

R²LLMs: Retrieval and Ranking with LLMs

Guido Zuccon¹, Shengyao Zhuang², Xueguang Ma³, Bevan Koopman^{1,2}

¹ ielab, The University of Queensland, Australia & Google Research Australia

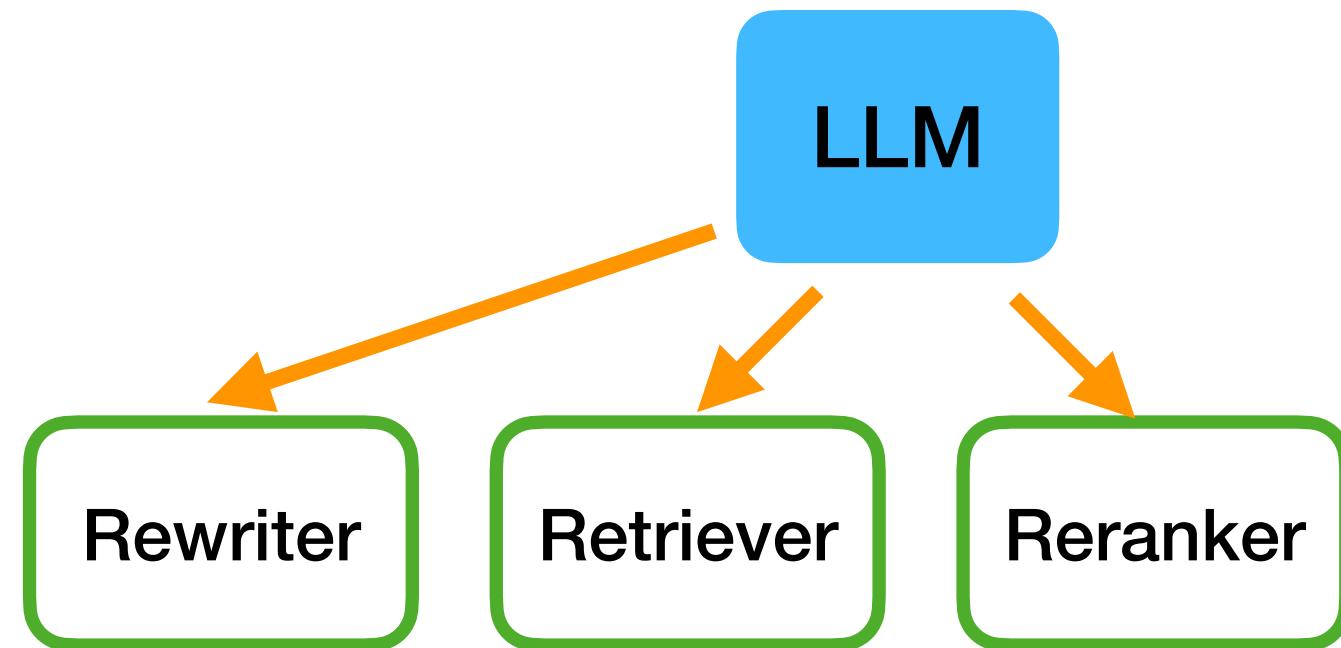
² ielab, CSIRO, Australia

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

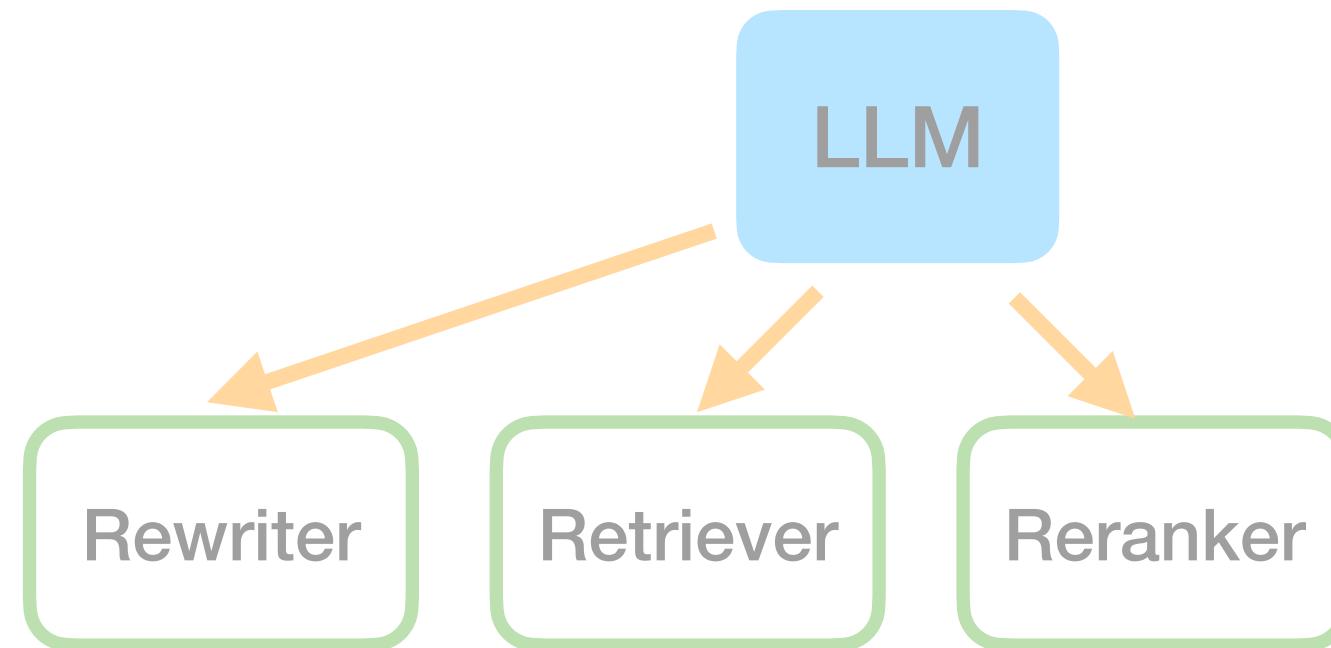


LLM-enhanced information retrieval

e.g. RankGPT, PromptReps

What this tutorial is not about

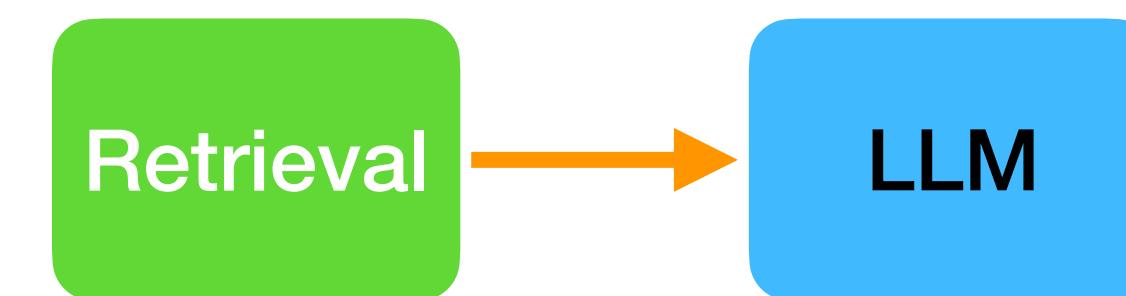
Prompting LLMs for Retrieval and Ranking



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

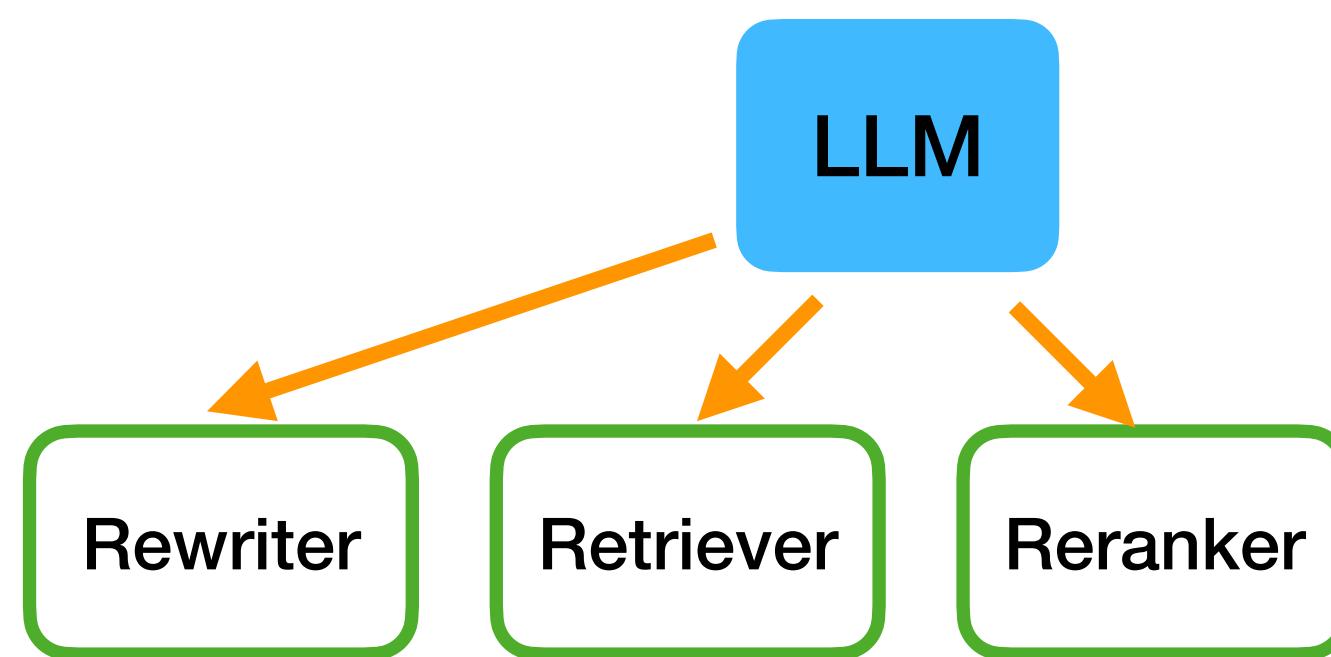


*LLM as Indexer and
Retriever, Generative
[document] Retrieval (GR)*

e.g. Differentiable Search Index (DSI)

