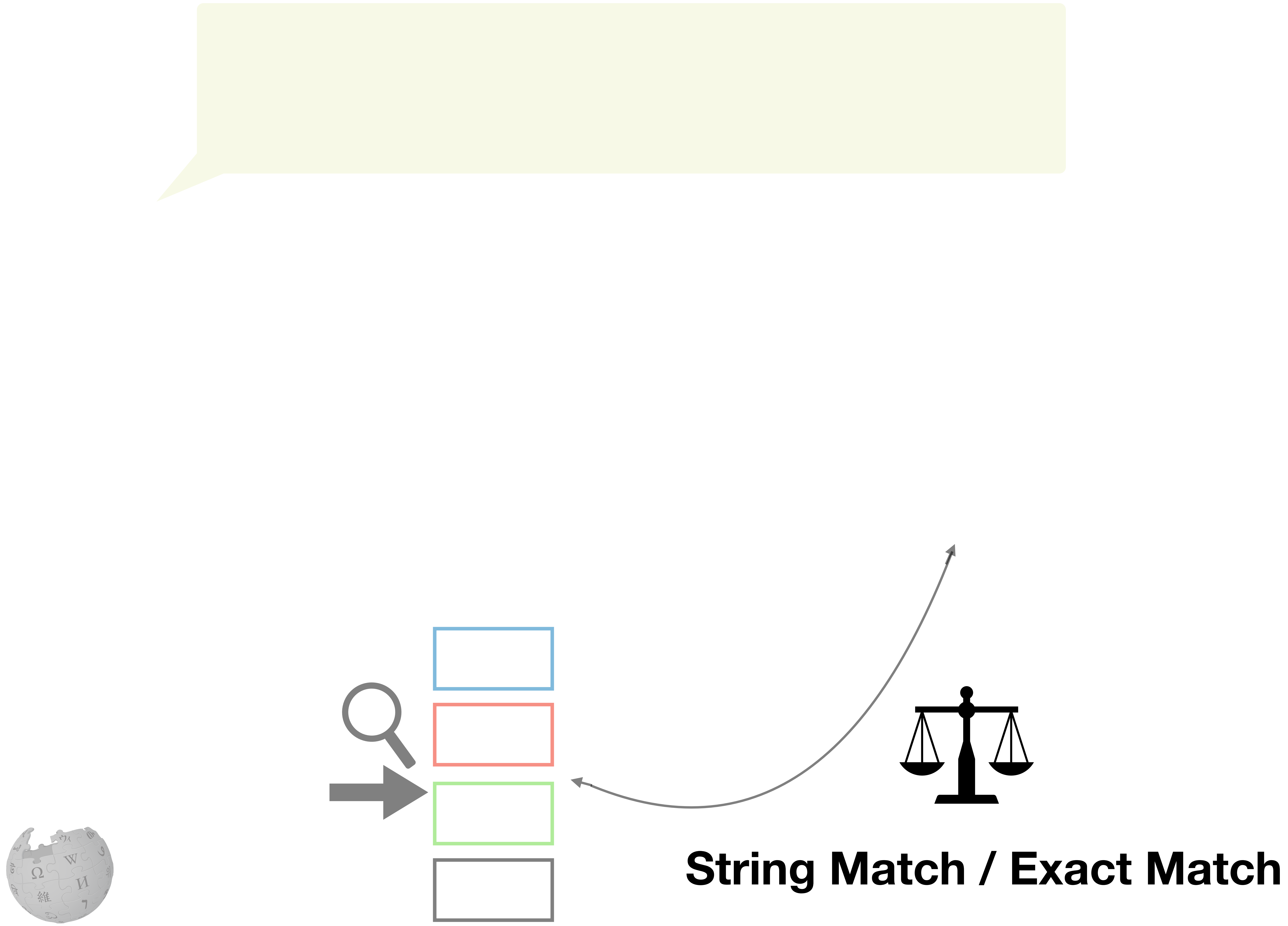
Who starred in Barefoot in the Park on broadway?

Answers	Gold Indices
a1* = Elizabeth Ashley;	y1* = 123
a2* = Kurt Kasznar;	y2* = 59021
a3* = Joseph Keating	y3* = 847230



