

# **R<sup>2</sup>LLMs:**

# **Retrieval and Ranking with LLMs**

**Guido Zuccon<sup>1</sup>, Shengyao Zhuang<sup>2</sup>, Xueguang Ma<sup>3</sup>, Bevan Koopman<sup>1,2</sup>**

<sup>1</sup> ielab, The University of Queensland, Australia & Google Research Australia

<sup>2</sup> ielab, CSIRO, Australia

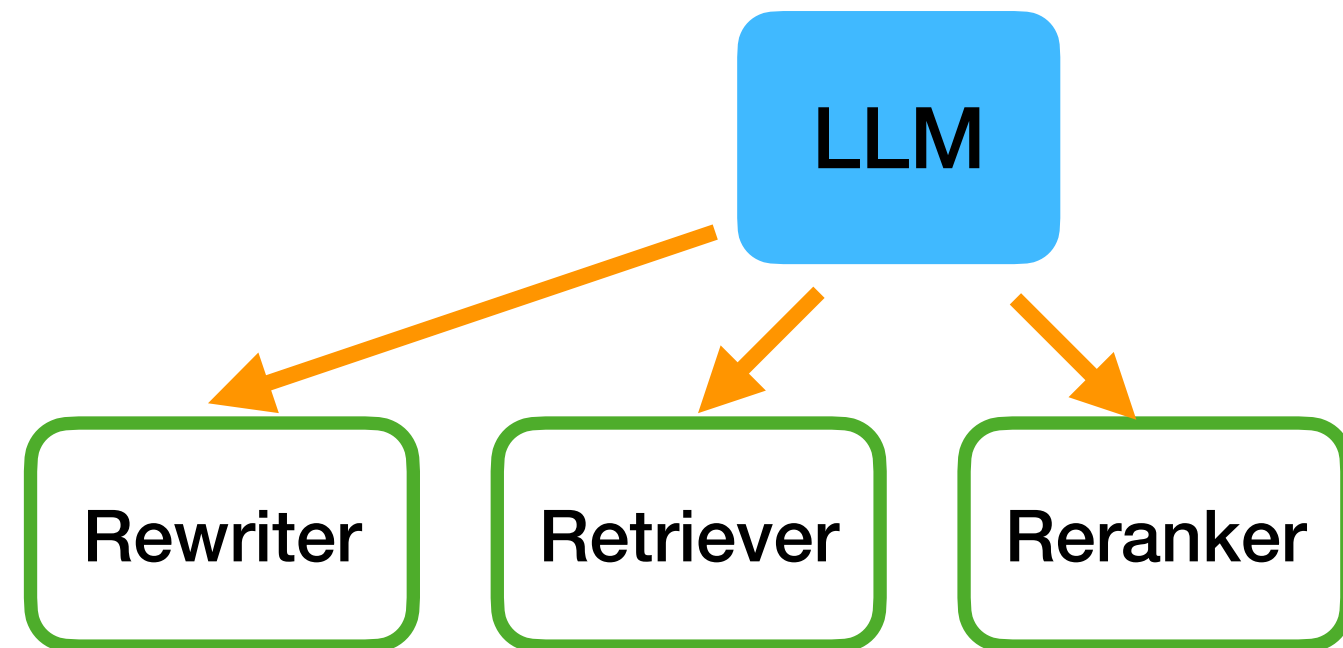
<sup>3</sup> The University of Waterloo

<https://ielab.io/tutorials/r2llms.html>

# Focus of this tutorial:

## Using LLMs for Retrieval and Ranking

### Prompting LLMs for Retrieval and Ranking

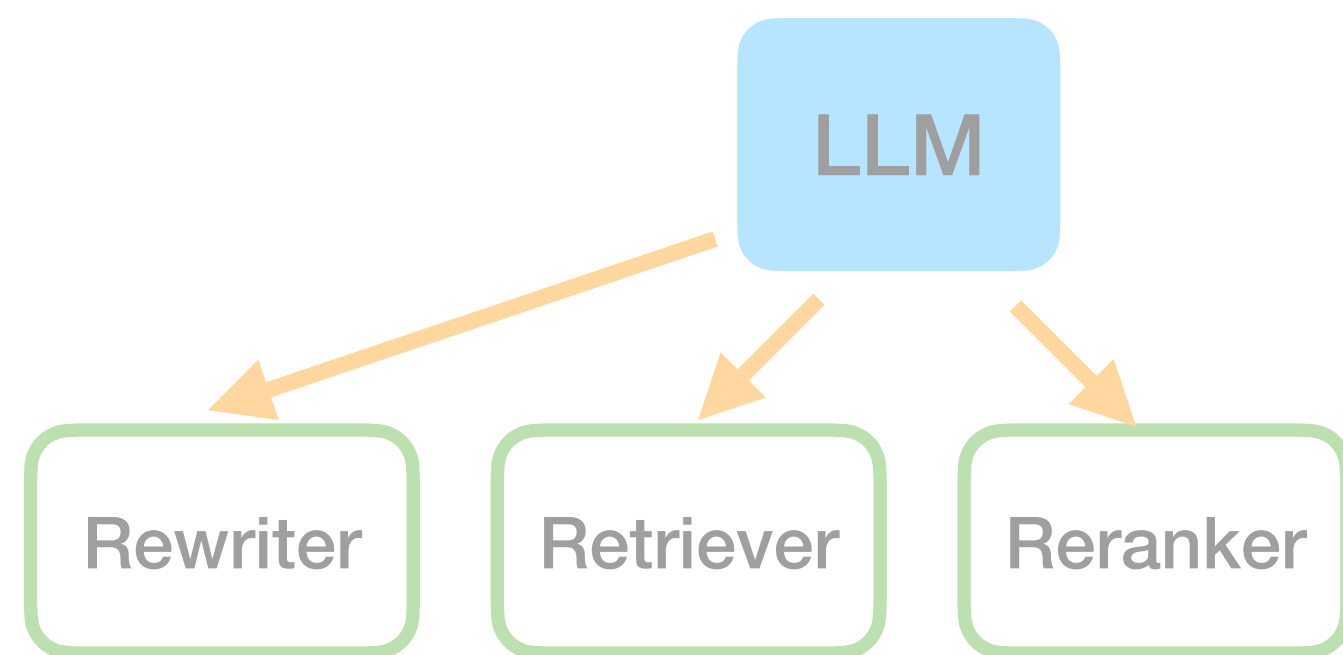


*LLM-enhanced information retrieval*

e.g. RankGPT, PromptReps

# What this tutorial is not about

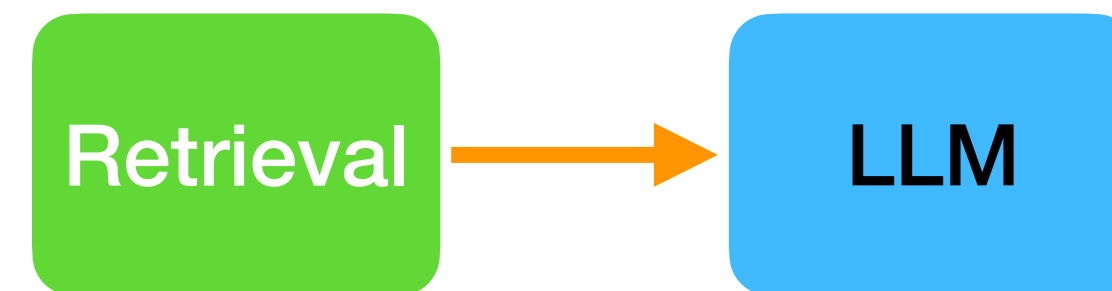
## Prompting LLMs for Retrieval and Ranking



*LLM-enhanced information retrieval*

e.g. RankGPT, PromptReps

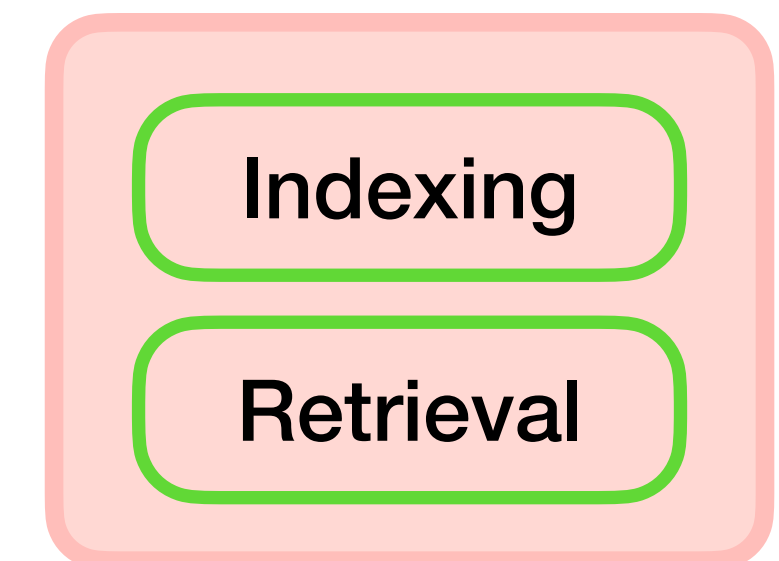
## Retrieval Augmented Generation



*IR4LLM,  
Retrieval-augmented LLM,  
Reliable response generation,  
Search Agents*

e.g. Self-RAG, BlendFilter, etc.

## Model-based IR



*LLM as Indexer and  
Retriever, Generative  
[document] Retrieval (GR)*

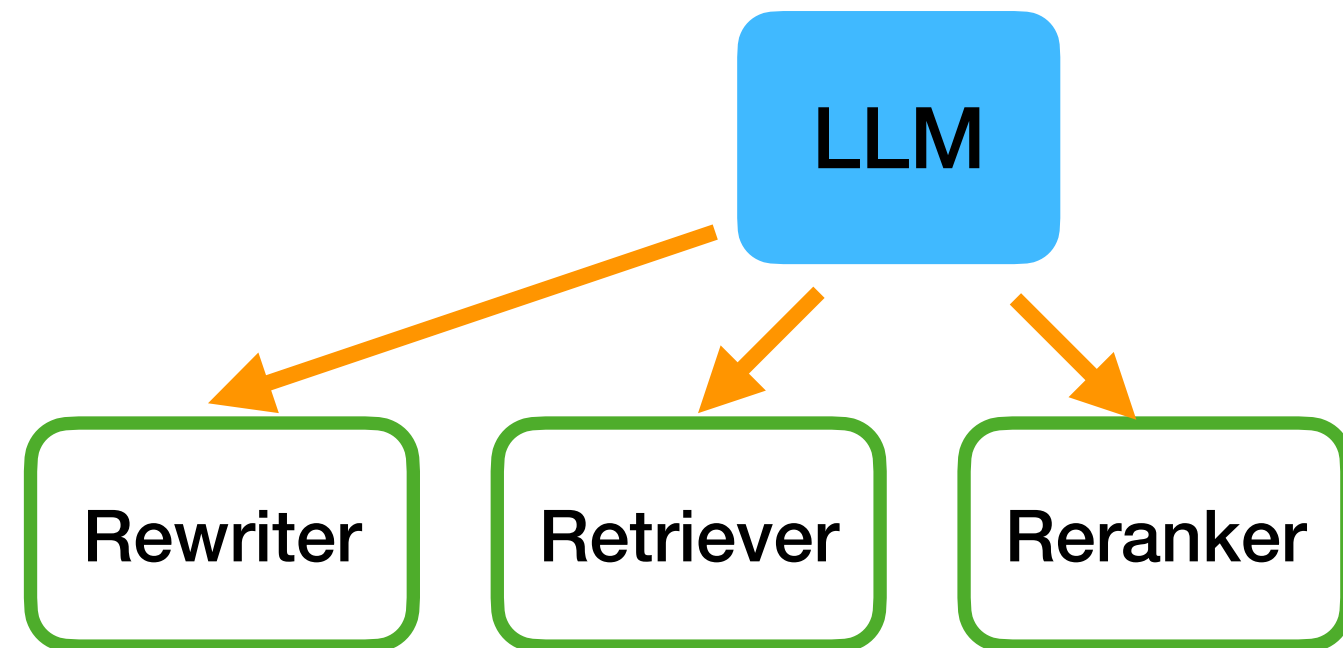
e.g. Differentiable Search Index  
(DSI)



# Focus of this tutorial:

## Using LLMs for Retrieval and Ranking

### Prompting LLMs for Retrieval and Ranking



*LLM-enhanced information retrieval*

e.g. RankGPT, PromptReps

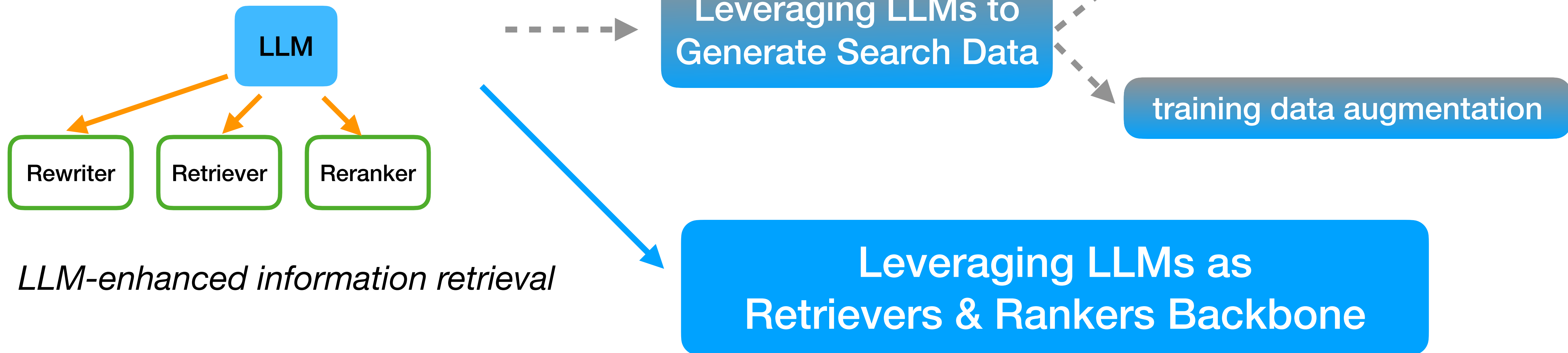




# Focus of this tutorial:

## Using LLMs for Retrieval and Ranking

### Prompting LLMs for Retrieval and Ranking



e.g. RankGPT, PromptReps

# Why still focusing on retrieval and ranking? Isn't search dead?

Search core function of information (seeking) systems, if we want them reliable, up to date