Focus of this tutorial: Using LLMs for Retrieval and Ranking

Prompting LLMs for Retrieval and Ranking

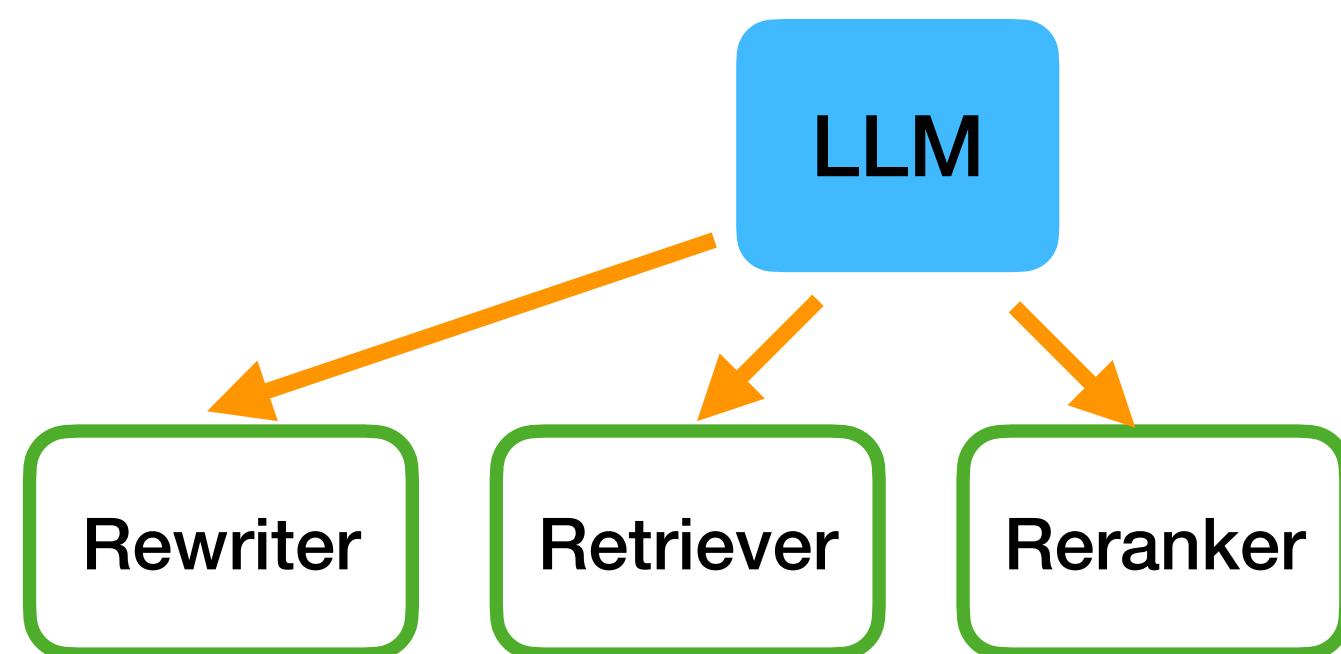


LLM-enhanced information retrieval

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

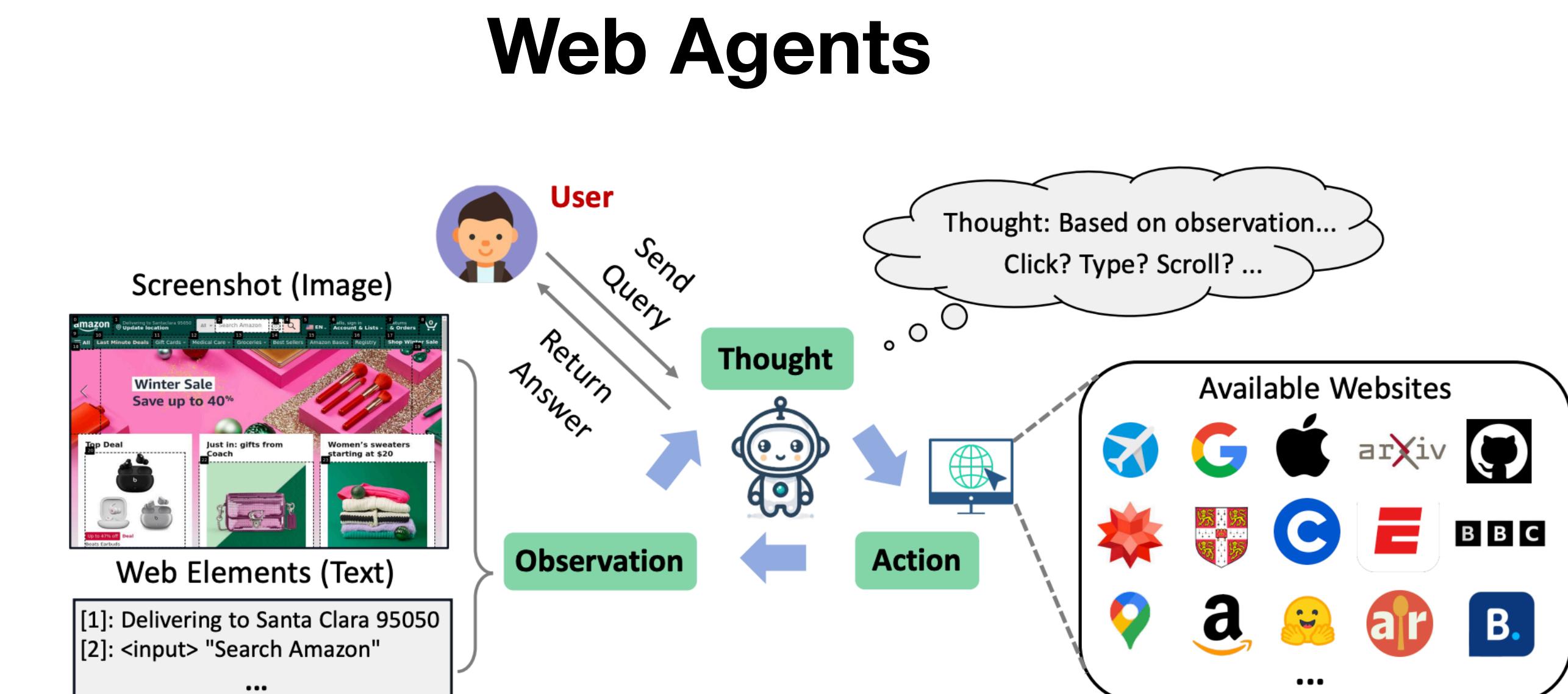
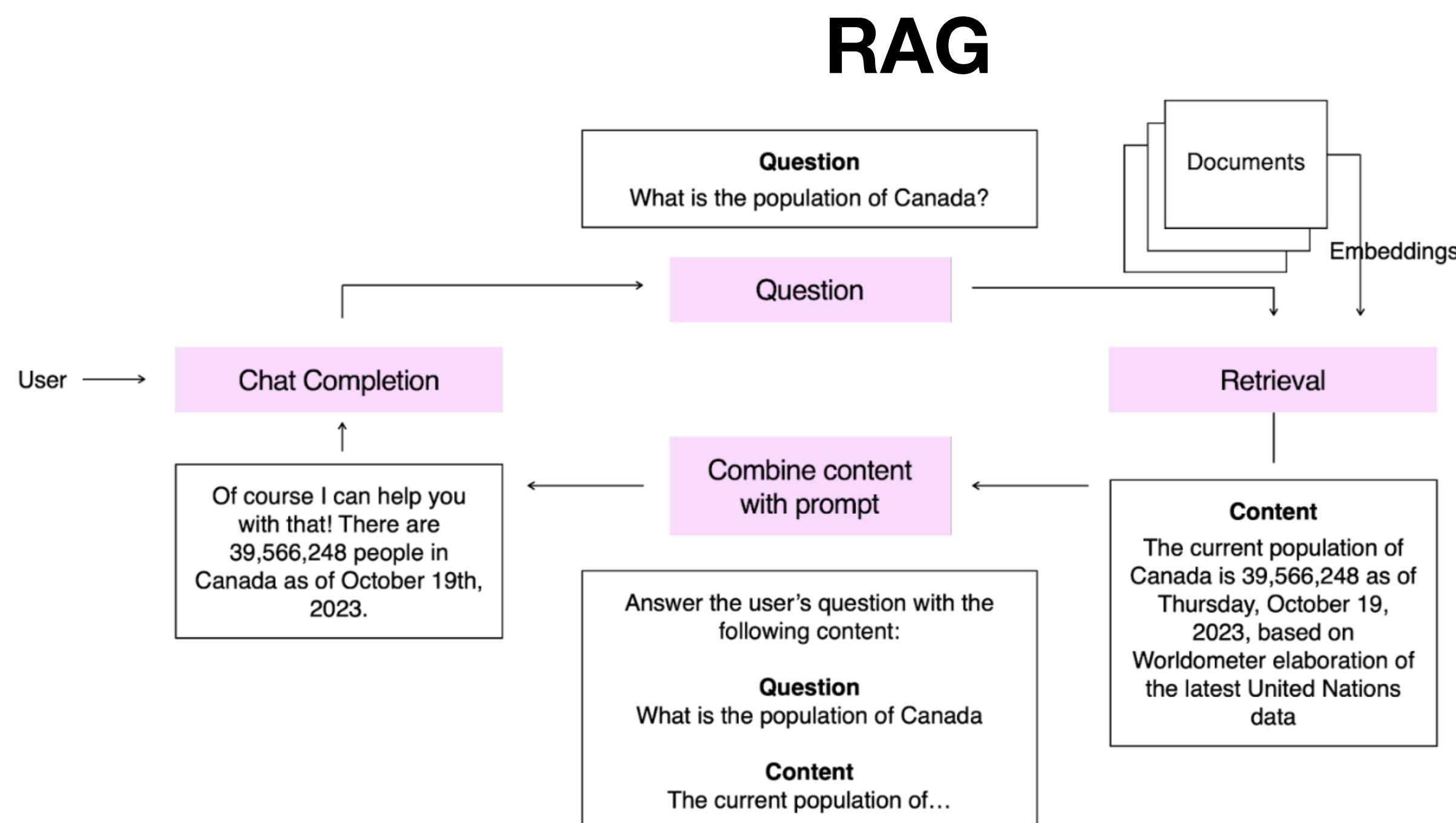
query reformulation

training data augmentation

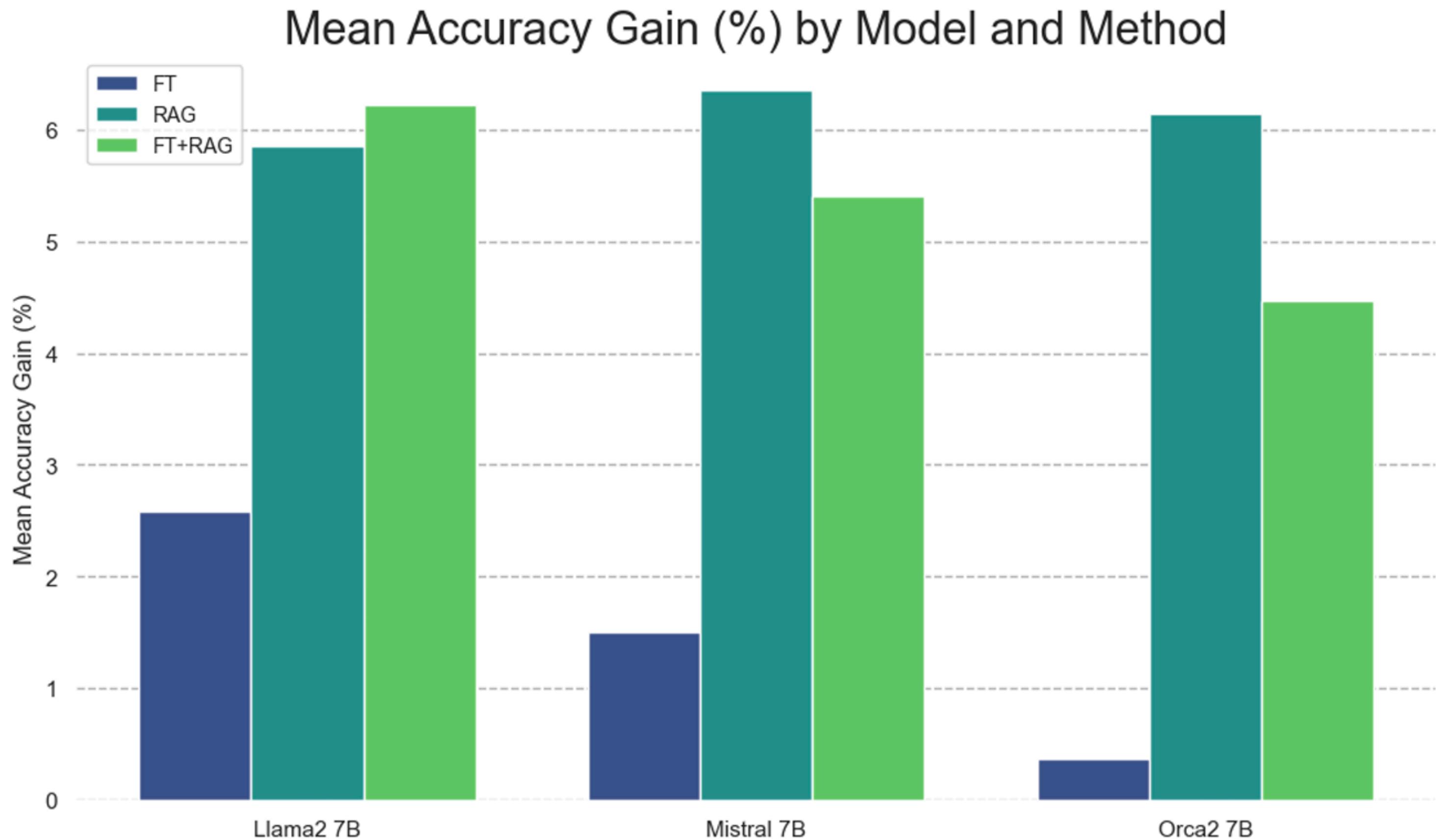
Leveraging LLMs as
Retrievers & Rankers Backbone

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



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

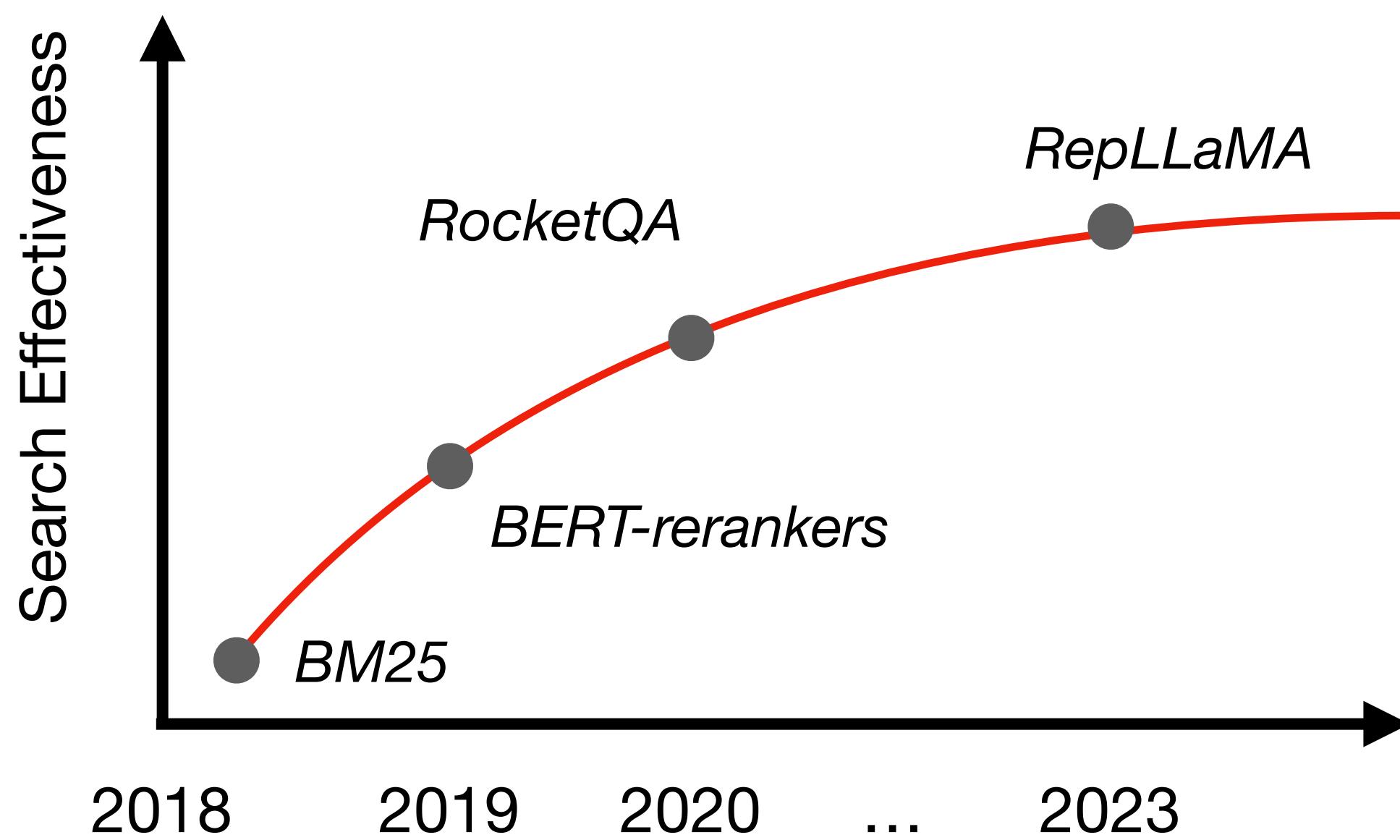


- Replacing search with LLMs is not the solution – integration is the way to go
- Ovadia et al.: in QA, unsupervised fine-tuning offers some improvement, but RAG consistently outperforms it, both for existing knowledge encountered during training and entirely new knowledge
- LLMs struggle to learn new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

