R²LLMs: Retrieval and Ranking with LLMs

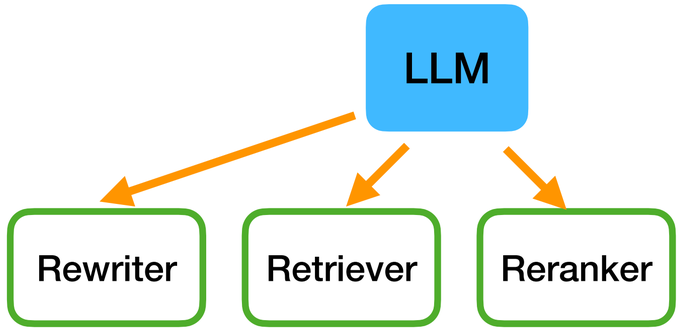
Guido Zuccon[1], **Shengyao** Zhuang[2], **Xueguang** Ma[3]

[1] ielab, The University of Queensland, Australia & Google Research Australia
[2] ielab, CSIRO, Australia
[3] 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

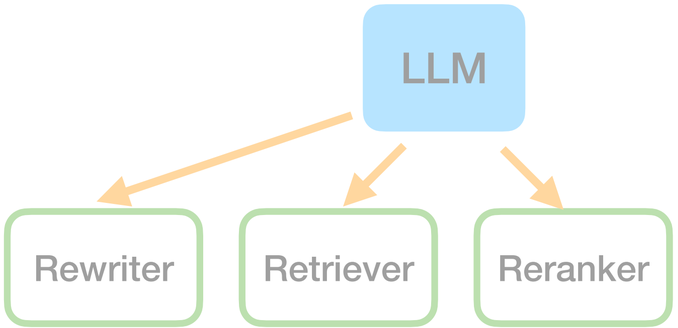


LLM-enhanced information retrieval

e.g. RankGPT, PromptReps

What this tutorial is not about

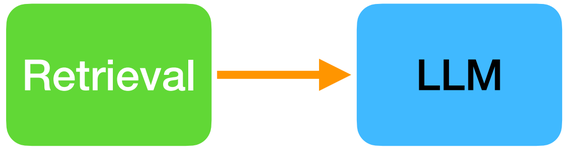
Prompting LLMs for Retrieval and Ranking



LLM-enhanced information retrieval

e.g. RankGPT, PromptReps

Retrieval Augmented Generation

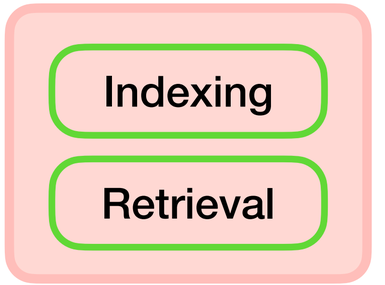


IR4LLM,
Retrieval-augmented LLM,
Reliable response generation,

Search Agents

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

Model-based IR

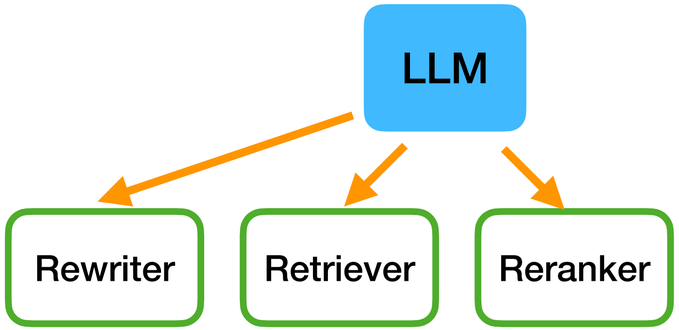


LLM as Indexer and Retriever, Generative [document] Retrieval (GR)

e.g. Differentiable Search Index (DSI)

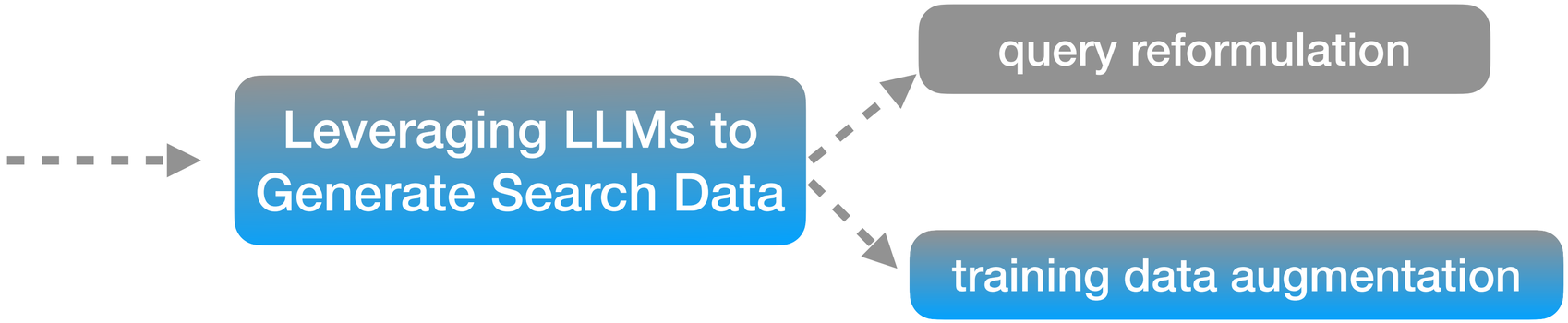
Focus of this tutorial: Using LLMs for Retrieval and Ranking

Prompting LLMs for Retrieval and Ranking



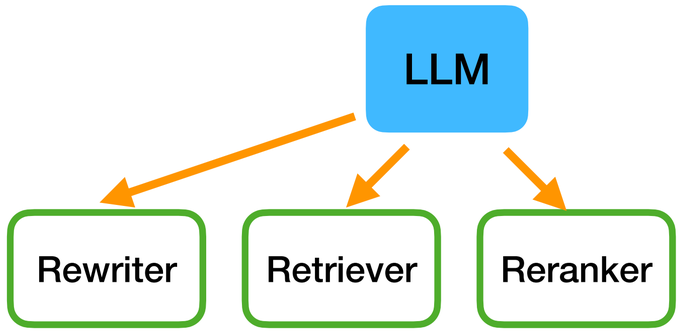
LLM-enhanced information retrieval

e.g. RankGPT, PromptReps



Focus of this tutorial: Using LLMs for Retrieval and Ranking

Prompting LLMs for Retrieval and Ranking



LLM-enhanced information retrieval

e.g. RankGPT, PromptReps

Leveraging LLMs to Generate Search Data

- query reformulation
- training data augmentation

Leveraging LLMs as Retrievers & Rankers Backbone

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

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

SIRP Panel @ Monday July 14 10:30-11:30

Is Search Dead? The Rise or Demise of Search in the Era of LLMs

Chirag Shah, Suzan Verberne, Yubin Kim, Neil Shah, Tracy Holloway King





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

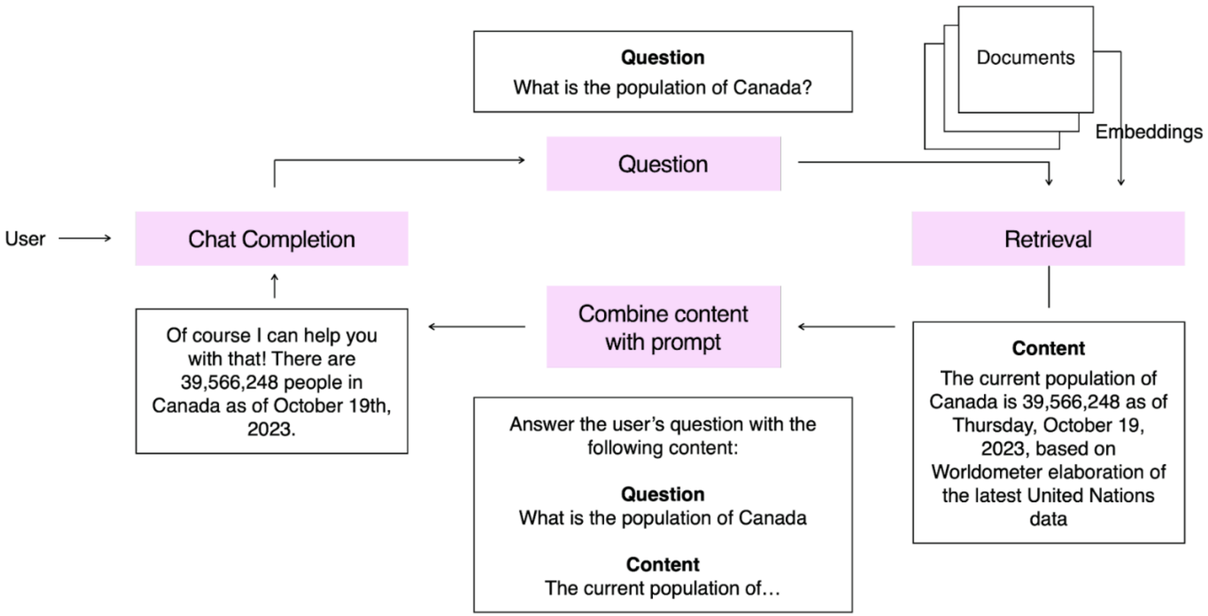


Image Source: https://platform.openai.com/docs/guides/optimizing-llm-accuracy/retrieval-augmented-generation-rag