Min et al., EMNLP 2021

Background: Evaluation for Multi-Answer QA



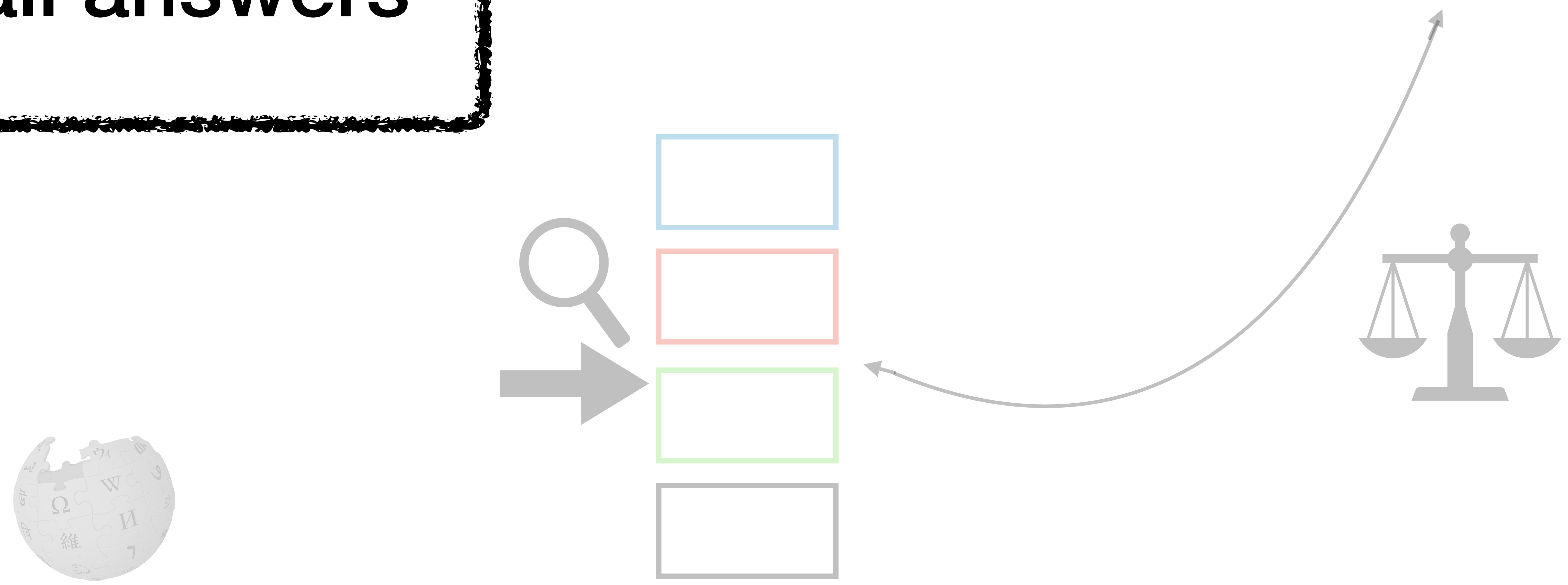
String Match / Exact Match

Min et al., EMNLP 2021

Background: Evaluation for Multi-Answer QA

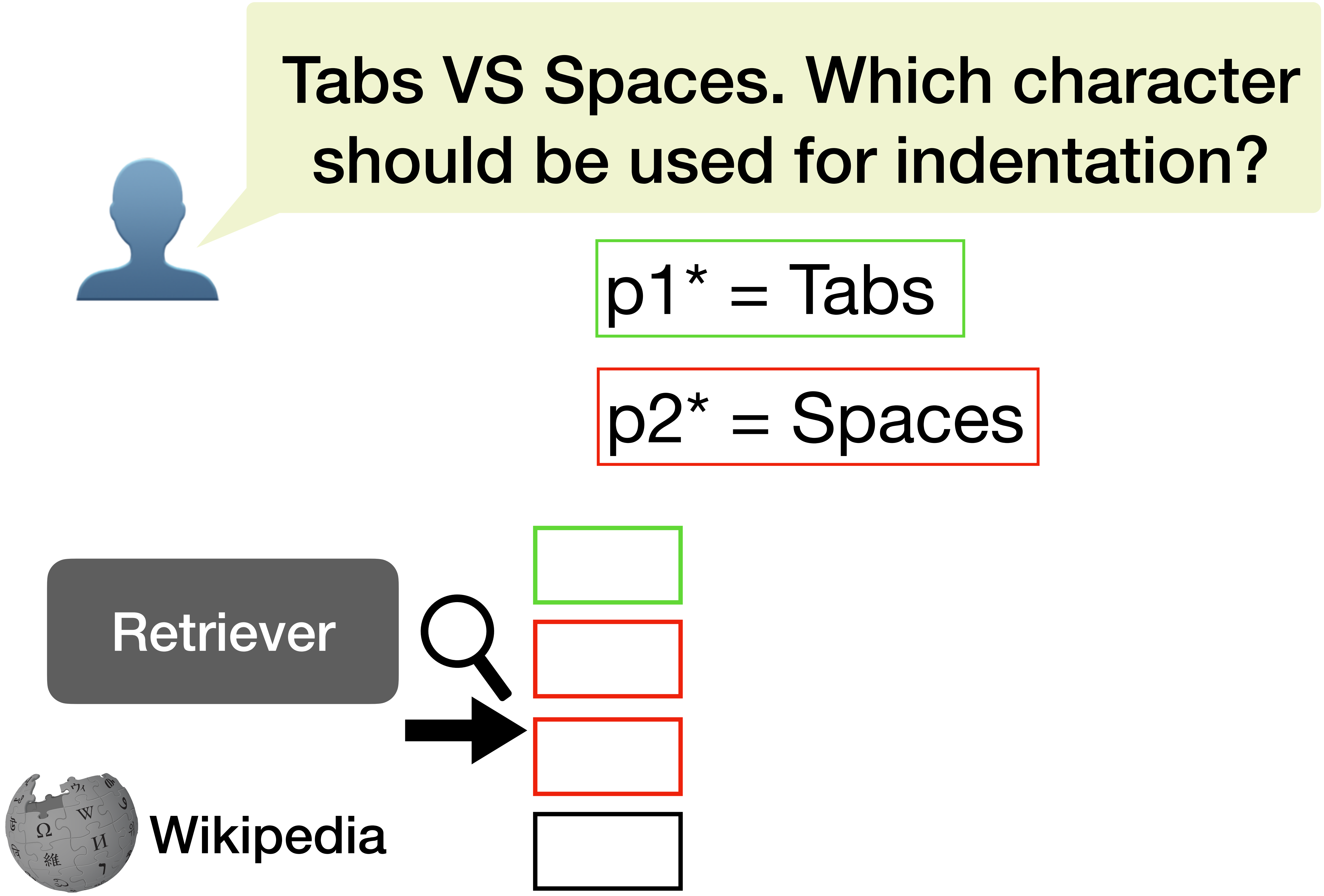
Metric: Precision, MRecall @ k

“MRecall @ k” =1 if top k document set contains all answers



Min et al., EMNLP 2021

Ours: Evaluation for Subjective Questions



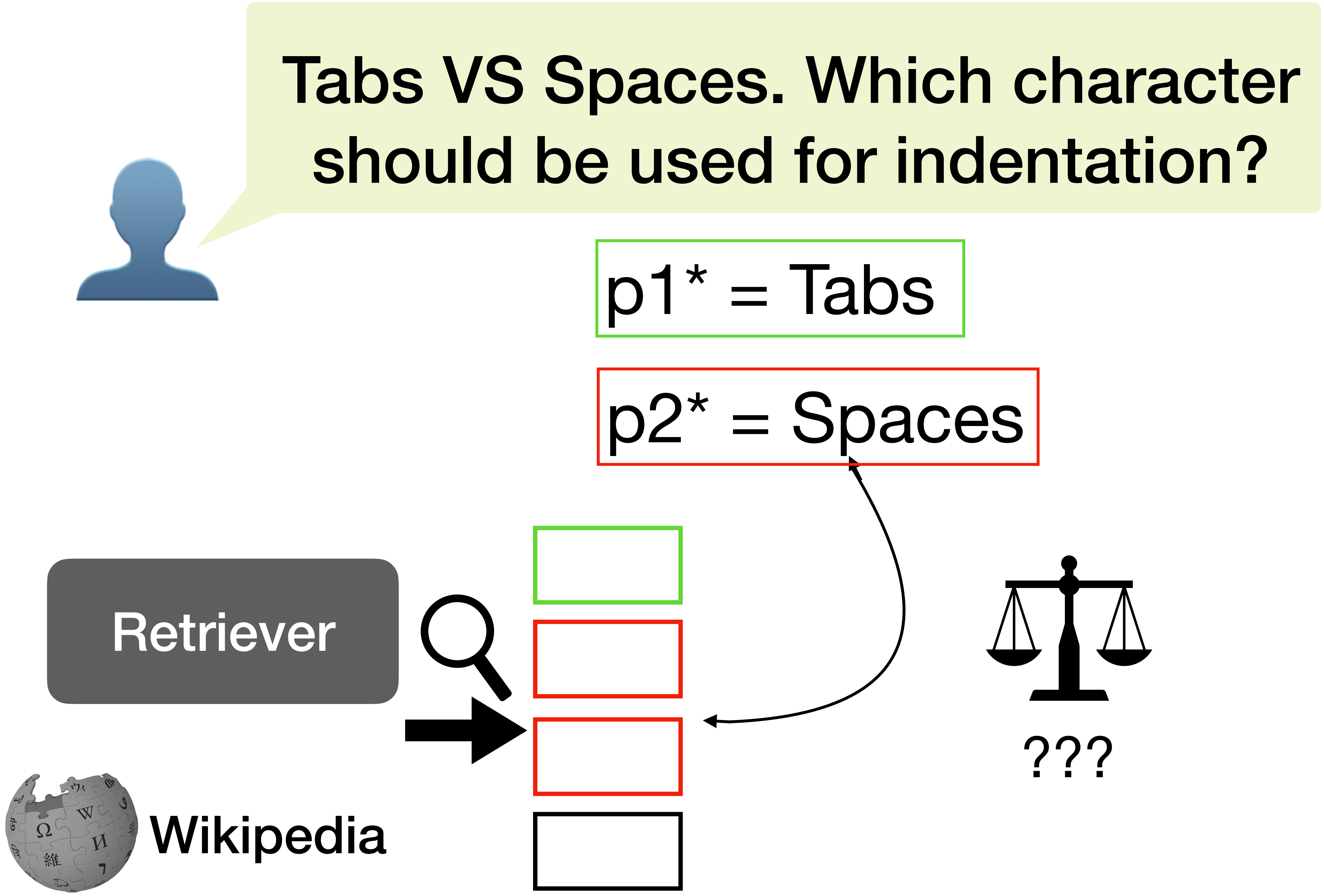
Tabs VS Spaces. Which character should be used for indentation?

$p1^*$ = Tabs

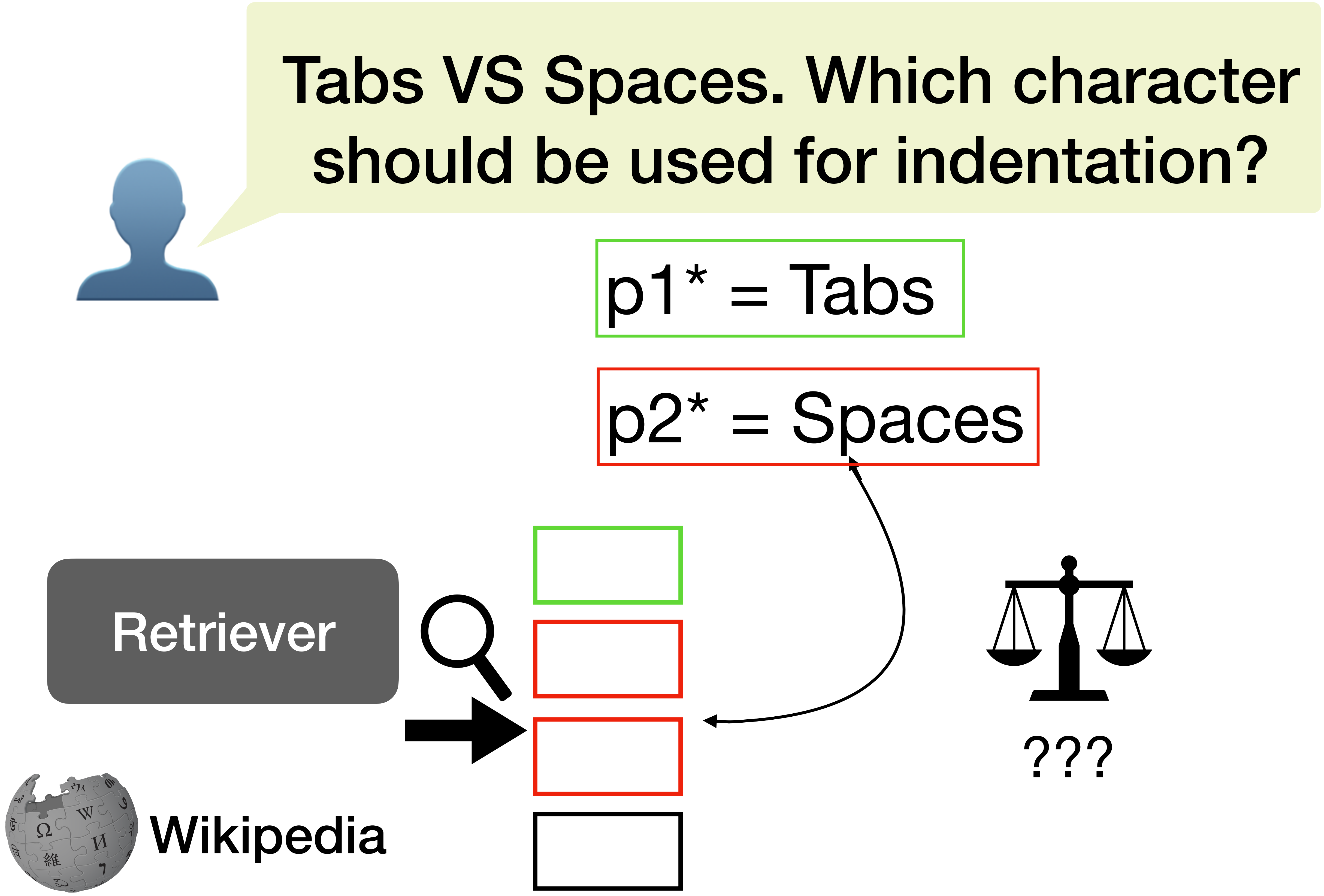
$p2^*$ = Spaces

We Can’t Re-Use Existing Evaluation Setup

1. No ground truth indices
2. String match doesn’t work


Tabs VS Spaces. Which character should be used for indentation?
p1* = Tabs
p2* = Spaces
Retriever
Wikipedia
???

How do we evaluate?



Tabs VS Spaces. Which character should be used for indentation?

p1* = Tabs

p2* = Spaces

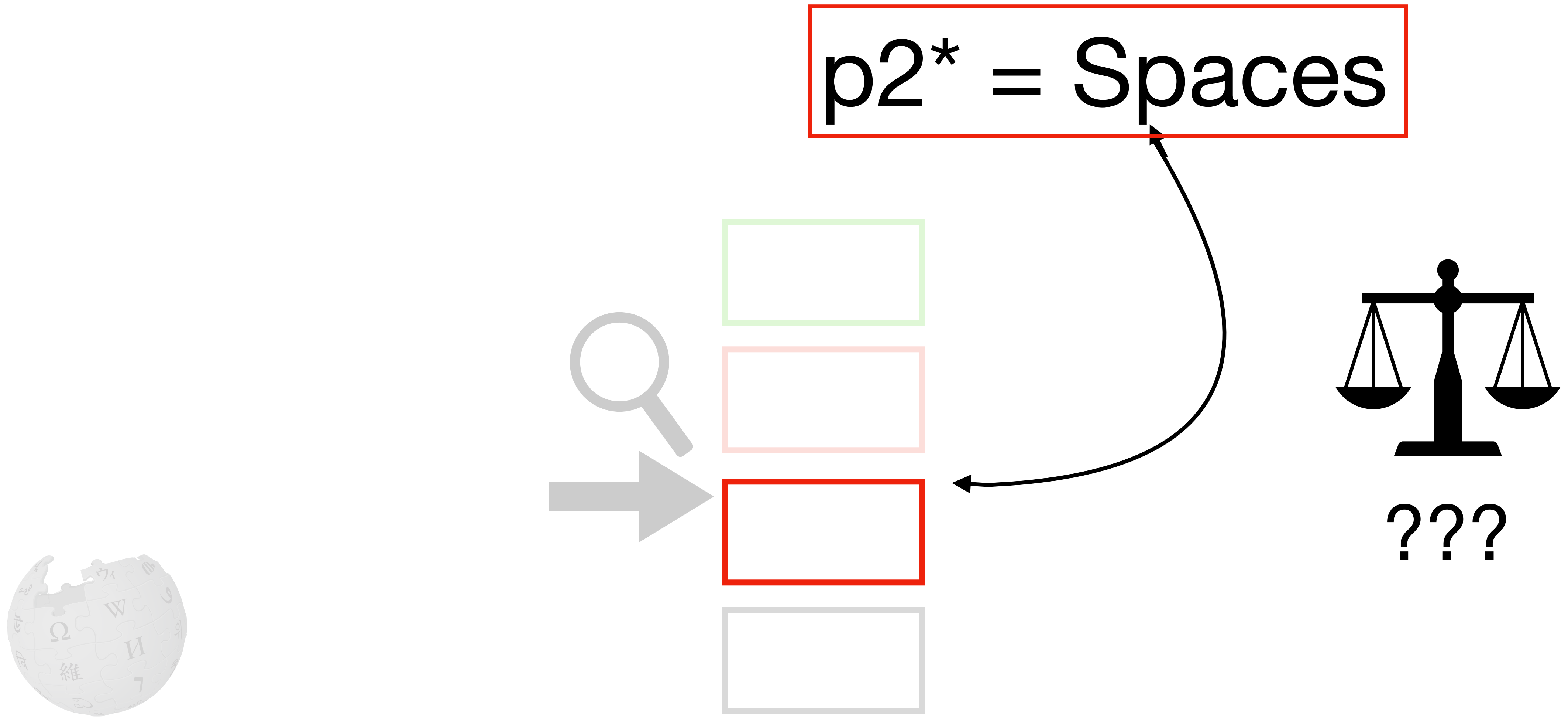
Retriever

Wikipedia

???

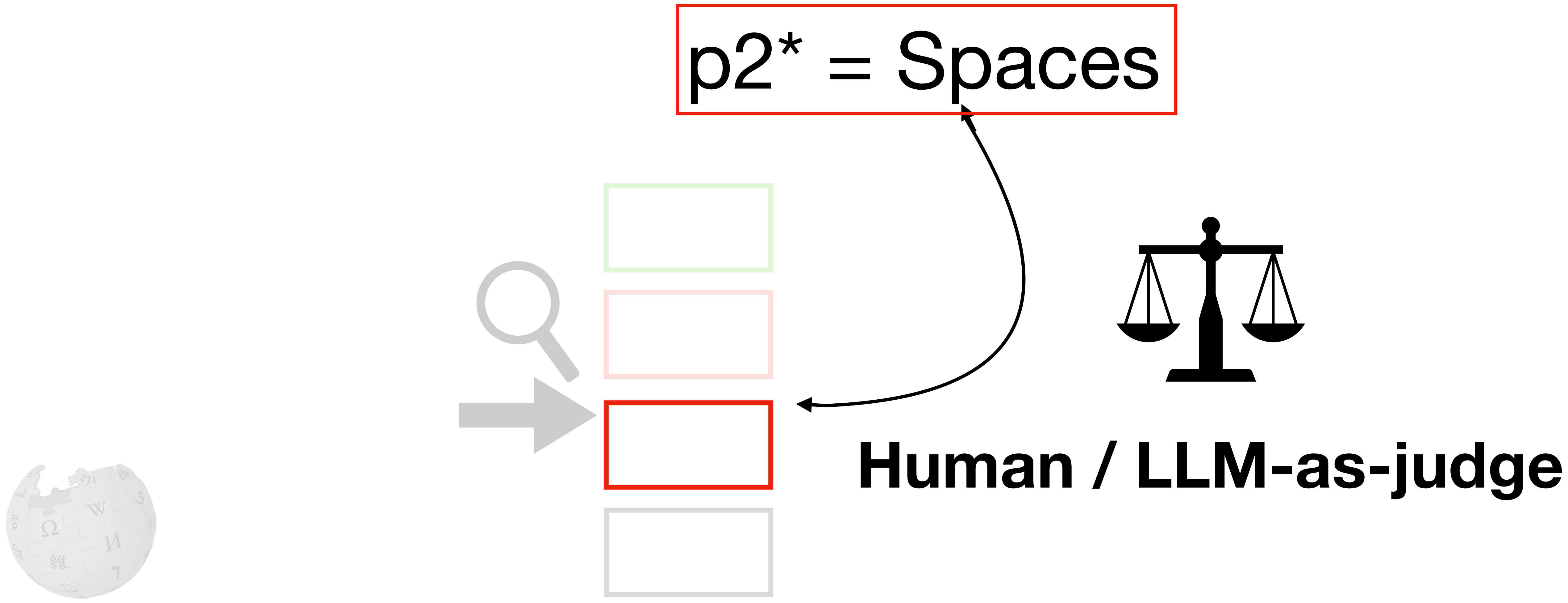
How do we evaluate?