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

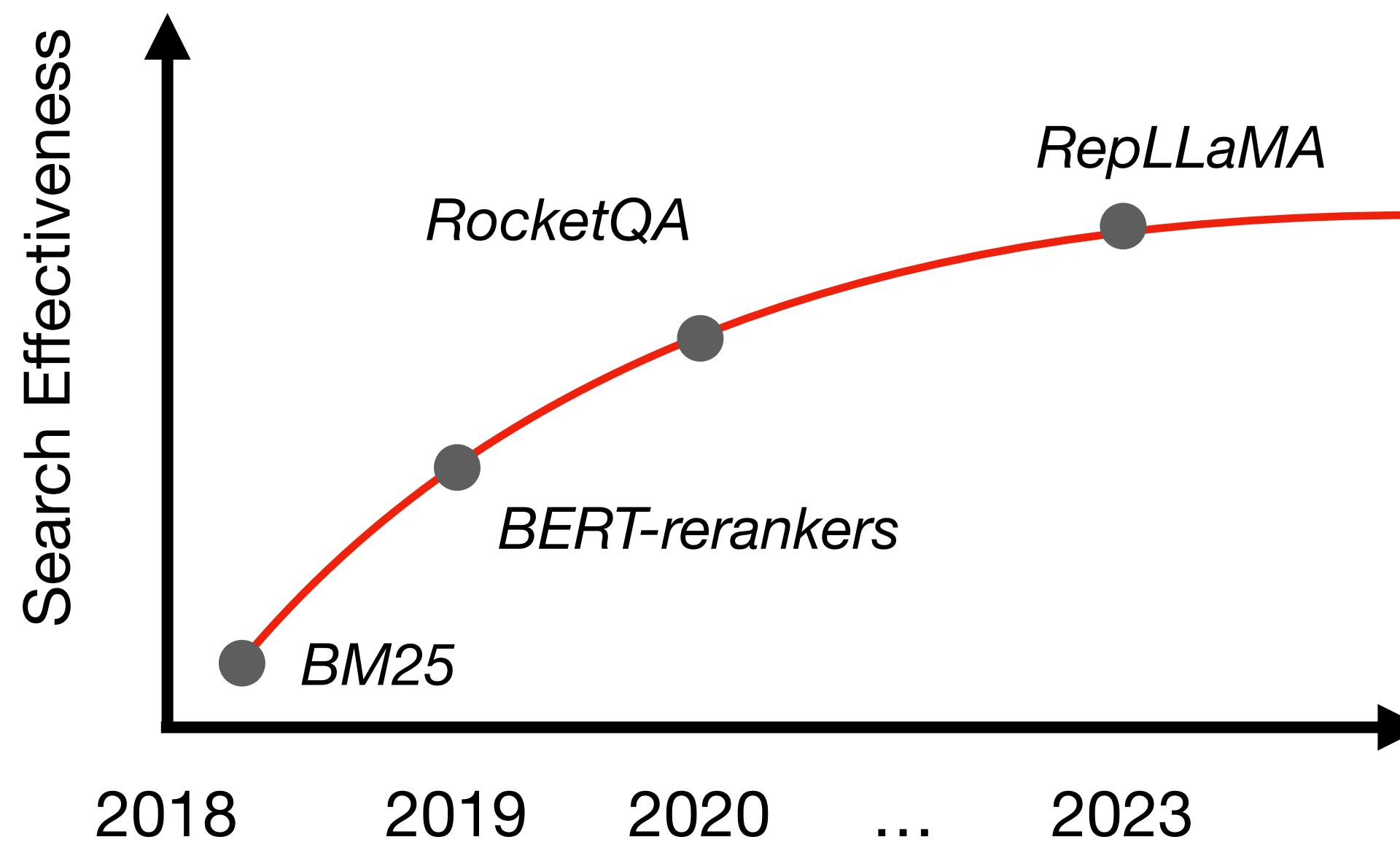
Massive progress in search quality in recent years

MS MARCO (in-domain data)

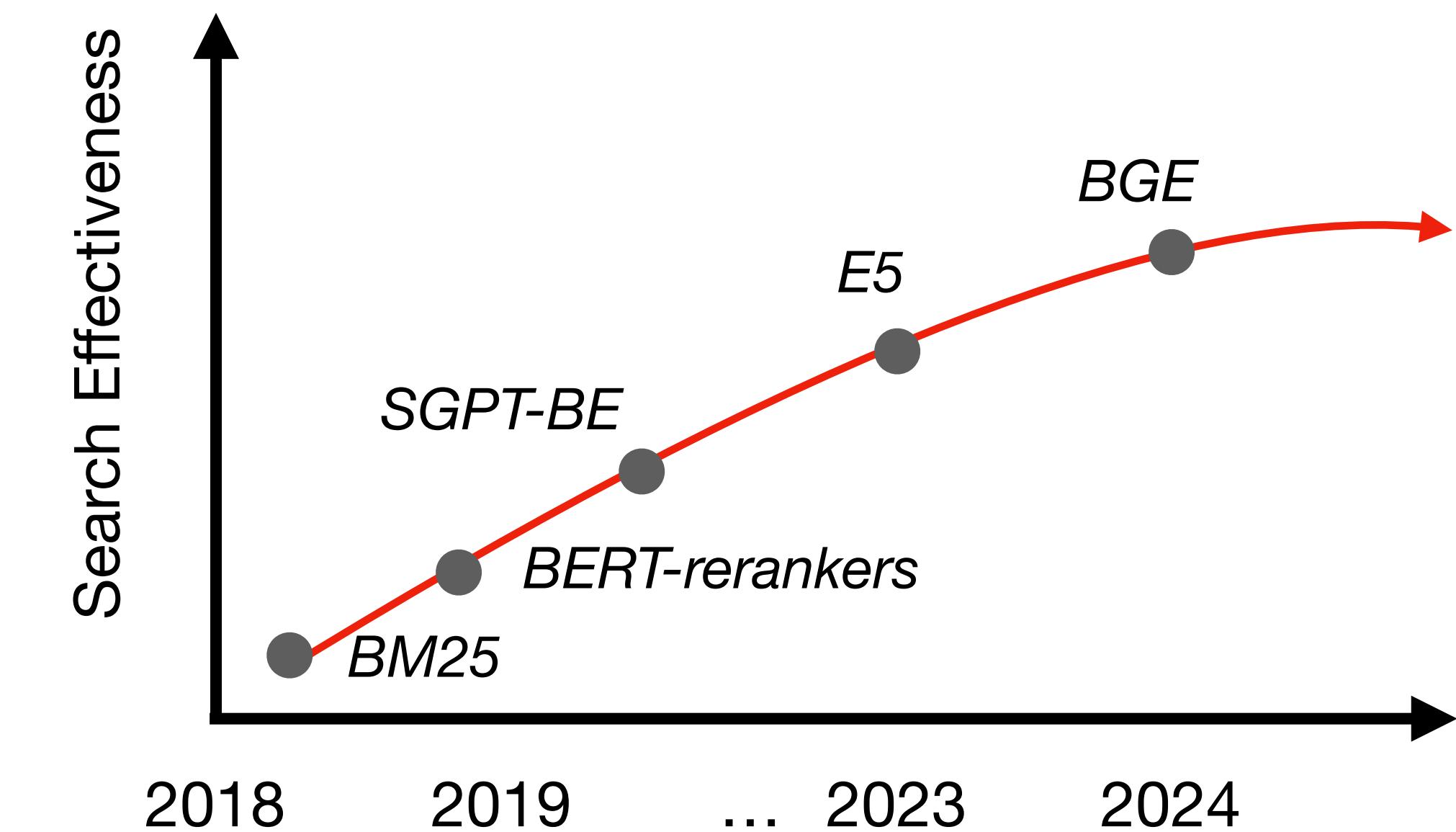


Massive progress in search quality in recent years

MS MARCO (in-domain data)



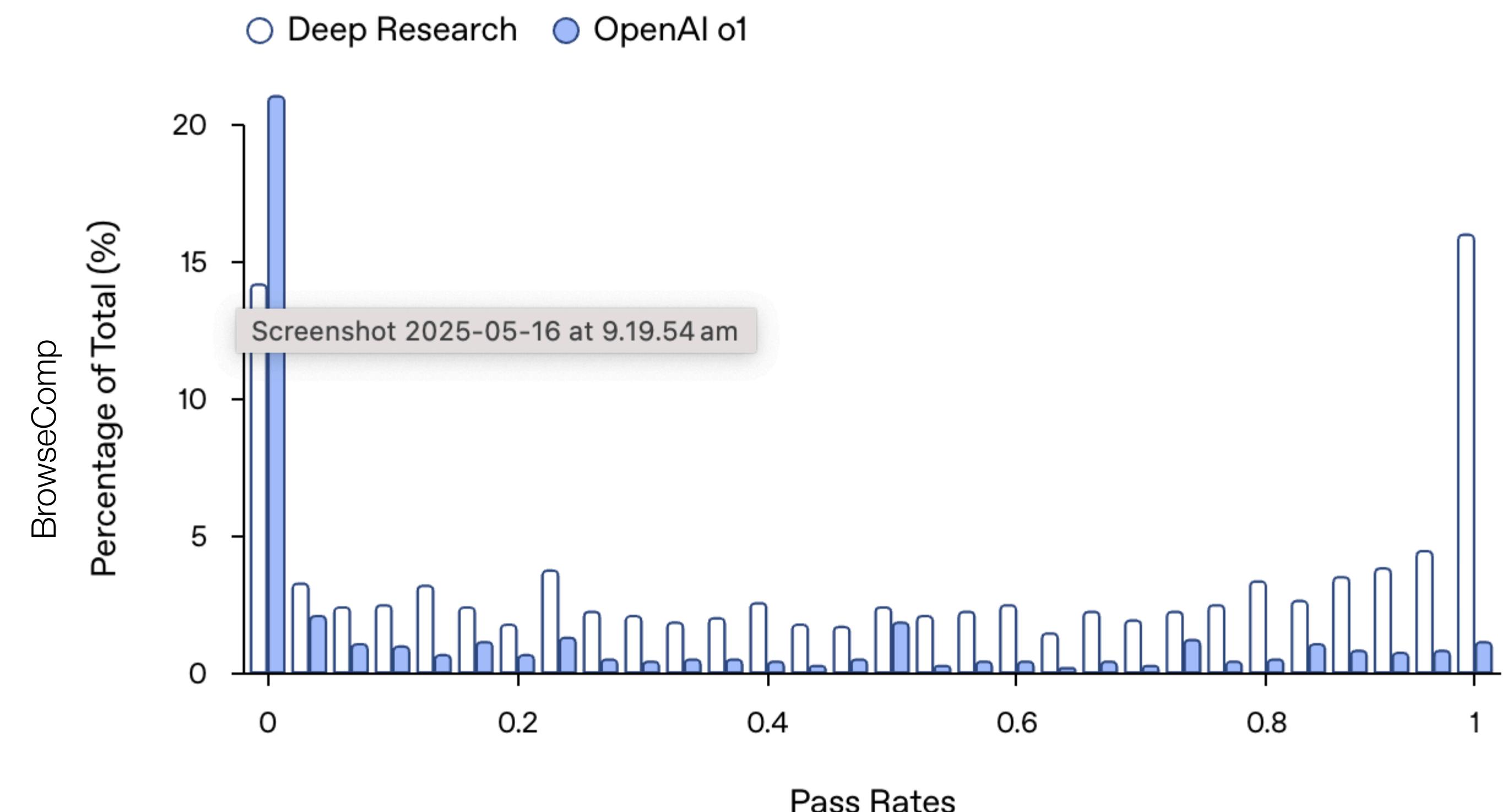
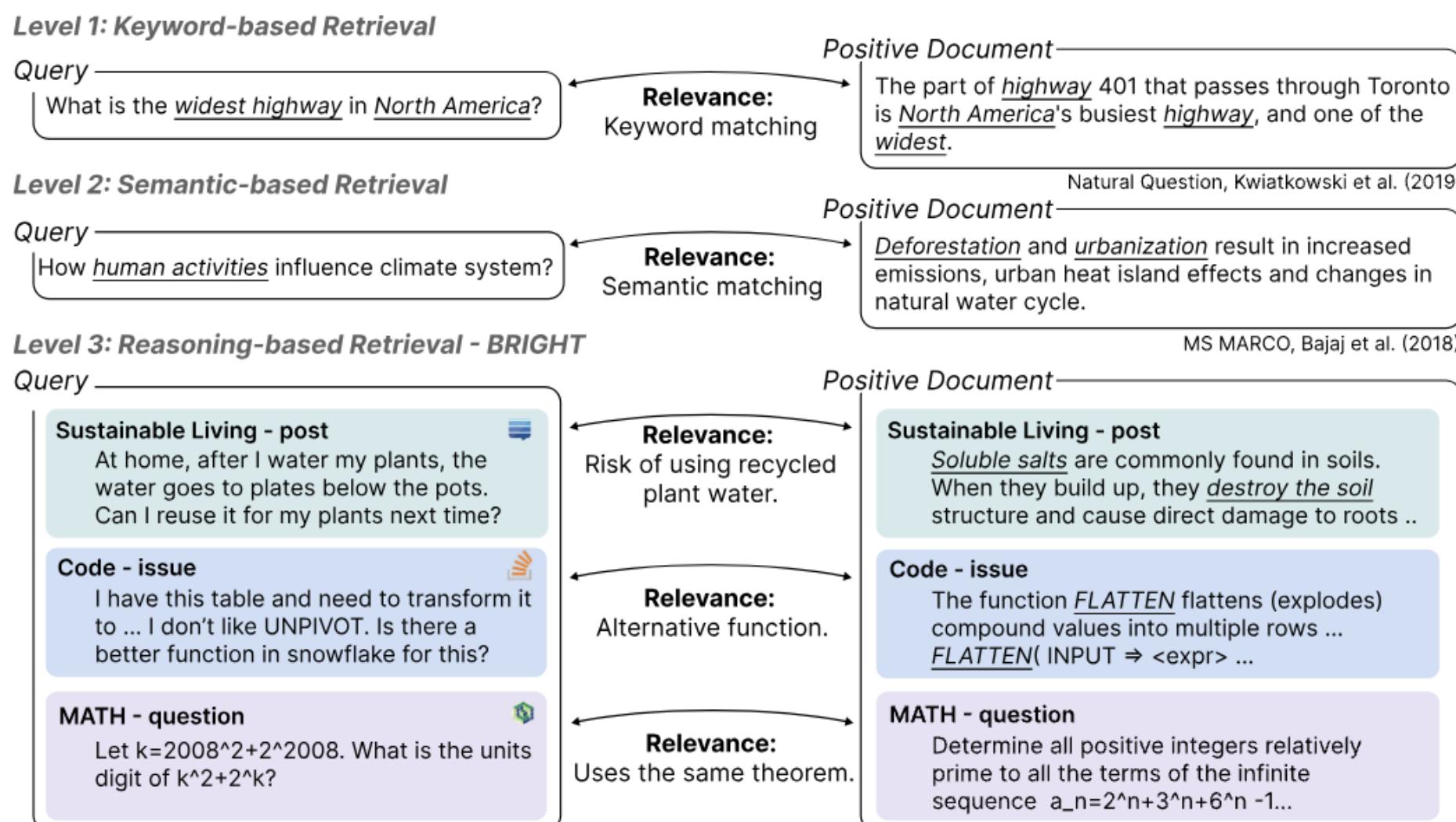
BEIR (out-domain data/~task)