## RAG

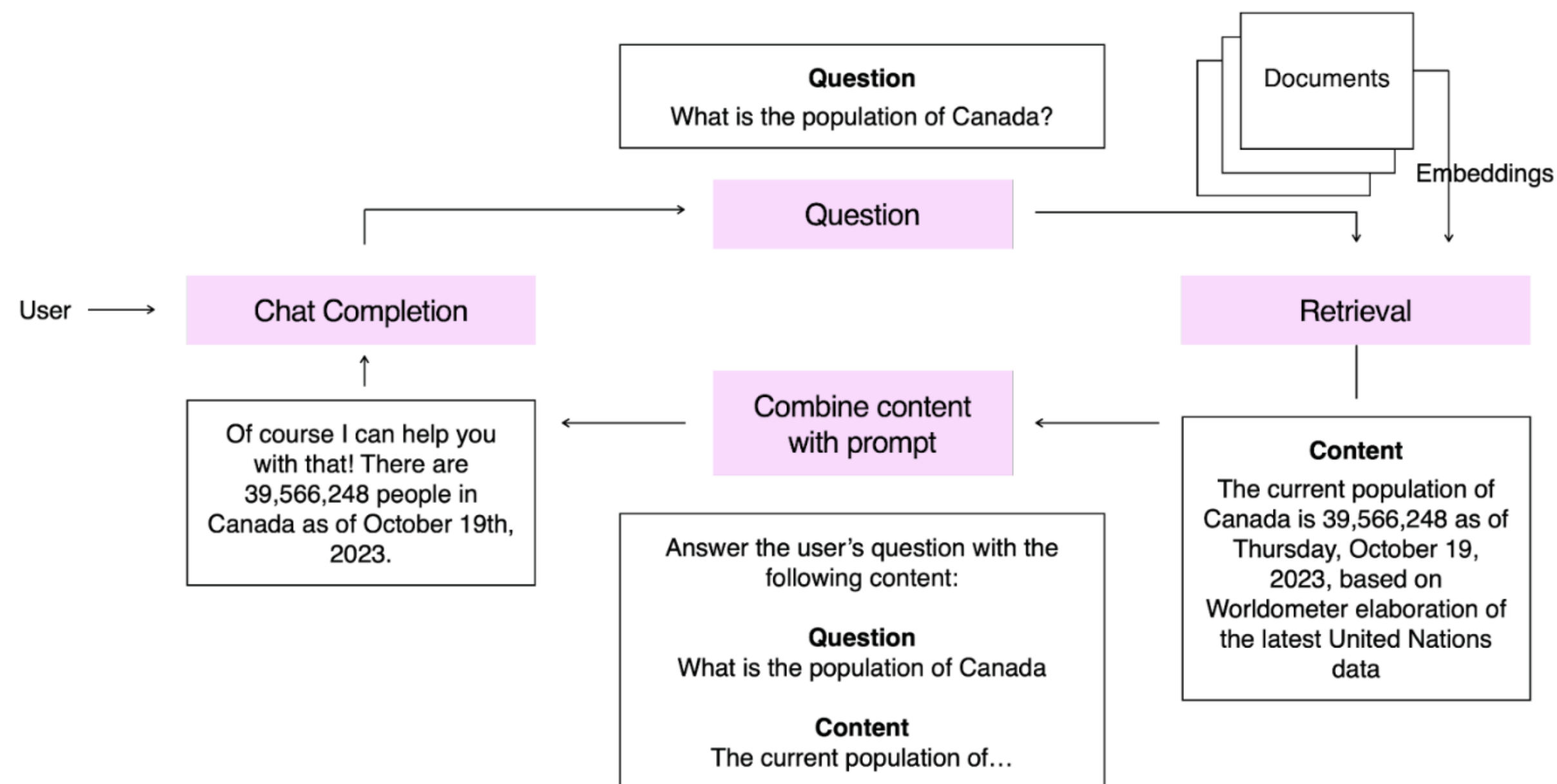


Image Source: <https://platform.openai.com/docs/guides/optimizing-llm-accuracy/retrieval-augmented-generation-rag>

## Web Agents

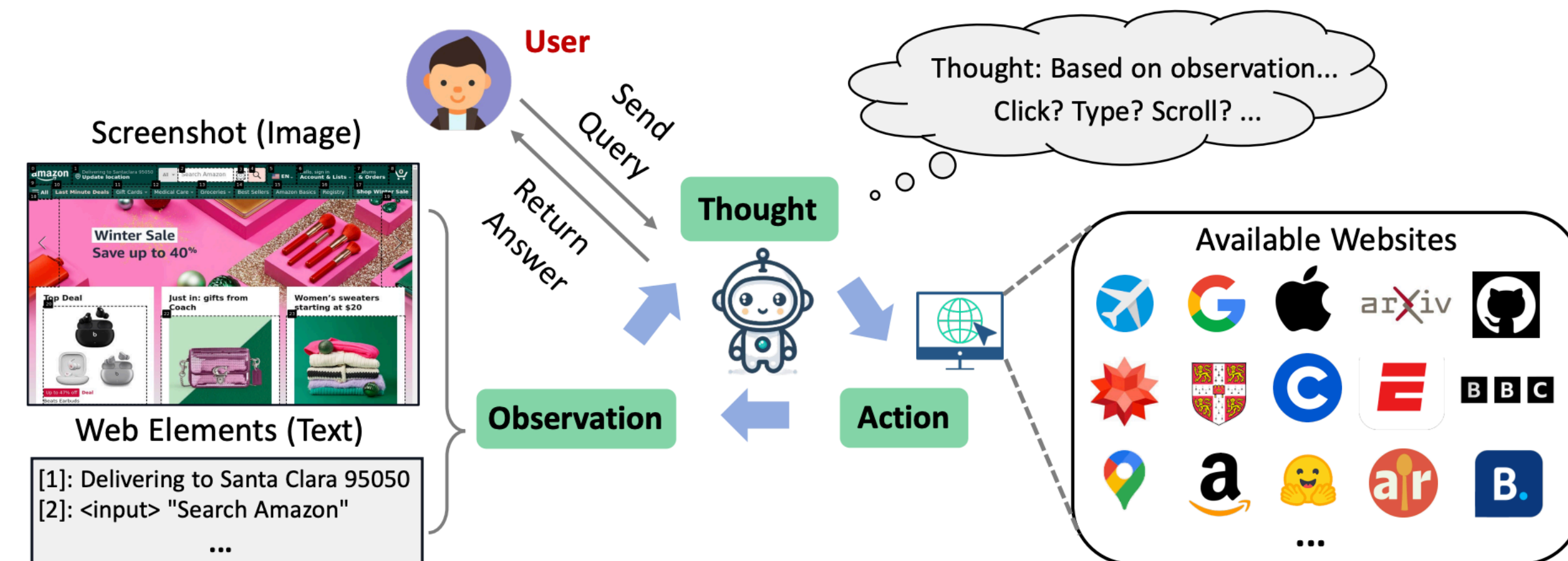
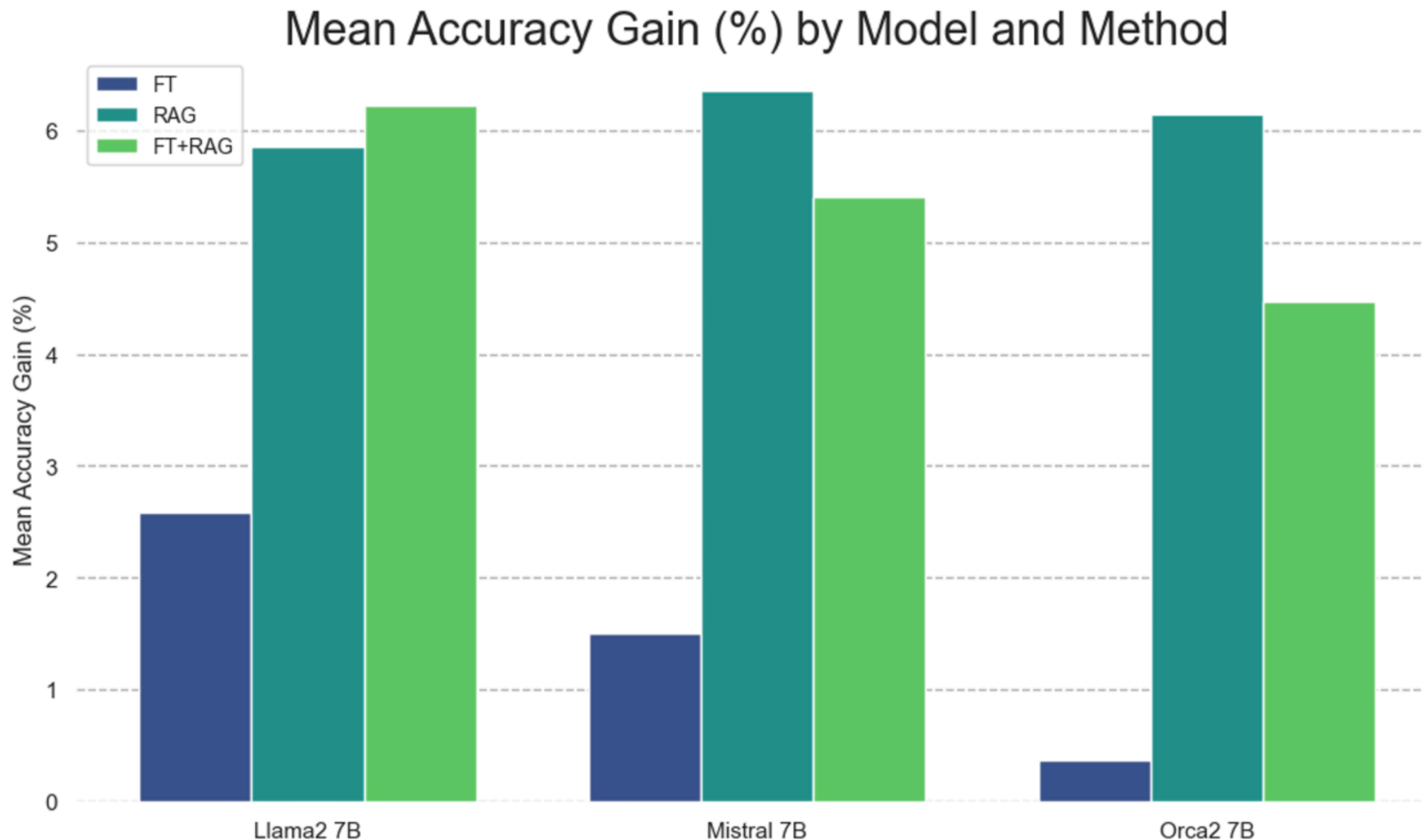


Image Source: H. He, et al. "WebVoyager: Building an End-to-End Web Agent with Large Multimodal Models", ACL 2024

# Why still focusing on retrieval and ranking? Isn't search dead?



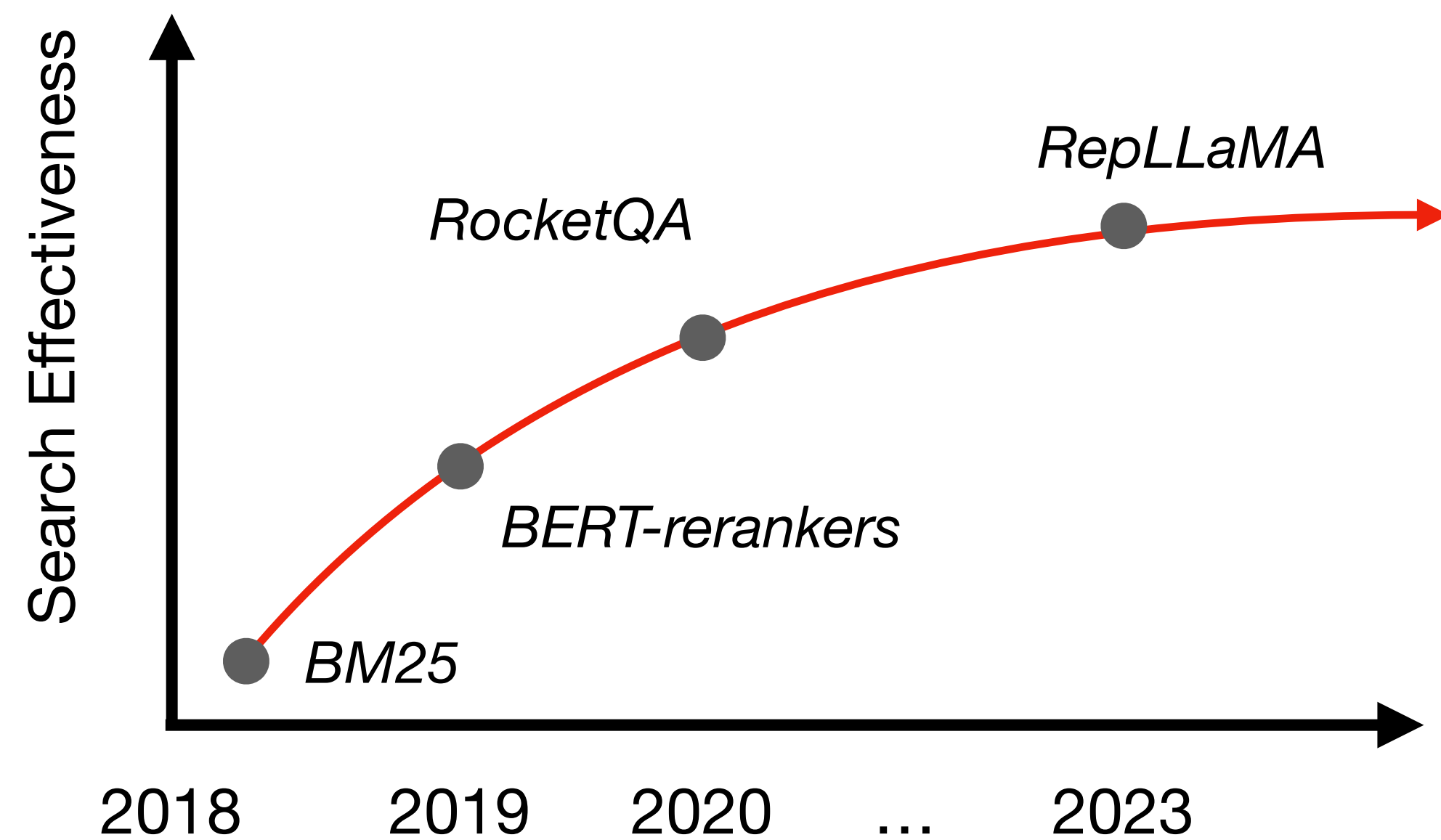
- Replacing search with LLMs is not the solution — integration is the way to go
- Ovadia et al.: in QA, unsupervised fine-tuning offers some improvement, but RAG consistently outperforms it, both for existing knowledge encountered during training and entirely new knowledge
- LLMs struggle to learn new factual information through unsupervised fine-tuning