We need to check whether each document contains a perspective.
=> Non-trivial



Human or LLM-as-Judge

Human Evaluation / LLM evaluator



Evaluation Procedure

- Given a (retrieved) document and a perspective, decide whether the document contains the perspective.

(

d_2: … You wouldn’t use a TAB between words in a sentence because that’s the wrong character. Similarly, you shouldn’t use spaces to indent your code, because that’s the wrong character.

d

,

Tab should be used for indentation.

$p*$

)

Evaluation Procedure

- Given a (retrieved) document and a perspective, decide whether the document contains the perspective.

Scoring Function: human or LLM

(

d_2: … You wouldn’t use a TAB between words in a sentence because that’s the wrong character. Similarly, you shouldn’t use spaces to indent your code, because that’s the wrong character.

d

,

Tab should be used for indentation.

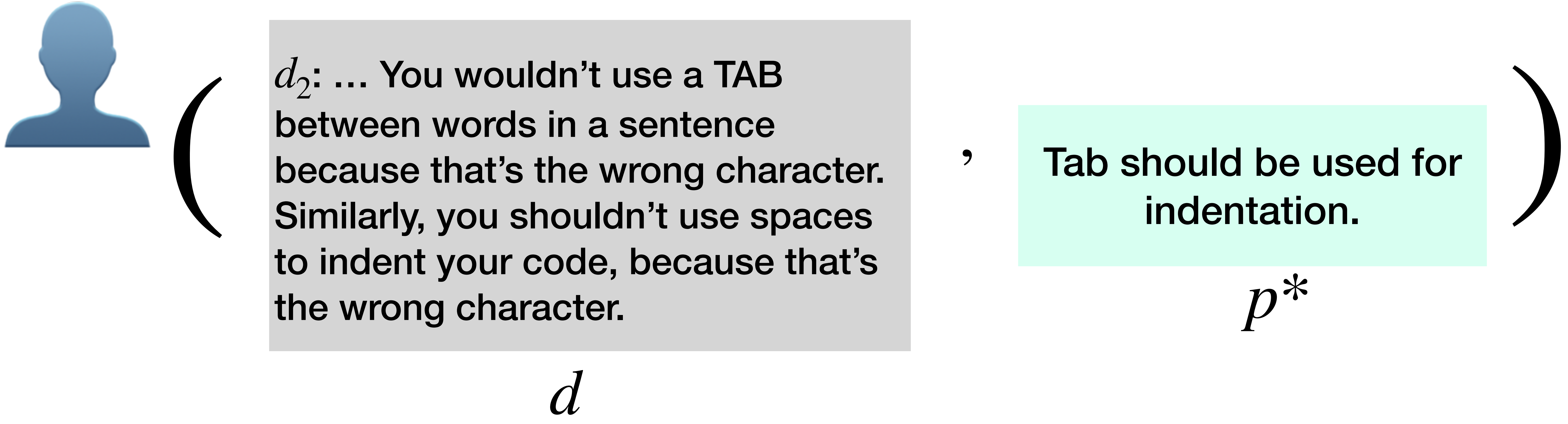
$p*$

)

Evaluation Procedure

- Given a (retrieved) document and a perspective, decide whether the document contains the perspective.

Scoring Function: human or LLM



d_2: … You wouldn't use a TAB between words in a sentence because that's the wrong character. Similarly, you shouldn't use spaces to indent your code, because that's the wrong character.

d

Tab should be used for indentation.

p^*

Evaluation Procedure

- Given a (retrieved) document and a perspective, decide whether the document contains the perspective.

Scoring Function: human or LLM

👤 🤖 (

d_2: … You wouldn't use a TAB between words in a sentence because that's the wrong character. Similarly, you shouldn't use spaces to indent your code, because that's the wrong character.

d

,

Tab should be used for indentation.

p^*

)

Evaluation Procedure

- Given a (retrieved) document and a perspective, decide whether the document contains the perspective.

Scoring Function: human or LLM

👤
🤖 (

d_2: … You wouldn’t use a TAB between words in a sentence because that’s the wrong character. Similarly, you shouldn’t use spaces to indent your code, because that’s the wrong character.

d

,

Tab should be used for indentation.

p^*

)

= 1

Evaluation Procedure

- Given a (retrieved) document and a perspective, decide whether the document contains the perspective.

Scoring Function: human or LLM

👤 🤖 $\Big($

d_2: ... You wouldn't use a TAB between words in a sentence because that's the wrong character. Similarly, you shouldn't use spaces to indent your code, because that's the wrong character.

d

,

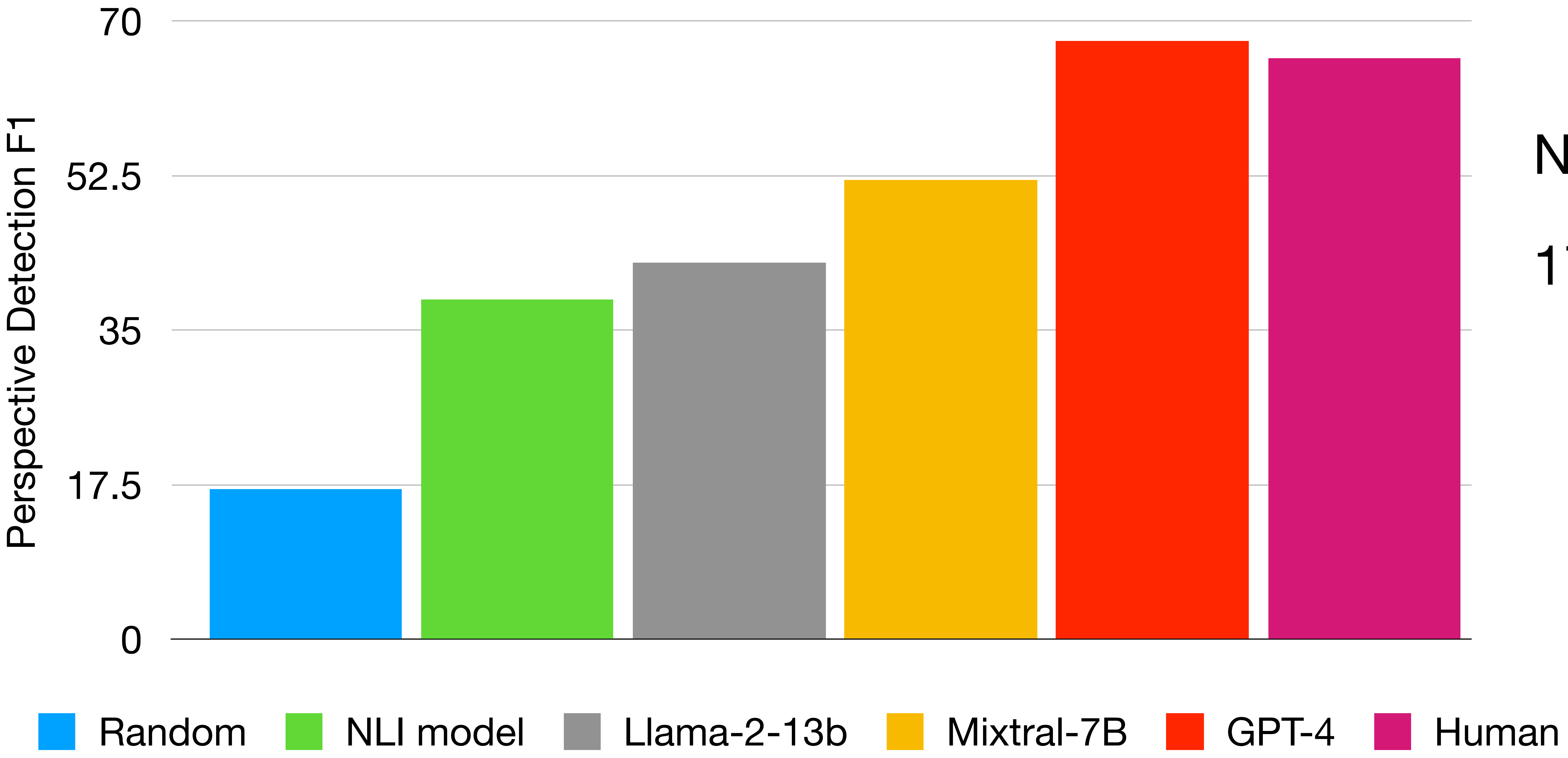
Tab should be used for indentation.