Web Agents

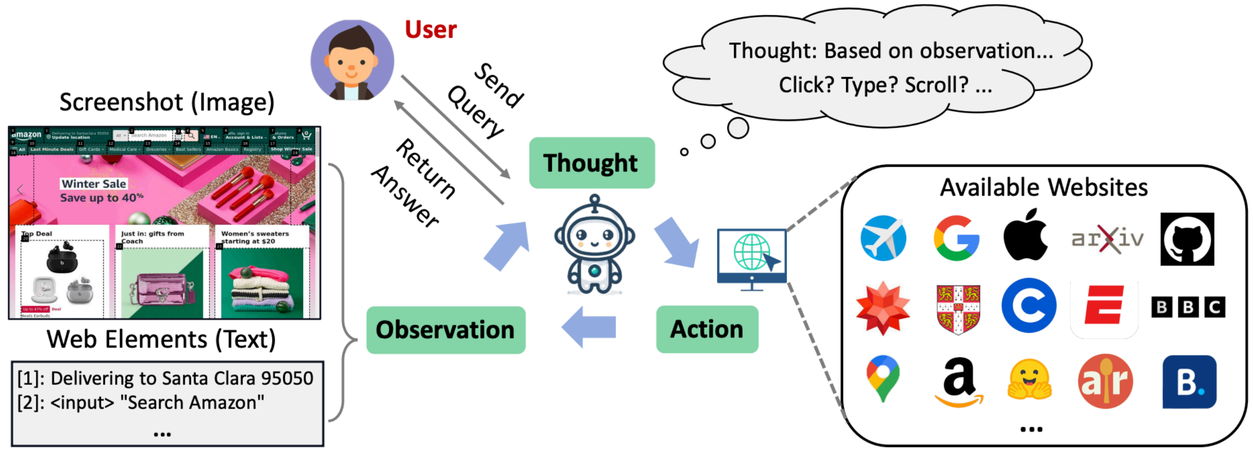
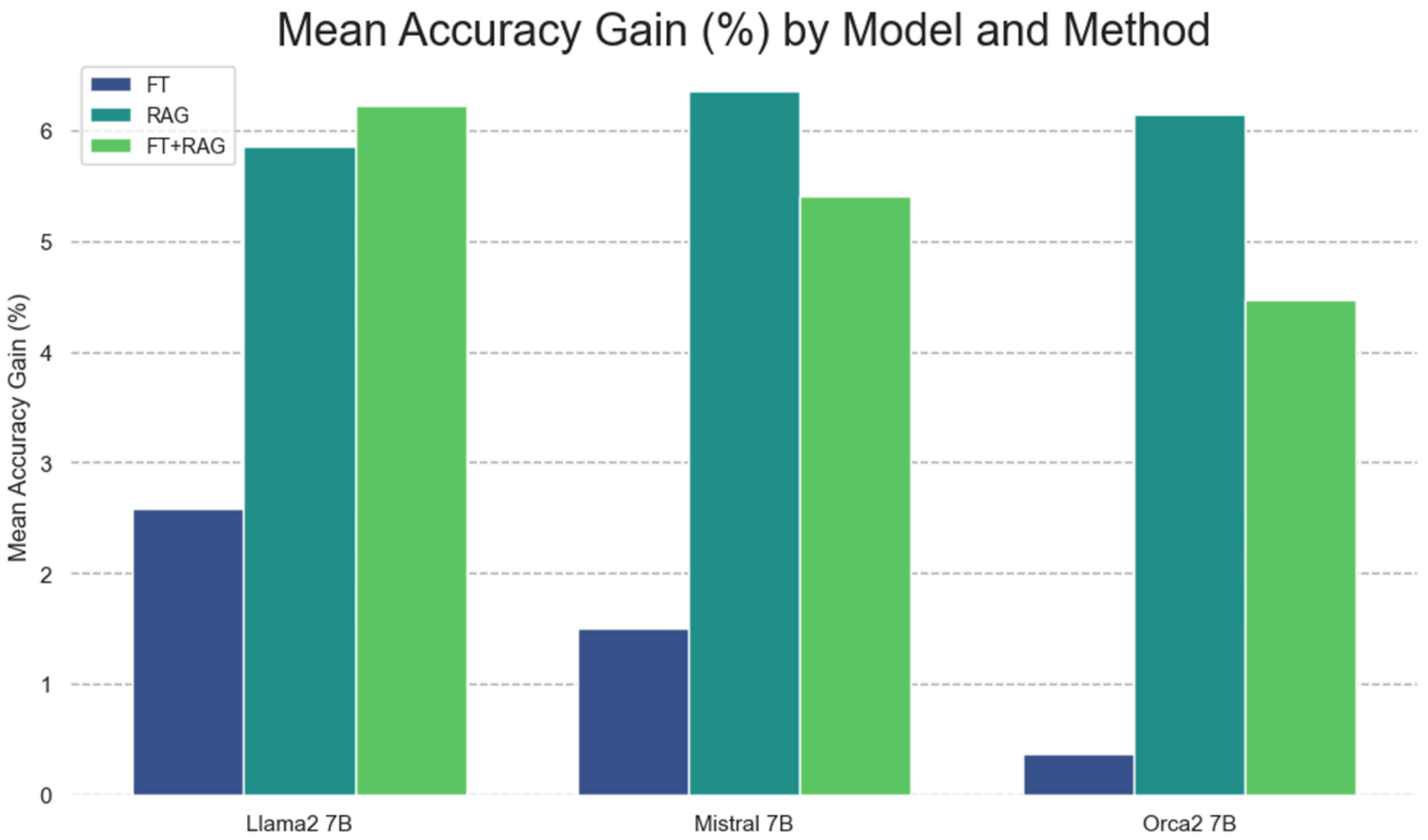


Image Source: H. He, et al. "WebVoyager: Building an End-to-End Web Agent with Large Multimodal Models", ACL 2024

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



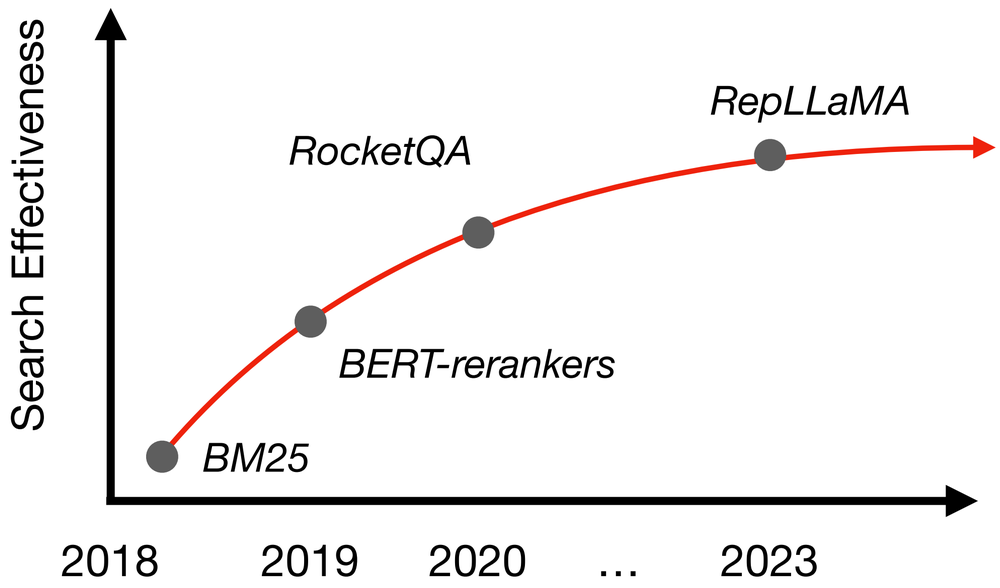
- Replacing search with LLMs is not the solution — integration is the way to go
- Ovadia et al.: in QA, unsupervised fine-tuning offers some improvement, but RAG consistently outperforms it, both for existing knowledge encountered during training and entirely new knowledge
- LLMs struggle to learn new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

Ovadia, O., Brief, M., Mishaeli, M. and Elisha, O., 2023. Fine-Tuning or Retrieval? Comparing Knowledge Injection in LLMs. *EMNLP 2024*

Ok, but why doing search with LLMs?

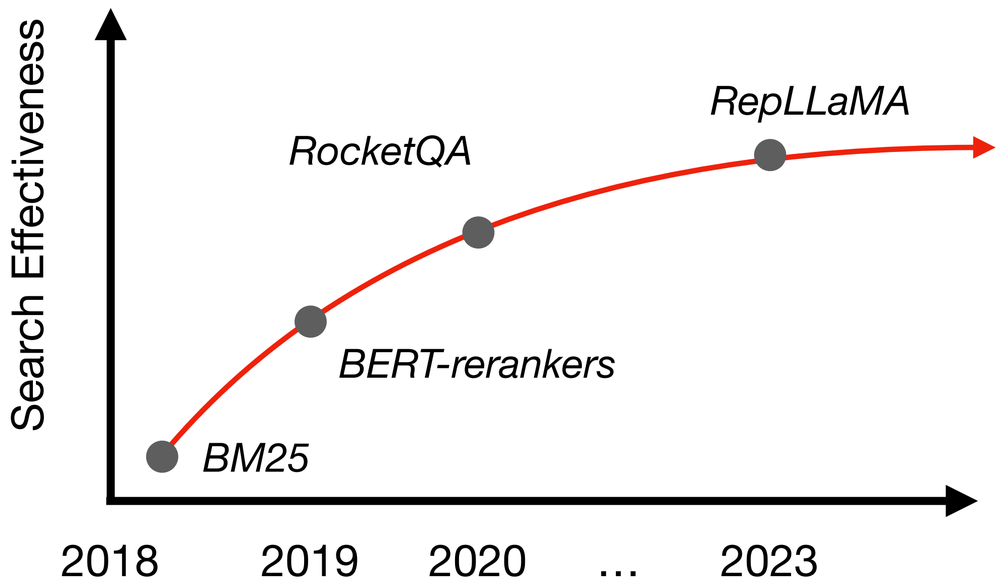
Massive progress in search quality in recent years

MS MARCO (in-domain data)

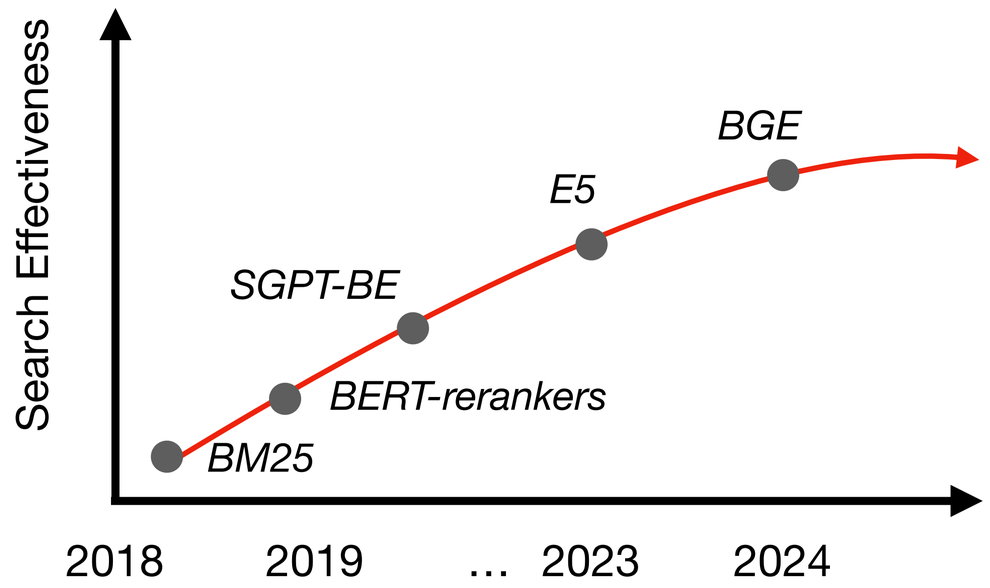


Massive progress in search quality in recent years

MS MARCO (in-domain data)



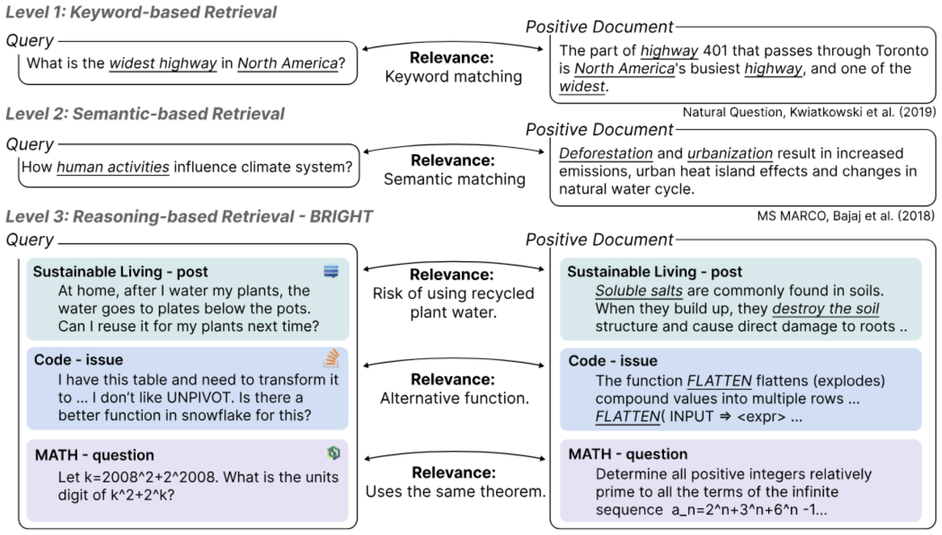
BEIR (out-domain data/~task)