However, search still challenging in many situations

Intensive Reasoning Datasets

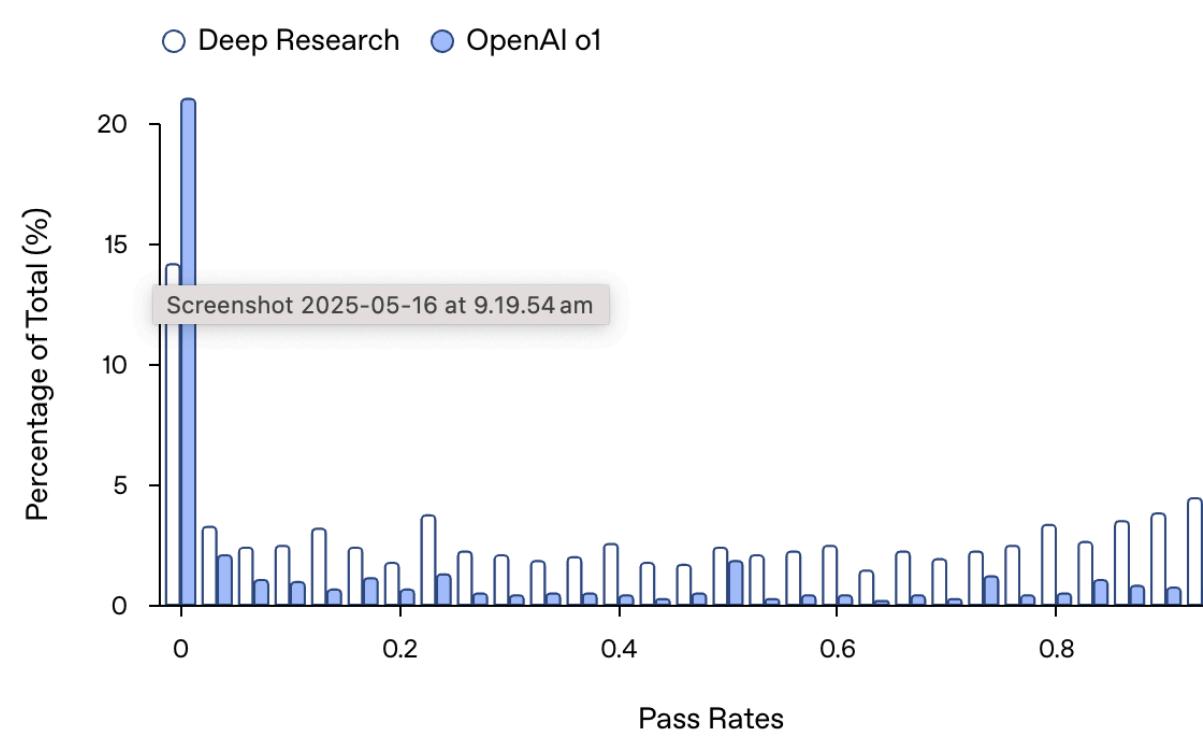
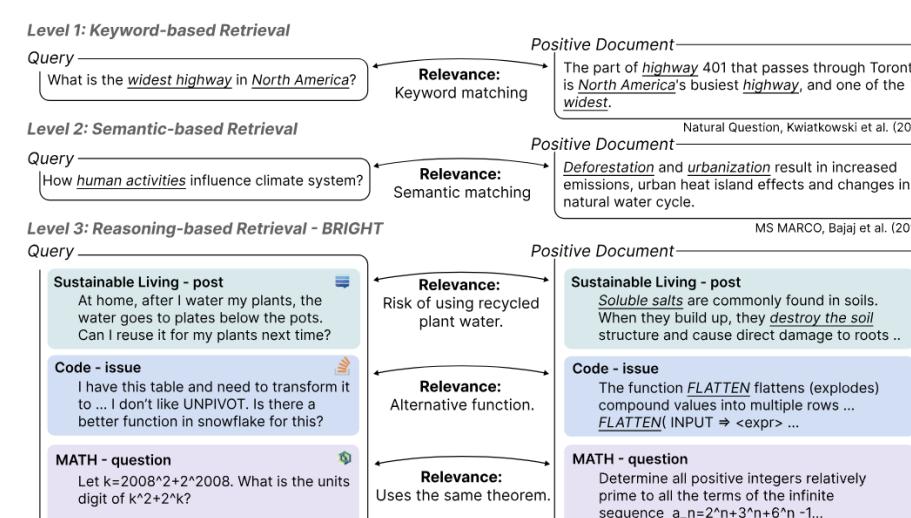
BRIGHT, BrowseComp



However, search still challenging in many situations

Intensive Reasoning Datasets

BRIGHT, BrowseComp



Multi/Cross Lingual

ECLeKTic, TREC NeuCLIR

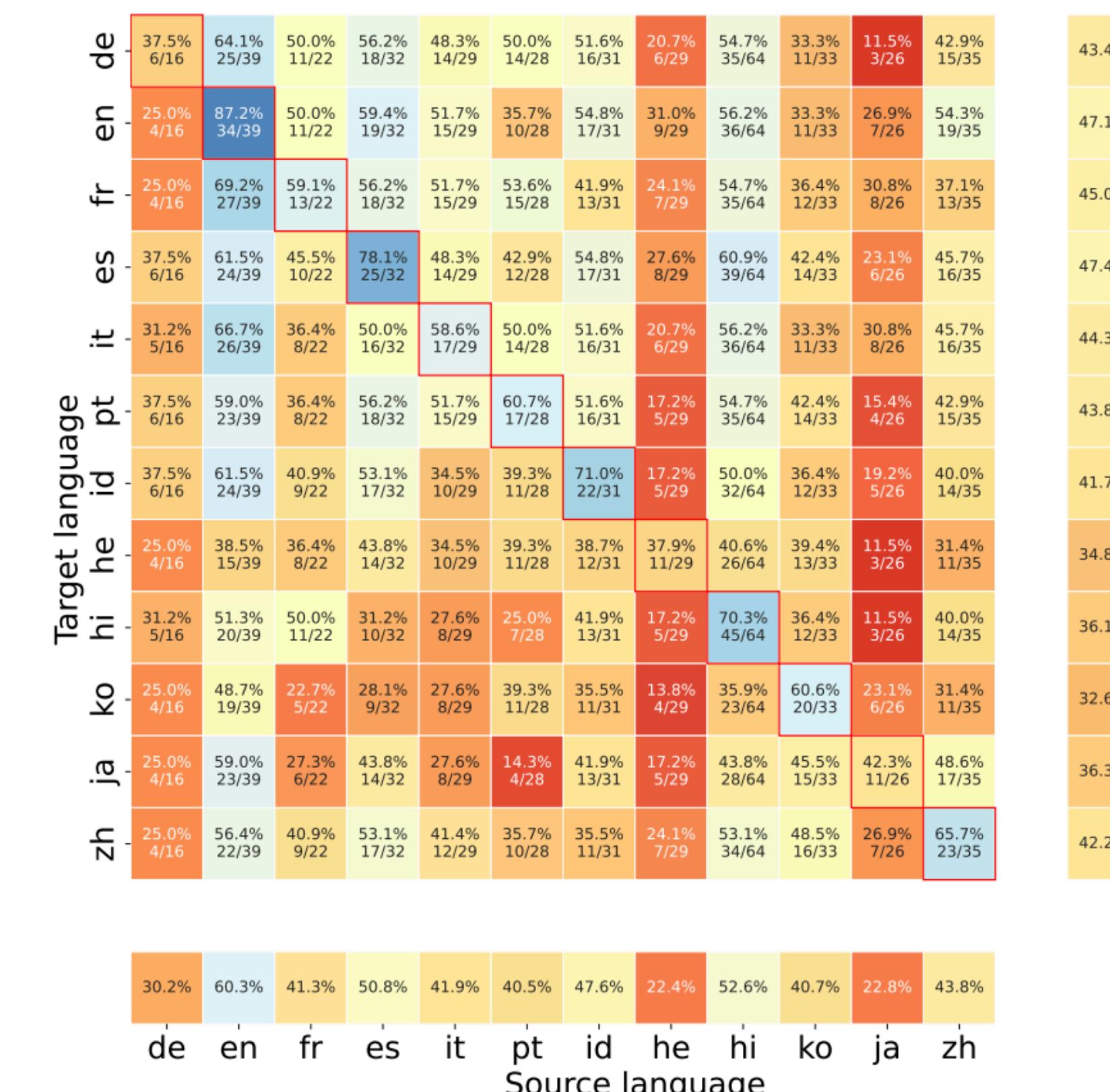
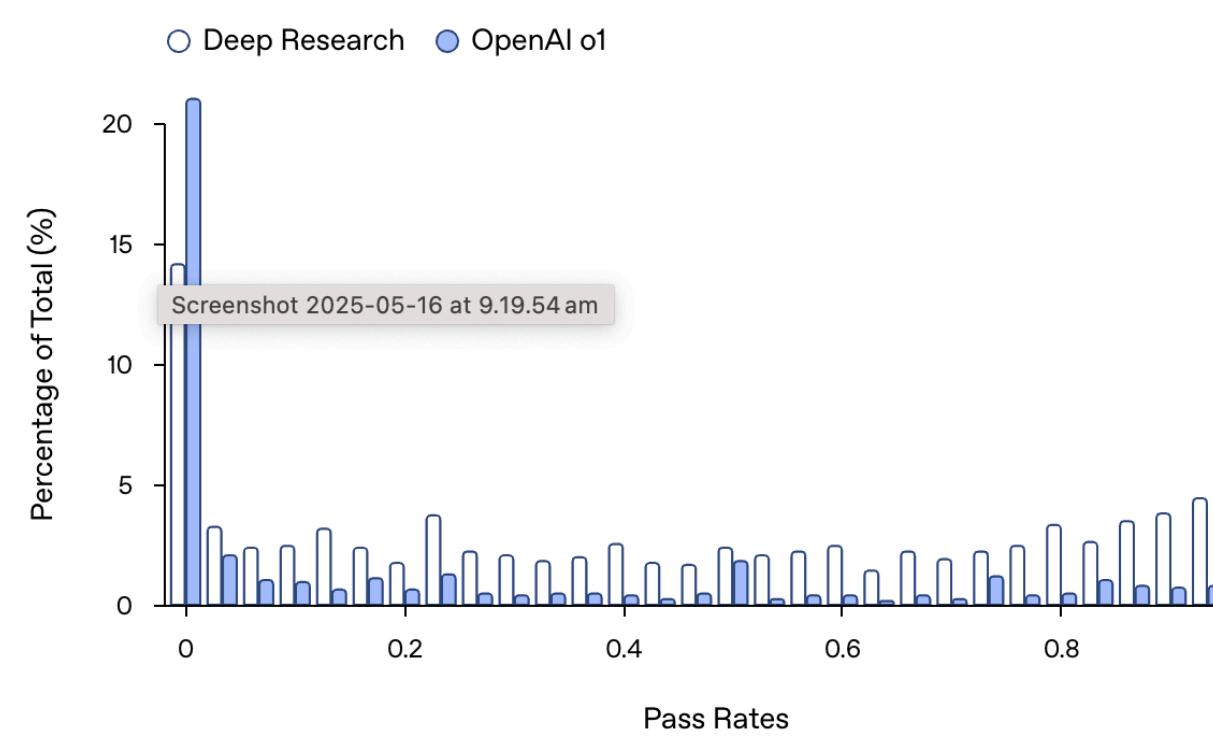
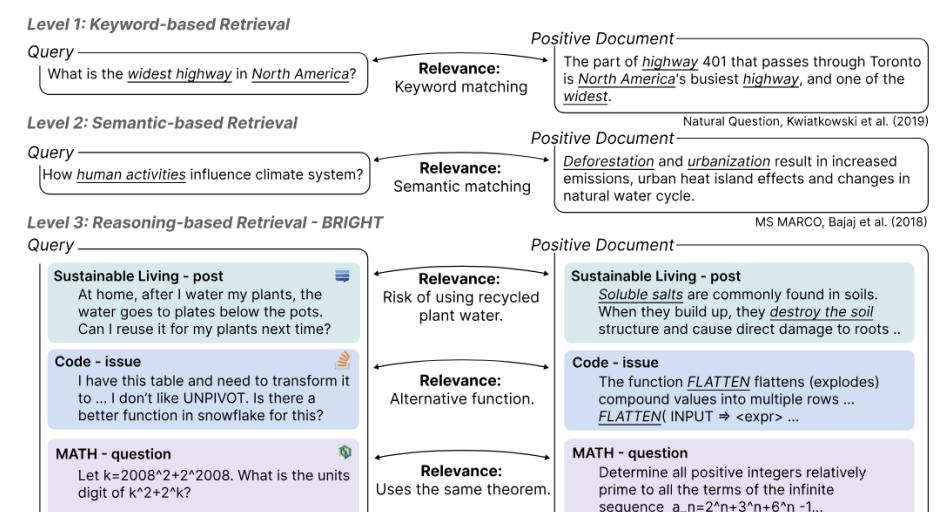


Image Source: O. Goldman, et al.
"Eclectic: a novel challenge set for evaluation of cross-lingual knowledge transfer.", arXiv:2502.21228. 2025.