p^*

$\Big)$ **= 1**

- Evaluation is somewhat costly:
  |# retrieved documents| * |# perspectives|

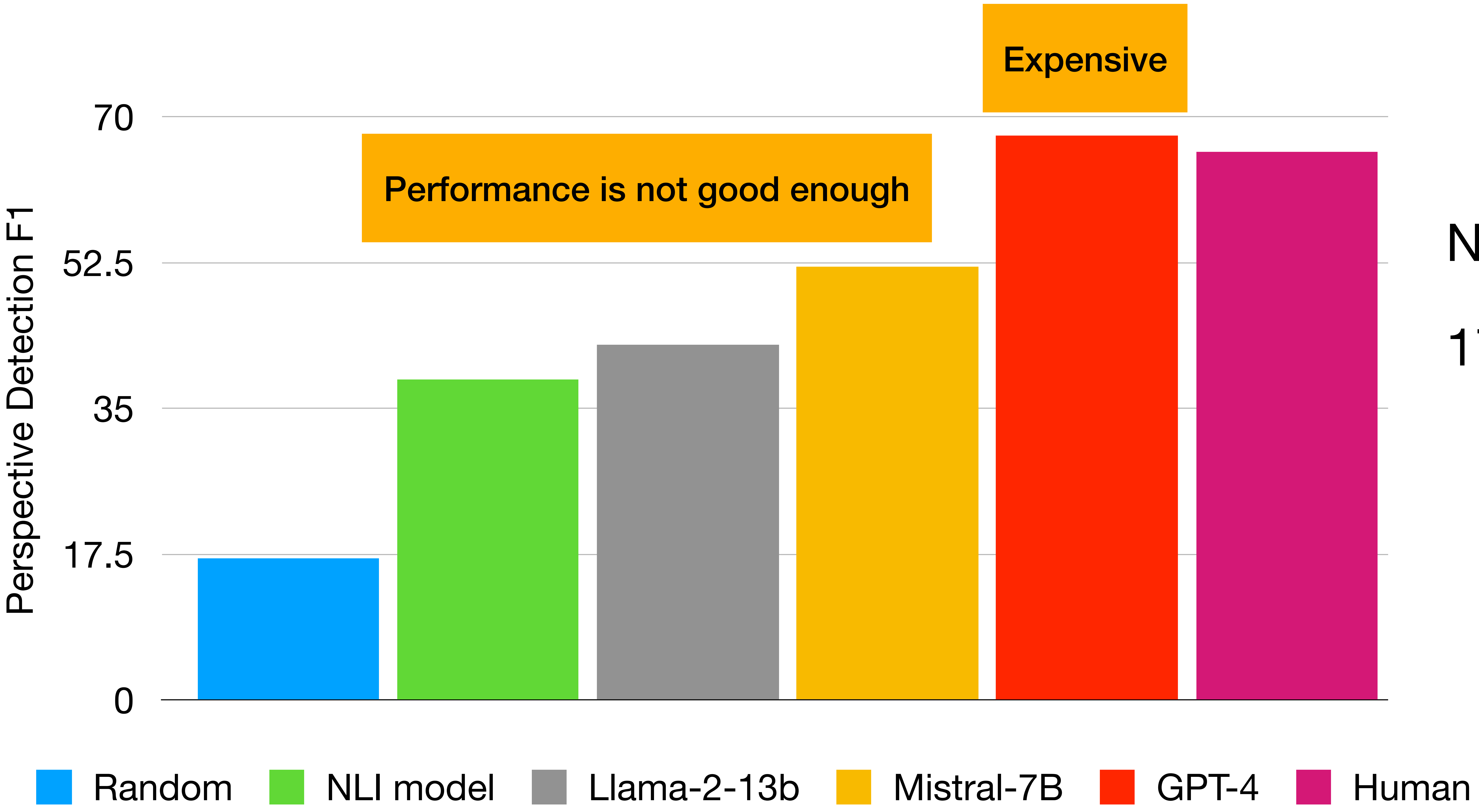
Evaluator Performance



N = 542

17% positive label

Evaluator Performance

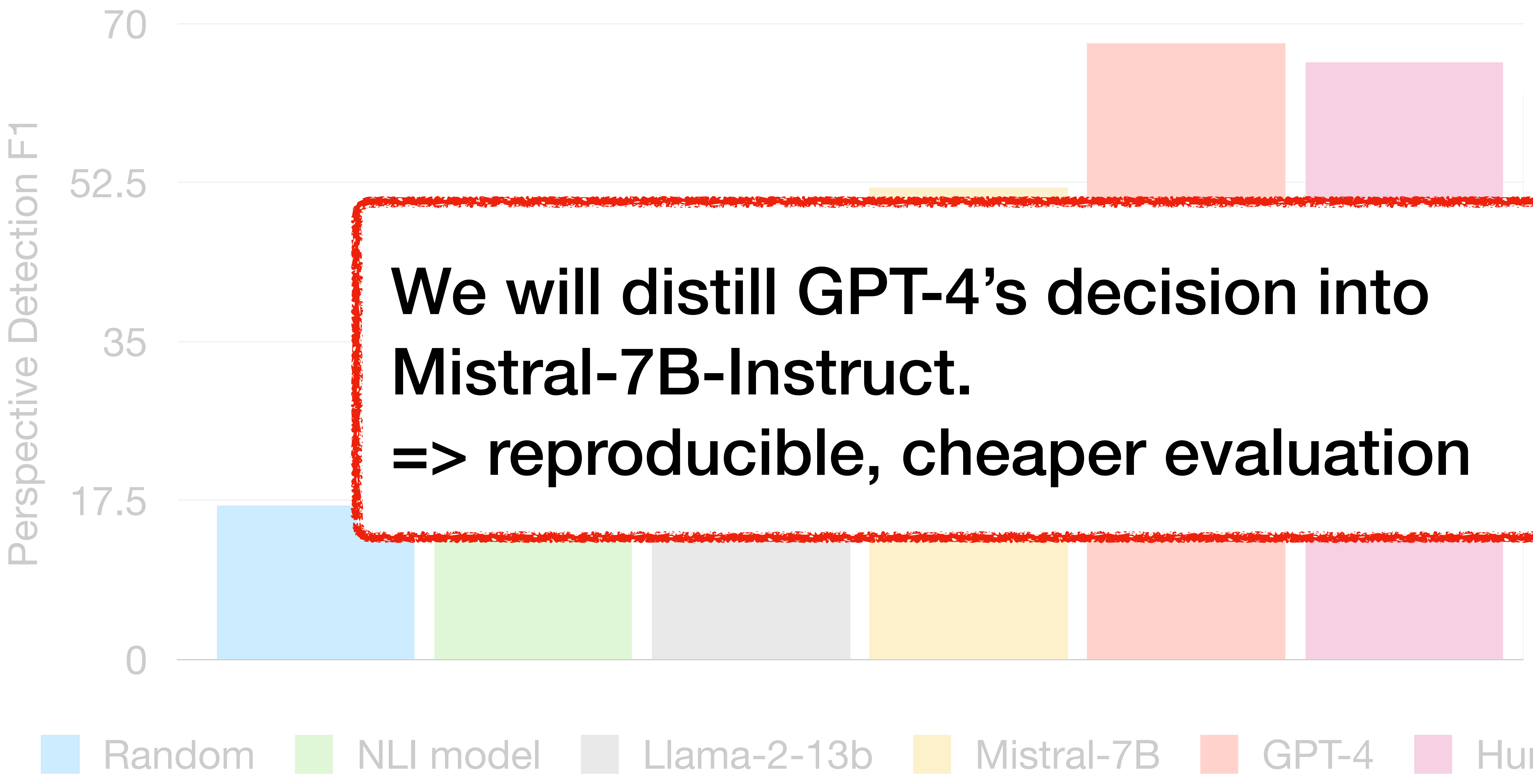


Expensive

Performance is not good enough

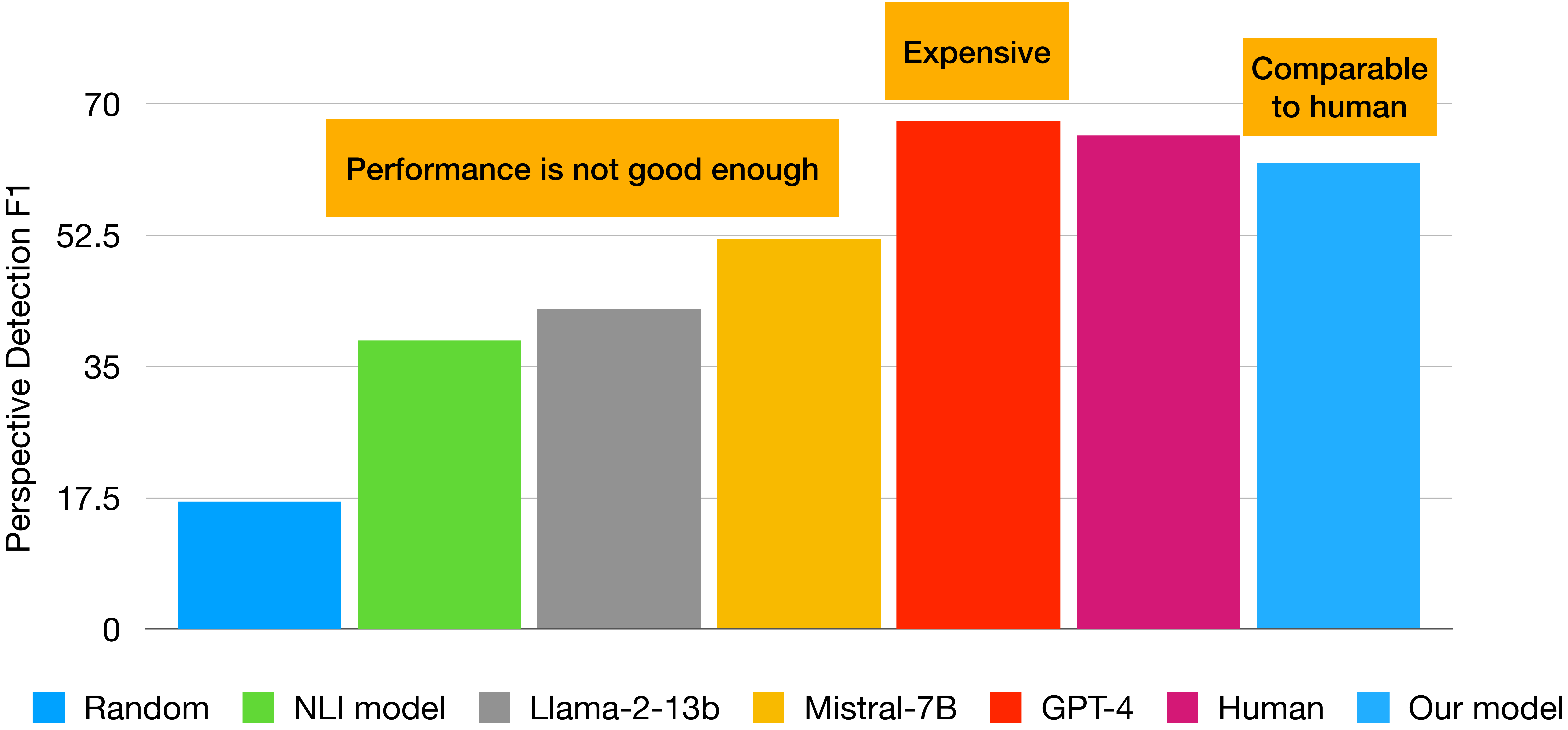
N = 542

17% positive label



We will distill GPT-4's decision into Mistral-7B-Instruct.
=> reproducible, cheaper evaluation

Evaluator Performance



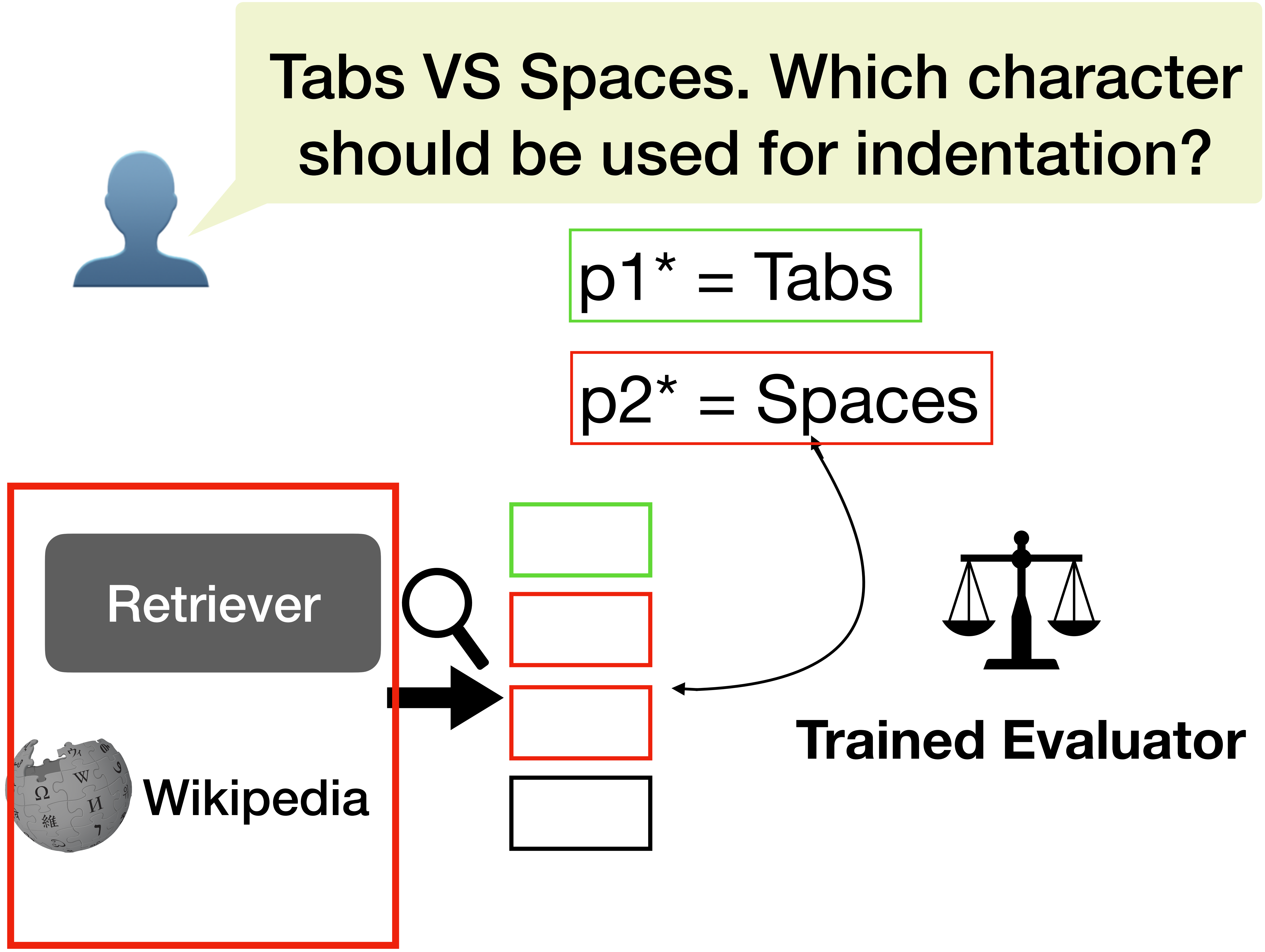
Performance is not good enough

Expensive

Comparable to human

What are we evaluating?