**Ok, but why do search with  
LLMs?**

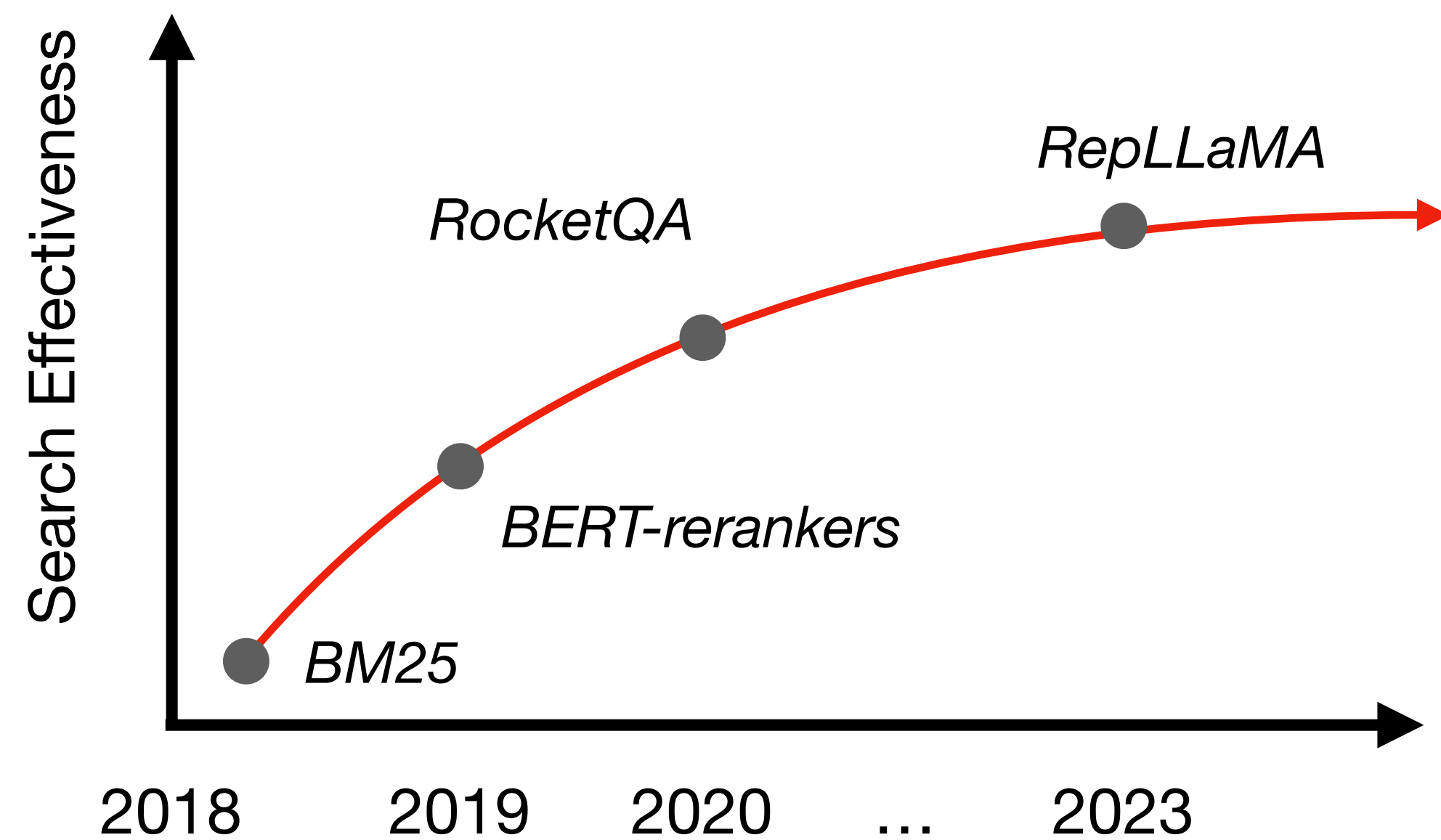
# Massive progress in search quality in recent years

MS MARCO (in-domain data)

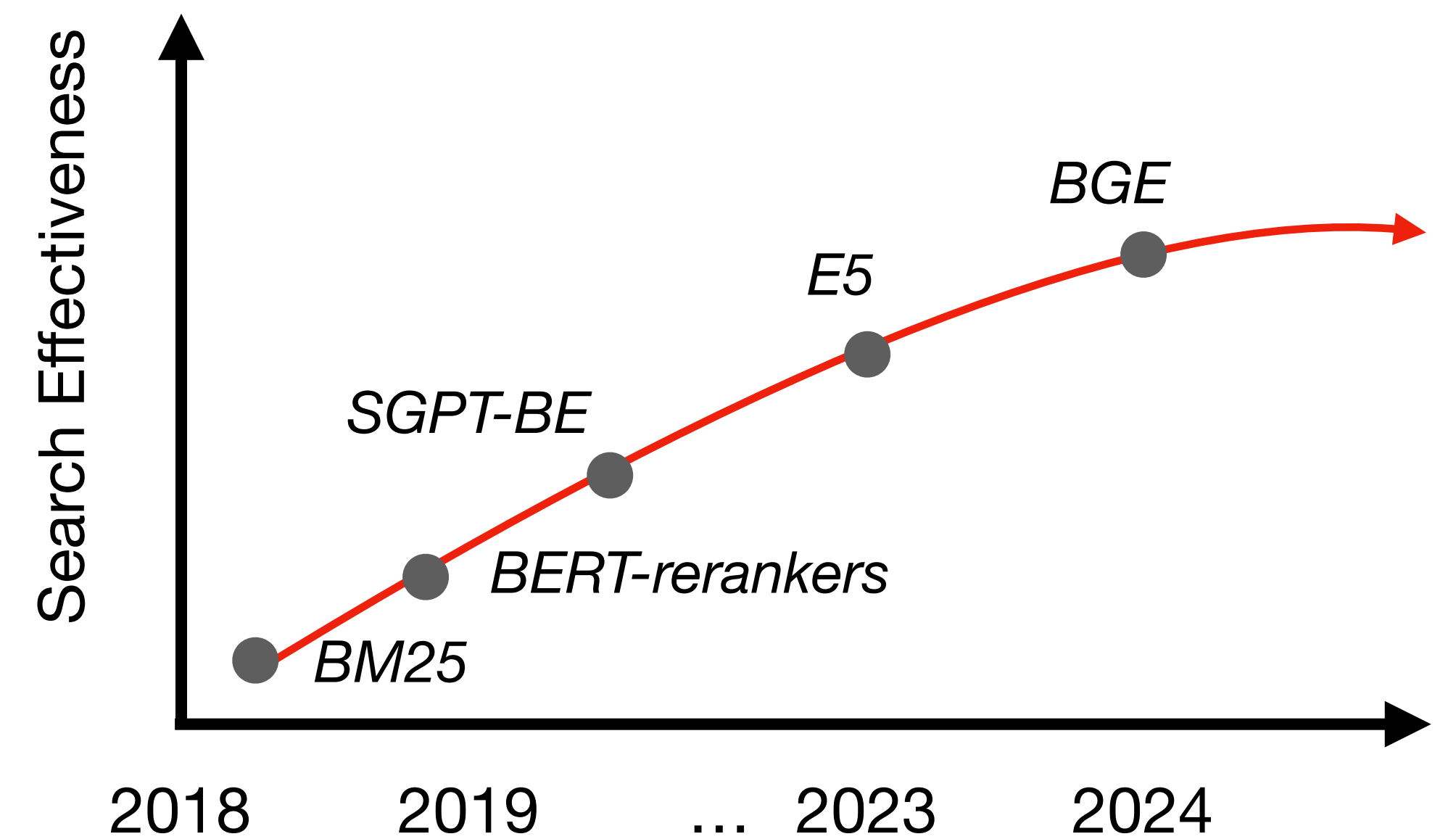


# Massive progress in search quality in recent years

MS MARCO (in-domain data)



BEIR (out-domain data/~task)

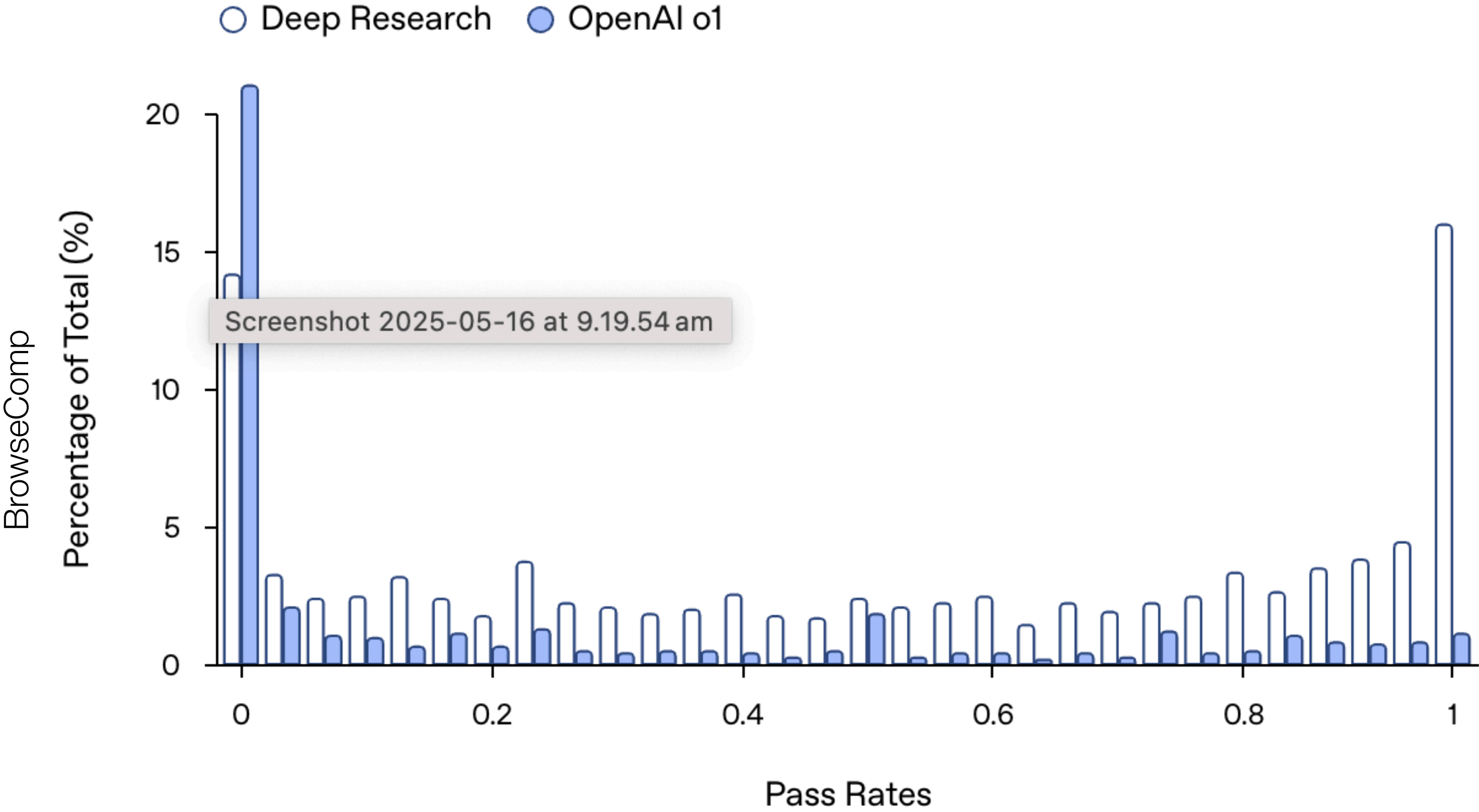
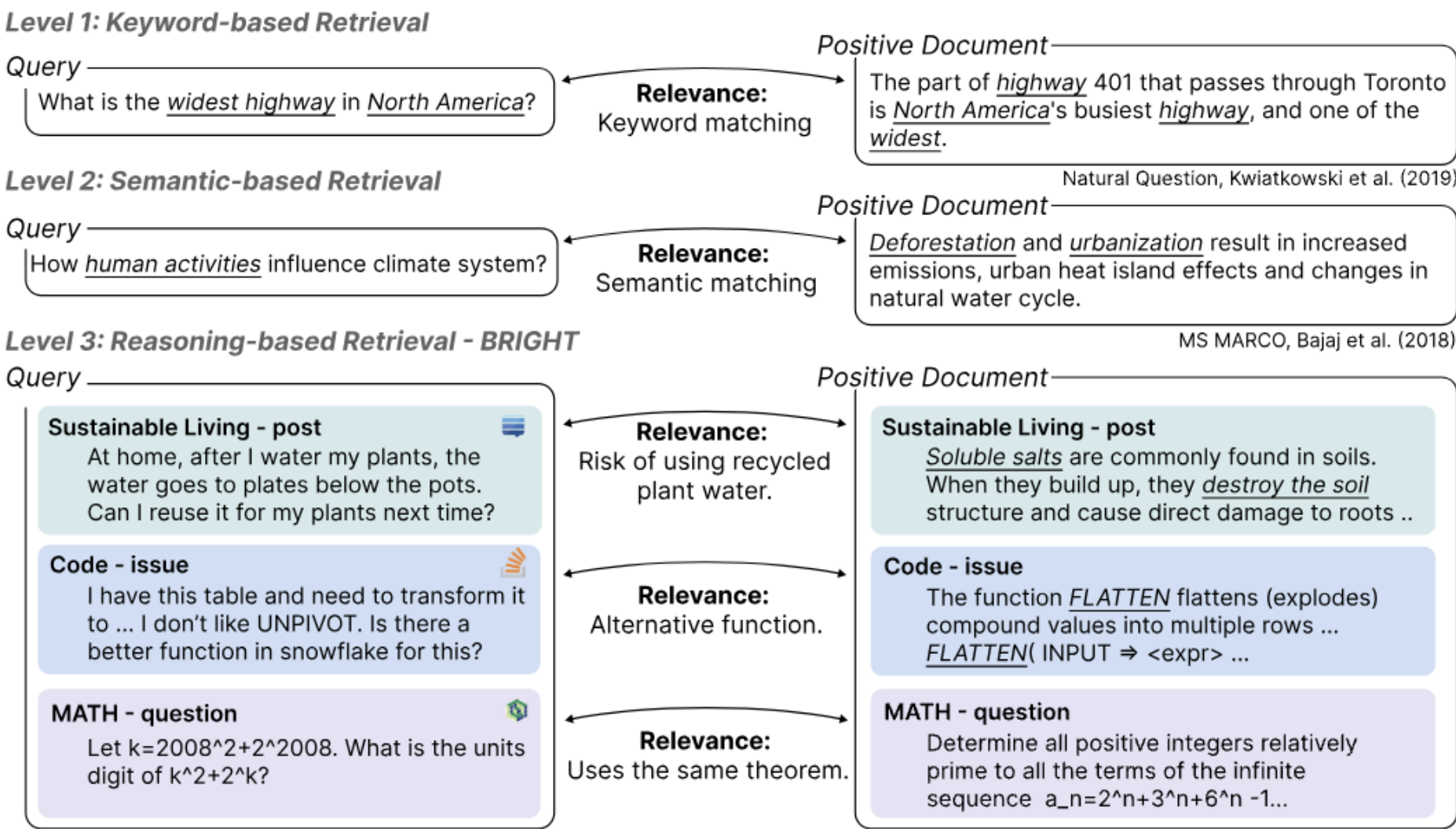