However, search still challenging in many situations

Intensive Reasoning Datasets

BRIGHT, BrowseComp



Multi/Cross Lingual

ECLeKTic, TREC NeuCLIR

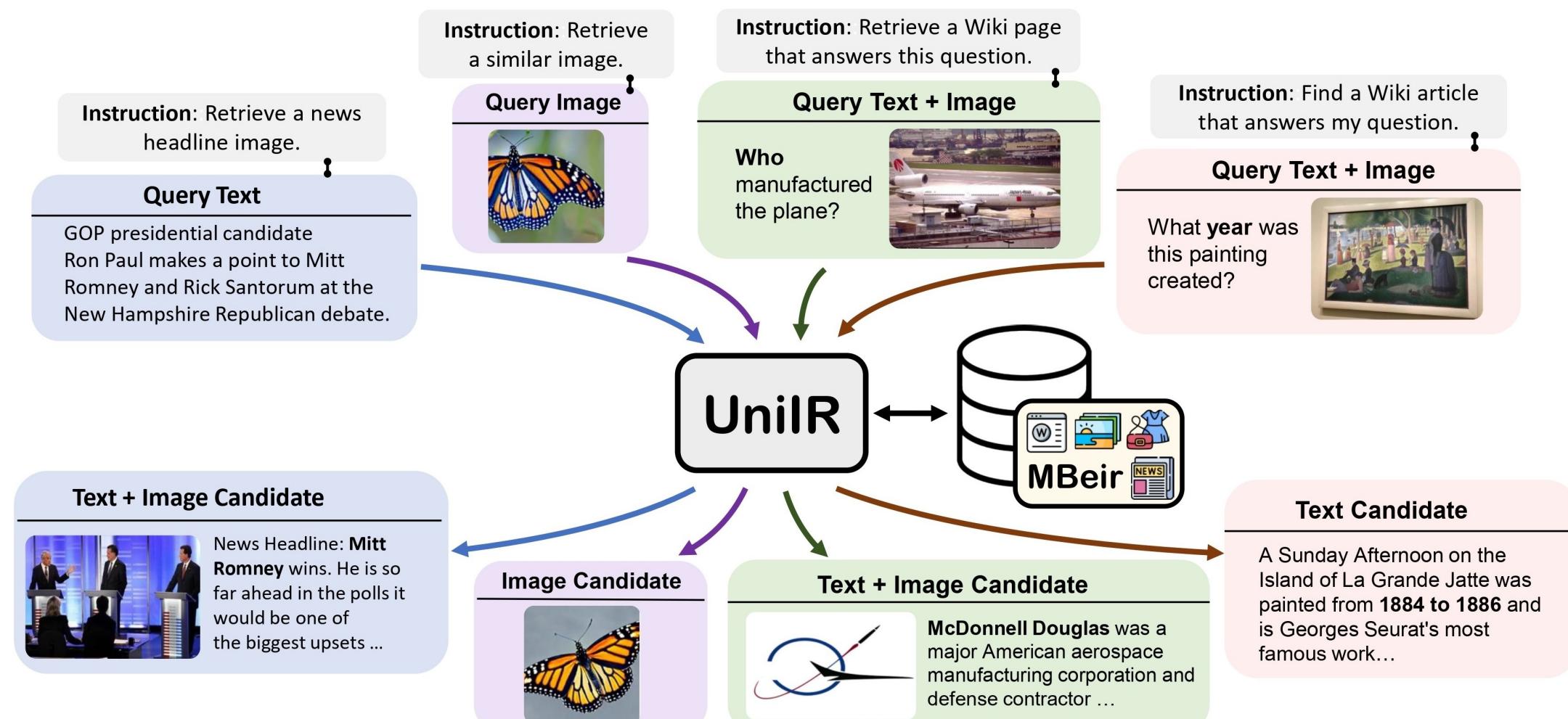
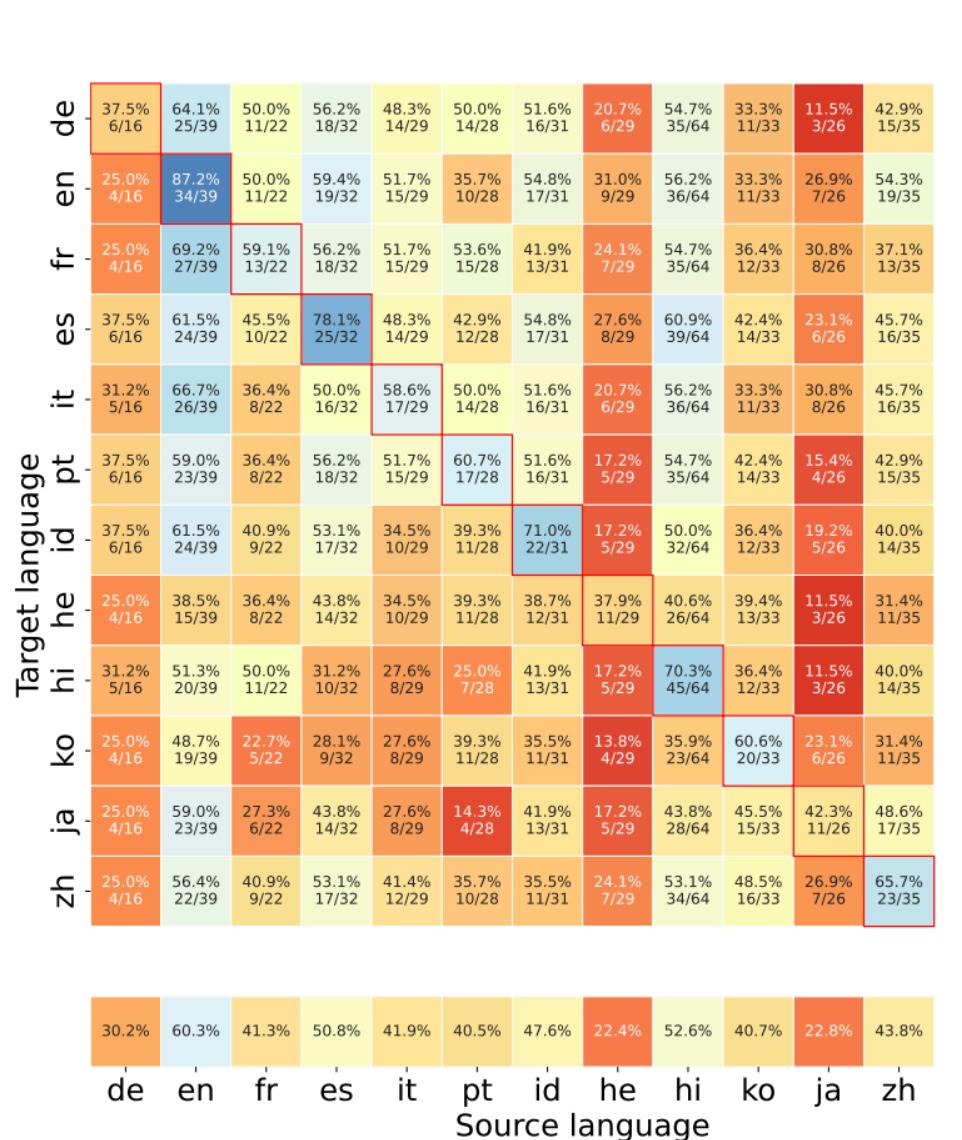
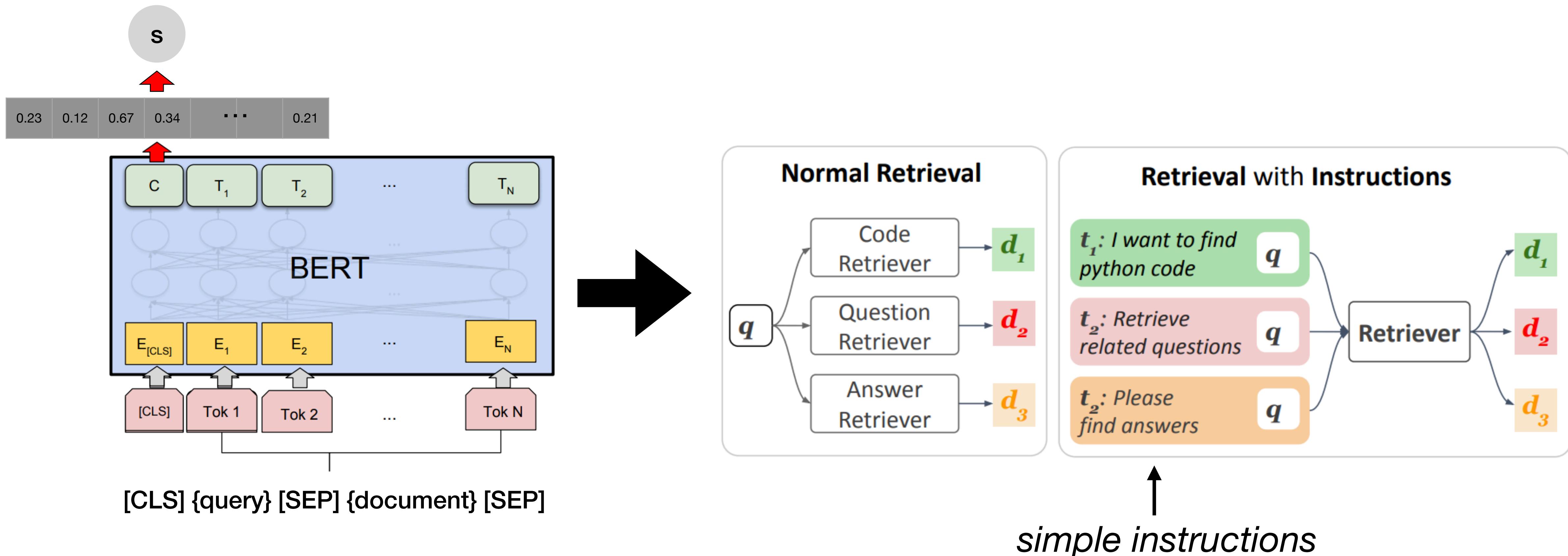


Image Source: C. Wei, et al. "Uniir: Training and benchmarking universal multimodal information retrievers.", ECCV 2024

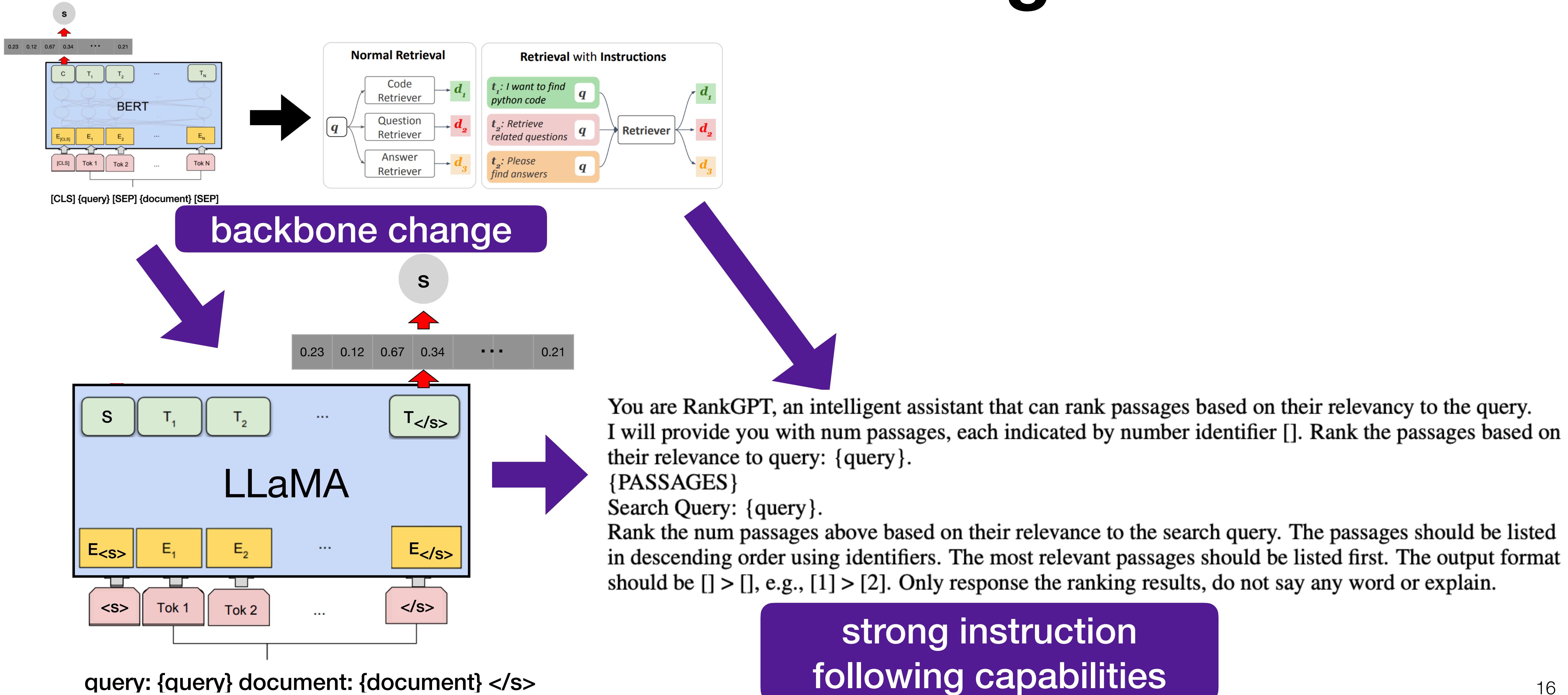
How did we get here?