We Evaluate Retriever + Retrieval Corpus

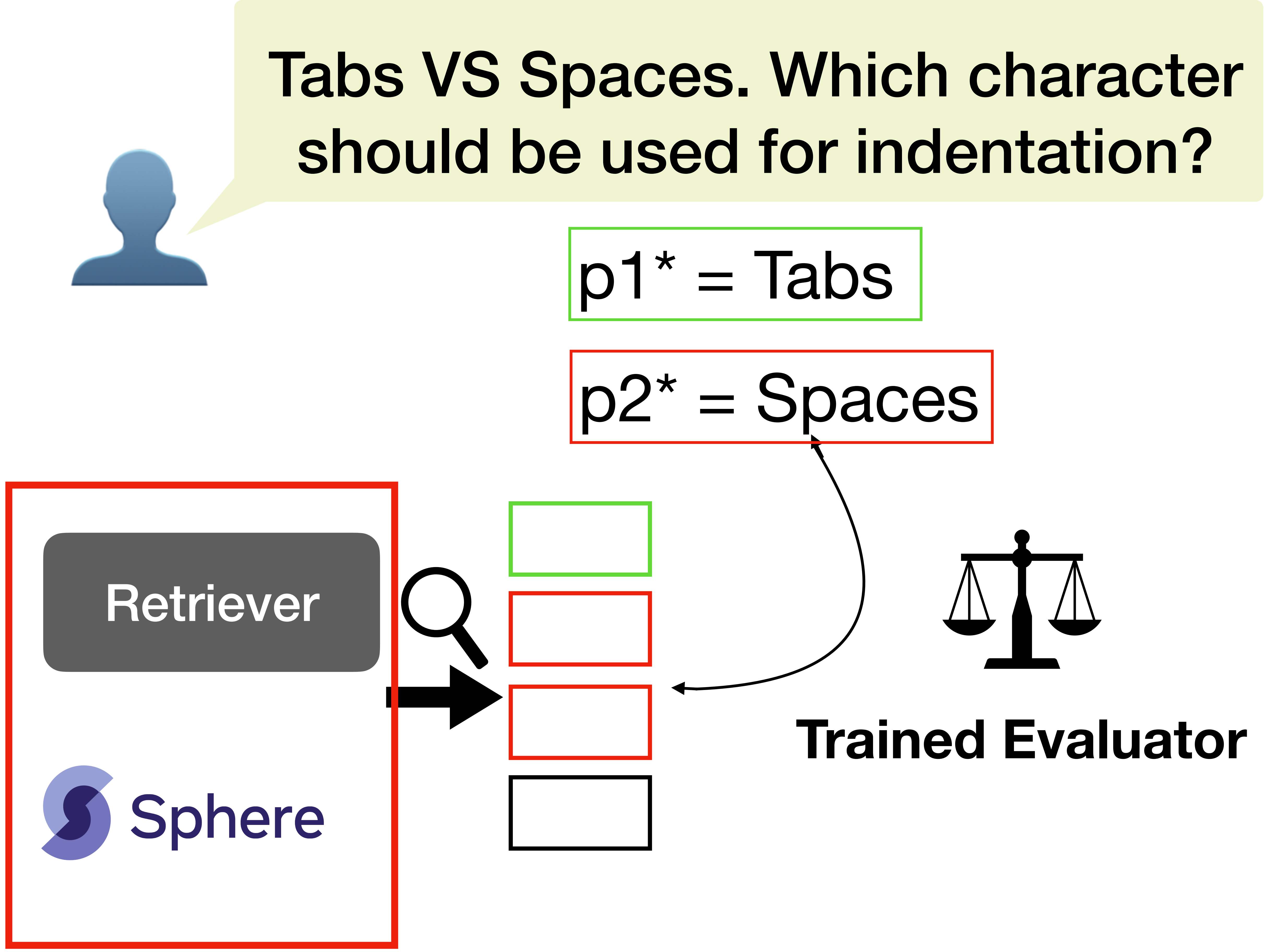


Tabs VS Spaces. Which character should be used for indentation?

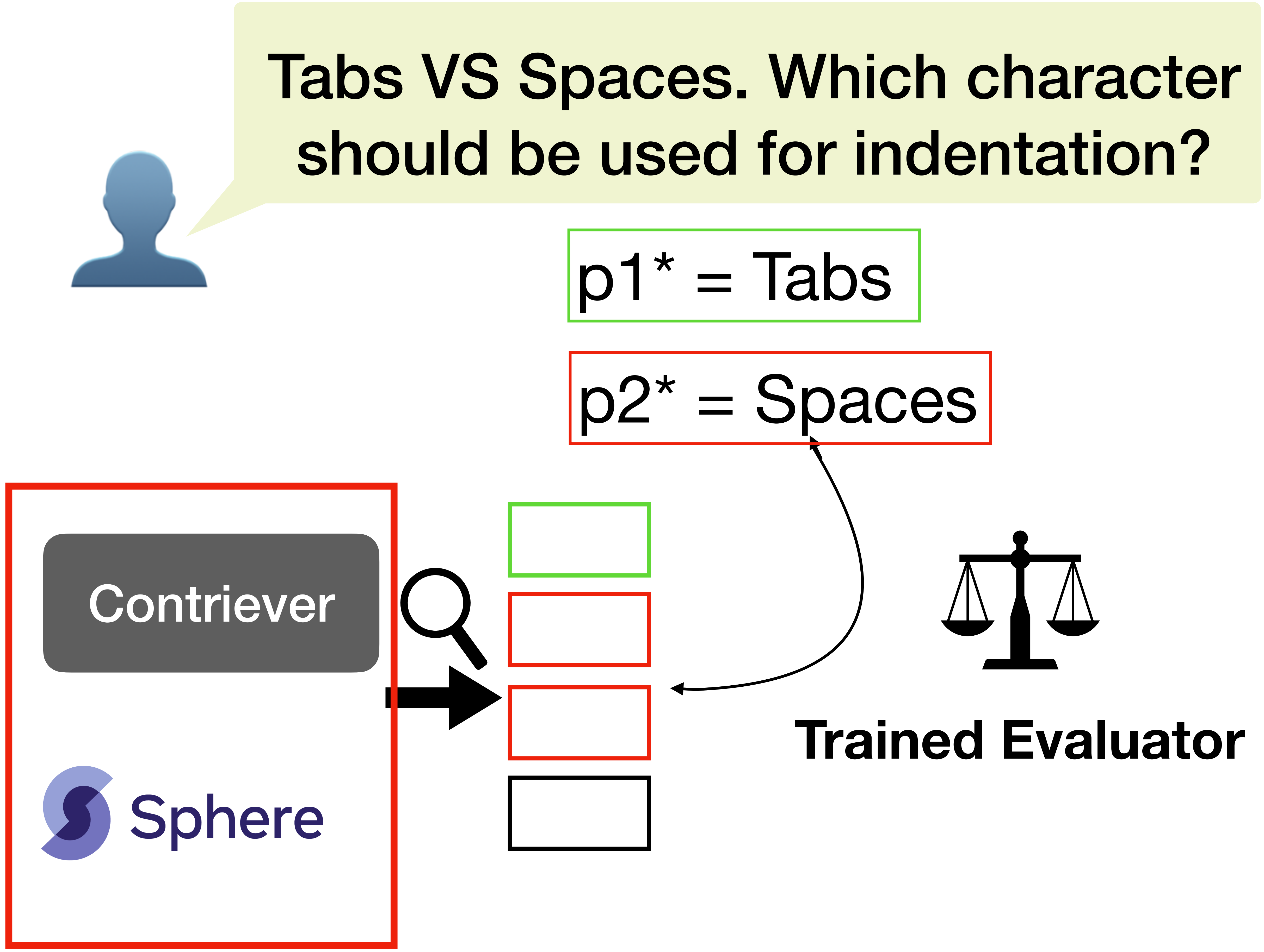
p1* = Tabs

p2* = Spaces

Eval Still Works if We Replace Wikipedia


Tabs VS Spaces. Which character should be used for indentation?
p1* = Tabs
p2* = Spaces
Retriever
Sphere
Trained Evaluator

Evaluate Different Retriever X Corpus Combinations



Tabs VS Spaces. Which character should be used for indentation?

