Adaptation of LLMs

https://adapt-llm.github.io/



Zixuan Ke Yifei Ming Shafiq Joty

Acknowledgement: Austin Xu

Minimal LLM Basics

Prerequisites

Training ML Models

- **Learning algorithms related:**
  - SGD, Learning rate, AdamW, Batch size
- **Model architecture related:**
  - Cross and Self Attentions
  - Encoder-Decoder
  - Transformers

Basic LLM concepts

- Transformer decoder
- Next token prediction
- Tokenization, sequence/context length
- In-context learning:
  - Zero- and few-shot learning

This Tutorial

Goals

Build Foundational understanding for LLM Adaptation

- Evaluation methods
- Key concepts of LLM adaptation
- Key techniques for LLM adaptation
  - Data perspective
  - Model perspective
- Key trends

Table of contents

Introduction and Motivation ~ 40min

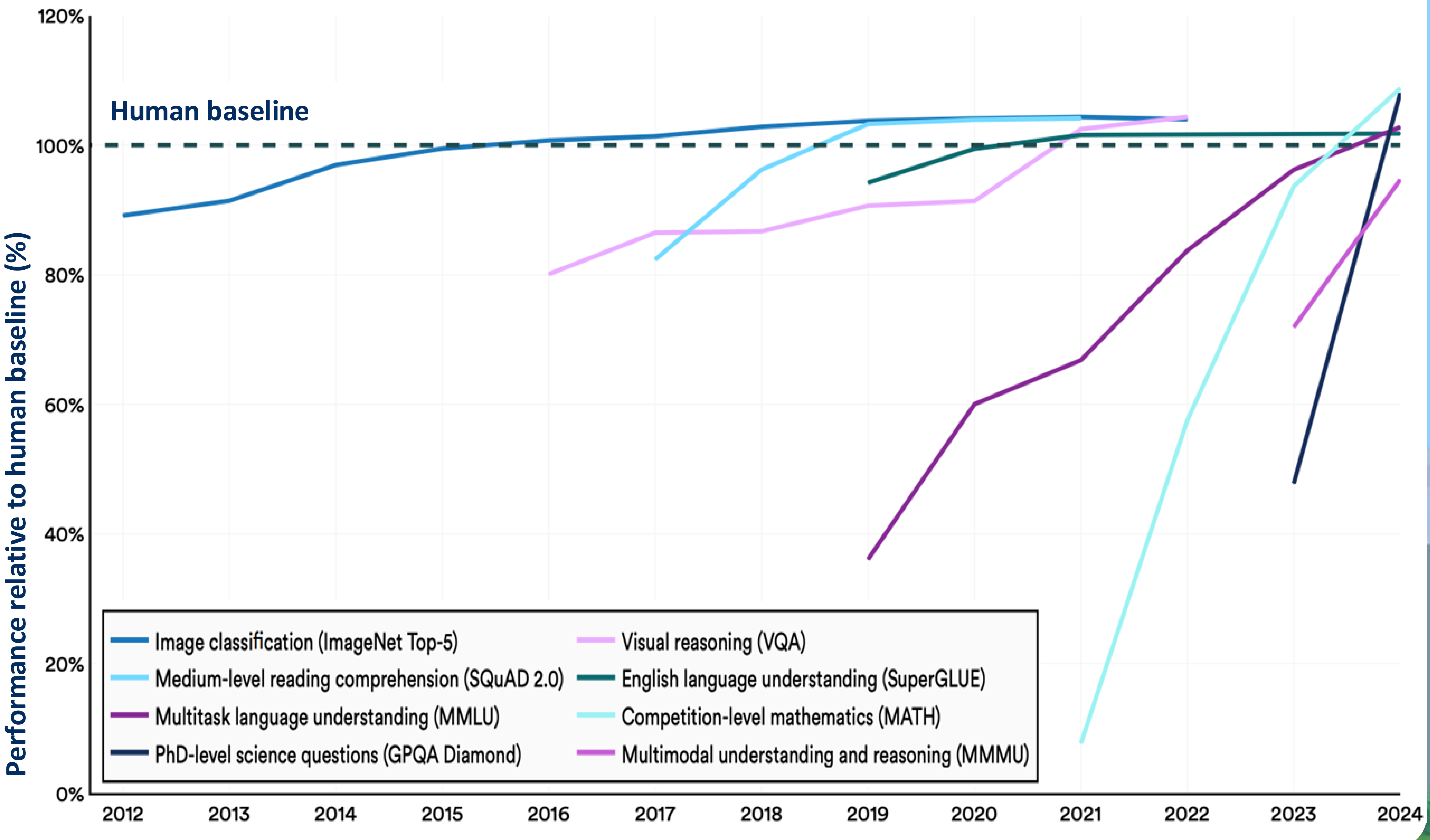
Evaluation and Benchmark ~20min

Parametric Knowledge Adaptation ~ 60min

Semi-Parametric Knowledge Adaptation ~ 30min

Summary, Discussion, QAs ~ 30min

LLM vs. human performance



The 2025 AI Index Report, Standford HAI, 2025

$$\overbrace{M_{adapt}}^{\text{Adapted LLM}} = \underbrace{F_{adapt}}_{\text{Adaptation Method}}(\overbrace{M_{base}}^{\text{Base LLM}})$$

Why We *Still* Need *Adaptation*

Adaptation → Performance↑

Domain

SaulLM-54B & SaulLM-141B: Scaling Up Domain Adaptation for the Legal Domain

Pierre Colombo equall Equall Telmo Pires equall Equall Malik Boudiaf equall Equall Rui Melo equall Equall

BioMedLM: A 2.7B Parameter Language Model Trained On Biomedical Text

Elliot Bolton[1†], Abhinav Venigalla[2], Michihiro Yasunaga[1], David Hall[1], Betty Xiong[1], Tony Lee[1], Roxana Daneshjou[1], Jonathan Frankle[2],

Demystifying Domain-adaptive Post-training for Financial LLMs