However, search still challenging in many situations

Intensive Reasoning Datasets

BRIGHT, BrowseComp



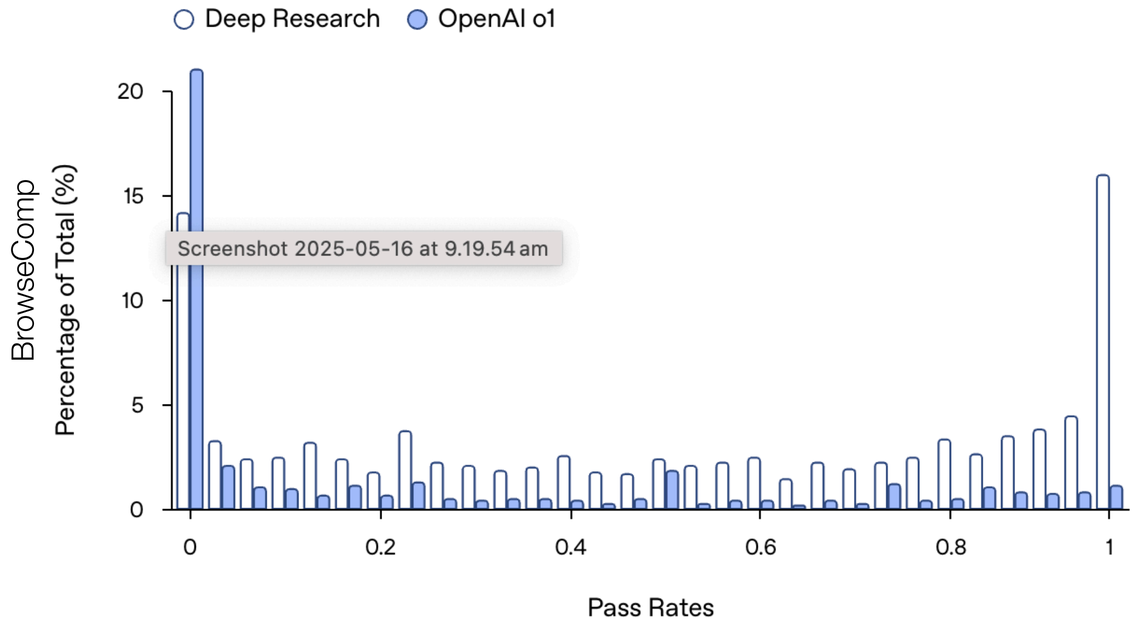
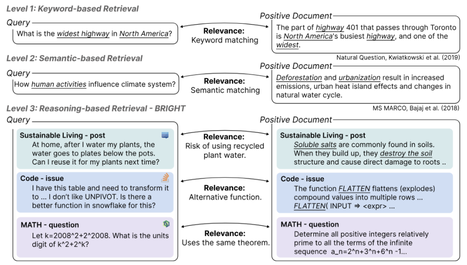


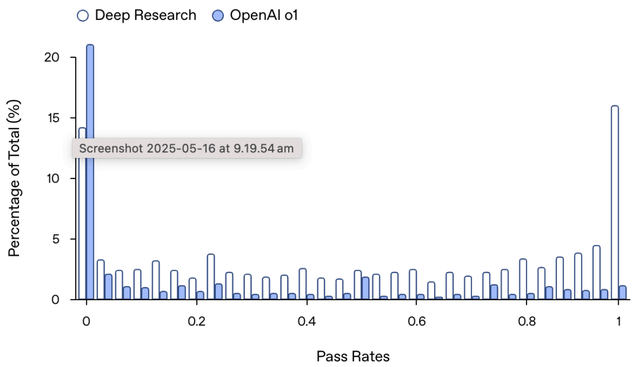
Image Source: https://openai.com/index/browsecomp/

However, search still challenging in many situations

Intensive Reasoning Datasets

BRIGHT, BrowseComp





Multi/Cross Lingual

ECLeKTic, TREC NeuCLIR

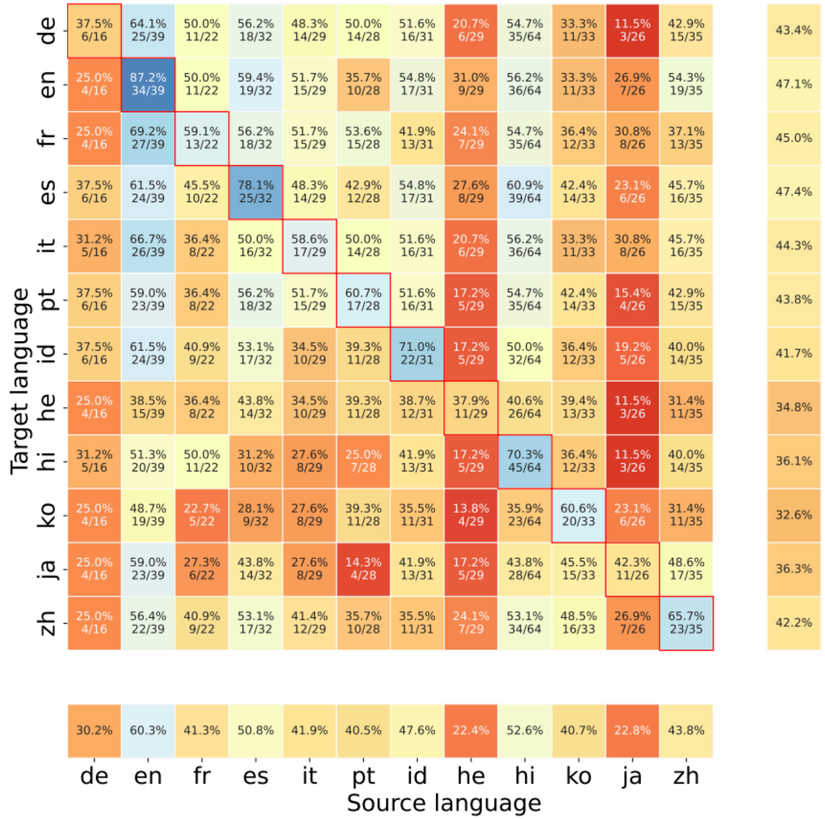
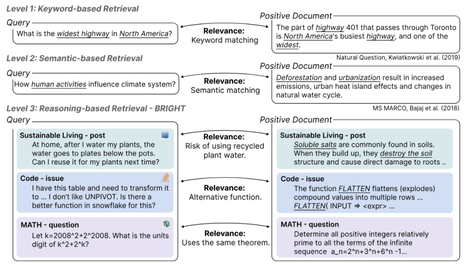


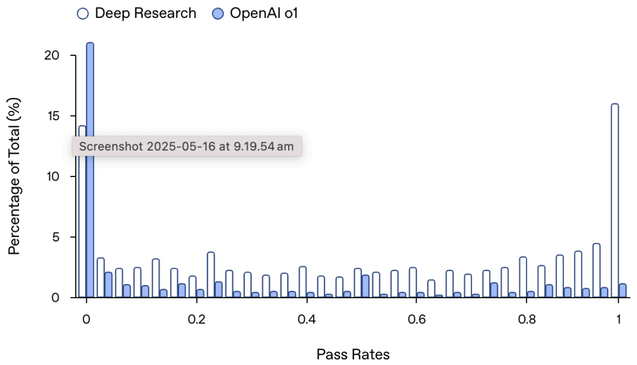
Image Source: O. Goldman, et al. "Eclektic: a novel challenge set for evaluation of cross-lingual knowledge transfer.", arXiv:2502.21228. 2025.

However, search still challenging in many situations

Intensive Reasoning Datasets

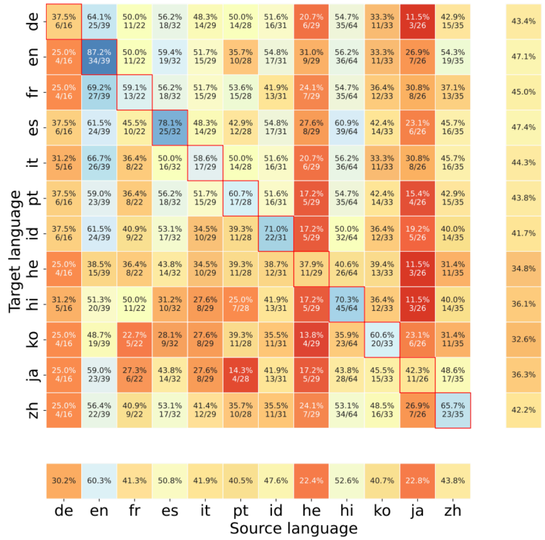
BRIGHT, BrowseComp





Multi/Cross Lingual

ECLeKTic, TREC NeuCLIR



Multi-Modality

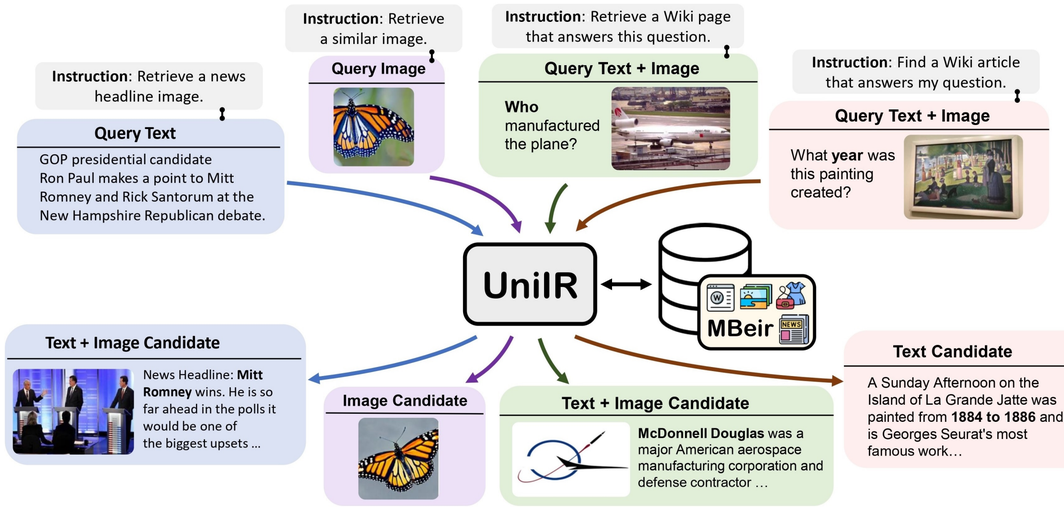
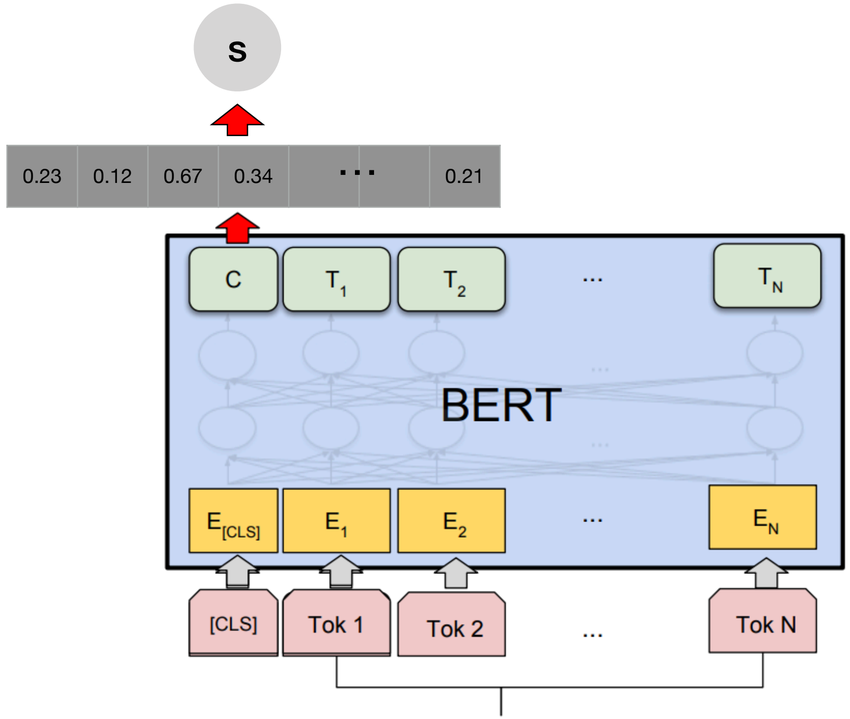
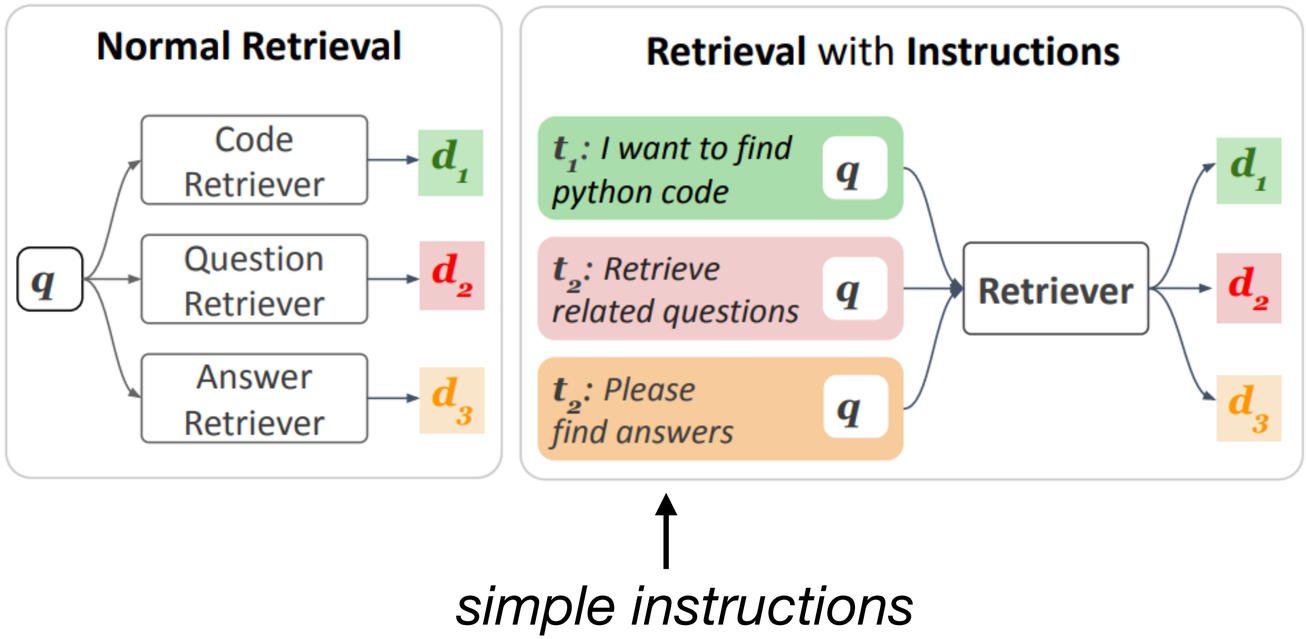