p1* = Tabs

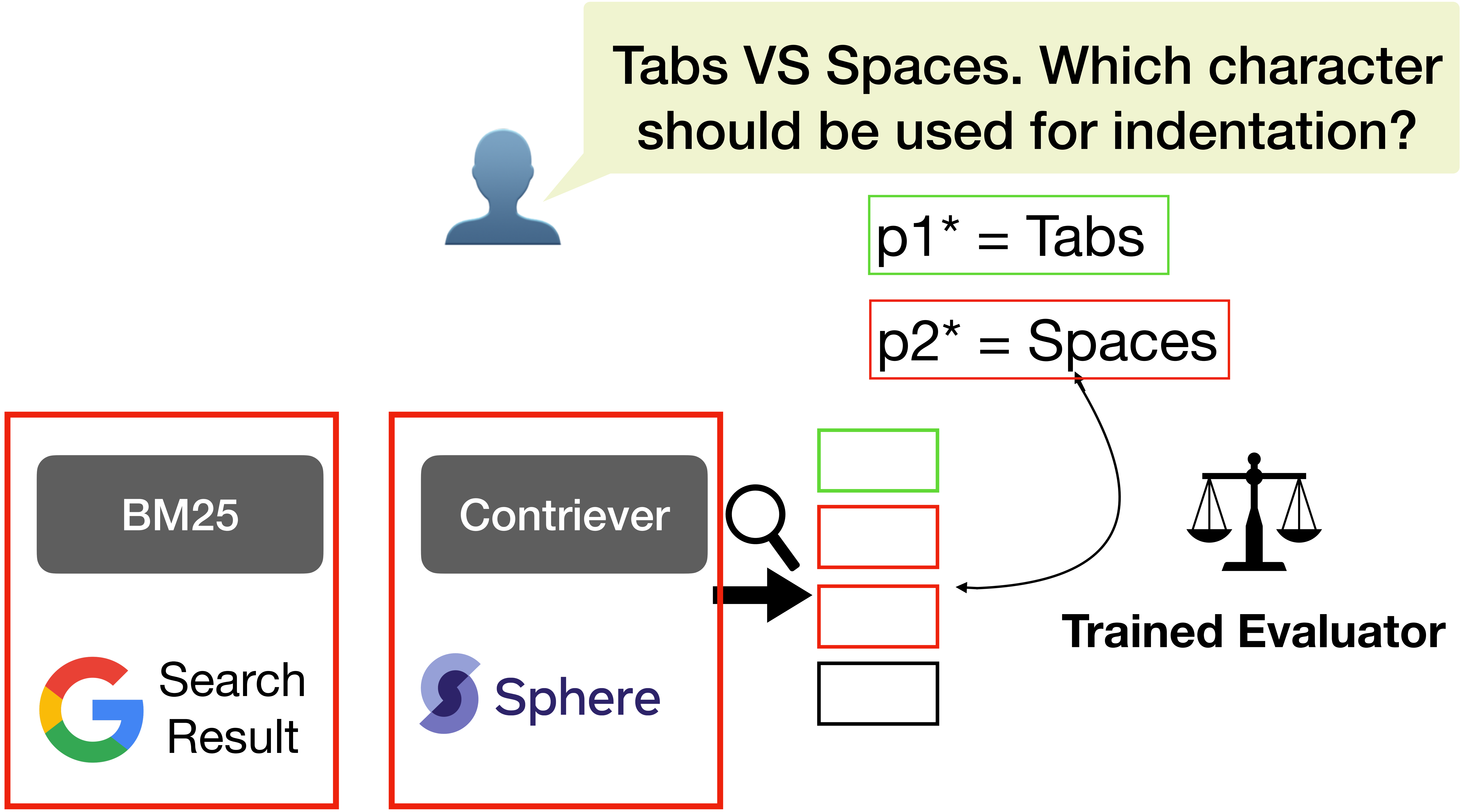
p2* = Spaces

Contriever

Sphere

Trained Evaluator

Evaluate Different Retriever X Corpus Combinations

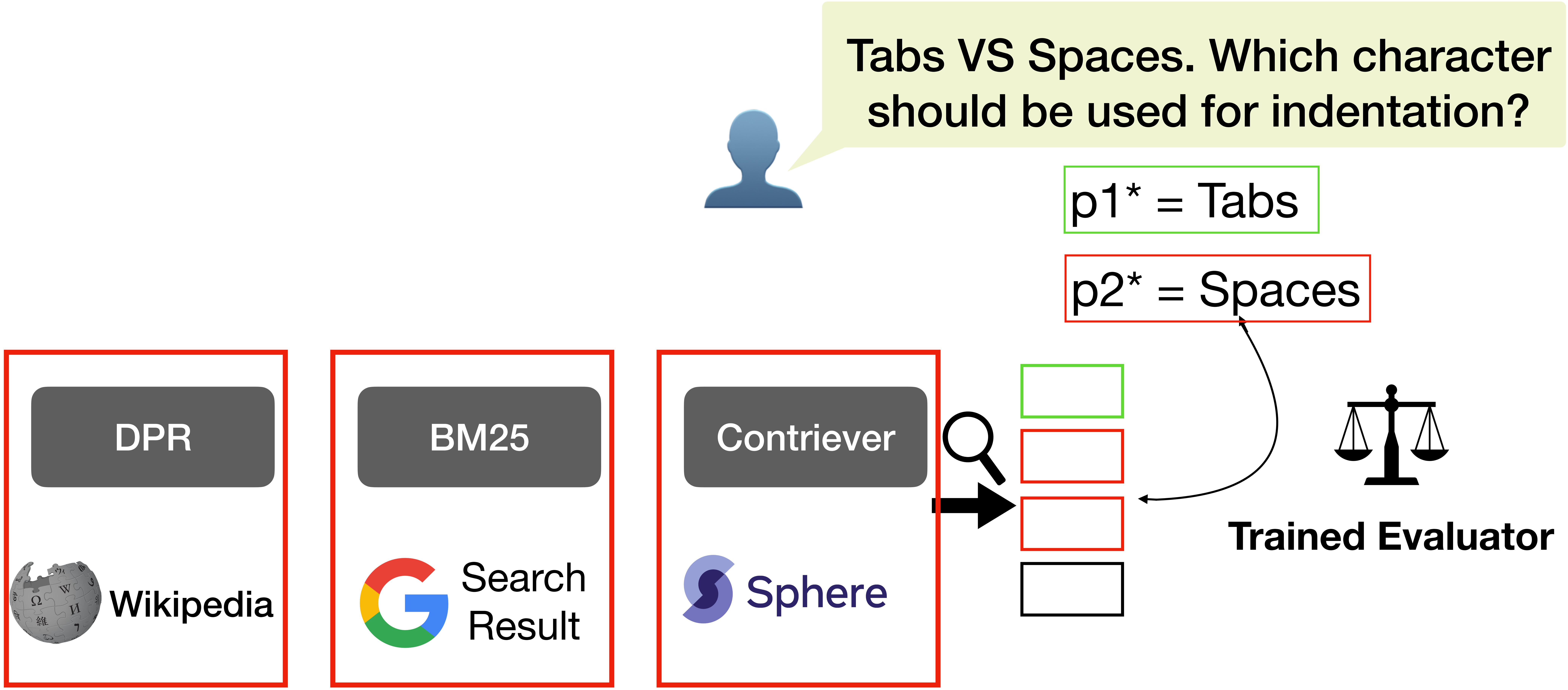


Tabs VS Spaces. Which character should be used for indentation?

p1* = Tabs

p2* = Spaces

Evaluate Different Retriever X Corpus Combinations



Tabs VS Spaces. Which character should be used for indentation?

p1* = Tabs

p2* = Spaces

DPR

Wikipedia

BM25

Search Result

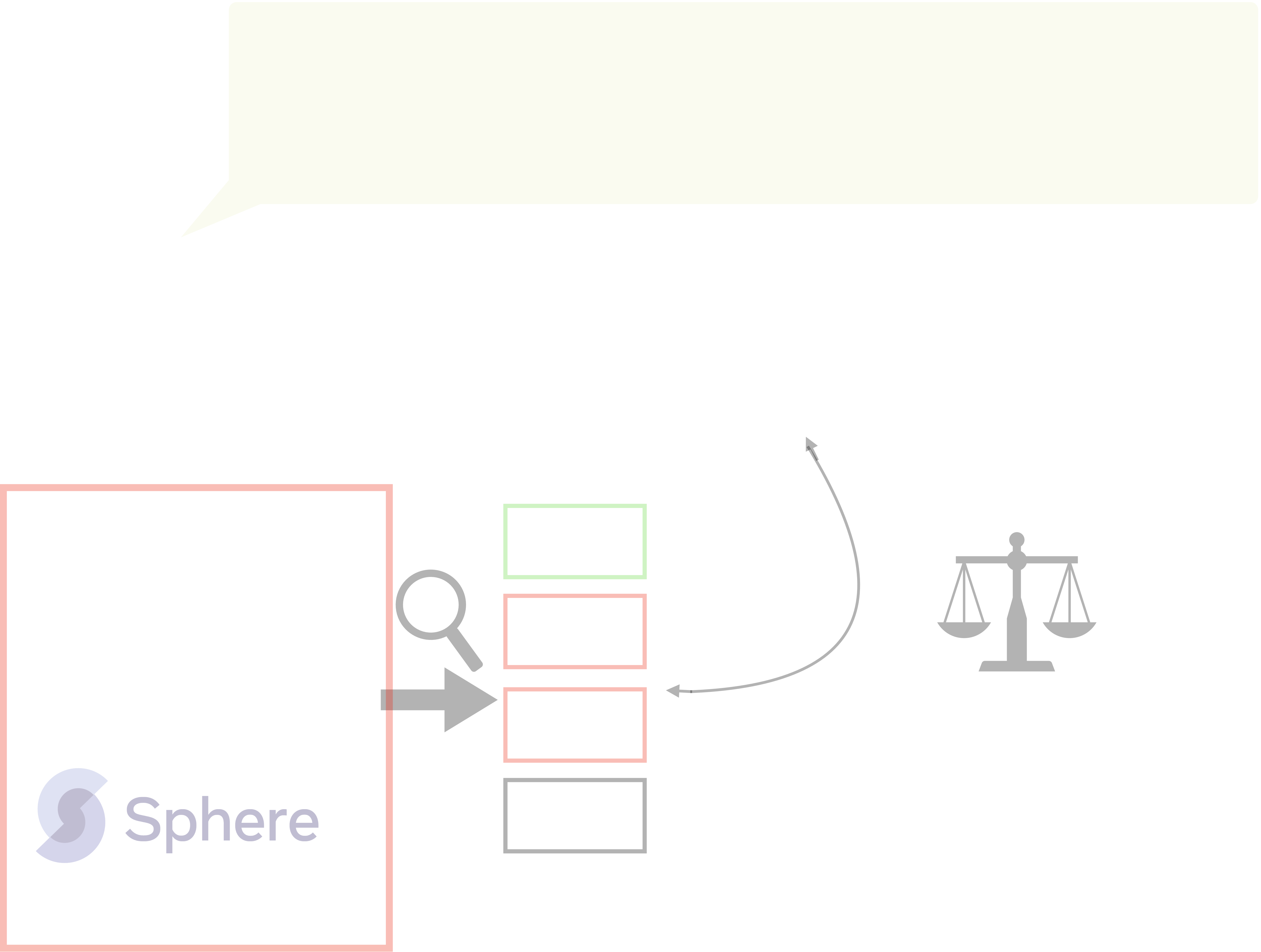
Contriever

Sphere

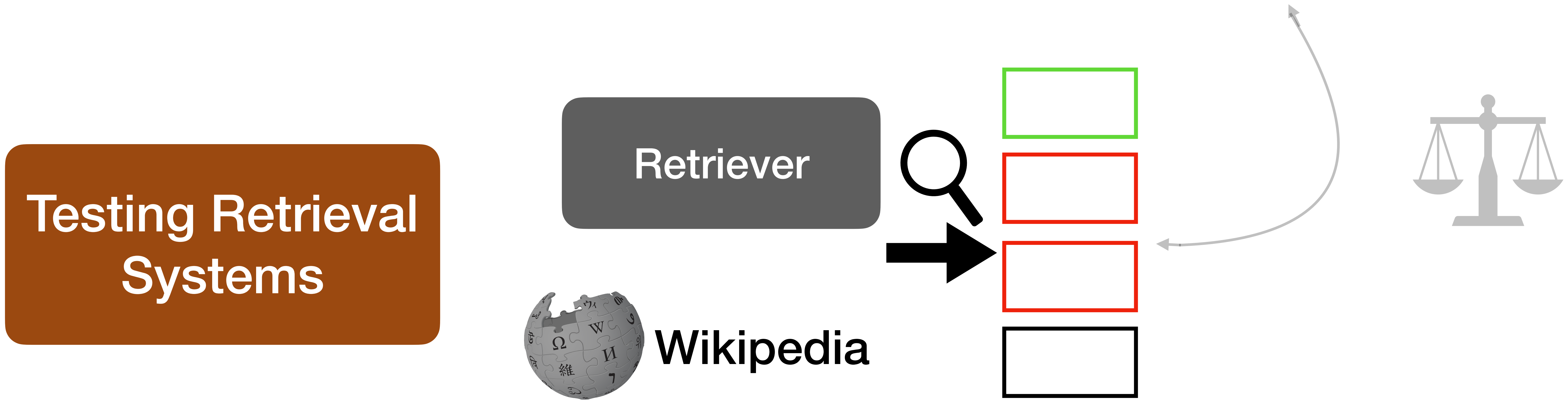
Trained Evaluator

We Evaluate Retriever + Retrieval Corpus

1. This more closely follows how users observe outputs in the real world.
2. There is no need to annotate the corpus when building retrieval datasets.



Evaluating Existing Retrieval Systems



Three Types of Corpus

Wikipedia

- Snapshot of English Wikipedia corpus, 5.9M documents, 22M passes

Sphere

- A subset of CCNet, 134M documents, 906M passages

Search Result

- Dynamically constructed per query by taking top 100 documents google search output, 4.5K passages