From PreTrained LMs to Instruction Following LLMs

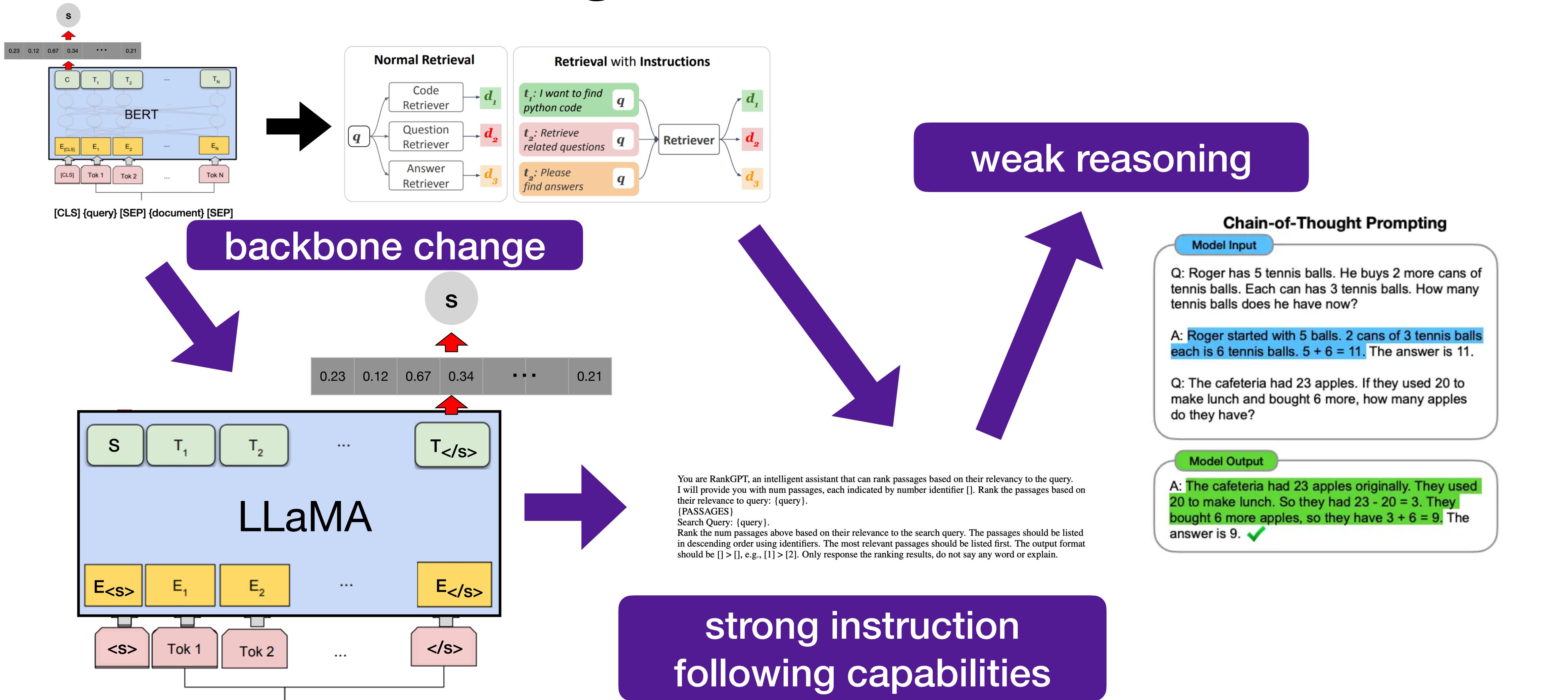


A. Asai, et al. "Task-aware Retrieval with Instructions." *ACL Findings* (2023).

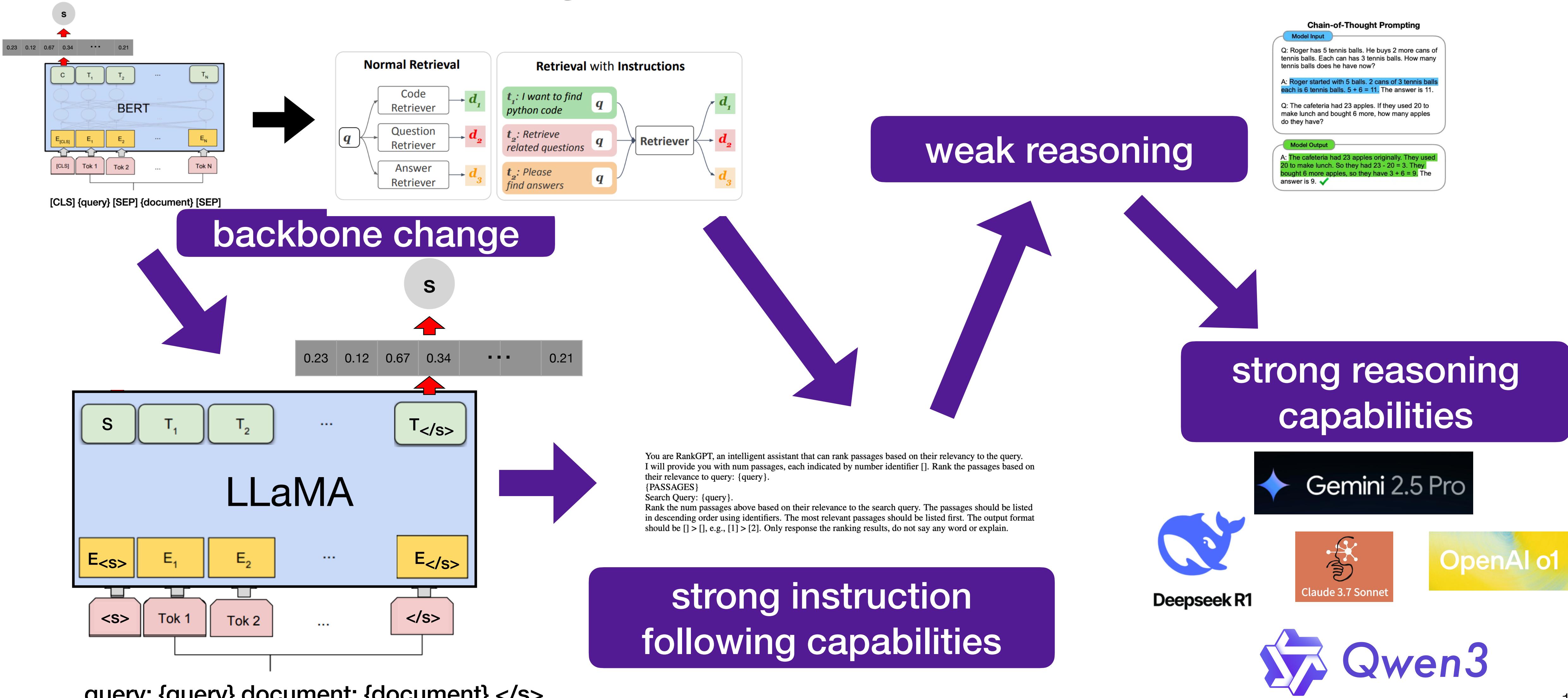
From PreTrained LMs to Instruction Following LLMs