Zixuan Ke, Yifei Ming, Xuan-Phi Nguyen, Caiming Xiong and Shafiq Joty
Salesforce AI Research
{zixuan.ke,yifei.ming,xnguyen,cxiong,sjoty}@salesforce.com

Project Page: https://github.com/SalesforceAIResearch/FinDAP
Datasets: https://huggingface.co/datasets/Salesforce/FinEval

Task

SFR-RAG: Towards Contextually Faithful LLMs

Foundational Autoraters: Taming Large Language Models for Better Automatic Evaluation

🔥 PROMETHEUS: INDUCING FINE-GRAINED EVALUATION CAPABILITY IN LANGUAGE MODELS

Seungone Kim[1,2*†] **Jamin Shin**[2,3*†] **Yejin Cho**[1*†] **Joel Jang**[4] **Shayne Longpre**[5] **Hwaran Lee**[2,3] **Sangdoo Yun**[2,3] **Seongjin Shin**[3] **Sungdong Kim**[1,2,3] **James Thorne**[1] **Minjoon Seo**[1†]

[1]KAIST AI [2]NAVER AI Lab [3]NAVER Cloud [4]University of Washington [5]MIT

Adaptation → Performance↑

salesforce

Domain/Language

Code Llama: Open Foundation Models for Code

Baptiste Rozière†, Jonas Gehring†, Fabian Gloeckle†,*, Sten Sootla†, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi◇, Jingyu Liu, Romain Sauvestre, Tal Remez, Jérémy Rapin, Artyom Kozhevnikov, Ivan Evtimov, Joanna Bitton, Manish Bhatt, Cristian Canton Ferrer, Aaron Grattafiori, Wenhan Xiong, Alexandre Défossez, Jade Copet, Faisal Azhar, Hugo Touvron, Louis Martin, Nicolas Usunier, Thomas Scialom, Gabriel Synnaeve†

Meta AI

CHIMED-GPT: A Chinese Medical Large Language Model with Full Training Regime and Better Alignment to Human Preferences

Yuanhe Tian♠♥*, Ruyi Gan♠♣*, Yan Song♠†, Jiaxing Zhang♣, Yongdong Zhang♠

ALLaM: Large Language Models for Arabic and English

NCAI
المركز الوطني
للذكاء الاصطناعي
National Center for AI

SDAIA
الهيئة السعودية للبيانات
والذكاء الاصطناعي
Saudi Data & AI Authority

Task

How to Train Long-Context Language Models (Effectively)