Evaluated Five Retrievers

Sparse Retriever

- **BM25**

Evaluated Five Retrievers

Sparse Retriever

- **BM25**

Dense Retriever

- **DPR**
- **Contriever (large-scale contrastive)**

Evaluated Five Retrievers

Sparse Retriever

- **BM25**

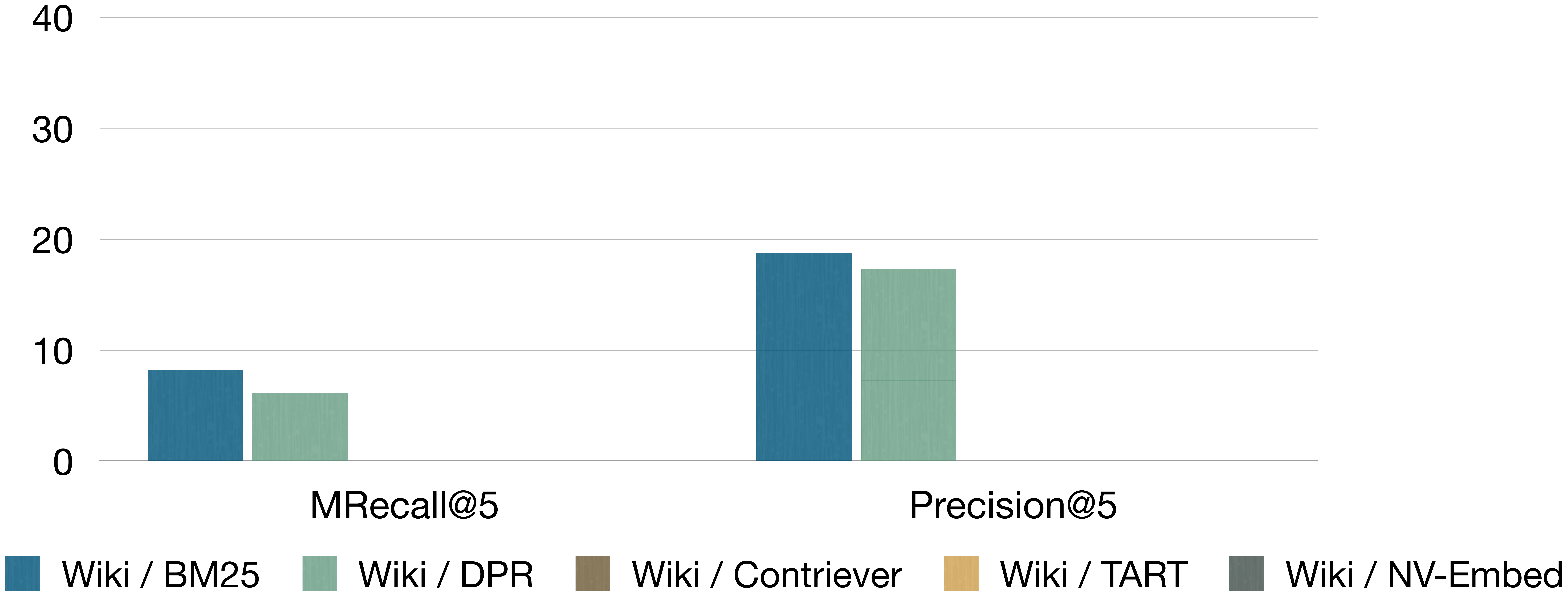
Dense Retriever

- **DPR**
- **Contriever (large-scale contrastive)**

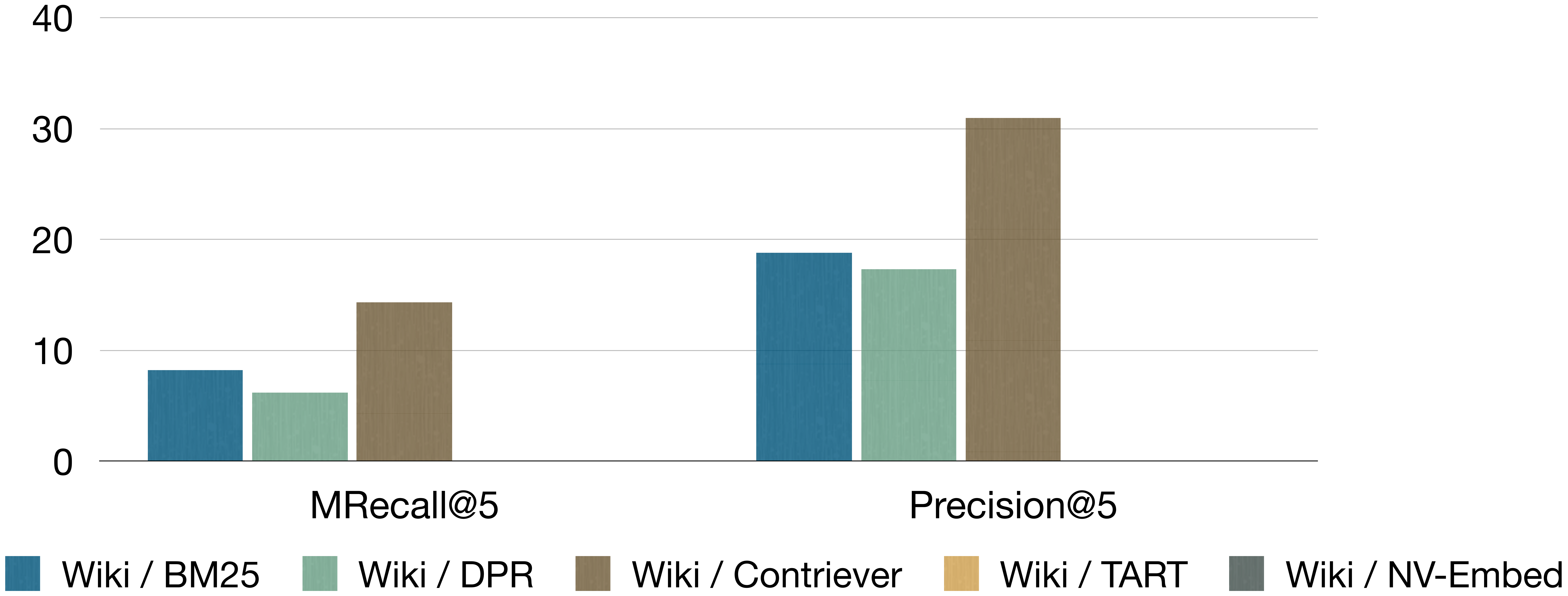
Instruction-Tuned Retriever

- **TART (instruction-aware)**
- **NV-Embed (SOTA then)**

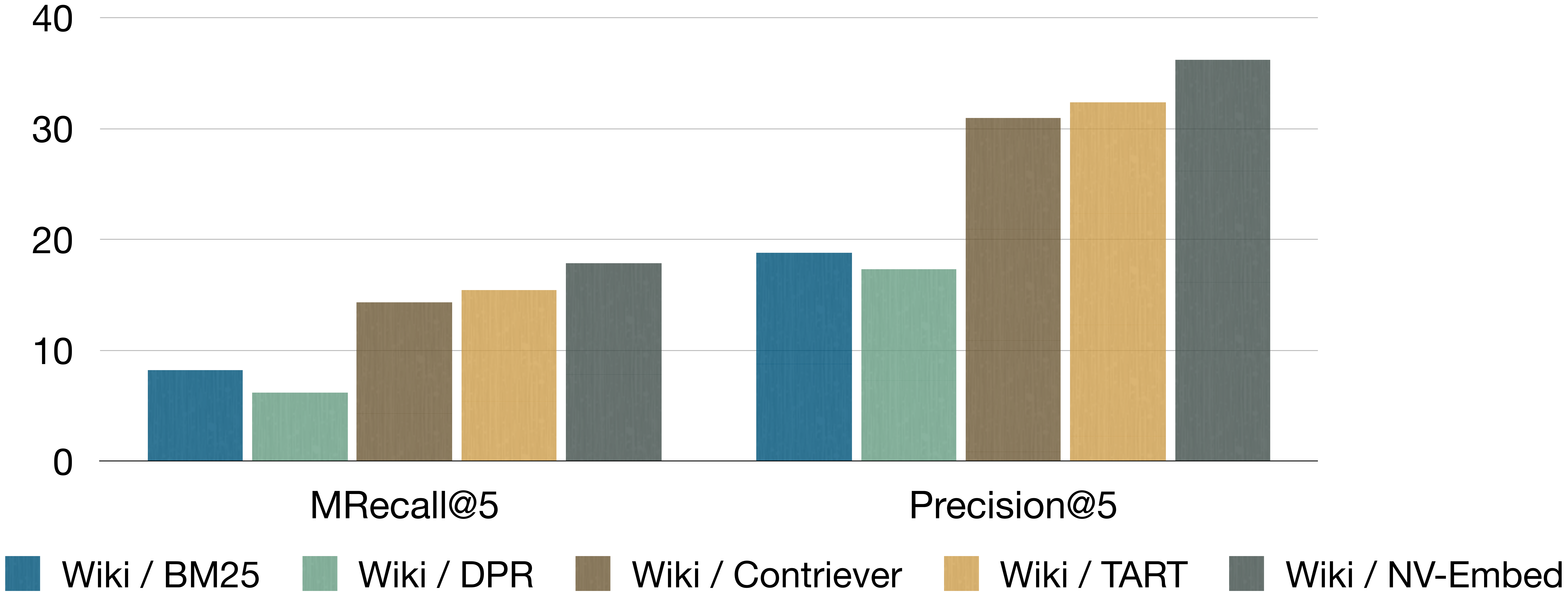
Results - Comparing Retriever (on Wikipedia)



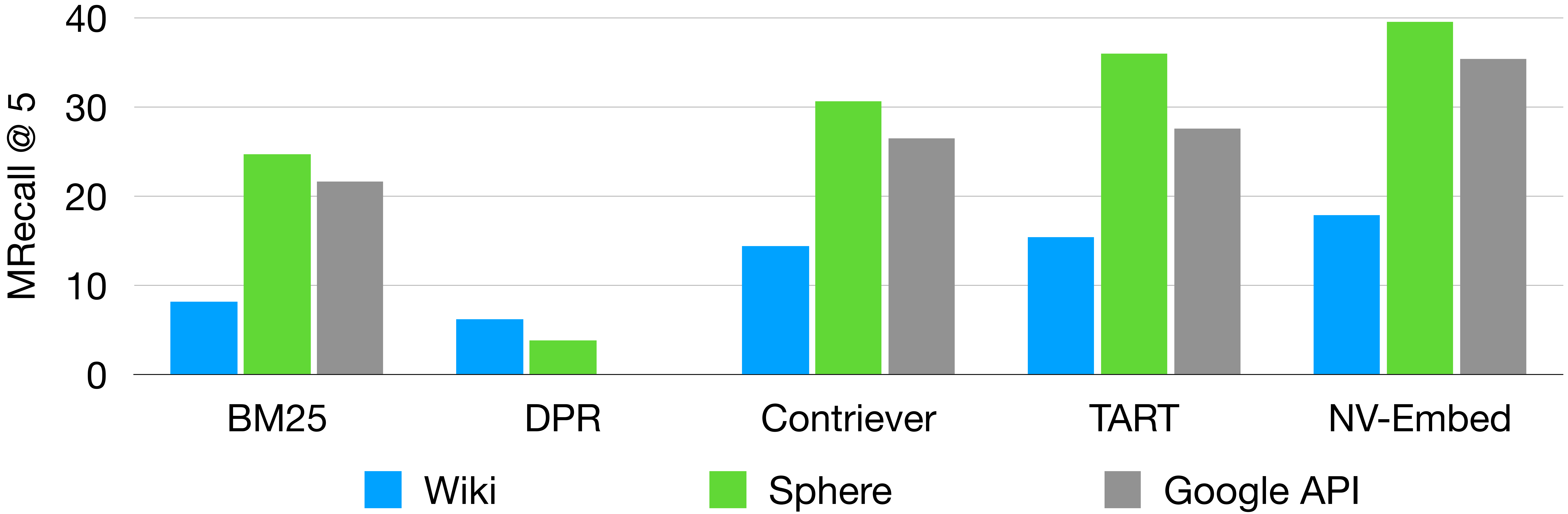
Results - Comparing Retriever (on Wikipedia)



Results - Comparing Retriever (on Wikipedia)

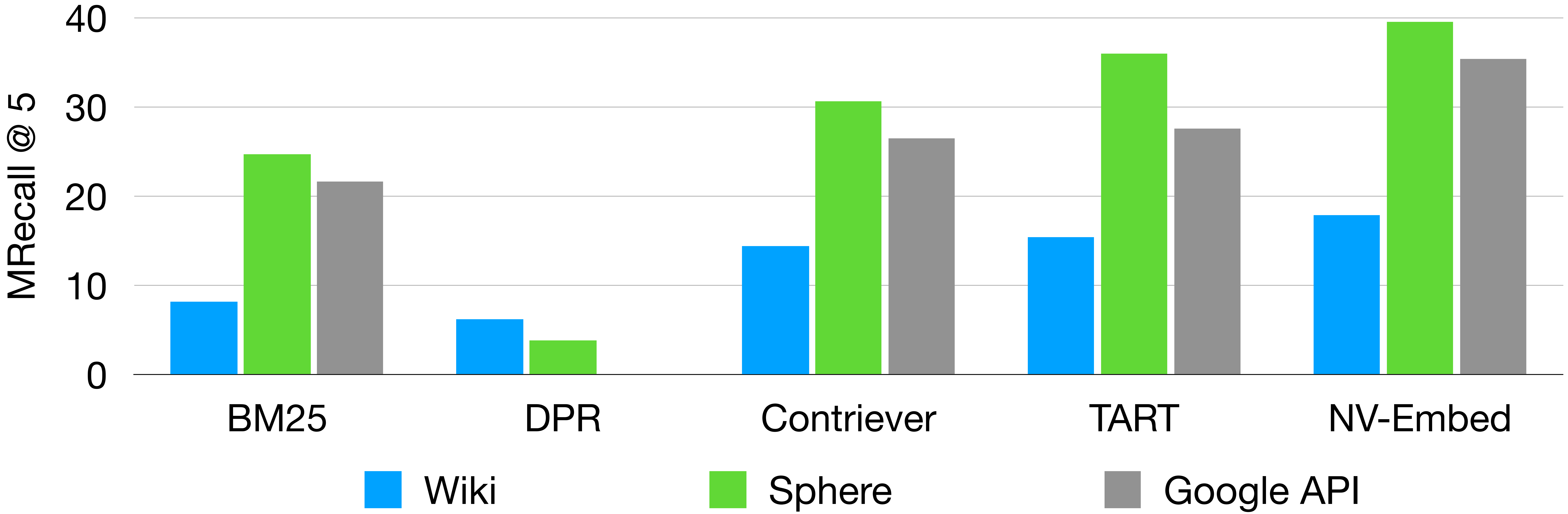


Results - Comparing Across Corpus



- **Main Findings:** Sphere > Google Search > Wikipedia

Results - Comparing Across Corpus



- **Main Findings:** Sphere > Google Search > Wikipedia
- DPR does not generalize to non-Wikipedia corpus