# However, search still challenging in many situations

## Intensive Reasoning Datasets

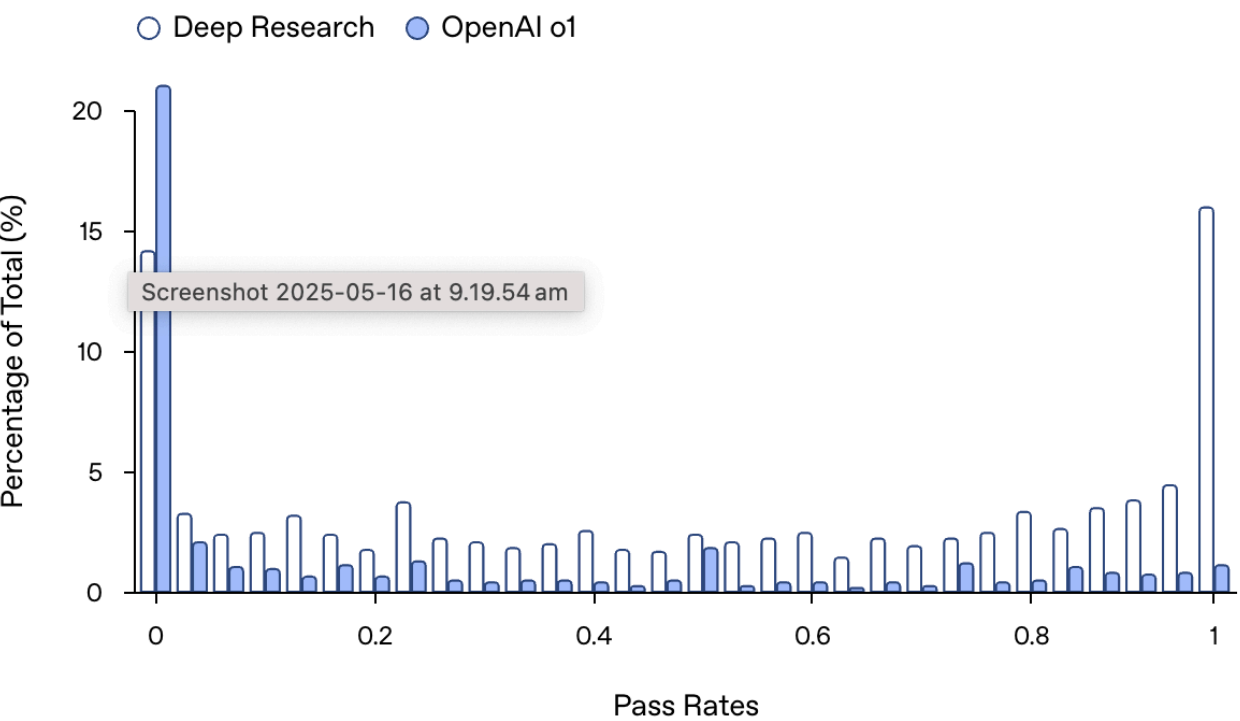
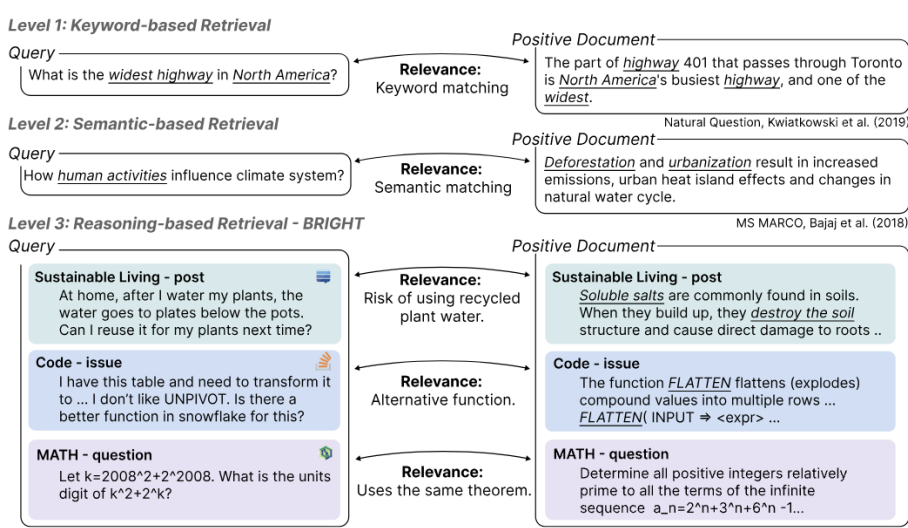
*BRIGHT, BrowseComp*



# However, search still challenging in many situations

## Intensive Reasoning Datasets

*BRIGHT, BrowseComp*



## Multi/Cross Lingual

*ECLeKTic, TREC NeuCLIR*

| Target language | de | 37.5%<br>6/16 | 64.1%<br>25/39 | 50.0%<br>11/22 | 56.2%<br>18/32 | 48.3%<br>14/29 | 50.0%<br>14/28 | 51.6%<br>16/31 | 20.7%<br>6/29  | 54.7%<br>35/64 | 33.3%<br>11/33 | 11.5%<br>3/26  | 42.9%<br>15/35 | 43.4%           |       |       |       |       |       |       |       |       |       |       |       |
|-----------------|----|---------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|-----------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
|                 | en | 25.0%<br>4/16 | 87.2%<br>34/39 | 50.0%<br>11/22 | 59.4%<br>19/32 | 51.7%<br>15/29 | 35.7%<br>10/28 | 54.8%<br>17/31 | 31.0%<br>9/29  | 56.2%<br>36/64 | 33.3%<br>11/33 | 26.9%<br>7/26  | 54.3%<br>19/35 | 47.1%           |       |       |       |       |       |       |       |       |       |       |       |
|                 | fr | 25.0%<br>4/16 | 69.2%<br>27/39 | 59.1%<br>13/22 | 56.2%<br>18/32 | 51.7%<br>15/29 | 53.6%<br>15/28 | 41.9%<br>13/31 | 24.1%<br>7/29  | 54.7%<br>35/64 | 36.4%<br>12/33 | 30.8%<br>8/26  | 37.1%<br>13/35 | 45.0%           |       |       |       |       |       |       |       |       |       |       |       |
|                 | es | 37.5%<br>6/16 | 61.5%<br>24/39 | 45.5%<br>10/22 | 78.1%<br>25/32 | 48.3%<br>14/29 | 42.9%<br>12/28 | 54.8%<br>17/31 | 27.6%<br>8/29  | 60.9%<br>39/64 | 42.4%<br>14/33 | 23.1%<br>6/26  | 45.7%<br>16/35 | 47.4%           |       |       |       |       |       |       |       |       |       |       |       |
|                 | it | 31.2%<br>5/16 | 66.7%<br>26/39 | 36.4%<br>8/22  | 50.0%<br>16/32 | 58.6%<br>17/29 | 50.0%<br>14/28 | 51.6%<br>16/31 | 20.7%<br>6/29  | 56.2%<br>36/64 | 33.3%<br>11/33 | 30.8%<br>8/26  | 45.7%<br>16/35 | 44.3%           |       |       |       |       |       |       |       |       |       |       |       |
|                 | pt | 37.5%<br>6/16 | 59.0%<br>23/39 | 36.4%<br>8/22  | 56.2%<br>18/32 | 51.7%<br>15/29 | 60.7%<br>17/28 | 51.6%<br>16/31 | 17.2%<br>5/29  | 54.7%<br>35/64 | 42.4%<br>14/33 | 15.4%<br>4/26  | 42.9%<br>15/35 | 43.8%           |       |       |       |       |       |       |       |       |       |       |       |
|                 | id | 37.5%<br>6/16 | 61.5%<br>24/39 | 40.9%<br>9/22  | 53.1%<br>17/32 | 34.5%<br>10/29 | 39.3%<br>11/28 | 71.0%<br>22/31 | 17.2%<br>5/29  | 50.0%<br>32/64 | 36.4%<br>12/33 | 19.2%<br>5/26  | 40.0%<br>14/35 | 41.7%           |       |       |       |       |       |       |       |       |       |       |       |
|                 | he | 25.0%<br>4/16 | 38.5%<br>15/39 | 36.4%<br>8/22  | 43.8%<br>14/32 | 34.5%<br>10/29 | 39.3%<br>11/28 | 38.7%<br>12/31 | 37.9%<br>11/29 | 40.6%<br>26/64 | 39.4%<br>13/33 | 11.5%<br>3/26  | 31.4%<br>11/35 | 34.8%           |       |       |       |       |       |       |       |       |       |       |       |
|                 | hi | 31.2%<br>5/16 | 51.3%<br>20/39 | 50.0%<br>11/22 | 31.2%<br>10/32 | 27.6%<br>8/29  | 25.0%<br>7/28  | 41.9%<br>13/31 | 17.2%<br>5/29  | 70.3%<br>45/64 | 36.4%<br>12/33 | 11.5%<br>3/26  | 40.0%<br>14/35 | 36.1%           |       |       |       |       |       |       |       |       |       |       |       |
|                 | ko | 25.0%<br>4/16 | 48.7%<br>19/39 | 22.7%<br>5/22  | 28.1%<br>9/32  | 27.6%<br>8/29  | 39.3%<br>11/28 | 35.5%<br>11/31 | 13.8%<br>4/29  | 35.9%<br>23/64 | 60.6%<br>20/33 | 23.1%<br>6/26  | 31.4%<br>11/35 | 32.6%           |       |       |       |       |       |       |       |       |       |       |       |
|                 | ja | 25.0%<br>4/16 | 59.0%<br>23/39 | 27.3%<br>6/22  | 43.8%<br>14/32 | 27.6%<br>8/29  | 14.3%<br>4/28  | 41.9%<br>13/31 | 17.2%<br>5/29  | 43.8%<br>28/64 | 45.5%<br>15/33 | 42.3%<br>11/26 | 48.6%<br>17/35 | 36.3%           |       |       |       |       |       |       |       |       |       |       |       |
|                 | zh | 25.0%<br>4/16 | 56.4%<br>22/39 | 40.9%<br>9/22  | 53.1%<br>17/32 | 41.4%<br>12/29 | 35.7%<br>10/28 | 35.5%<br>11/31 | 24.1%<br>7/29  | 53.1%<br>34/64 | 48.5%<br>16/33 | 26.9%<br>7/26  | 65.7%<br>23/35 | 42.2%           |       |       |       |       |       |       |       |       |       |       |       |
|                 |    |               |                |                |                |                |                |                |                |                |                |                |                |                 | 30.2% | 60.3% | 41.3% | 50.8% | 41.9% | 40.5% | 47.6% | 22.4% | 52.6% | 40.7% | 22.8% |
|                 |    |               |                |                |                |                |                |                |                |                |                |                |                | de              | en    | fr    | es    | it    | pt    | id    | he    | hi    | ko    | ja    | zh    |
|                 |    |               |                |                |                |                |                |                |                |                |                |                |                | Source language |       |       |       |       |       |       |       |       |       |       |       |

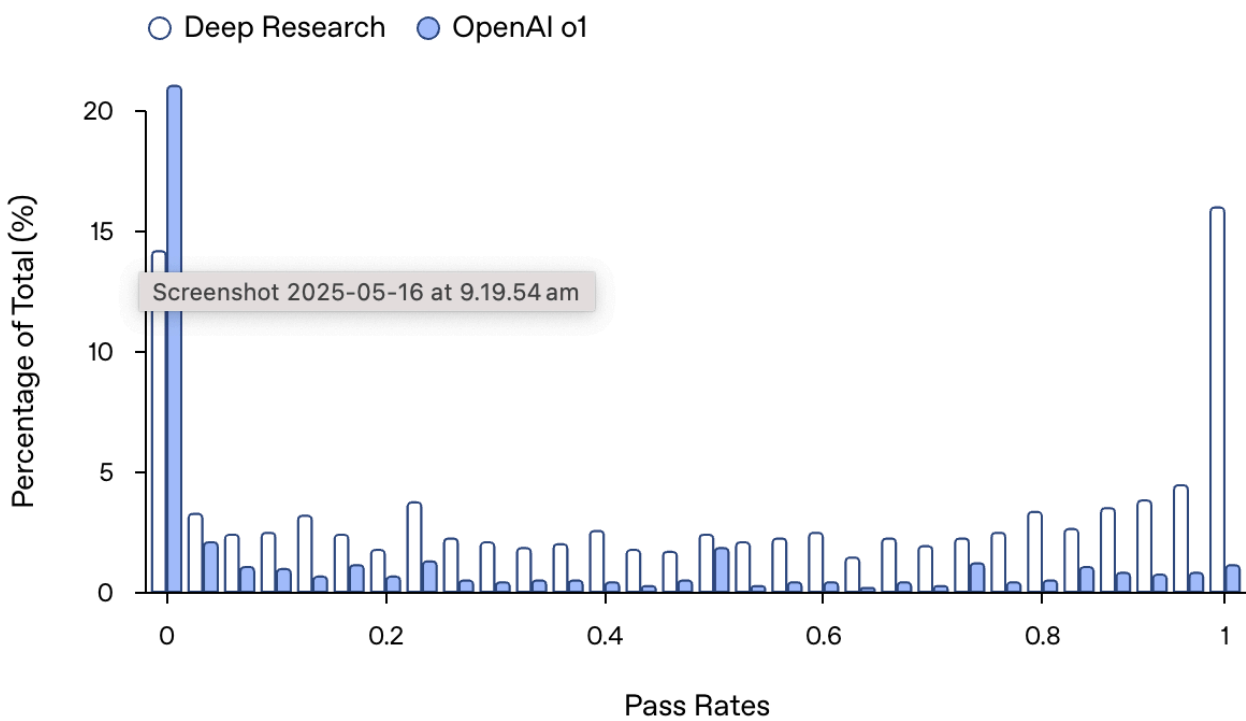
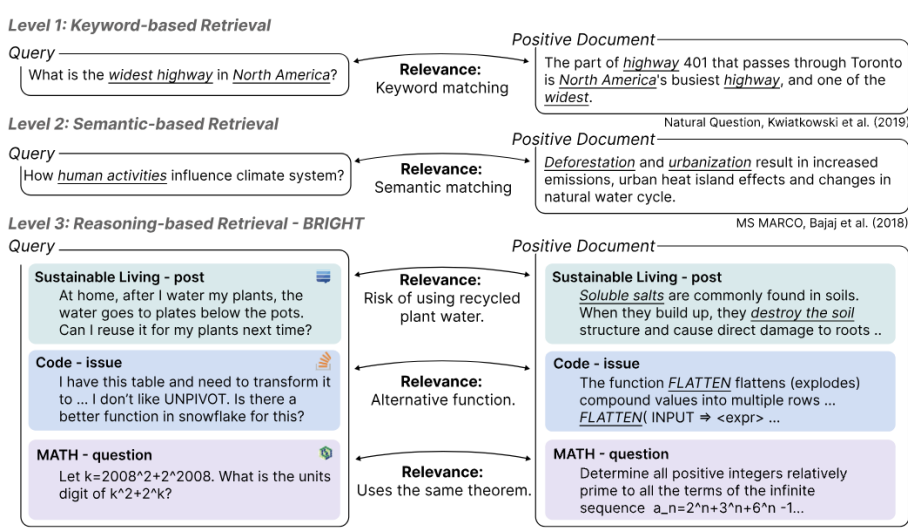
Image Source: O. Goldman, et al.  
“Eclektic: a novel challenge set for  
evaluation of cross-lingual knowledge  
transfer.”, arXiv:2502.21228. 2025.