Image Source: C. Wei, et al. "Uniir: Training and benchmarking universal multimodal information retrievers.", ECCV 2024

From PreTrained LMs to Instruction Following LLMs

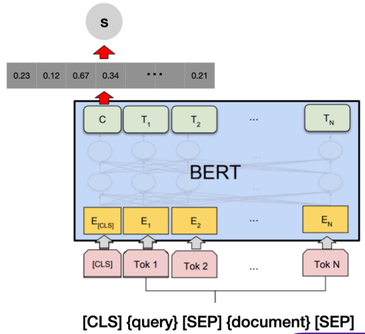


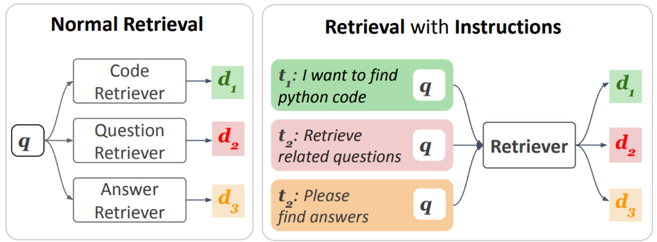
[CLS] {query} [SEP] {document} [SEP]



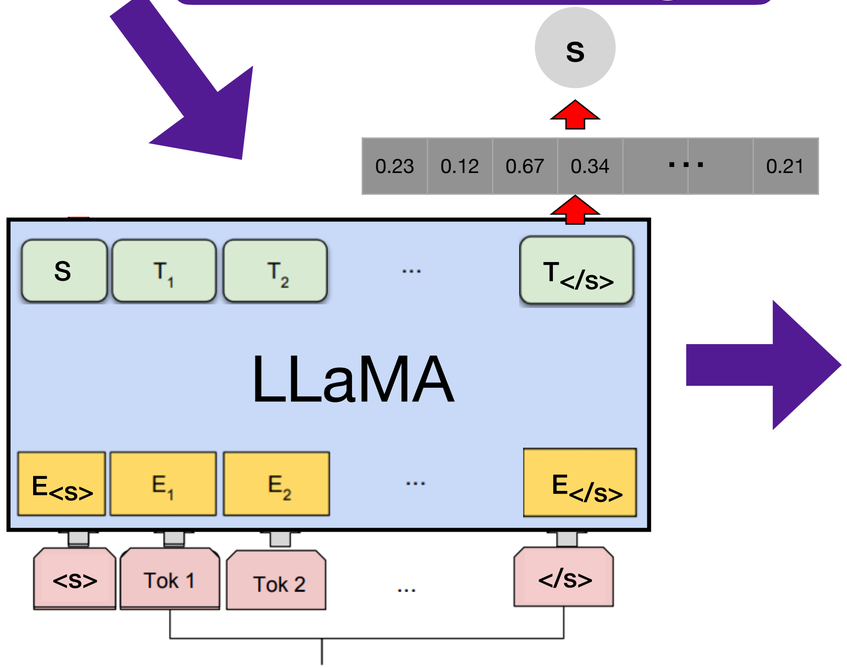
A. Asai, et al. "Task-aware Retrieval with Instructions." *ACL Findings* (2023).

From PreTrained LMs to Instruction Following LLMs





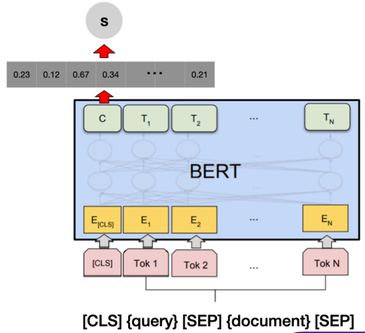
backbone change

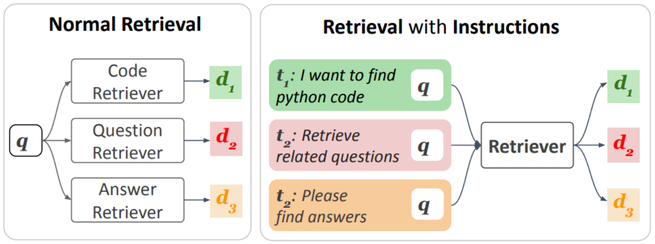


You are RankGPT, an intelligent assistant that can rank passages based on their relevancy to the query.
I will provide you with num passages, each indicated by number identifier []. Rank the passages based on their relevance to query: {query}.
{PASSAGES}
Search Query: {query}.
Rank the num passages above 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

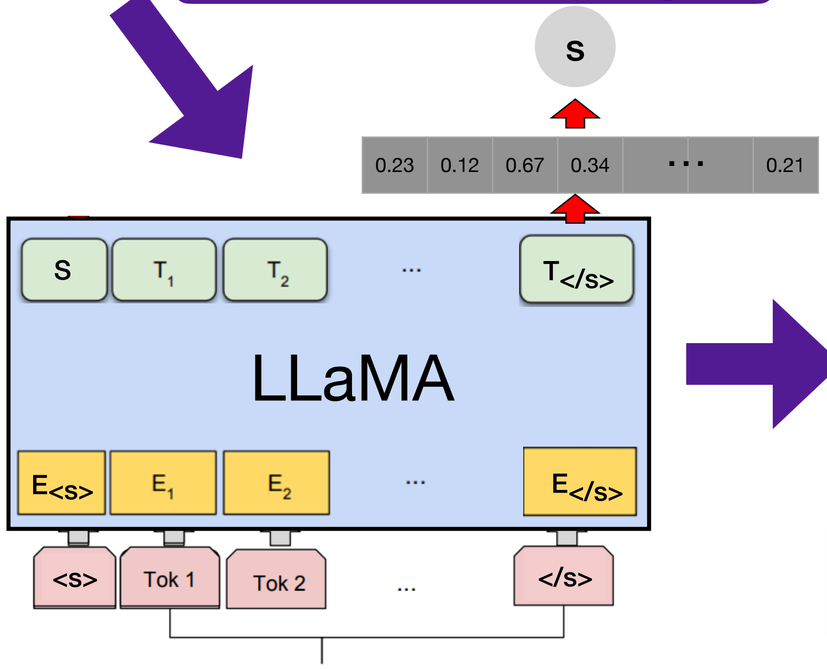
From PreTrained LMs to Instruction Following LLMs… to reasoners

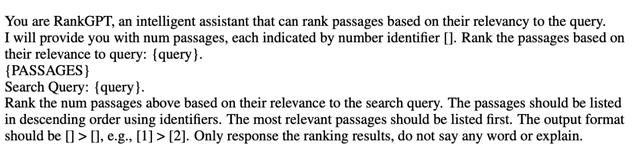




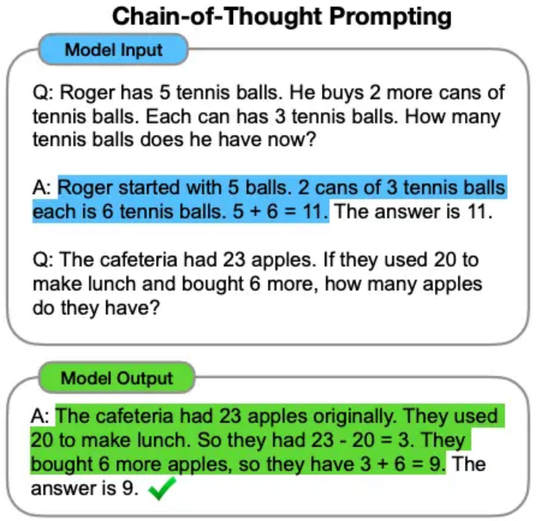
weak reasoning

backbone change

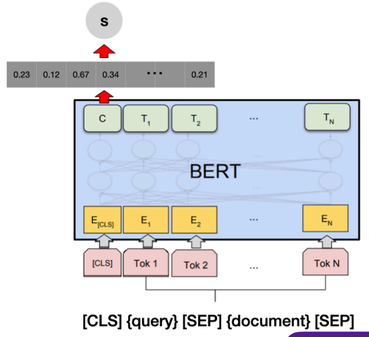


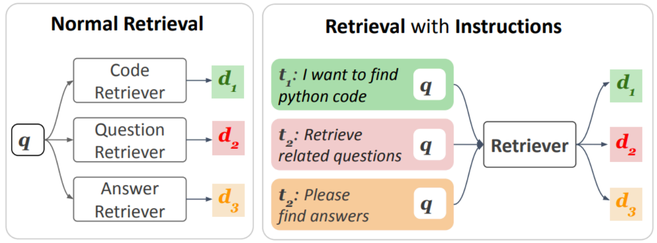


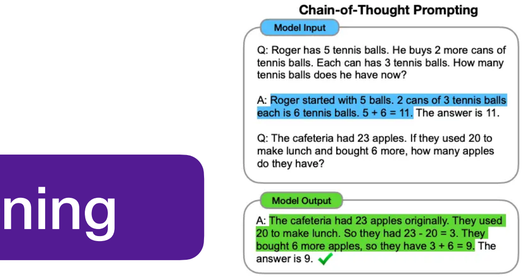
strong instruction following capabilities