Simple Methods for Improving Diversity

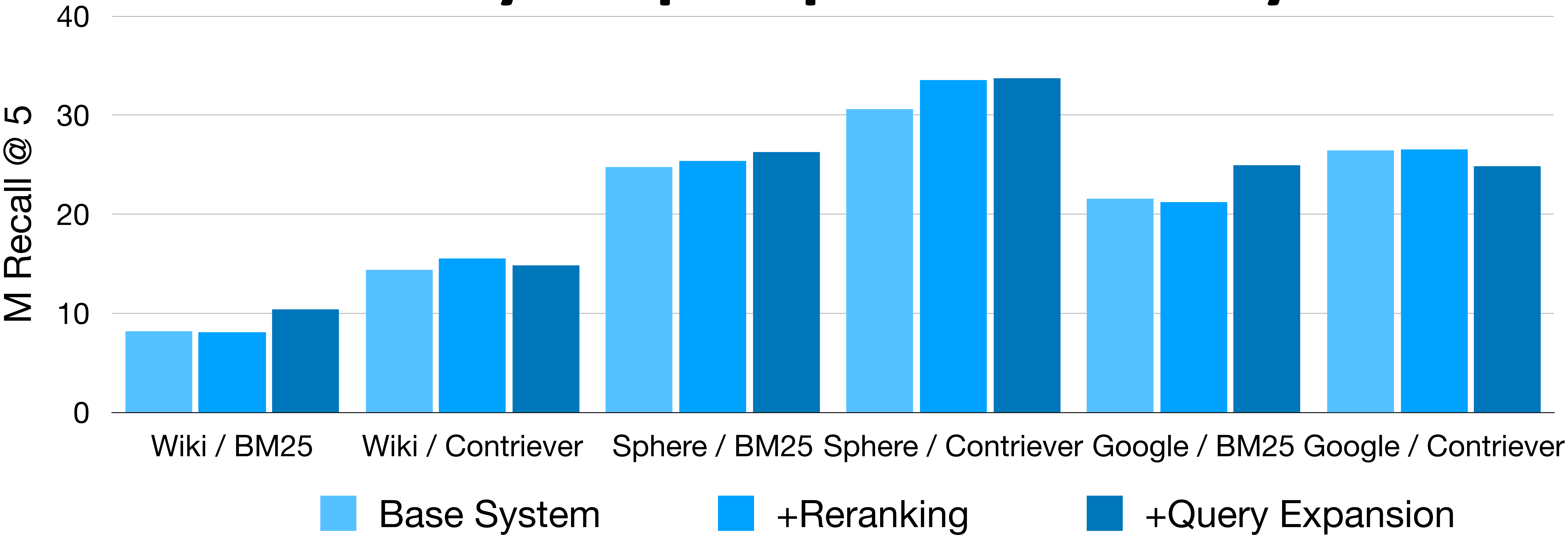
- **Reranking:**
  penalize similar documents via MMR (maximal marginal relevance) objective

Simple Methods for Improving Diversity

- **Reranking:**
  penalize similar documents via MMR (maximal marginal relevance) objective

- **Query Expansion:**
  - 1. Generate diverse new queries using GPT-4
  - 2. Query the retriever using generated queries

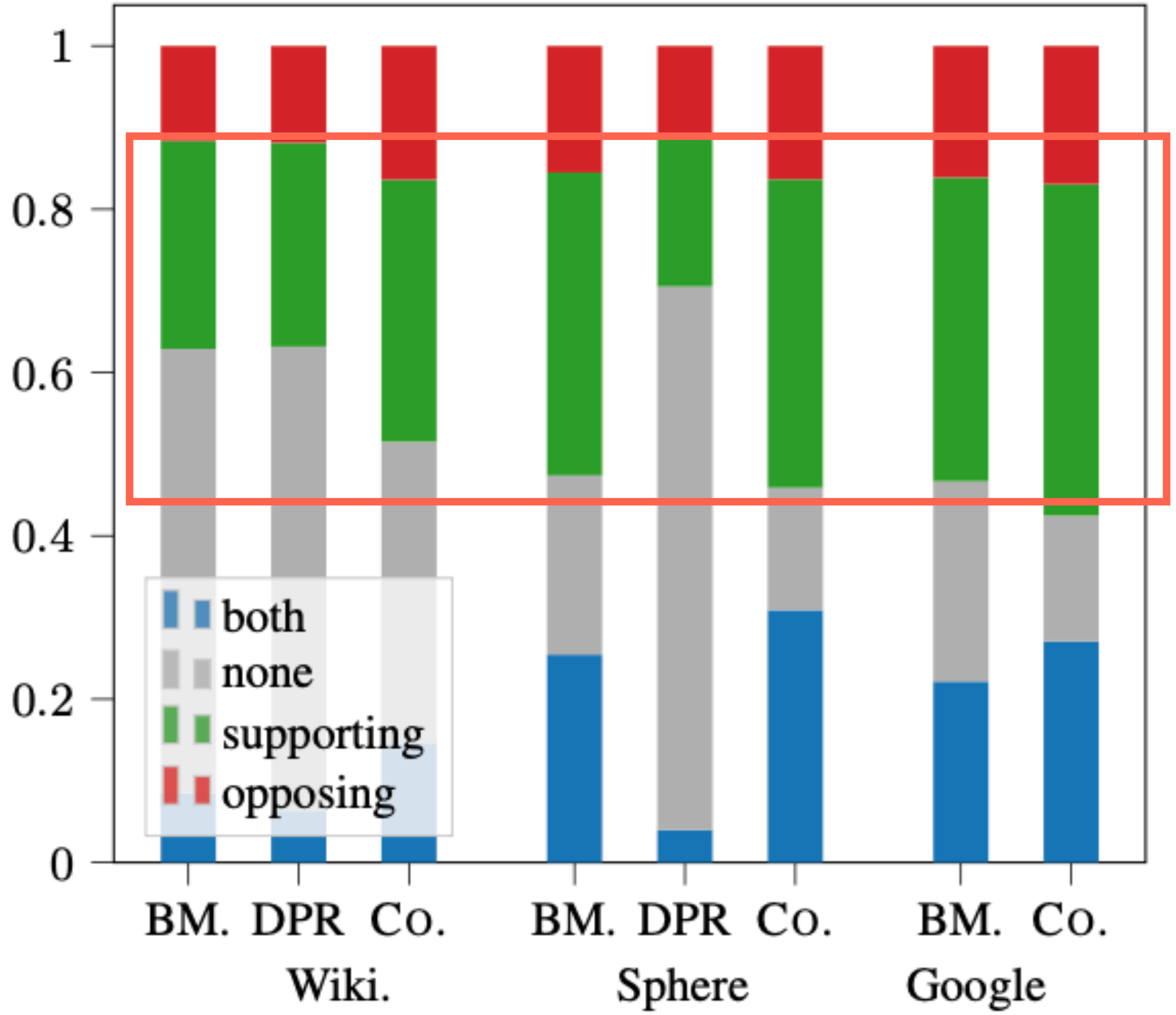
They Help Improve Diversity



- But the improvements are not consistent across retriever / corpus settings
- More work to be done to improve diversity!

Do Retrievers Prefer Supporting vs. Opposing?

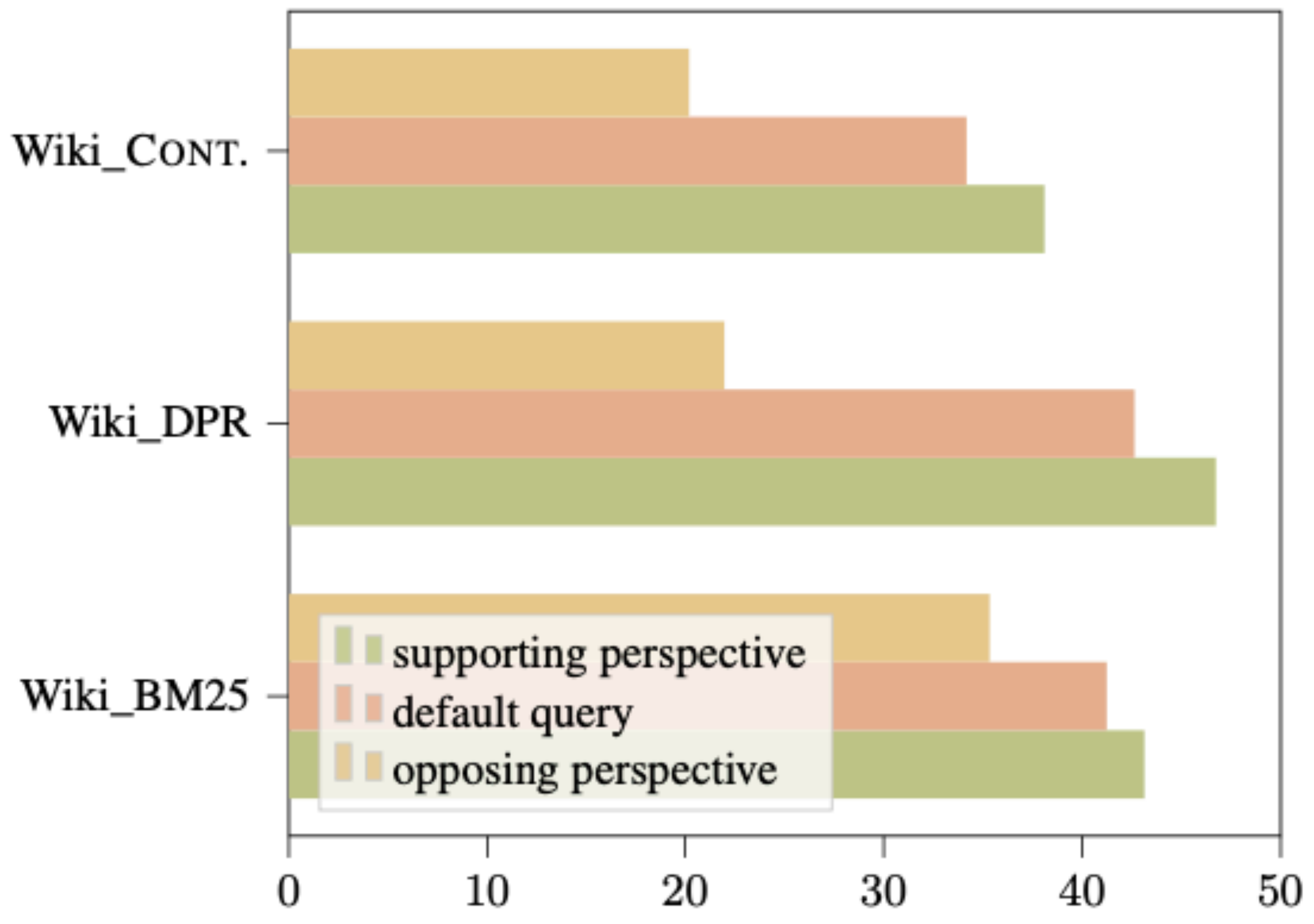
Supporting.



Retriever Sycophancy

- Does retriever returns documents that shares perspective with the query?

✓ Supporting perspective as query leads to more supporting documents



Positive leaning $\Delta = \frac{p-n}{p}$

Takeaways

- We construct the BERDS dataset & evaluation framework for studying retrieval diversity
- Retrieving a set of diverse opinion is challenging, MRecall @ 5 is at best below 40, even when perspectives are binary.
- Both query expansion and query reranking improves diversity, more research to be done!
- Considering LLM-based evaluation can open new exciting research opportunities for retrieval research
  - This can allow evaluating a combination of knowledge source and a retriever
  - No need to annotate the corpus for building a dataset