# However, search still challenging in many situations

Intensive Reasoning Datasets

*BRIGHT, BrowseComp*



Multi/Cross Lingual

*ECLeKTic, TREC NeuCLIR*

| Target language | de            | en             | fr             | es             | it             | pt             | id             | he            | hi             | ko             | ja             | zh             |
|-----------------|---------------|----------------|----------------|----------------|----------------|----------------|----------------|---------------|----------------|----------------|----------------|----------------|
| de              | 37.5%<br>6/16 | 64.1%<br>25/39 | 50.0%<br>11/22 | 56.2%<br>18/32 | 48.3%<br>14/29 | 50.0%<br>14/28 | 51.6%<br>16/31 | 20.7%<br>6/29 | 54.7%<br>35/64 | 33.3%<br>11/33 | 11.5%<br>3/26  | 42.9%<br>15/35 |
| en              | 25.0%<br>4/16 | 87.2%<br>34/39 | 50.0%<br>11/22 | 59.4%<br>19/32 | 51.7%<br>15/29 | 35.7%<br>10/28 | 54.8%<br>17/31 | 31.0%<br>9/29 | 56.2%<br>36/64 | 33.3%<br>11/33 | 26.9%<br>7/26  | 54.3%<br>19/35 |
| fr              | 25.0%<br>4/16 | 69.2%<br>27/39 | 59.1%<br>13/22 | 56.2%<br>18/32 | 51.7%<br>15/29 | 53.6%<br>15/28 | 41.9%<br>13/31 | 24.1%<br>7/29 | 54.7%<br>35/64 | 36.4%<br>12/33 | 30.8%<br>8/26  | 37.1%<br>13/35 |
| es              | 37.5%<br>6/16 | 61.5%<br>24/39 | 45.5%<br>10/22 | 78.1%<br>25/32 | 48.3%<br>14/29 | 42.9%<br>12/28 | 54.8%<br>17/31 | 27.6%<br>8/29 | 60.9%<br>39/64 | 42.4%<br>14/33 | 23.1%<br>6/26  | 45.7%<br>16/35 |
| it              | 31.2%<br>5/16 | 66.7%<br>26/39 | 36.4%<br>8/22  | 50.0%<br>16/32 | 58.6%<br>17/29 | 50.0%<br>14/28 | 51.6%<br>16/31 | 20.7%<br>6/29 | 56.2%<br>36/64 | 33.3%<br>11/33 | 30.8%<br>8/26  | 45.7%<br>16/35 |
| pt              | 37.5%<br>6/16 | 59.0%<br>23/39 | 36.4%<br>8/22  | 56.2%<br>18/32 | 51.7%<br>15/29 | 60.7%<br>17/28 | 51.6%<br>16/31 | 17.2%<br>5/29 | 54.7%<br>35/64 | 42.4%<br>14/33 | 15.4%<br>4/26  | 42.9%<br>15/35 |
| id              | 37.5%<br>6/16 | 61.5%<br>24/39 | 40.9%<br>9/22  | 53.1%<br>17/32 | 34.5%<br>10/29 | 39.3%<br>11/28 | 71.0%<br>22/31 | 17.2%<br>5/29 | 50.0%<br>32/64 | 36.4%<br>12/33 | 19.2%<br>5/26  | 40.0%<br>14/35 |
| he              | 25.0%<br>4/16 | 38.5%<br>15/39 | 36.4%<br>8/22  | 43.8%<br>14/32 | 34.5%<br>10/29 | 38.7%<br>11/28 | 37.9%<br>12/31 | 11/29         | 40.6%<br>26/64 | 39.4%<br>13/33 | 11.5%<br>3/26  | 31.4%<br>11/35 |
| hi              | 31.2%<br>5/16 | 51.3%<br>20/39 | 50.0%<br>11/22 | 31.2%<br>10/32 | 27.6%<br>8/29  | 25.0%<br>7/28  | 41.9%<br>13/31 | 17.2%<br>5/29 | 70.3%<br>45/64 | 36.4%<br>12/33 | 11.5%<br>3/26  | 40.0%<br>14/35 |
| ko              | 25.0%<br>4/16 | 48.7%<br>19/39 | 22.7%<br>5/22  | 28.1%<br>9/32  | 27.6%<br>8/29  | 39.3%<br>11/28 | 35.5%<br>11/31 | 13.8%<br>4/29 | 35.9%<br>23/64 | 60.6%<br>20/33 | 23.1%<br>6/26  | 31.4%<br>11/35 |
| ja              | 25.0%<br>4/16 | 59.0%<br>23/39 | 27.3%<br>6/22  | 43.8%<br>14/32 | 27.6%<br>8/29  | 14.3%<br>4/28  | 41.9%<br>13/31 | 17.2%<br>5/29 | 43.8%<br>28/64 | 45.5%<br>15/33 | 42.3%<br>11/26 | 48.6%<br>17/35 |
| zh              | 25.0%<br>4/16 | 56.4%<br>22/39 | 40.9%<br>9/22  | 53.1%<br>17/32 | 41.4%<br>12/29 | 35.7%<br>10/28 | 35.5%<br>11/31 | 24.1%<br>7/29 | 53.1%<br>34/64 | 48.5%<br>16/33 | 26.9%<br>7/26  | 65.7%<br>23/35 |
|                 | 30.2%         | 60.3%          | 41.3%          | 50.8%          | 41.9%          | 40.5%          | 47.6%          | 22.4%         | 52.6%          | 40.7%          | 22.8%          | 43.8%          |
| Source language | de            | en             | fr             | es             | it             | pt             | id             | he            | hi             | ko             | ja             | zh             |

Multi-Modality

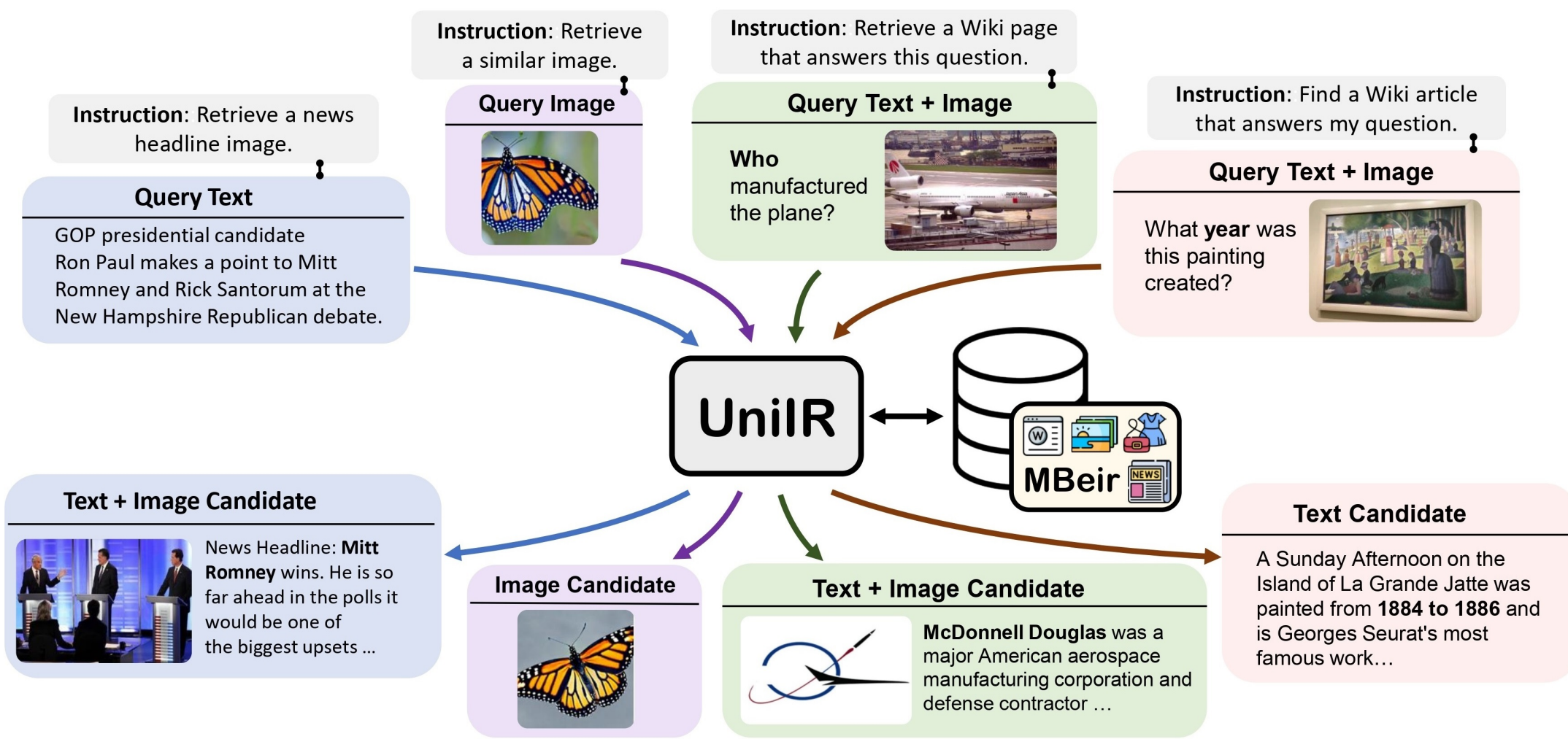
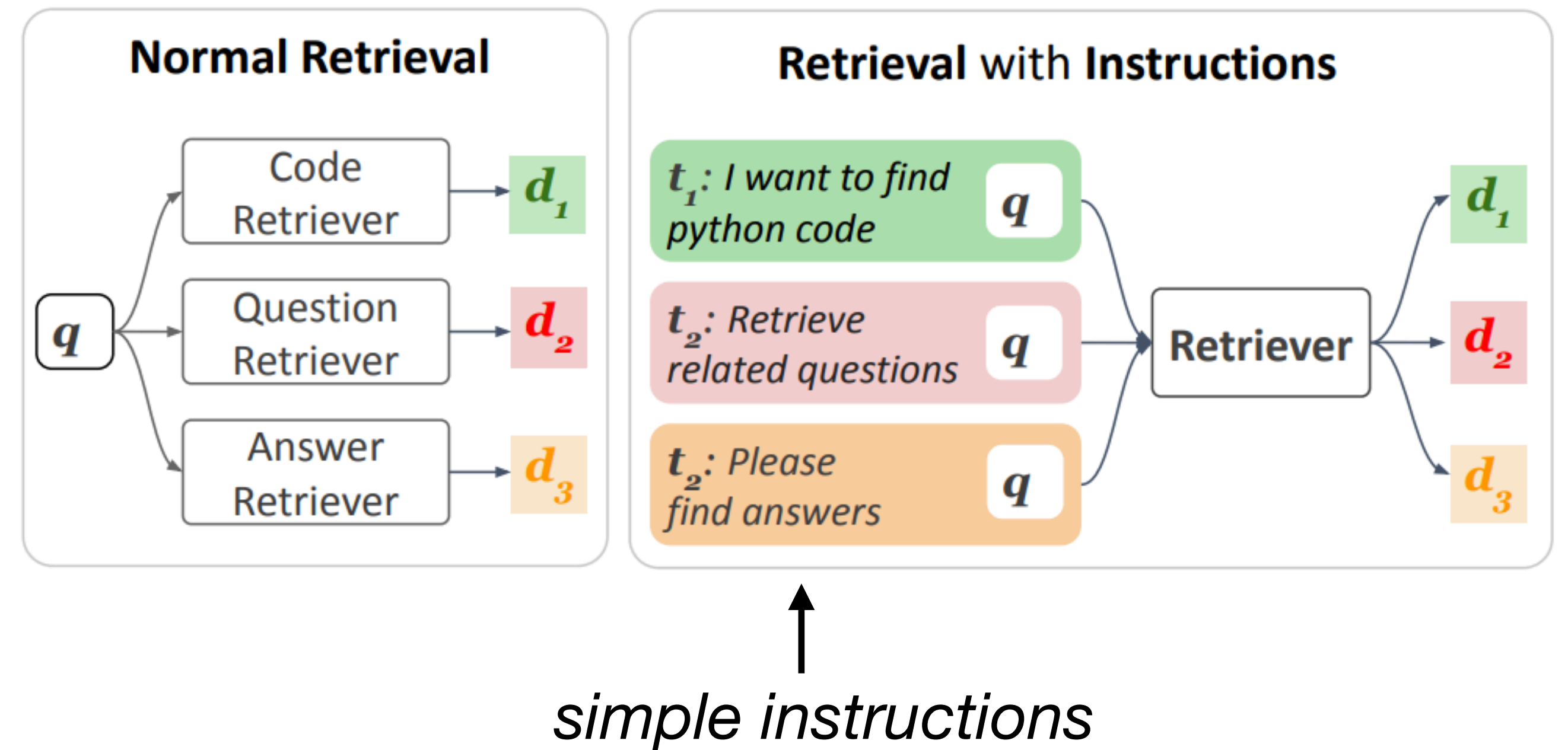
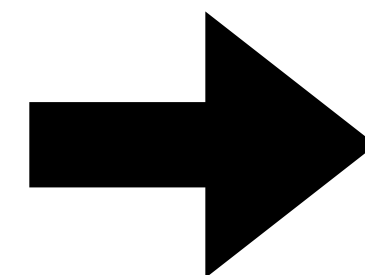
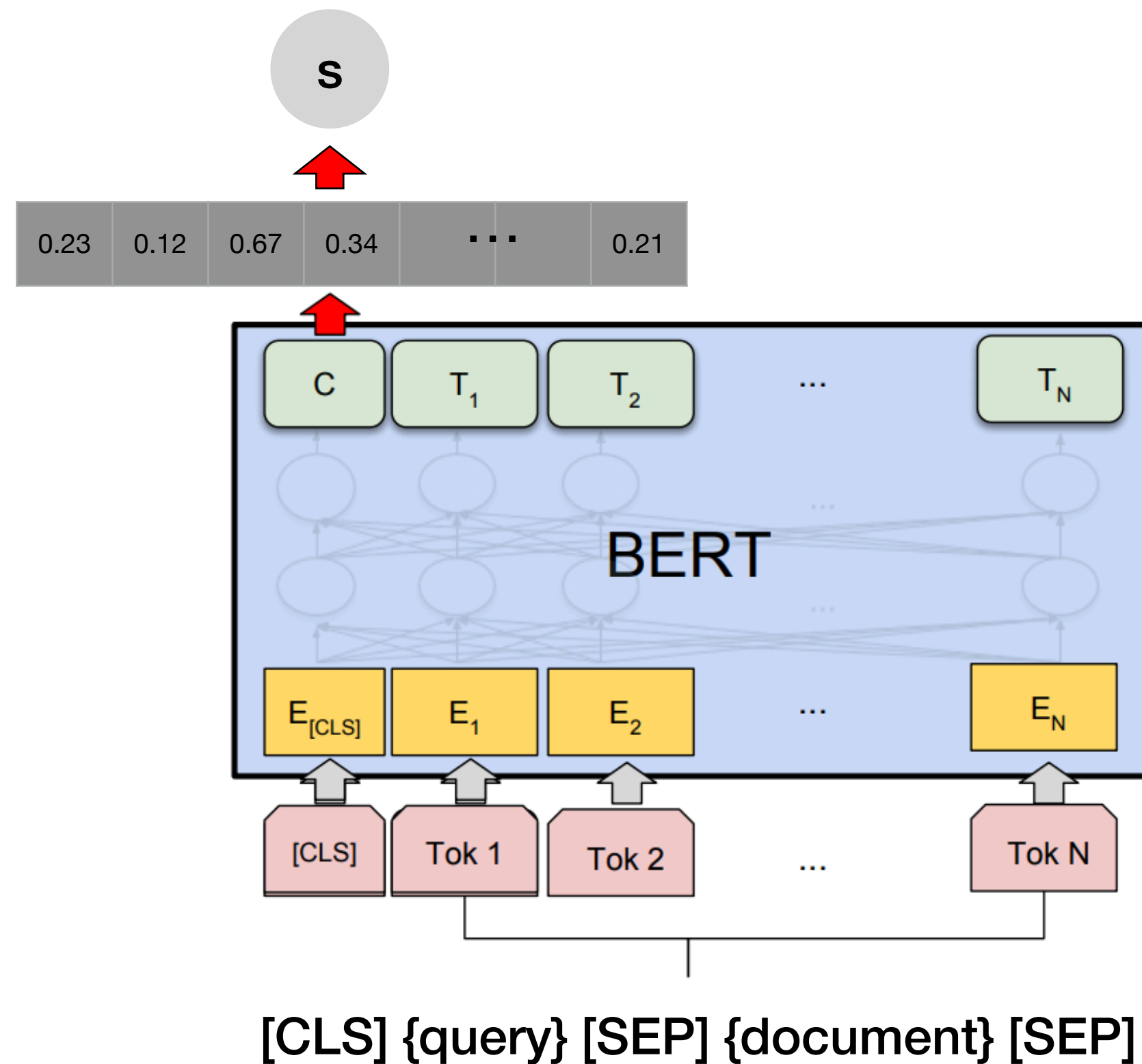