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

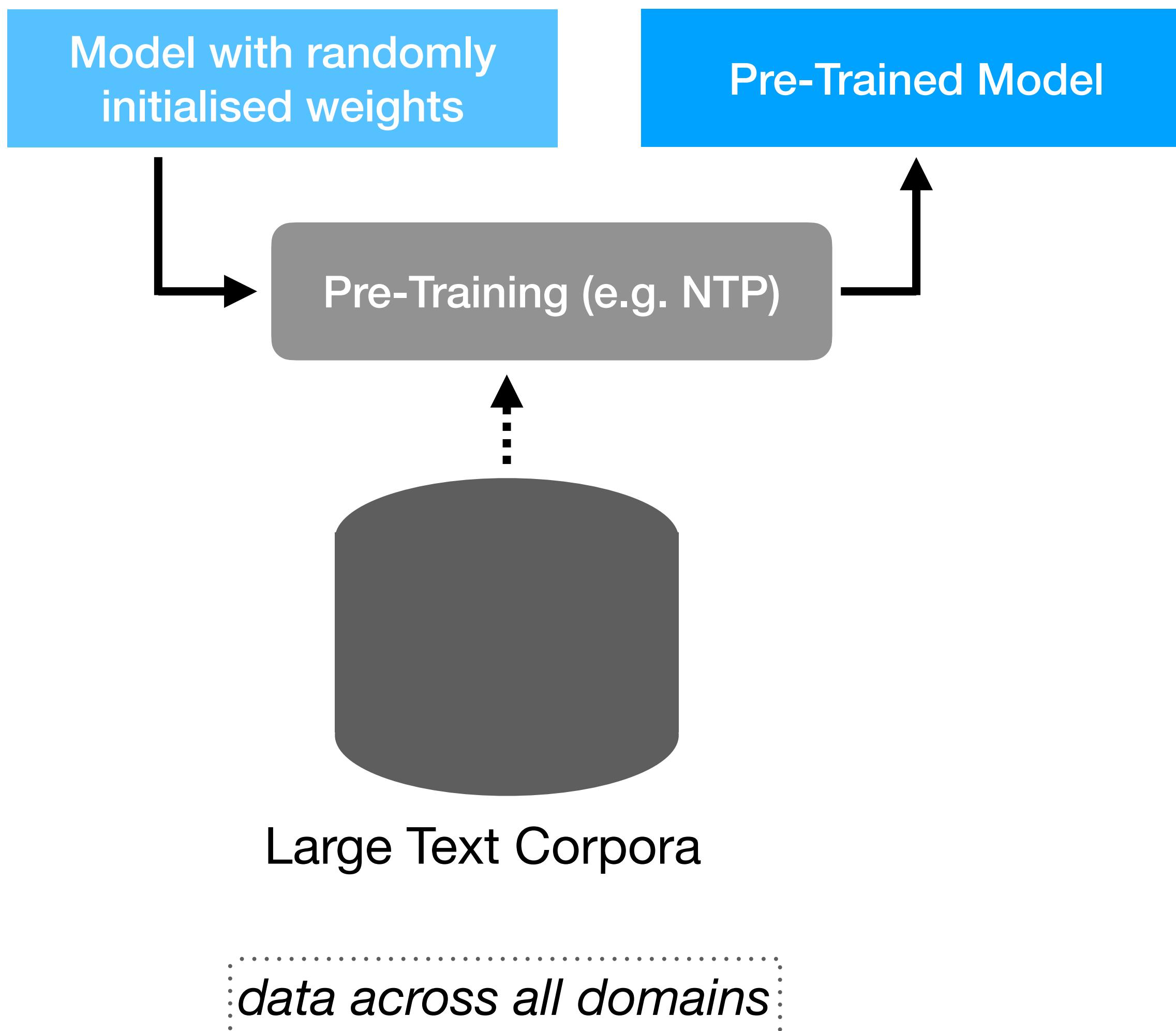


From PreTrained LMs to Instruction Following LLMs... to reasoners

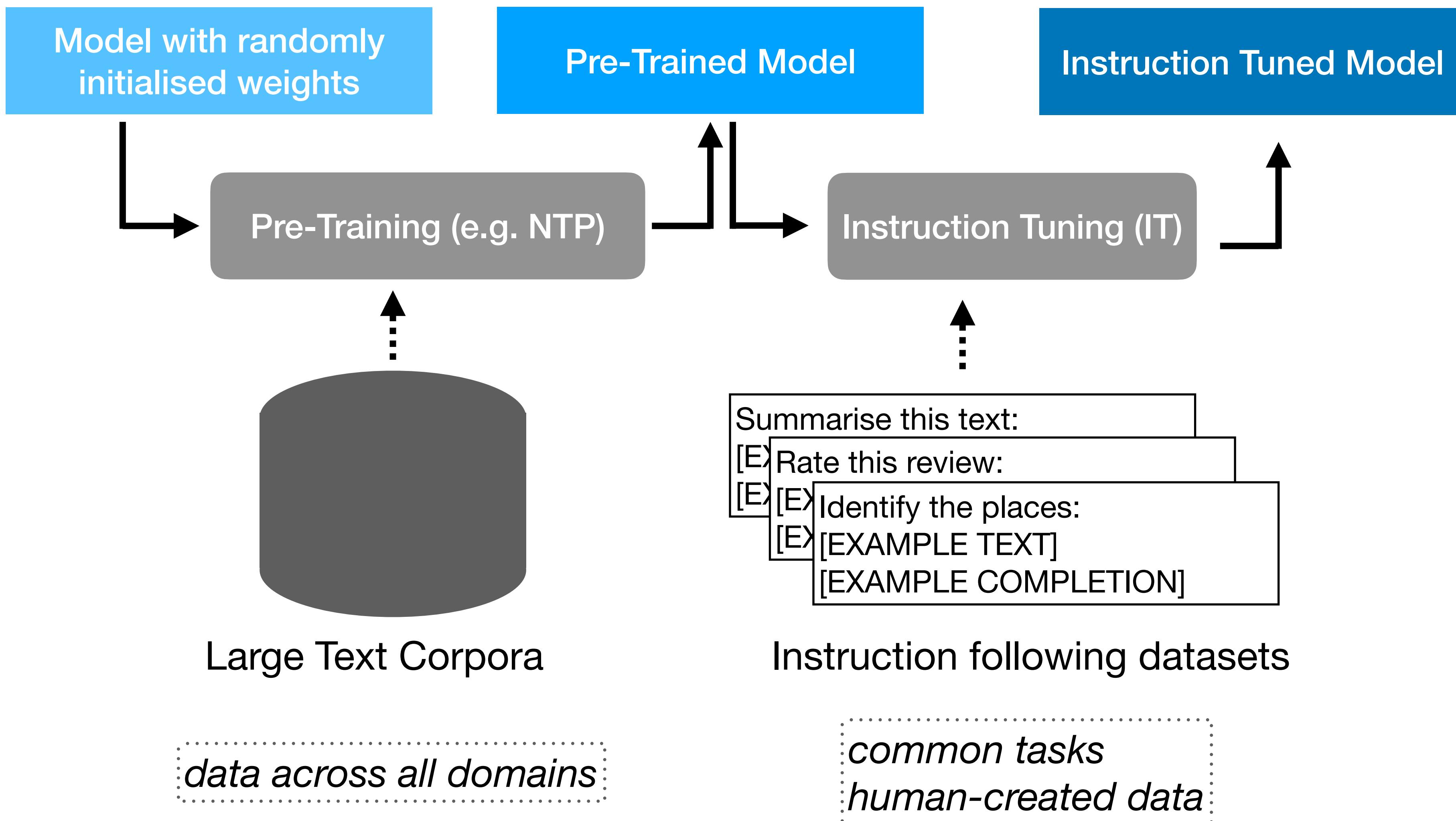


LLM training 101

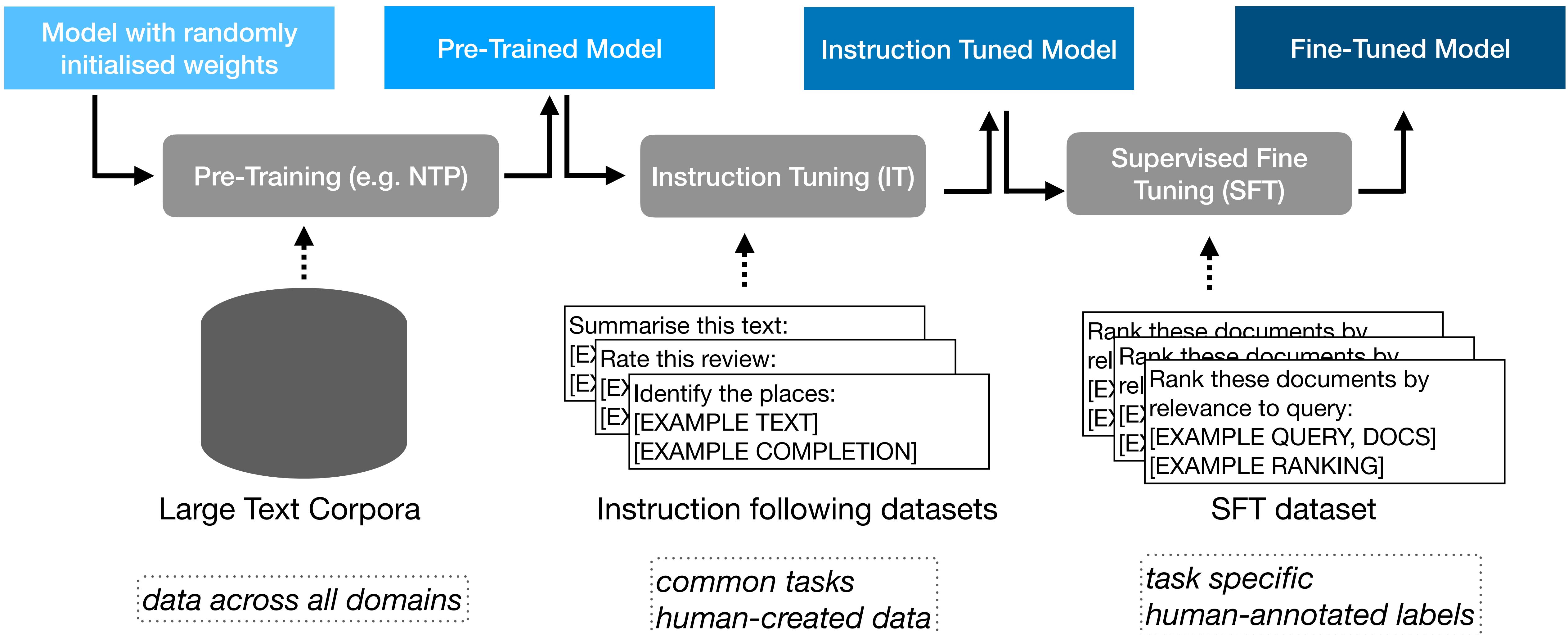
Pre-Training



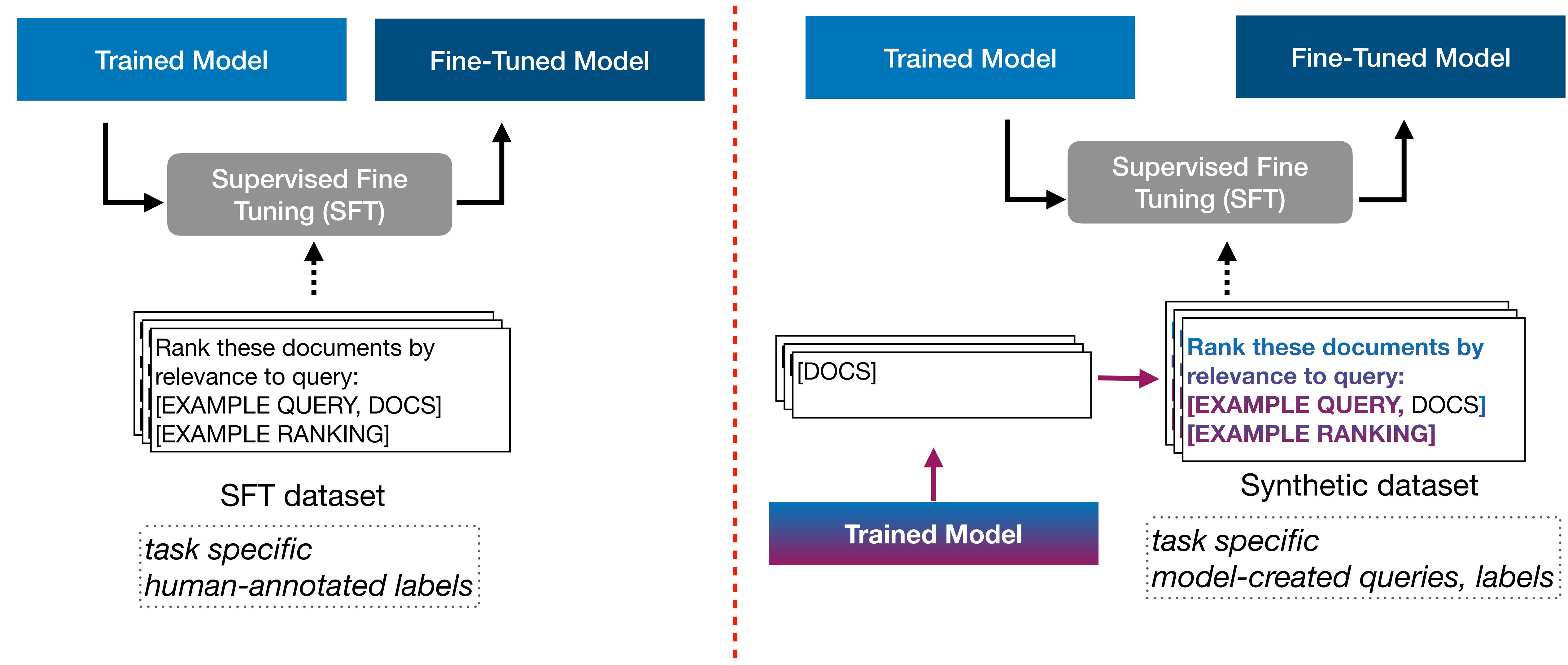
Instruction Tuning



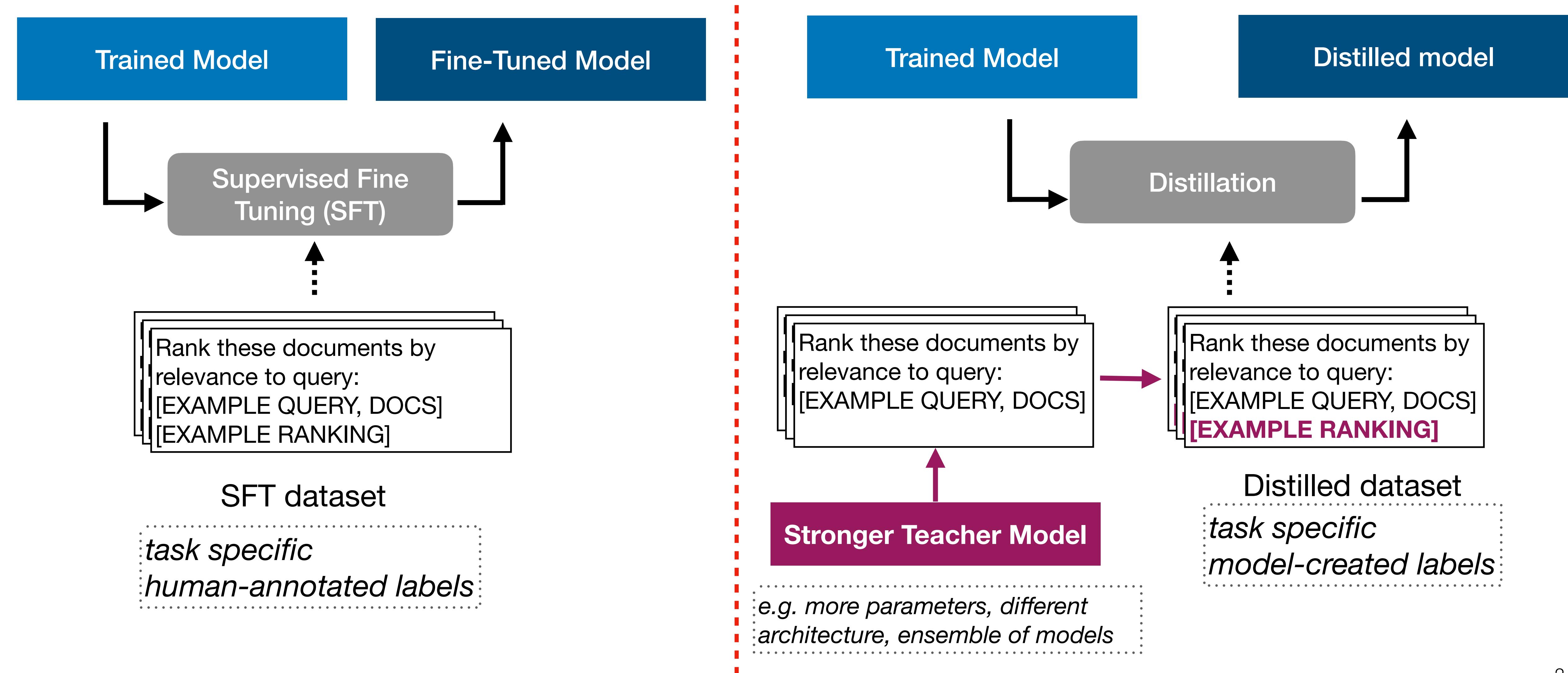
Supervised Fine-Tuning (SFT)