From PreTrained LMs to Instruction Following LLMs… to reasoners



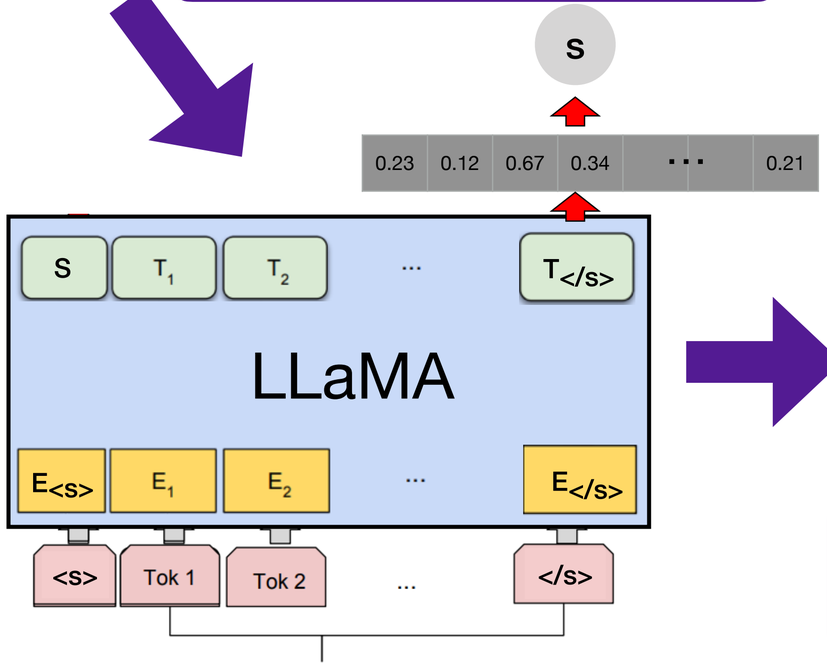




backbone change

weak reasoning

strong reasoning capabilities



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

Gemini 2.5 Pro

Deepseek R1

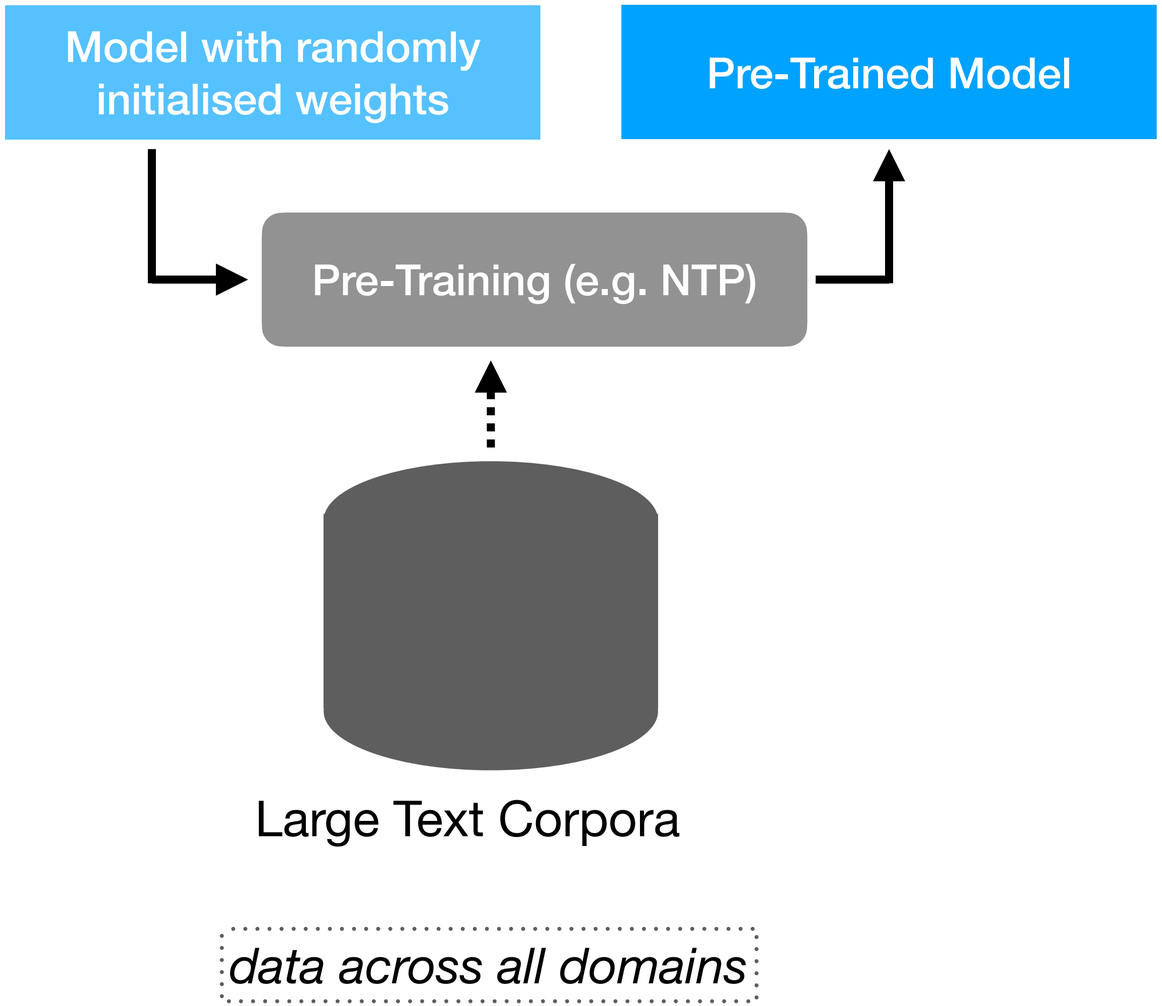