Image Source: C. Wei, et al. "Uniir: Training and benchmarking universal multimodal information retrievers.", ECCV 2024

# How did we get here?

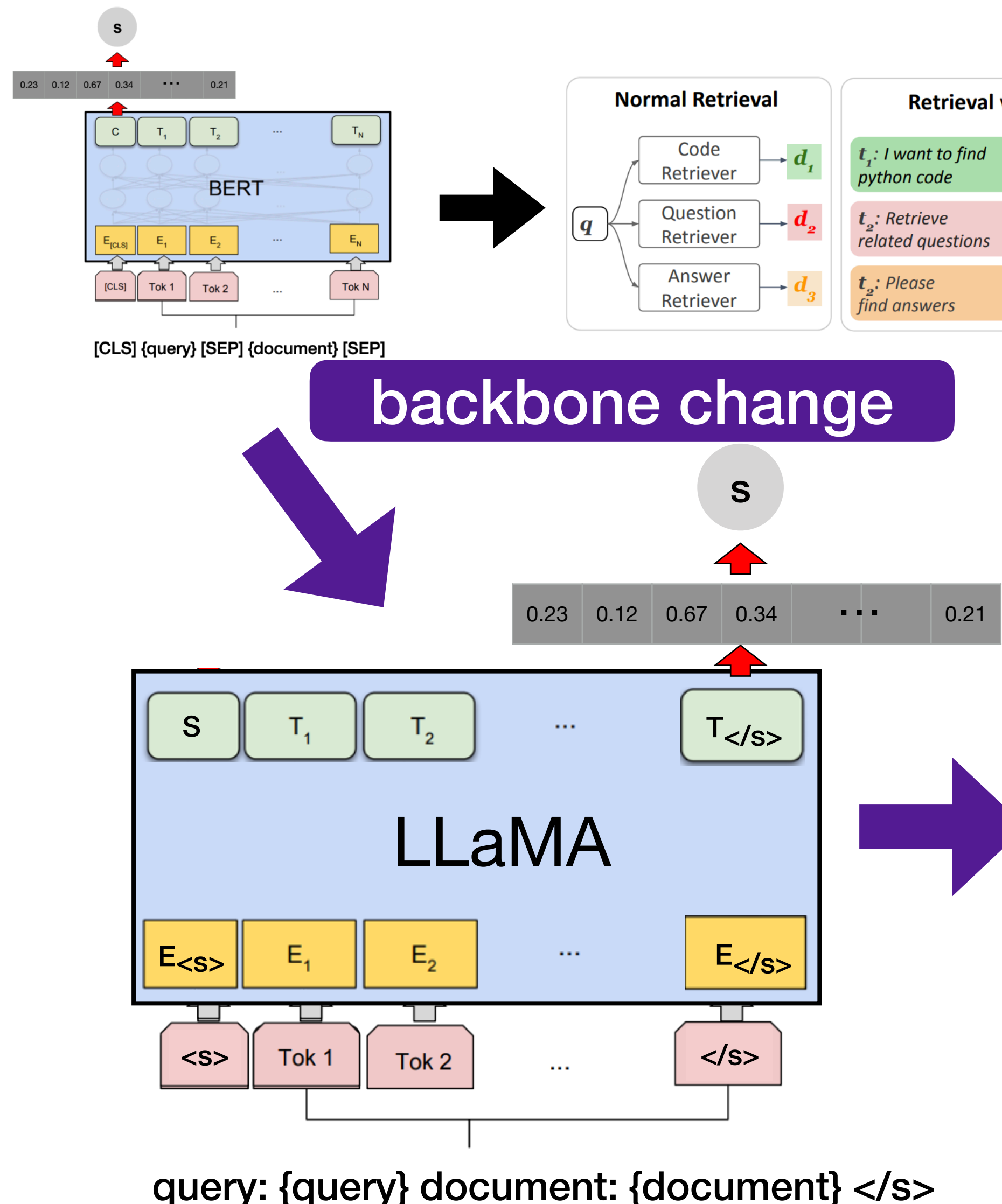
# From PreTrained LMs to Instruction Following LLMs



A. Asai, et al. "Task-aware Retrieval with Instructions." *ACL Findings* (2023).



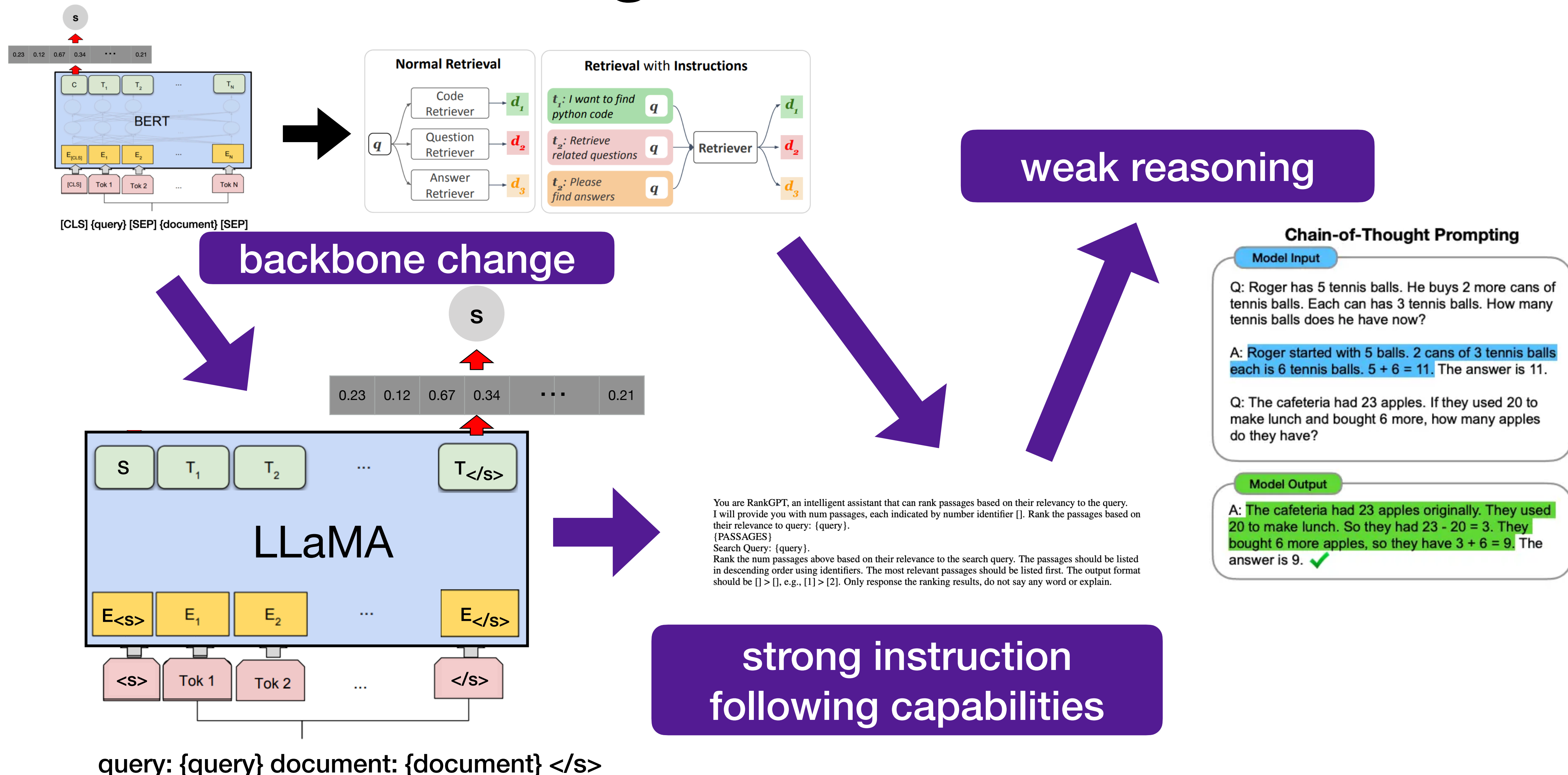
# From PreTrained LMs to Instruction Following LLMs



You are RankGPT, an intelligent assistant that can rank passages based on their relevancy to the query. I will provide you with num passages, each indicated by number identifier []. Rank the passages based on their relevance to query: {query}.  
{PASSAGES}  
Search Query: {query}.  
Rank the num passages above based on their relevance to the search query. The passages should be listed in descending order using identifiers. The most relevant passages should be listed first. The output format should be [] > [], e.g., [1] > [2]. Only response the ranking results, do not say any word or explain.

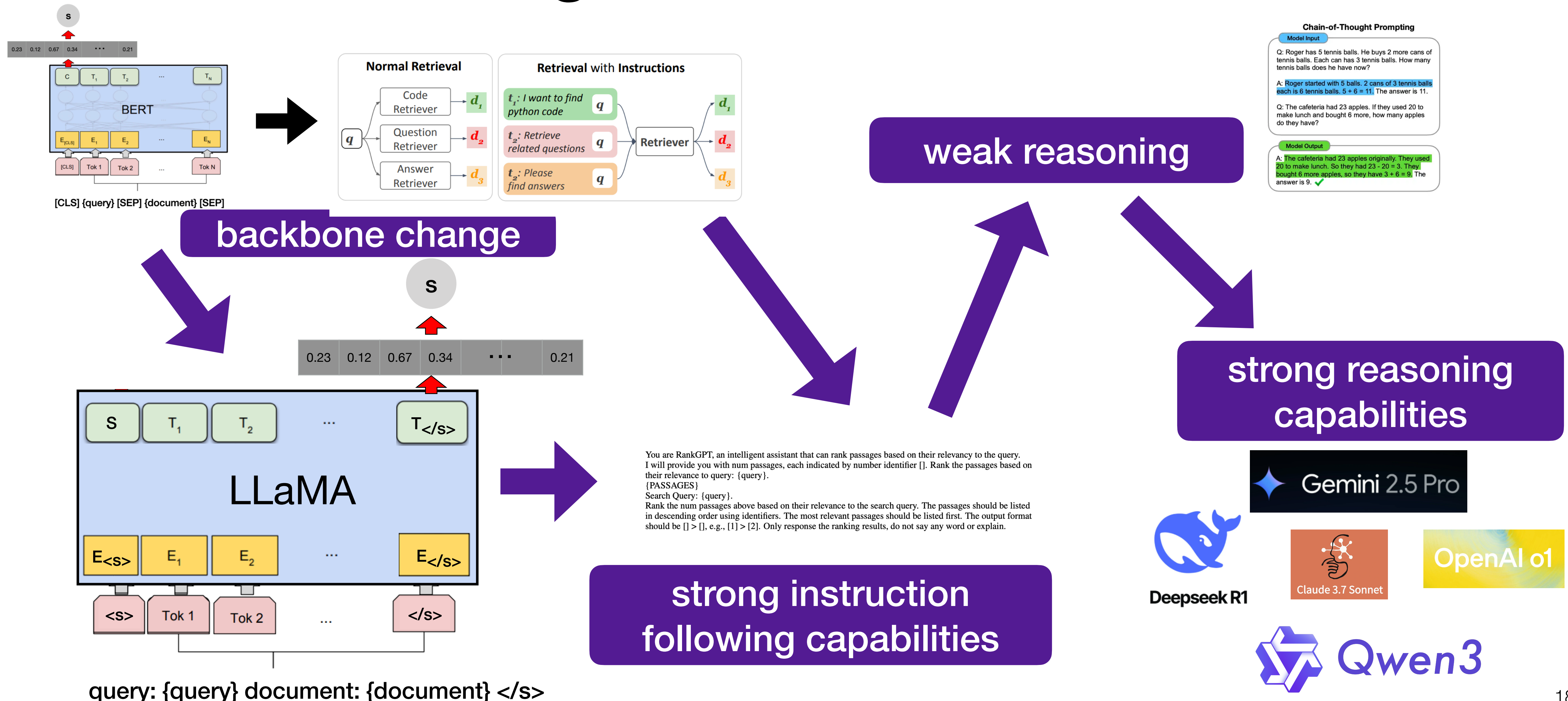
strong instruction following capabilities