Tianyu Gao* Alexander Wettig* Howard Yen Danqi Chen
Princeton Language and Intelligence, Princeton University
{tianyug,awettig,hyen,danqic}@cs.princeton.edu

DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning

DeepSeek-AI

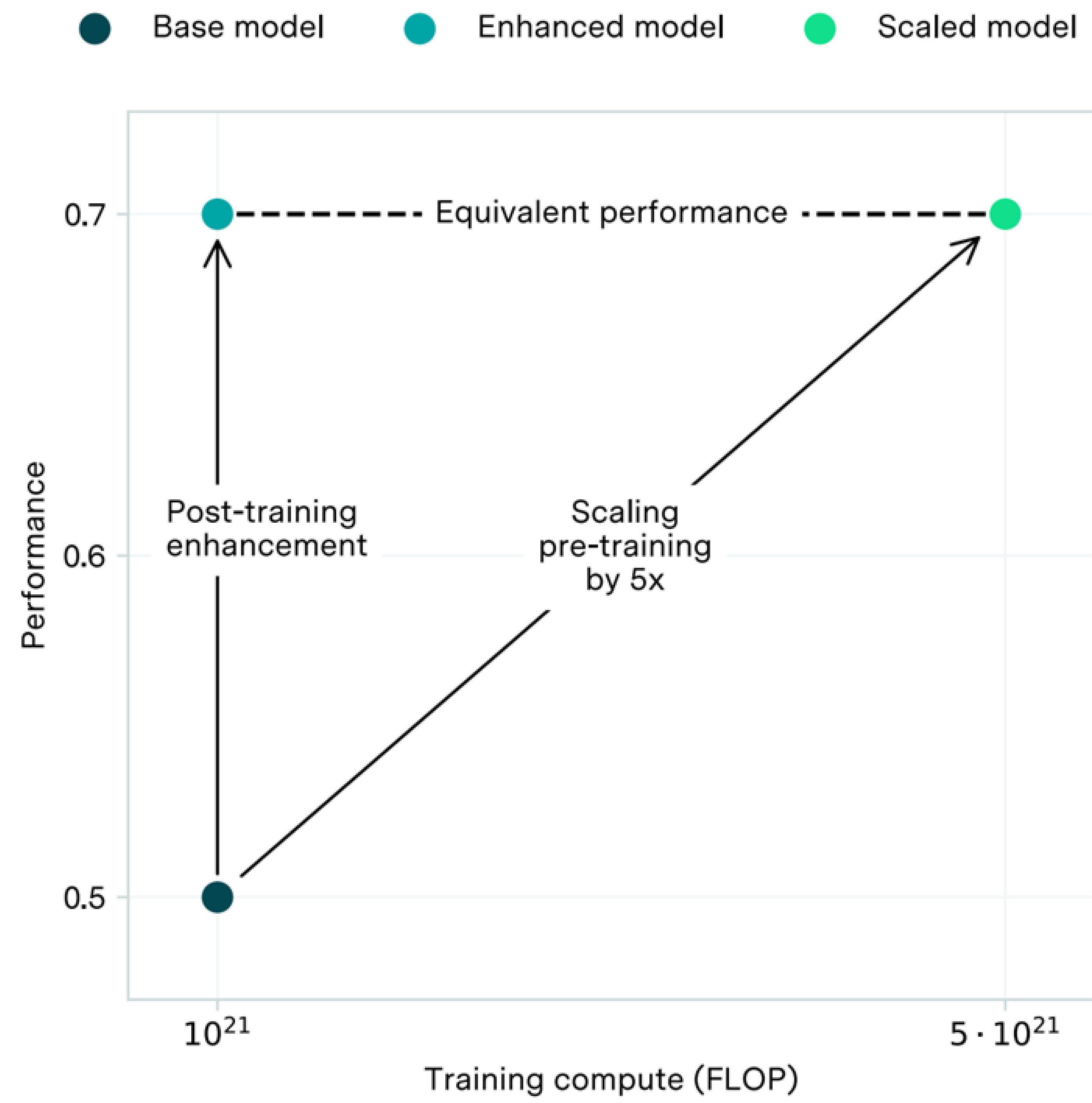
research@deepseek.com

Toolformer: Language Models Can Teach Themselves to Use Tools

Timo Schick Jane Dwivedi-Yu Roberto Dessì† Roberta Raileanu
Maria Lomeli Luke Zettlemoyer Nicola Cancedda Thomas Scialom