Claude 3.7 Sonnet

OpenAI o1

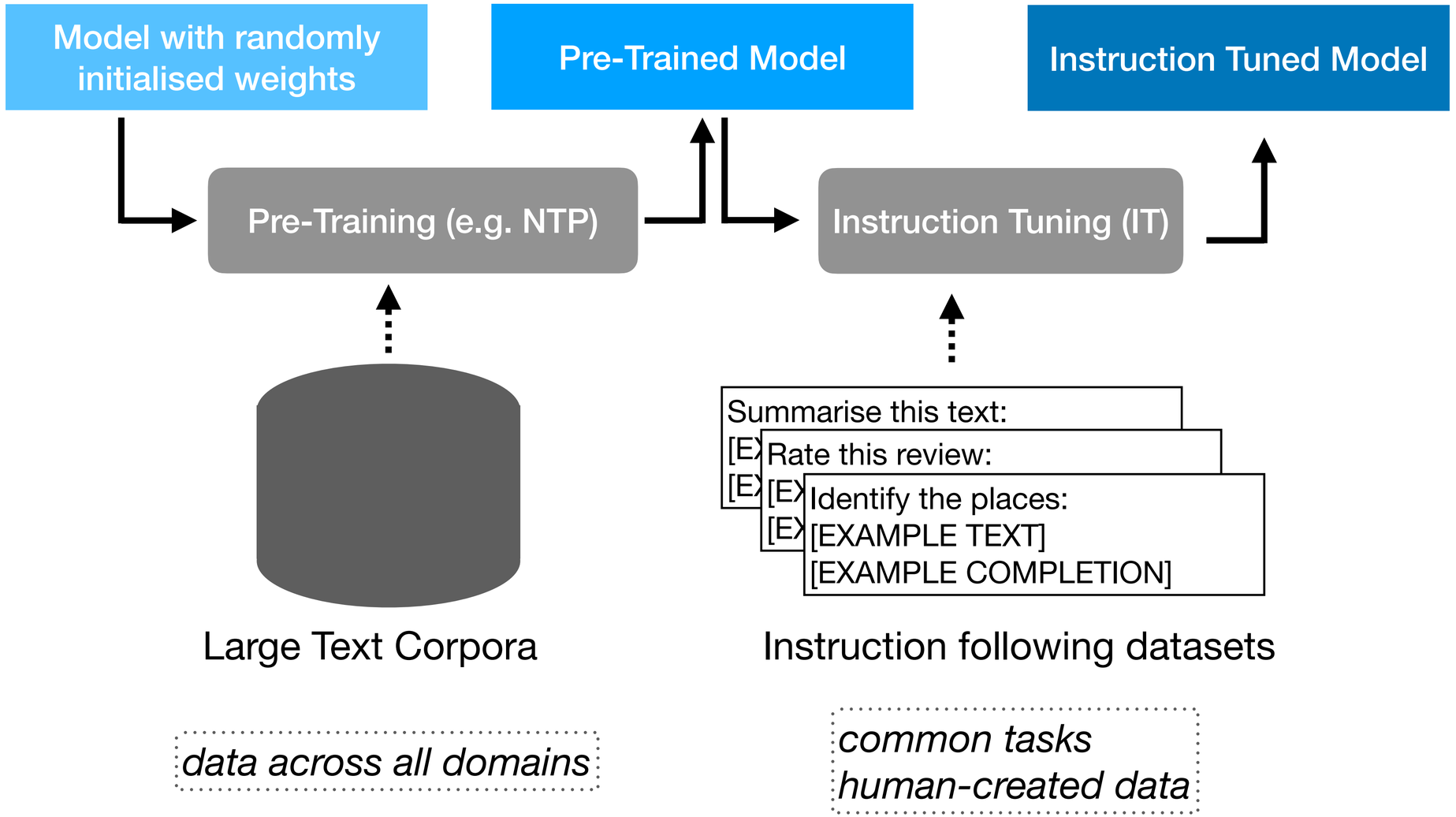
Qwen3

LLM training 101

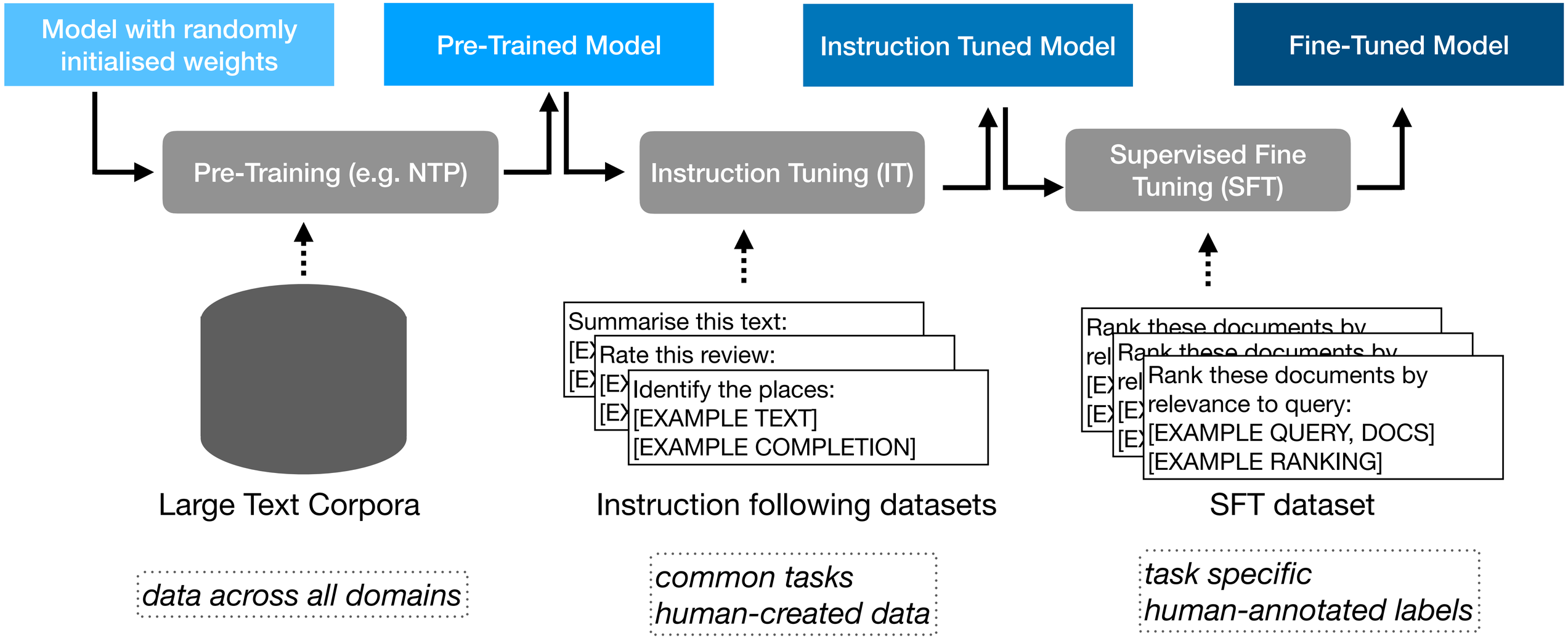
Pre-Training



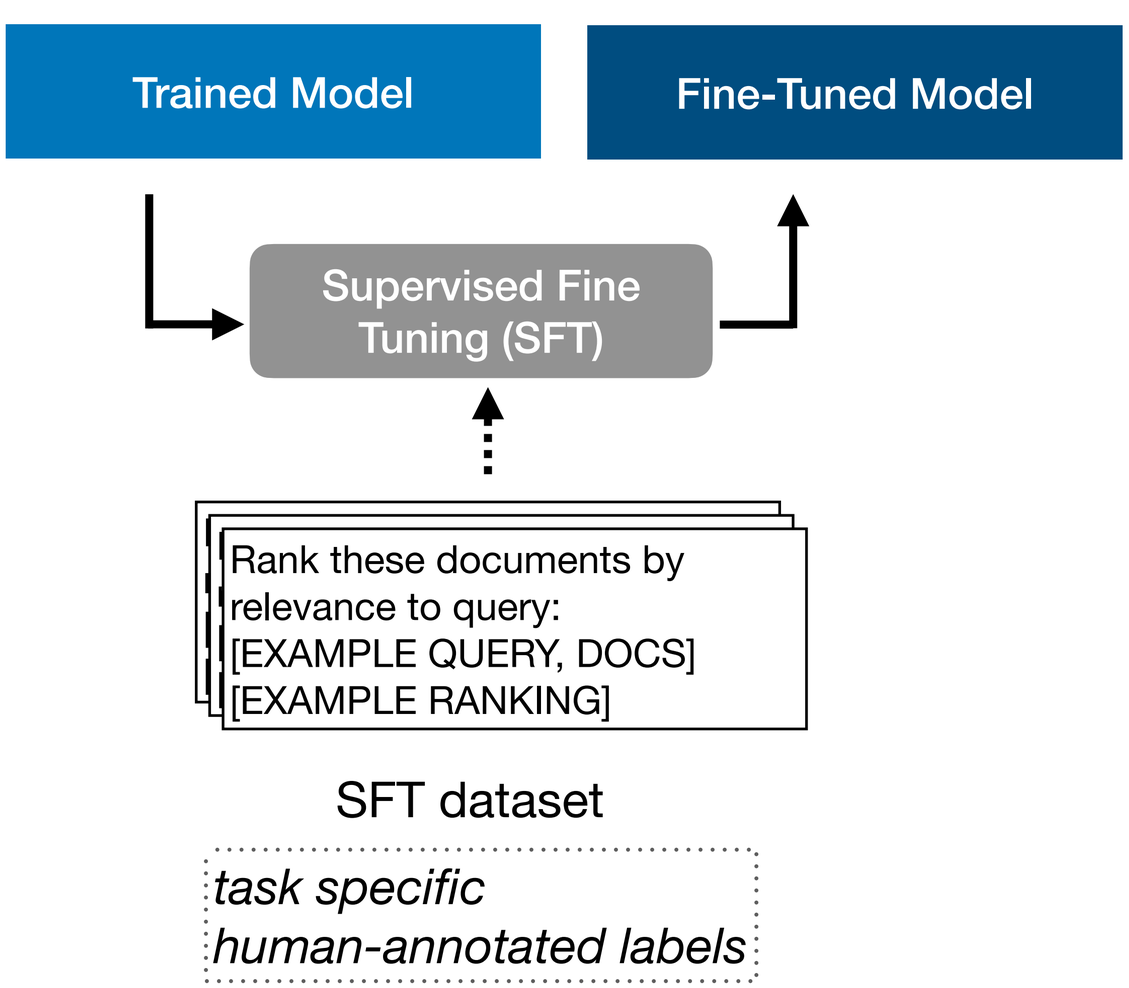
Instruction Tuning

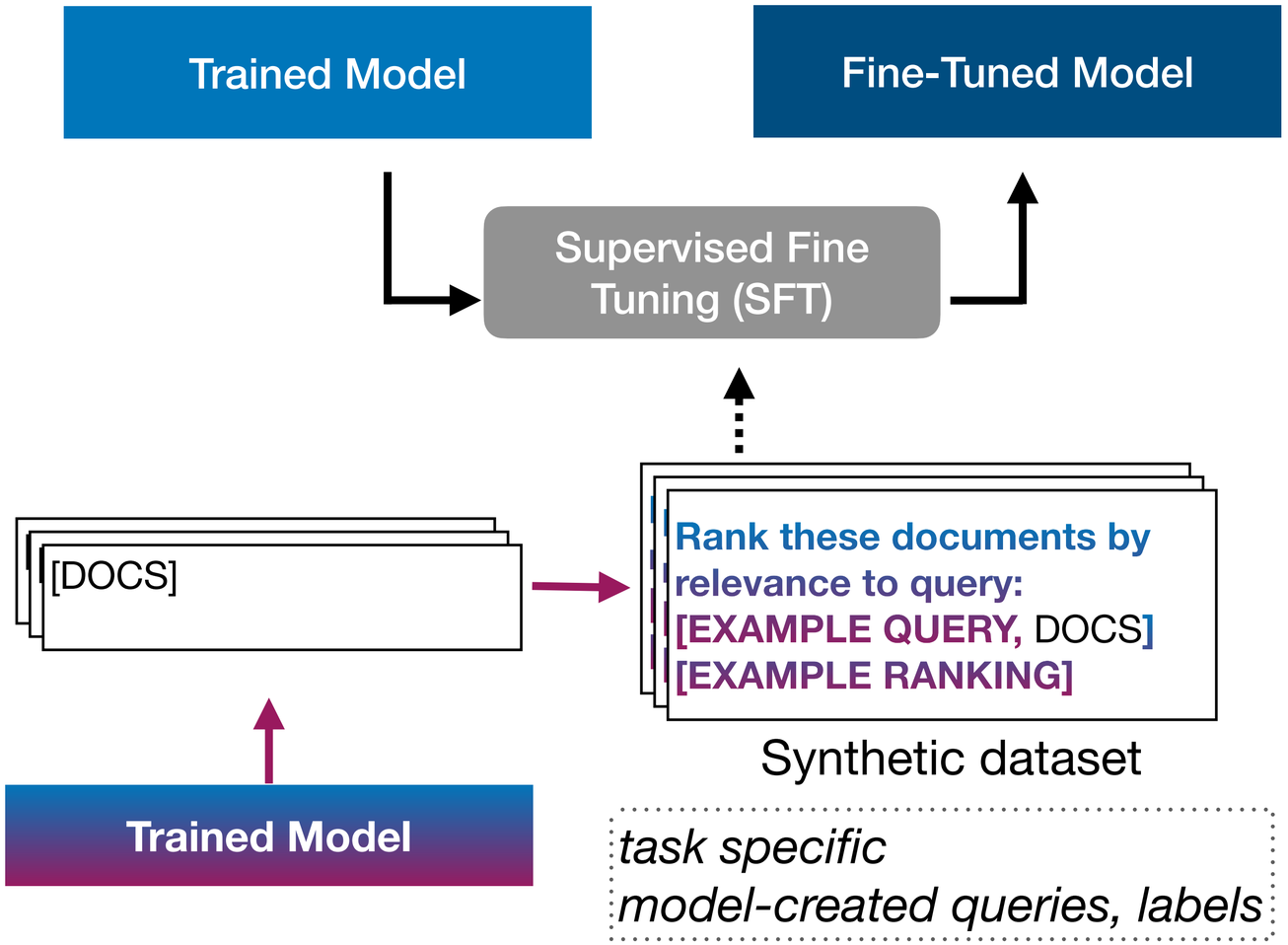


Supervised Fine-Tuning (SFT)