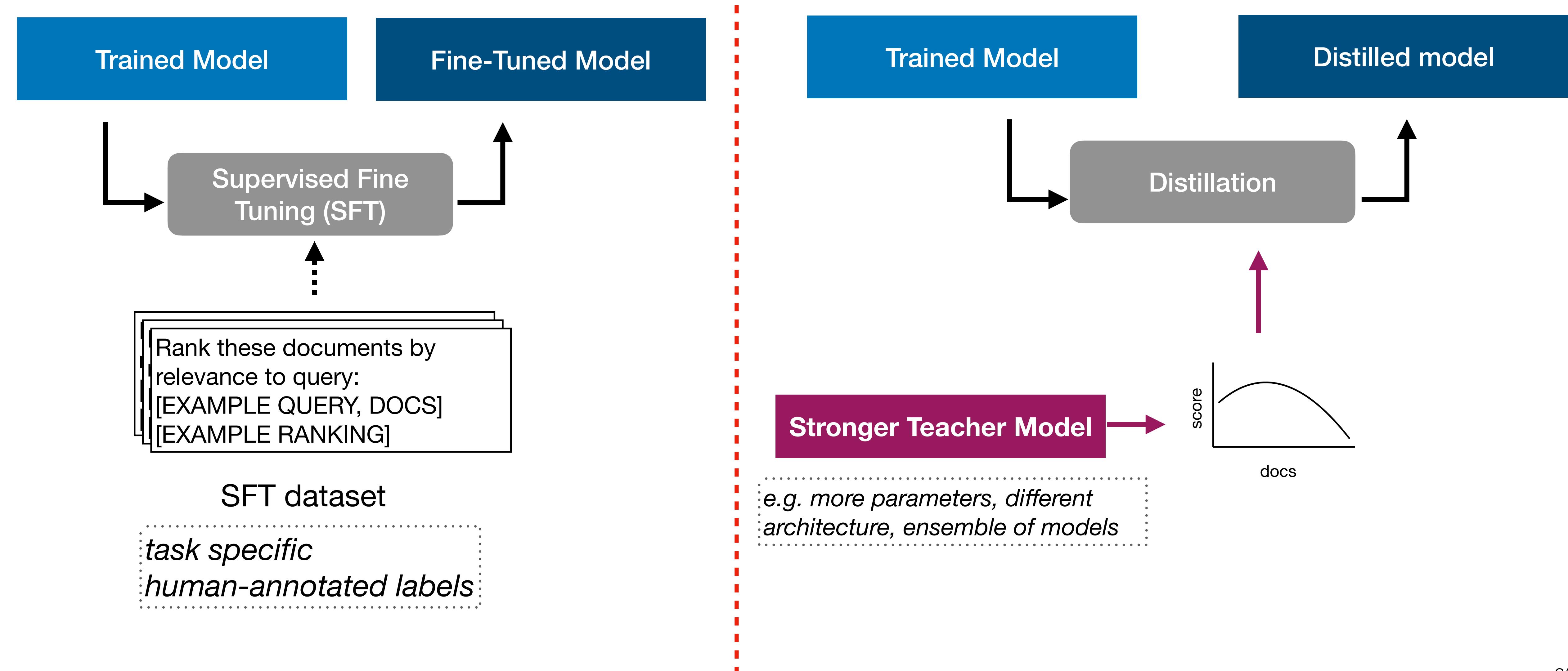
Synthetic Training Data Generation



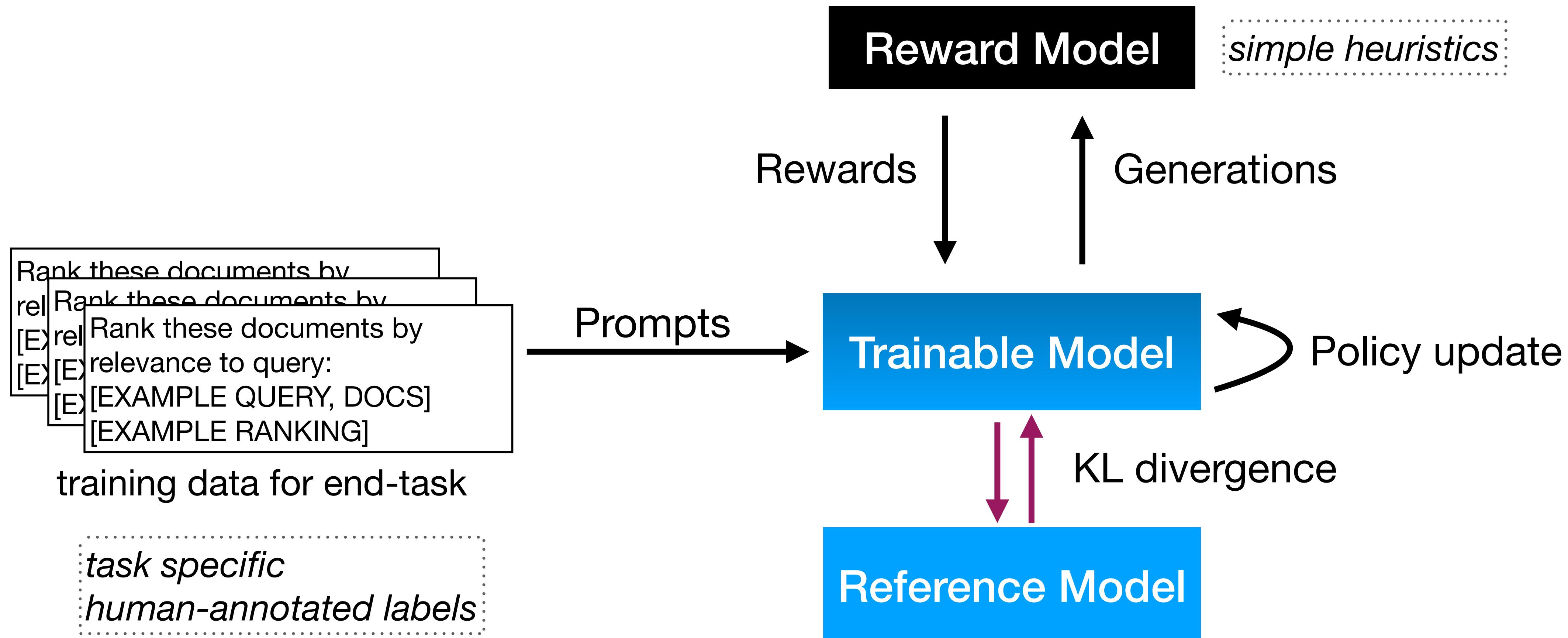
Distillation (of labels)



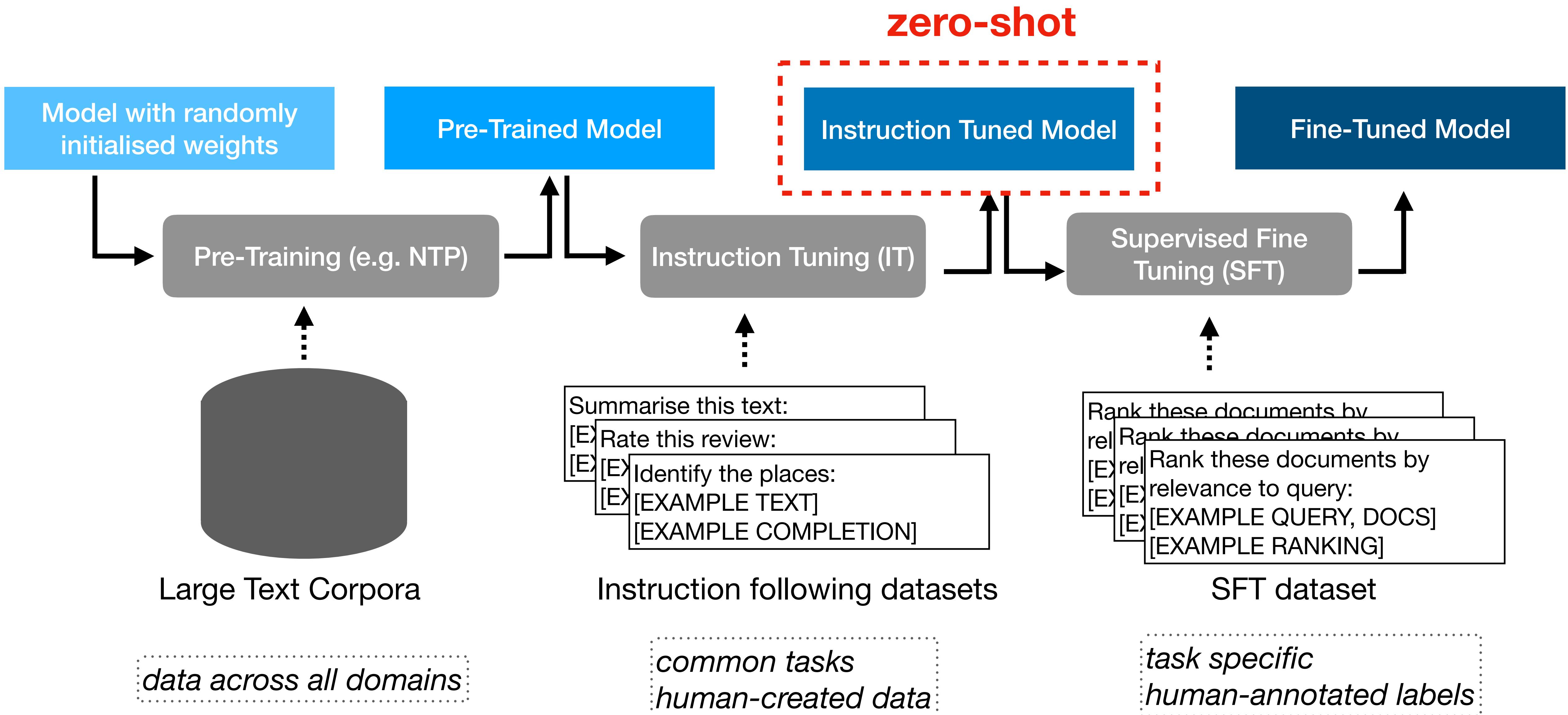
Distillation (of scores)



Reinforcement Learning (with GRPO)

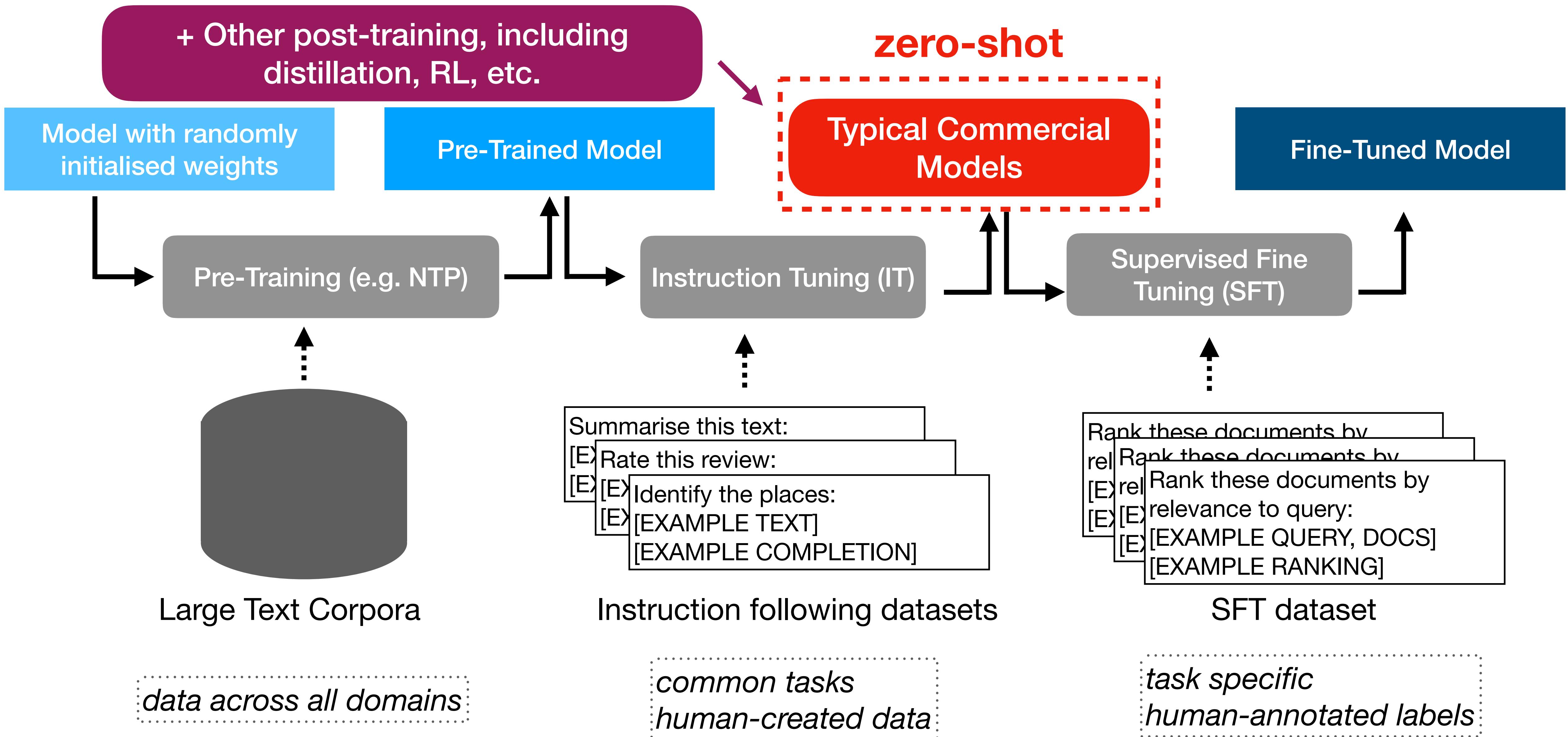


At Inference: (ideal) Zero-shot vs. Fine-Tuned



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

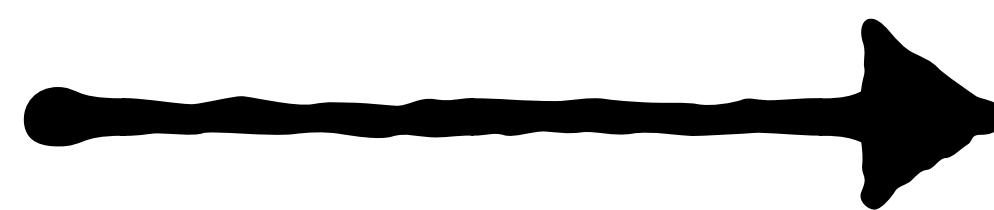
+ Other post-training, including distillation, RL, etc.



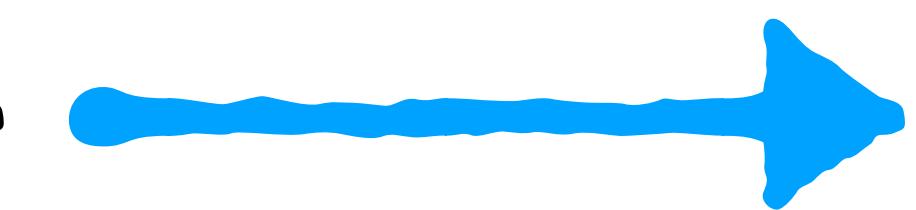
What's in the next 2 hours

Tutorial Parts

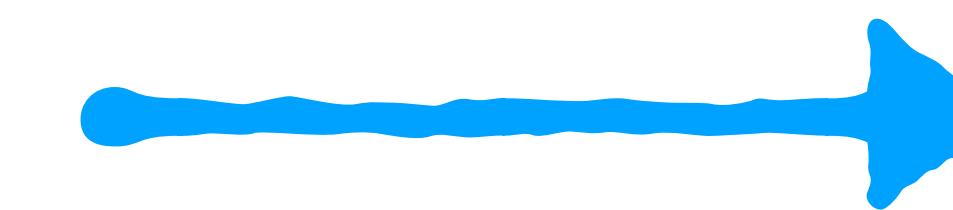
Part 1:
Intro



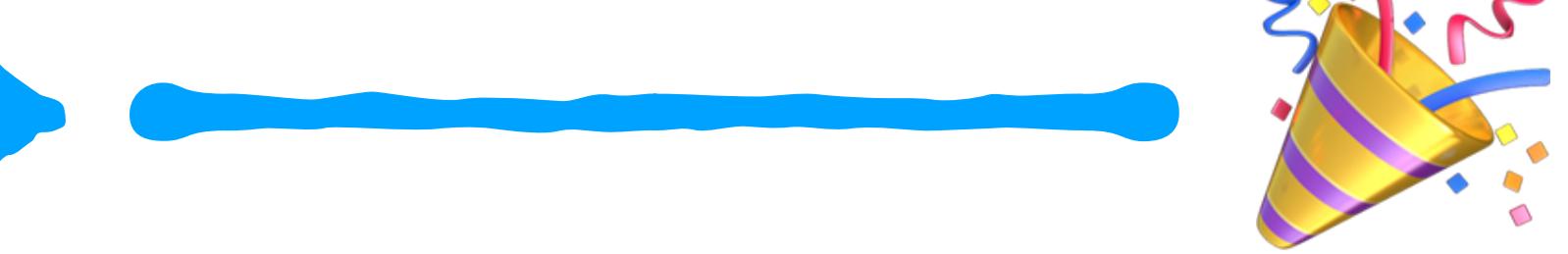
Part 2:
Retrieval



Part 3:
Ranking



Part 4:
Challenges/
Opportunities

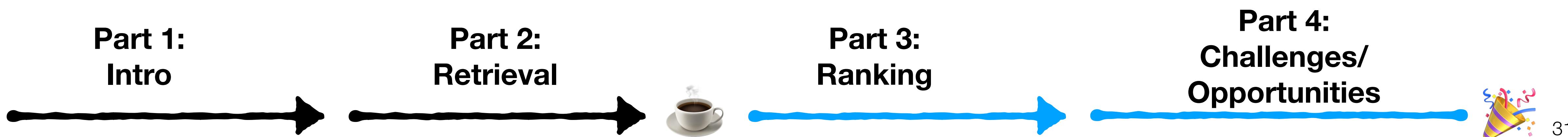


Part 2: Retrieval

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

...