# From PreTrained LMs to Instruction Following LLMs... to reasoners



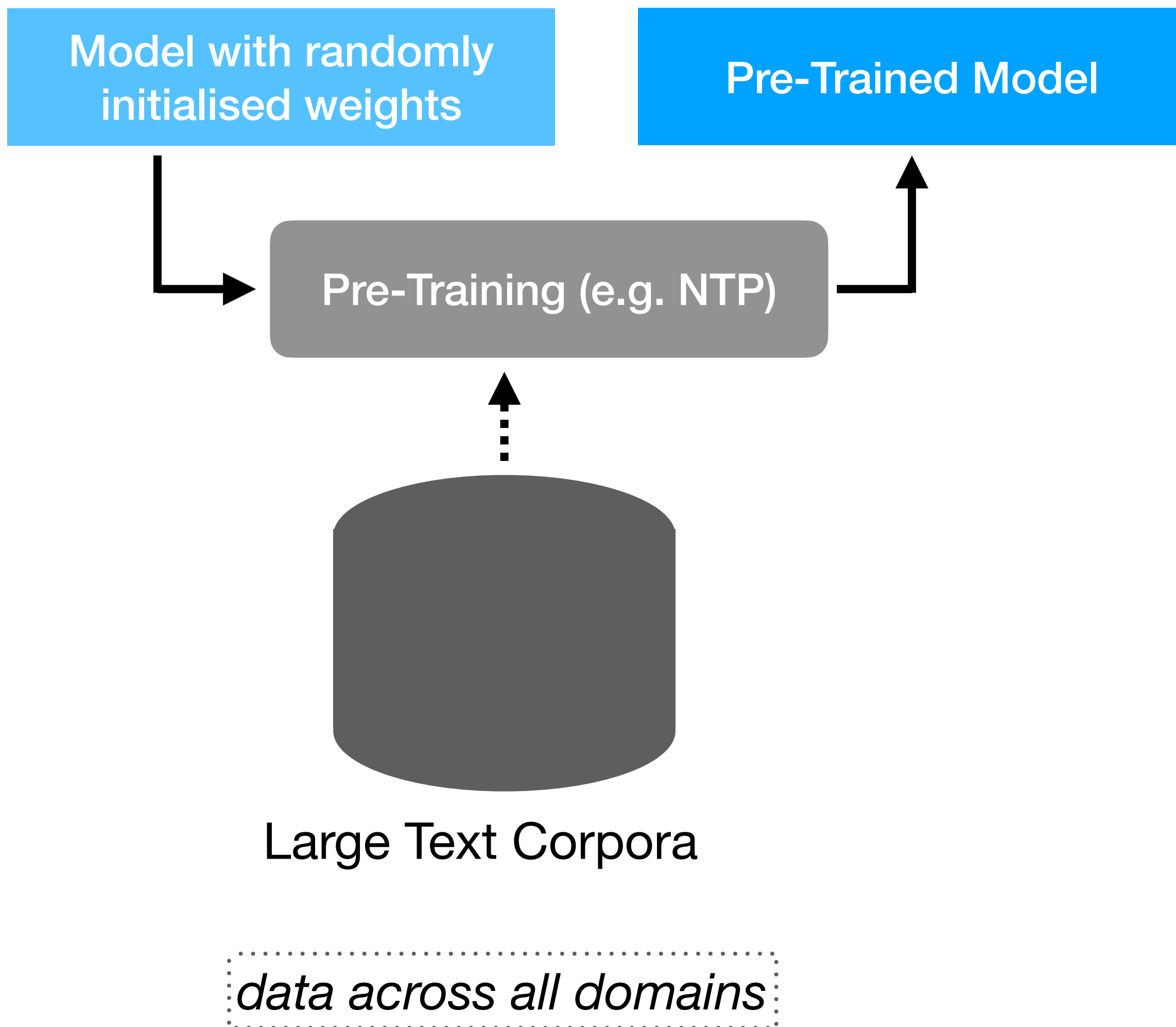


# From PreTrained LMs to Instruction Following LLMs... to reasoners



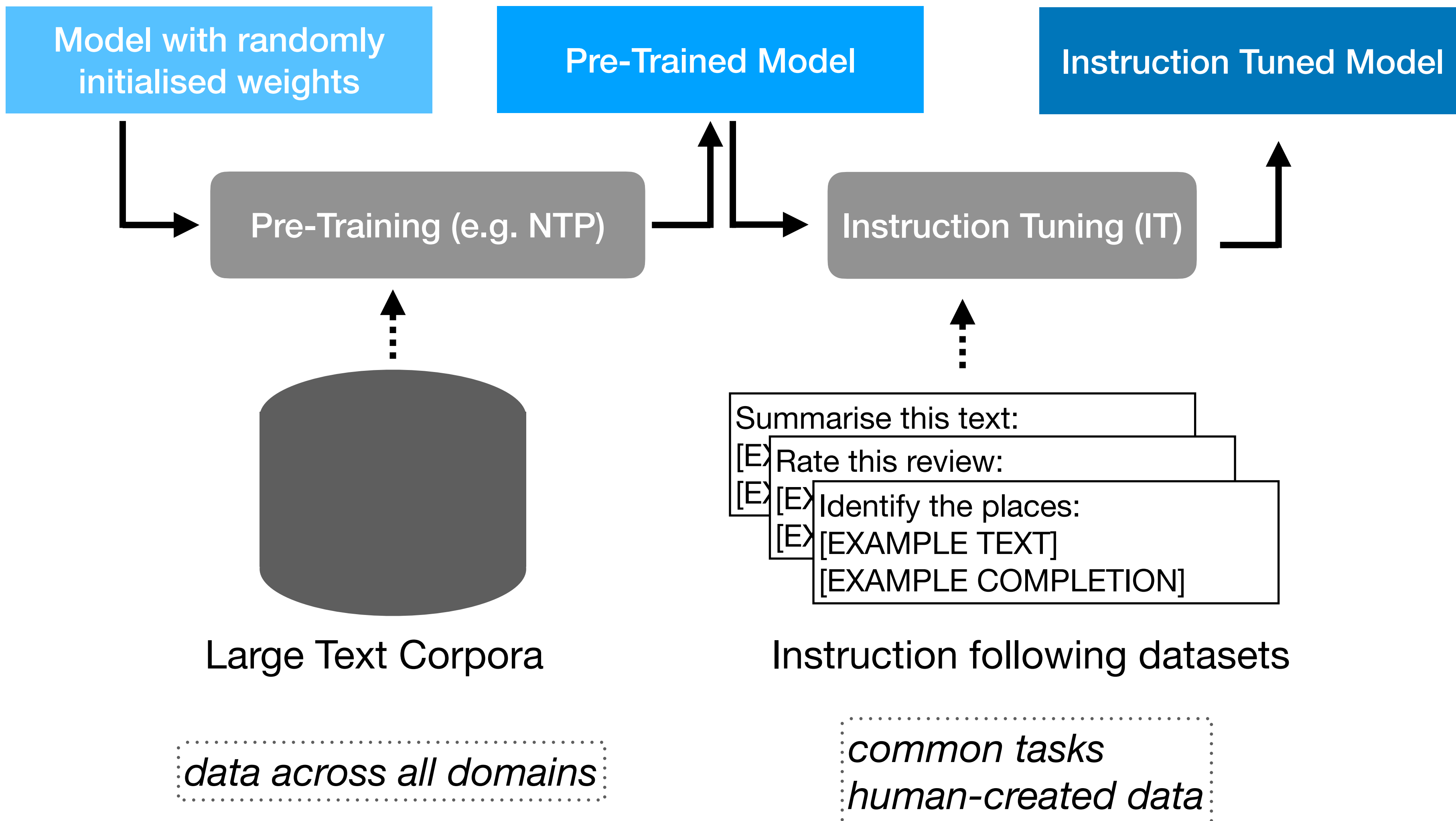
# LLM training 101

# Pre-Training



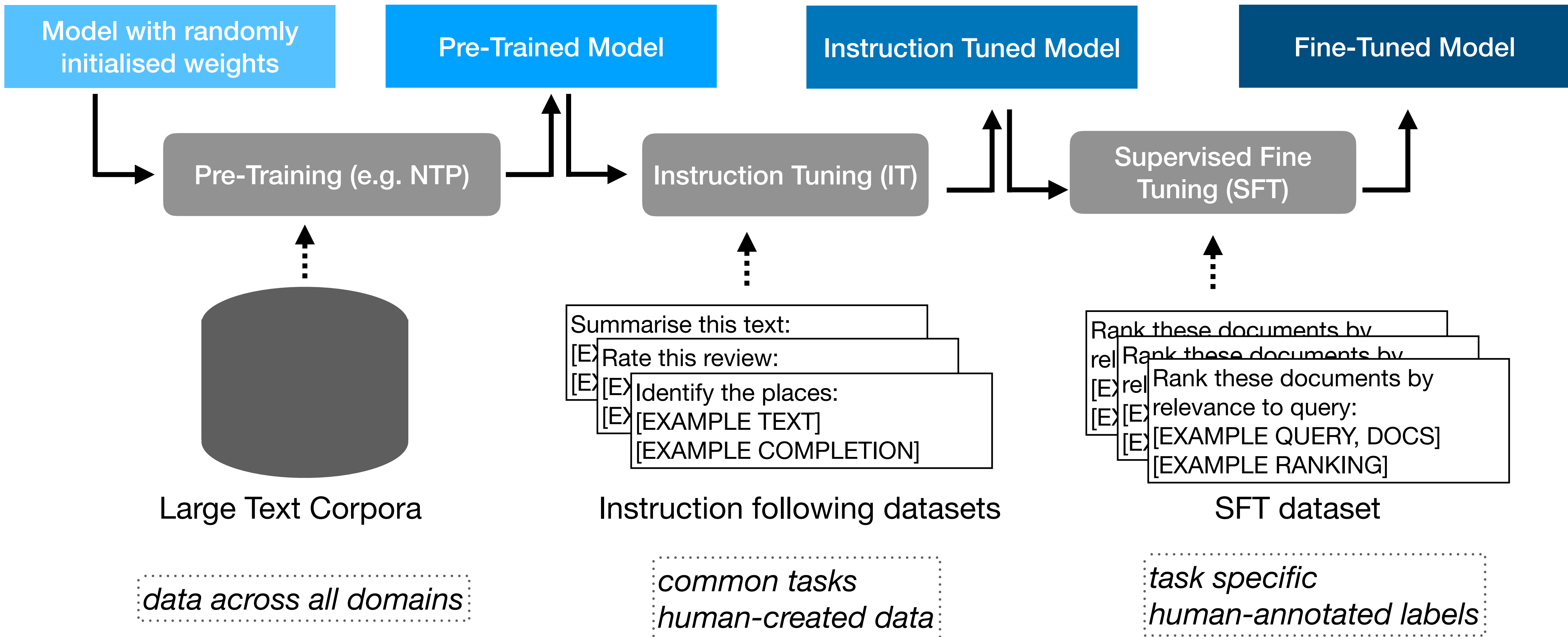


# Instruction Tuning

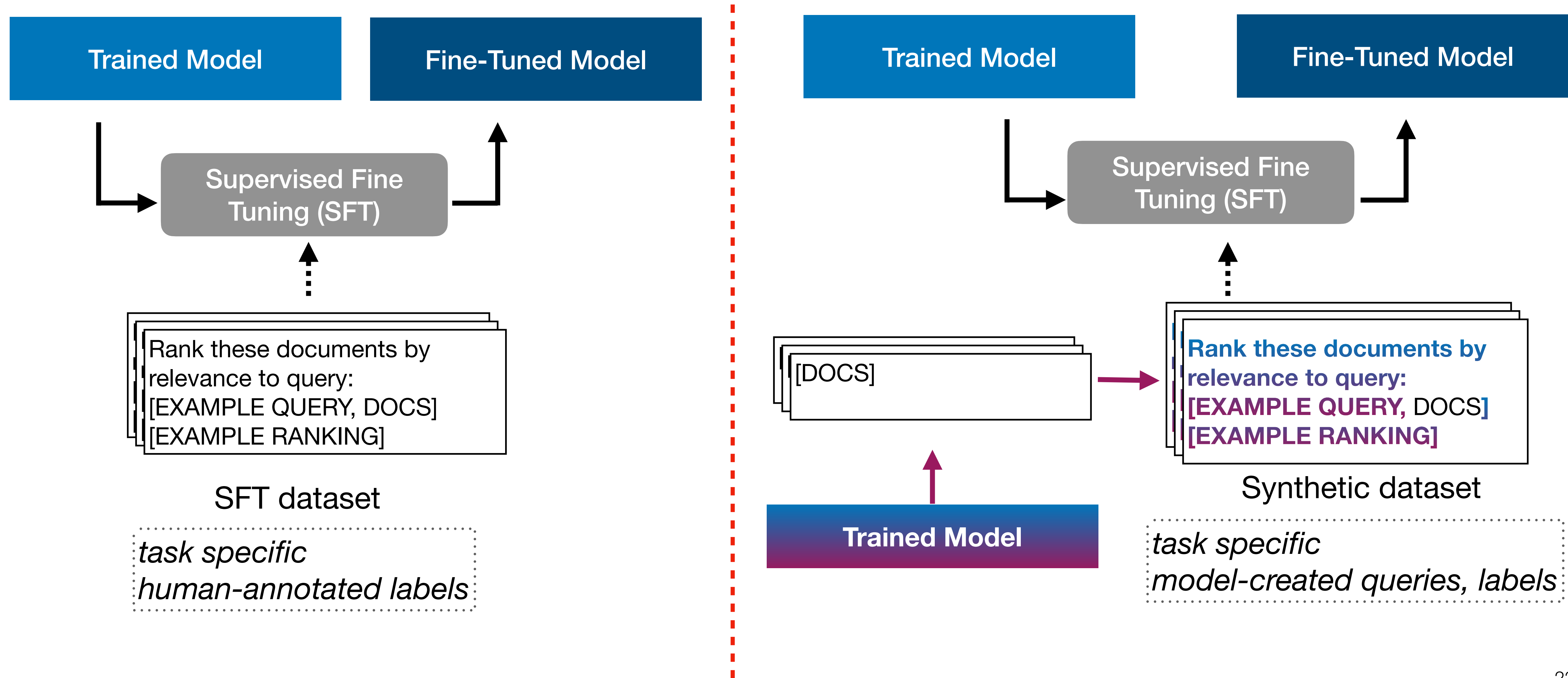




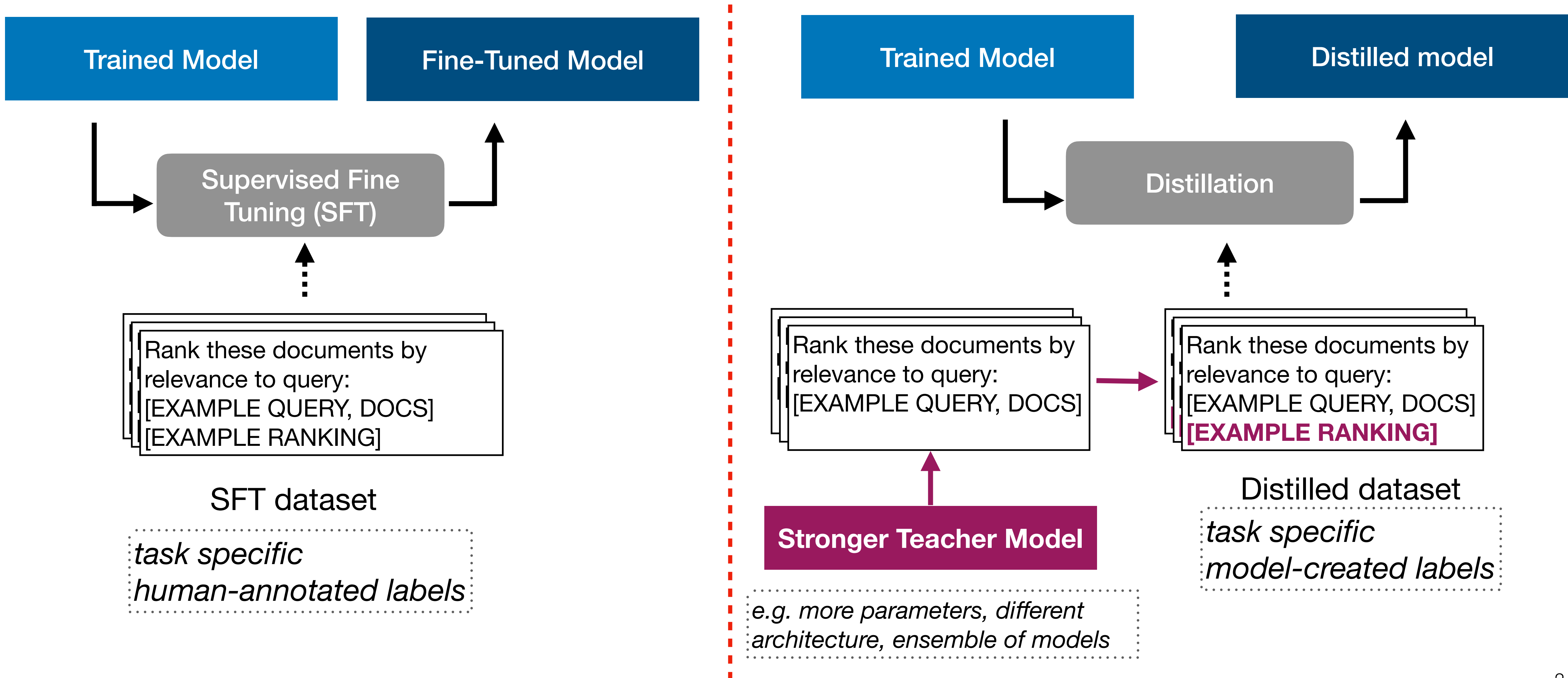
# Supervised Fine-Tuning (SFT)



# Synthetic Training Data Generation



# Distillation (of labels)





# Distillation (of scores)

Trained Model

Fine-Tuned Model

Supervised Fine  
Tuning (SFT)

Rank these documents by  
relevance to query:  
[EXAMPLE QUERY, DOCS]  
[EXAMPLE RANKING]