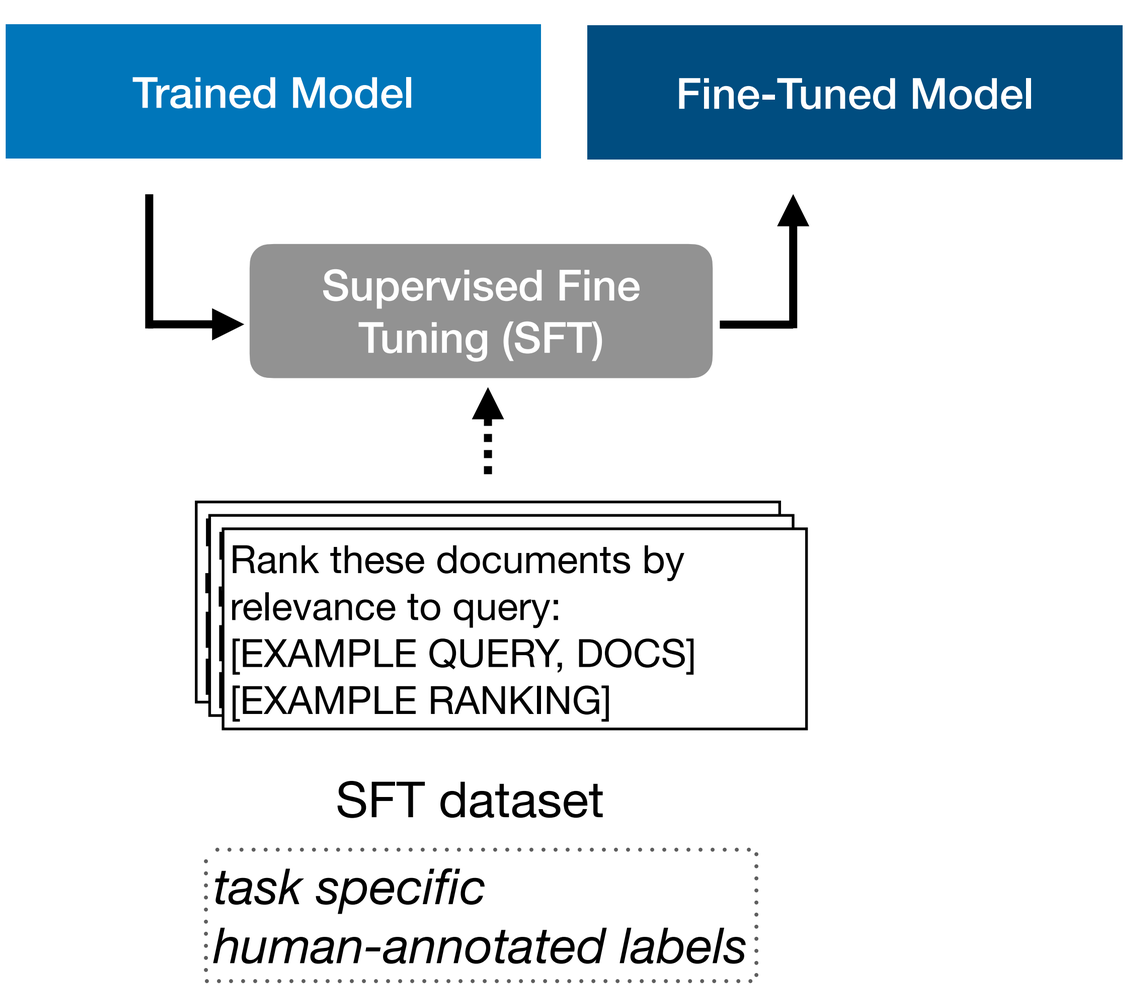


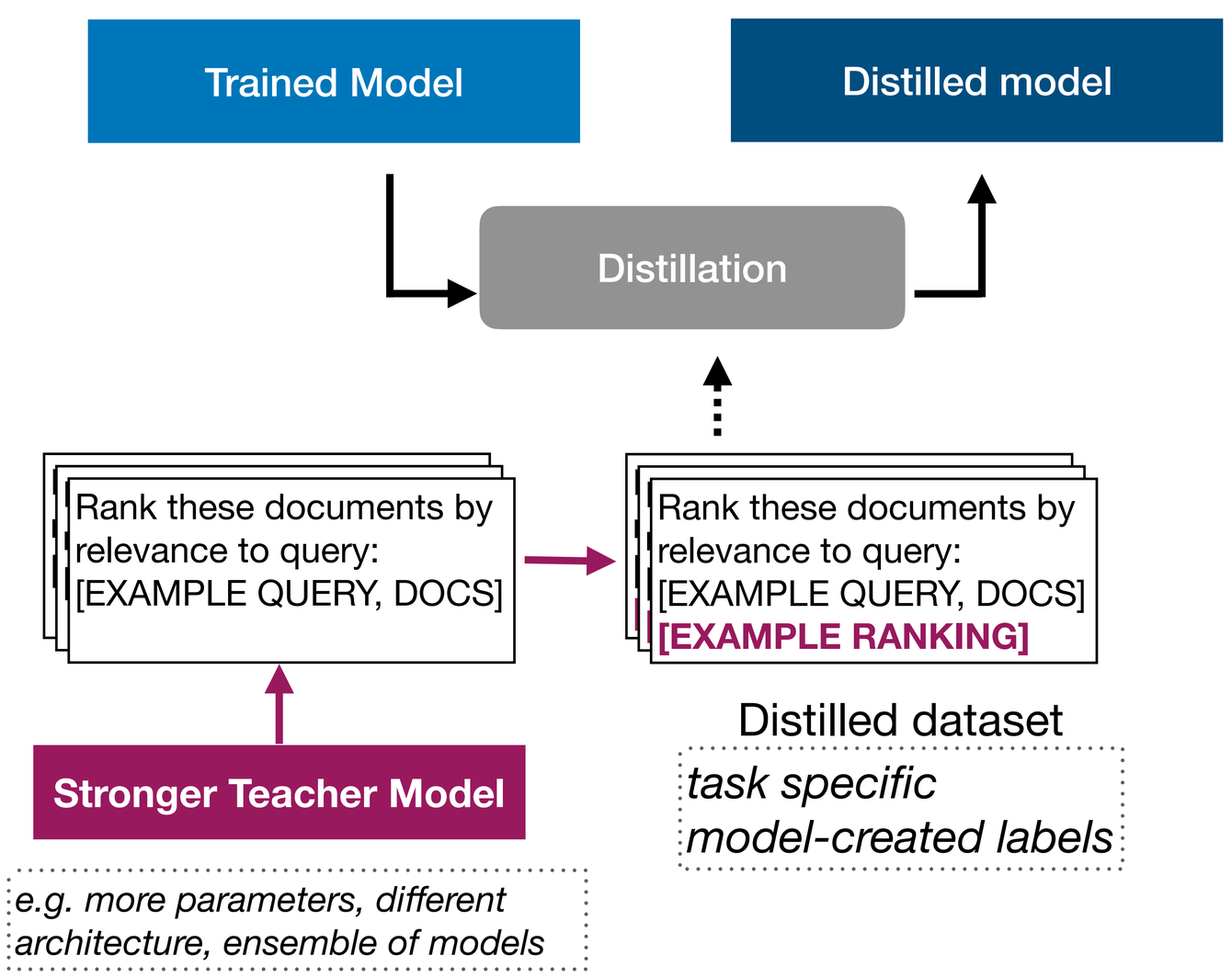
Synthetic Training Data Generation



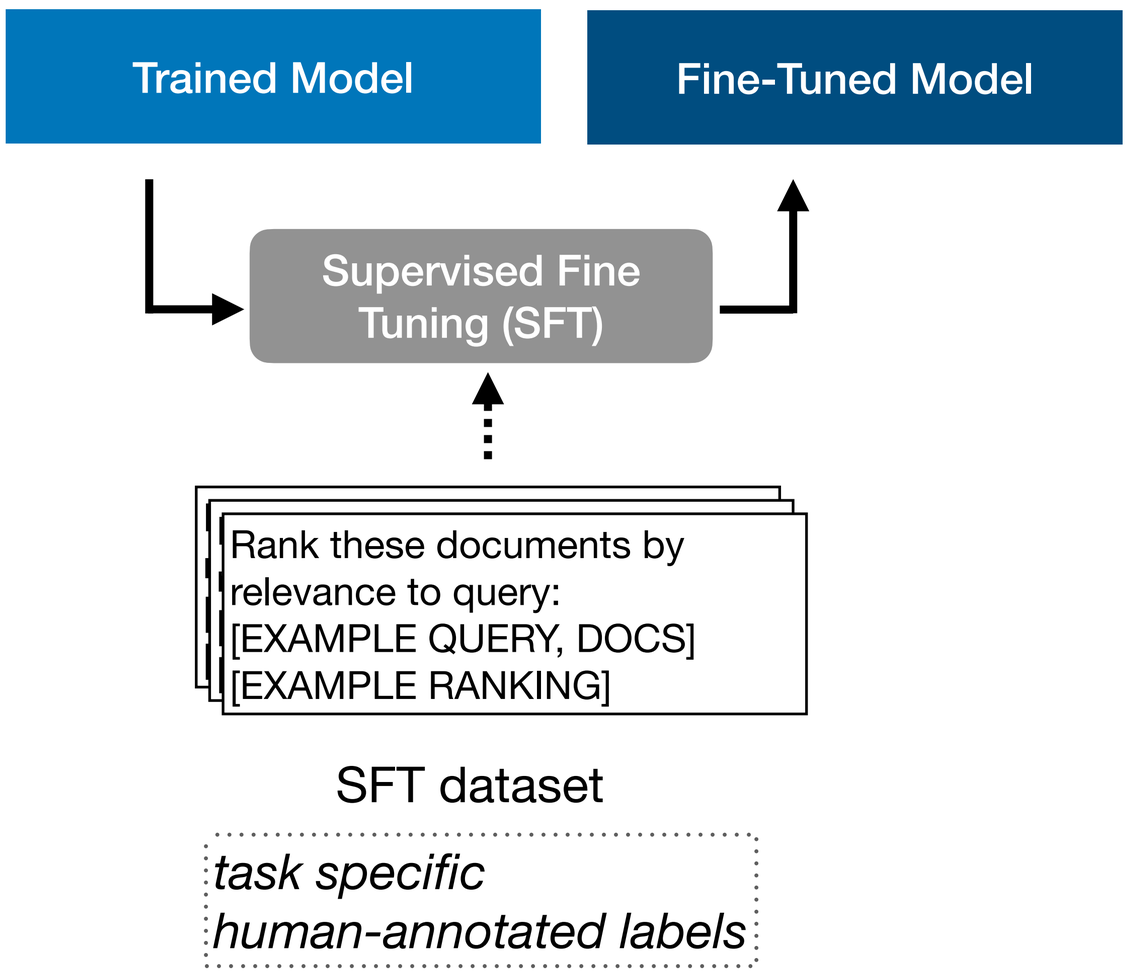


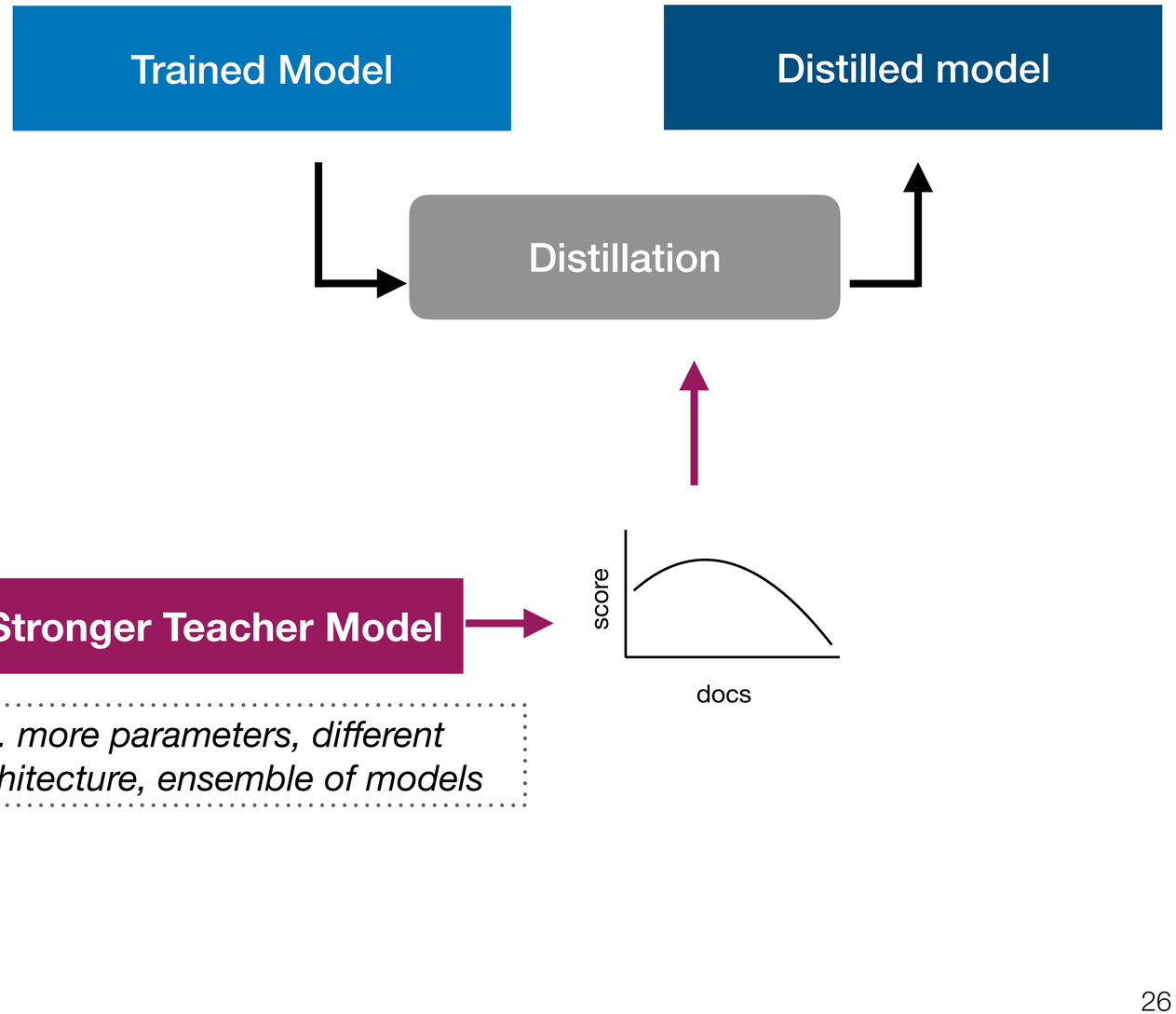
Distillation (of labels)



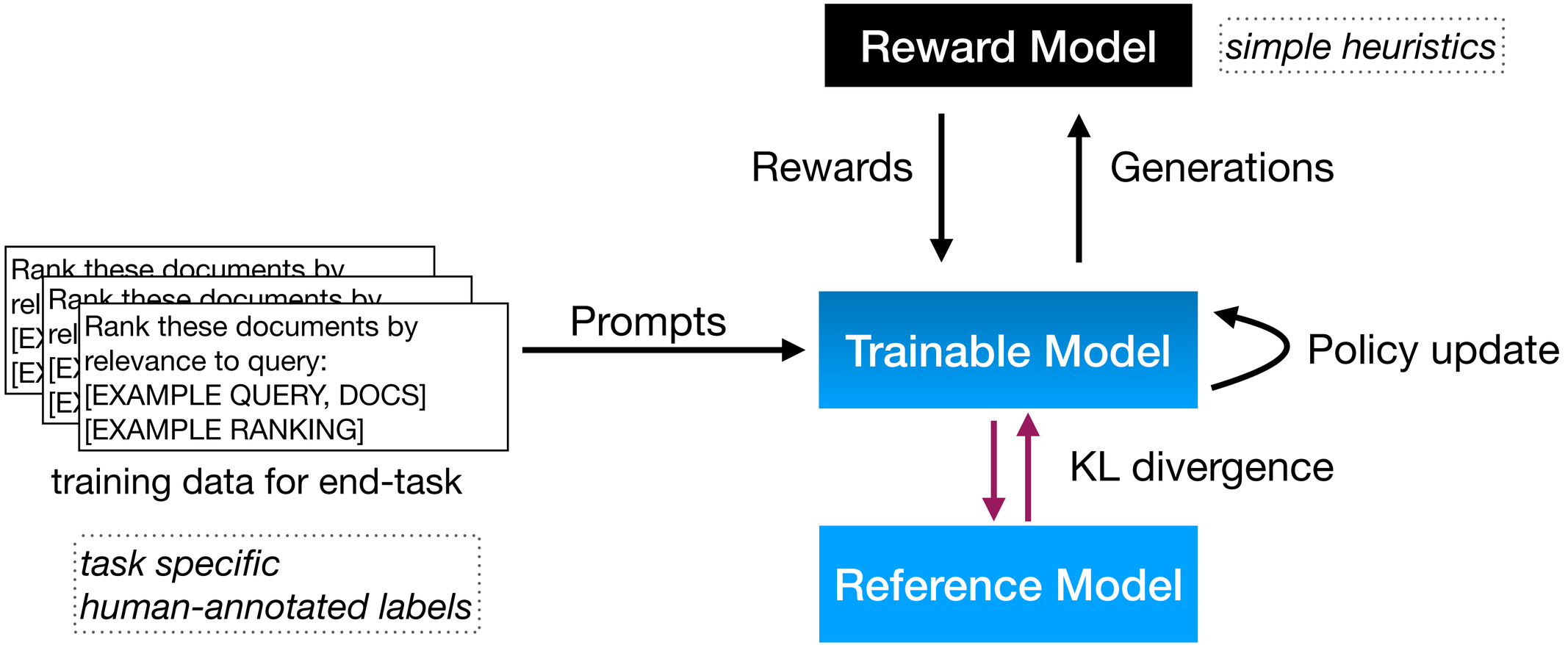


Distillation (of scores)

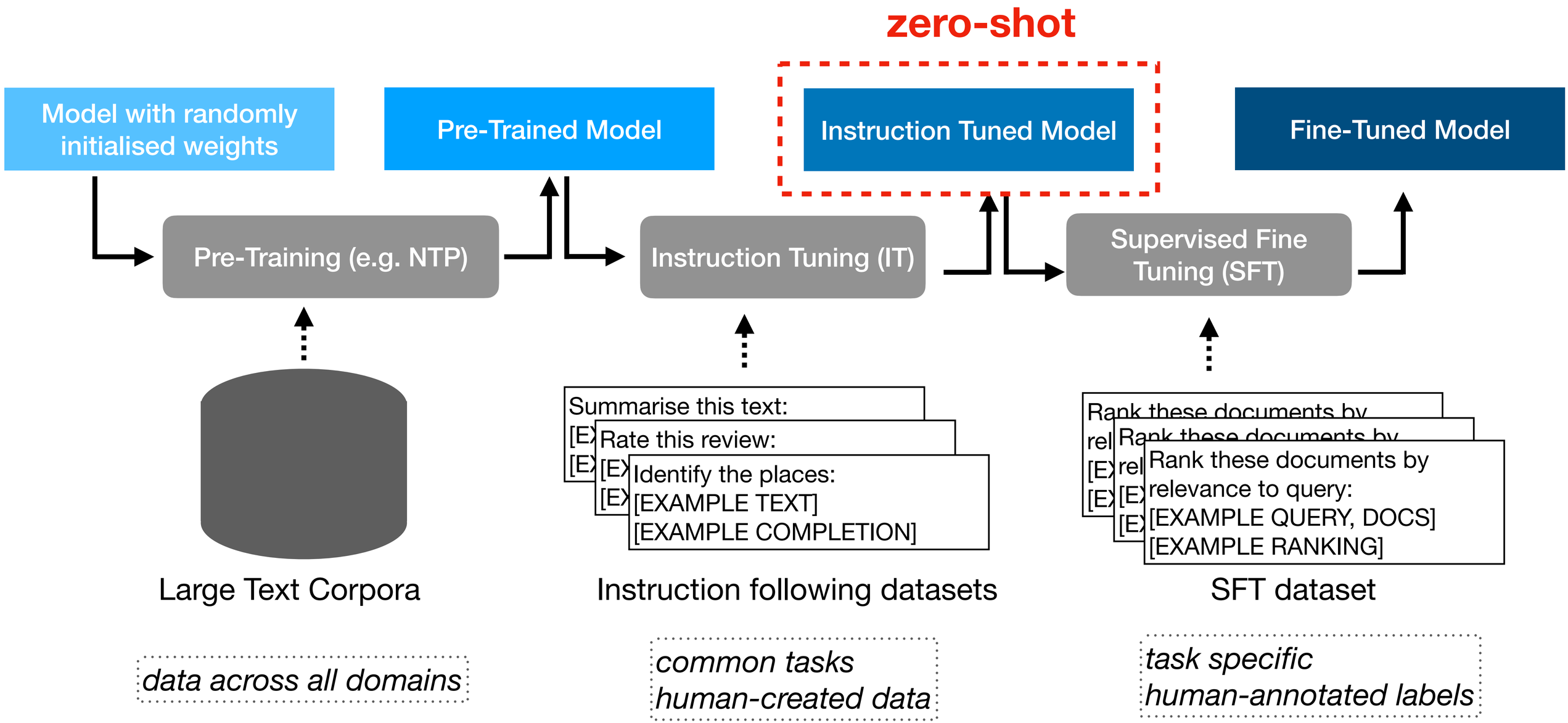




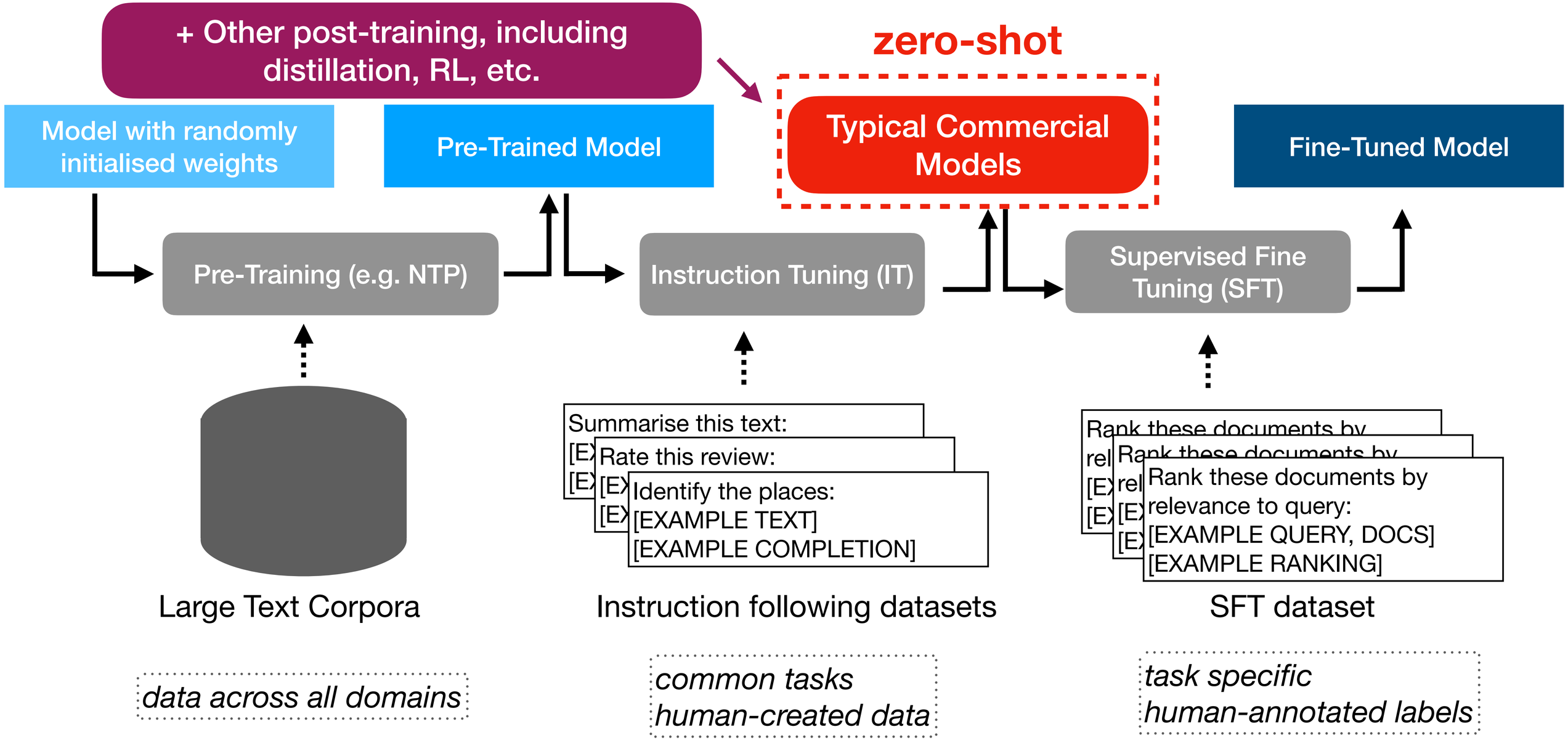
Reinforcement Learning (*with GRPO*)



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



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



What’s in the next 2 hours

Part 2: Retrieval (to coffee break)

Prelude: Neural Retrieval Background

Large Language Models Retriever

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

Part 3: Ranking (after coffee break)

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 (up to lunch)

Connections to LLM4Eval

Challenges and Opportunities

Pointers and Resources

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

- PostDoctoral Research Fellow at CSIRO, adjunct Research 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

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

End of Part 1

...