SFT dataset

*task specific*  
*human-annotated labels*

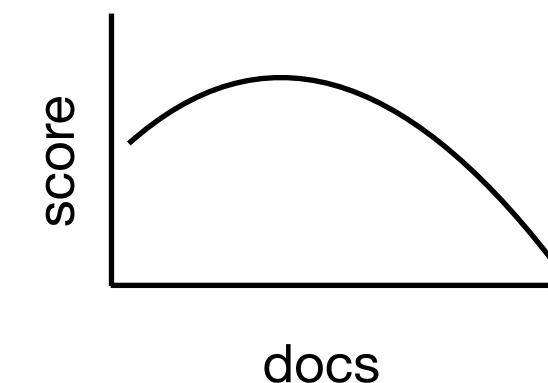
Trained Model

Distilled model

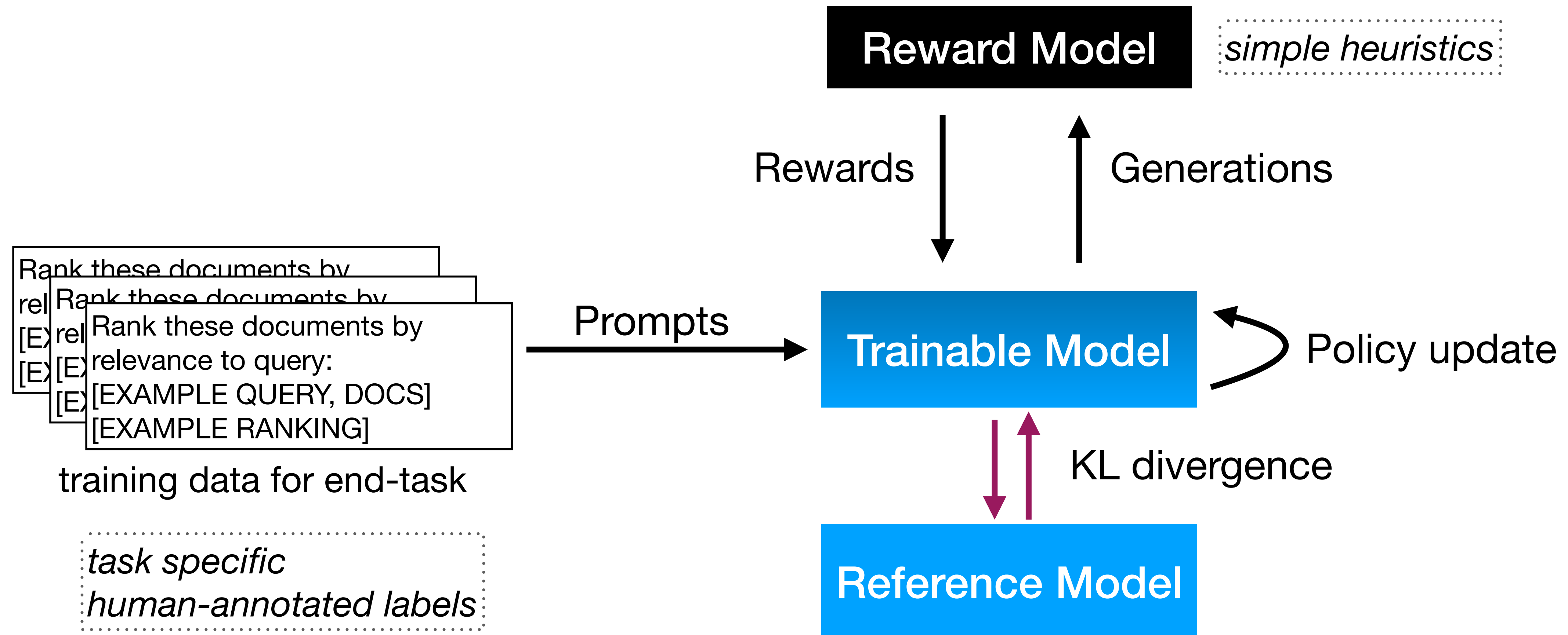
Distillation

Stronger Teacher Model

*e.g. more parameters, different  
architecture, ensemble of models*

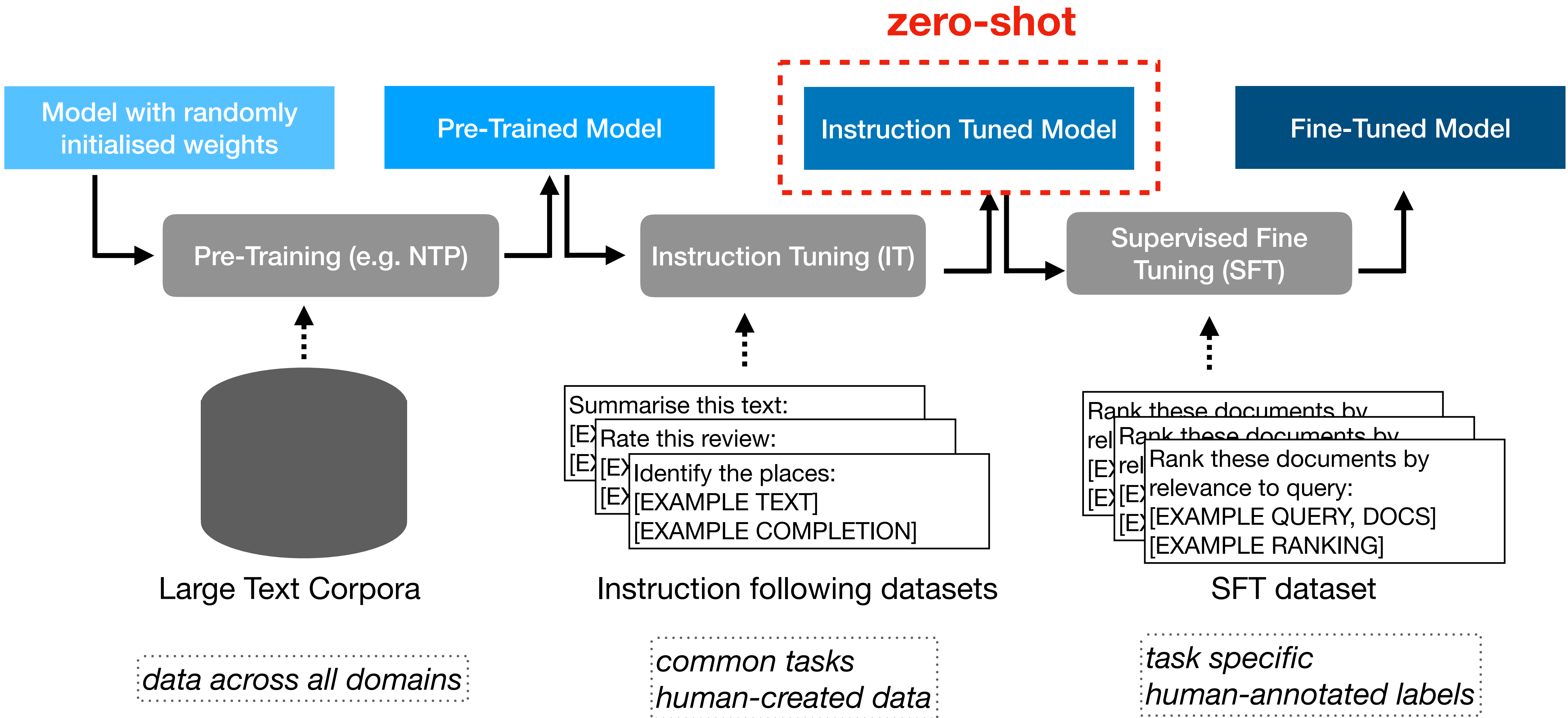


# Reinforcement Learning (*with GRPO*)

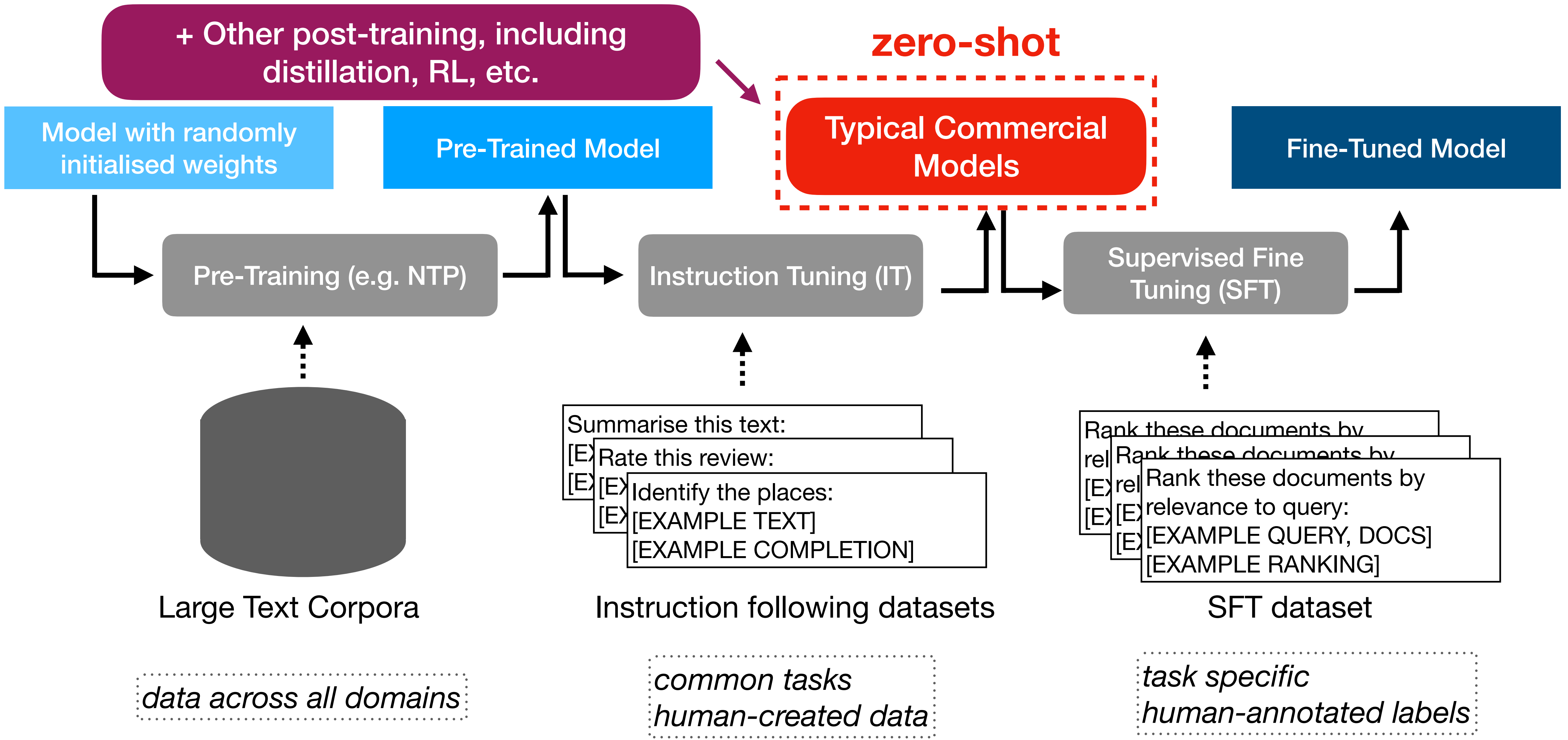




# At Inference: (ideal) Zero-shot vs. Fine-Tuned



# At Inference: (nowadays) Zero-shot vs. Fine-Tuned



**What's in the next 2 hours**



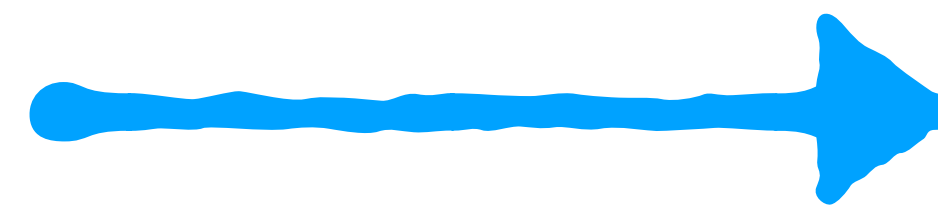
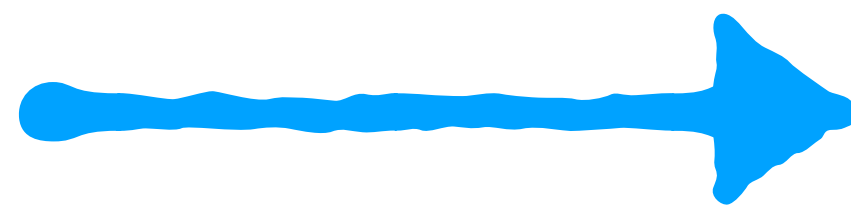
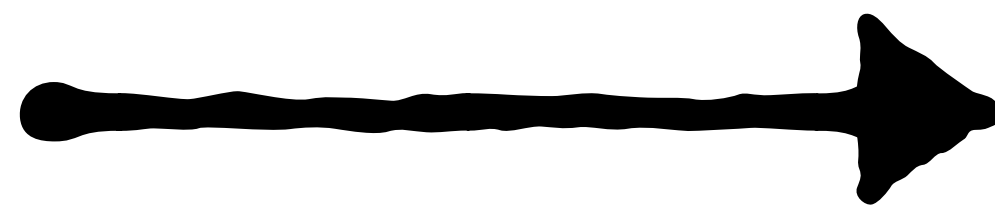
# Tutorial Parts

**Part 1:  
Intro**

**Part 2:  
Retrieval**

**Part 3:  
Ranking**

**Part 4:  
Challenges/  
Opportunities**



# Part 2: Retrieval

## Prelude: Neural Retrieval Background

### Large Language Models Retriever

- Backbone
- Training Data
- Deep Retrieval
- Beyond English & Text:  
Multi-modality, Multi-linguality





# Part 3: Ranking

## Prelude: From LTR to Transformer-based Ranking

### Large Language Models Ranking

- Fine-tuning
- Zero-shot: prompting, sorting methods
- Distillation
- Reasoning for ranking



# Part 4: Challenges and Opportunities

- Connections to LLM4Eval
- Challenges and Opportunities
- Pointers and Resources



# Who are we?

## Bevan Koopman



- Associate Professor at The University of Queensland, Principal Researcher at CSIRO Australia
- LLM rankers for searching in high stakes decision making domains (e.g., health)
- Information retrieval for human decision making
- Retriever-generator interplay
- Putting neural IR into practise



# Who are we?

## Guido Zuccon



- Professor at The University of Queensland, Visiting Researcher at Google Research Australia
- Developed LLM-based retrieval and ranking approaches
  - Setwise
  - Rank-R1
  - PromptReps
  - Methods to leverage LLMs for relevance feedback
- focus on efficiency and trade-offs in terms of computational resources and latency, and of training data



# Who are we?

## Xueguang Ma



- PhD Candidate at the University of Waterloo
- Extensive expertise in effective training of neural rankers and retrievers, including based on generative LLMs.
  - RankLLaMa (re-ranking) & RepLLaMa (dense retrieval)
  - LRL (zero-shot listwise re-ranker),
  - HyDE (relevance feedback)
  - DRAMA
- multimodal capabilities of retrieval methods based on generative LLM backbones
- Creator of Tevatron3: popular open-source LLM-based ranker training toolkit



# Who are we?

## Shengyao Zhuang



- Applied Scientist @ Amazon. Adjunct Fellow at The University of Queensland.
- Developed LLM-based retrieval and ranking approaches
  - Setwise
  - Rank-R1
  - PromptReps
  - Methods to leverage LLMs for relevance feedback
- Creator of Tevatron + open-source codebase for LLM-based rerankers

# End of Part 1

...