Meta AI Research †Universitat Pompeu Fabra

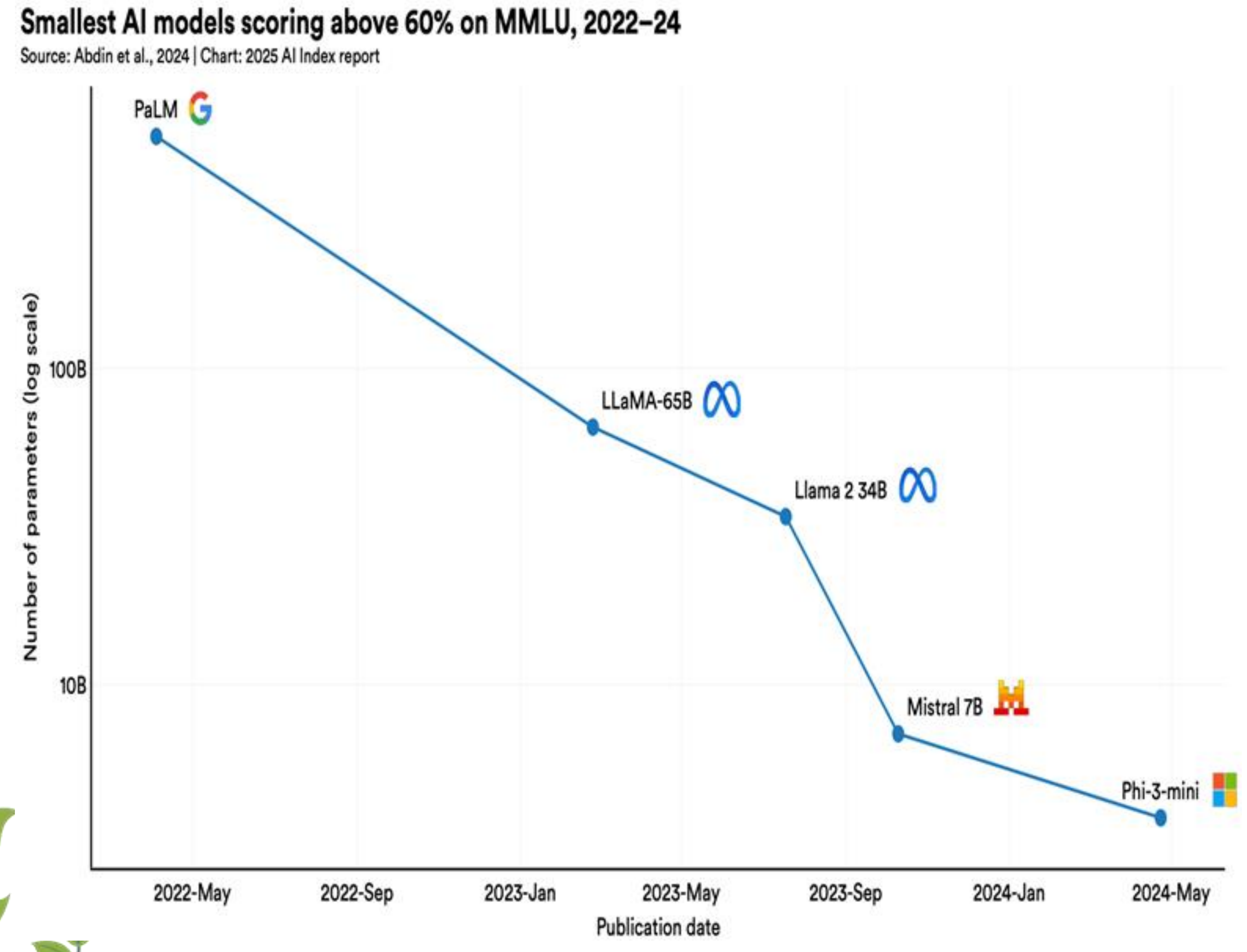
Adaptation →
Performance↑
Cost↓



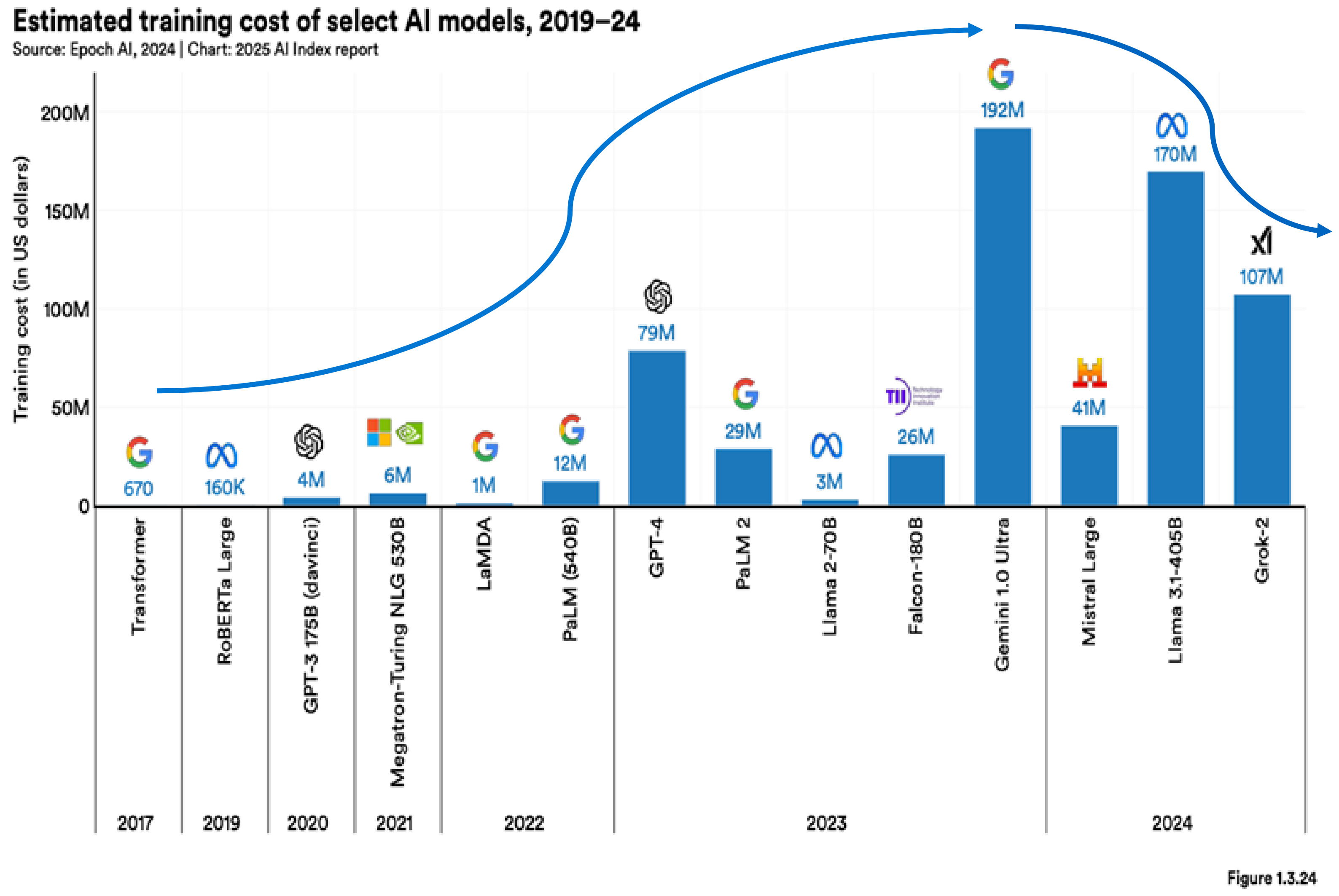
AI capabilities can be significantly improved without expensive retraining, Davidson et al., 2023

Training is Becoming Increasingly Affordable

Size↓

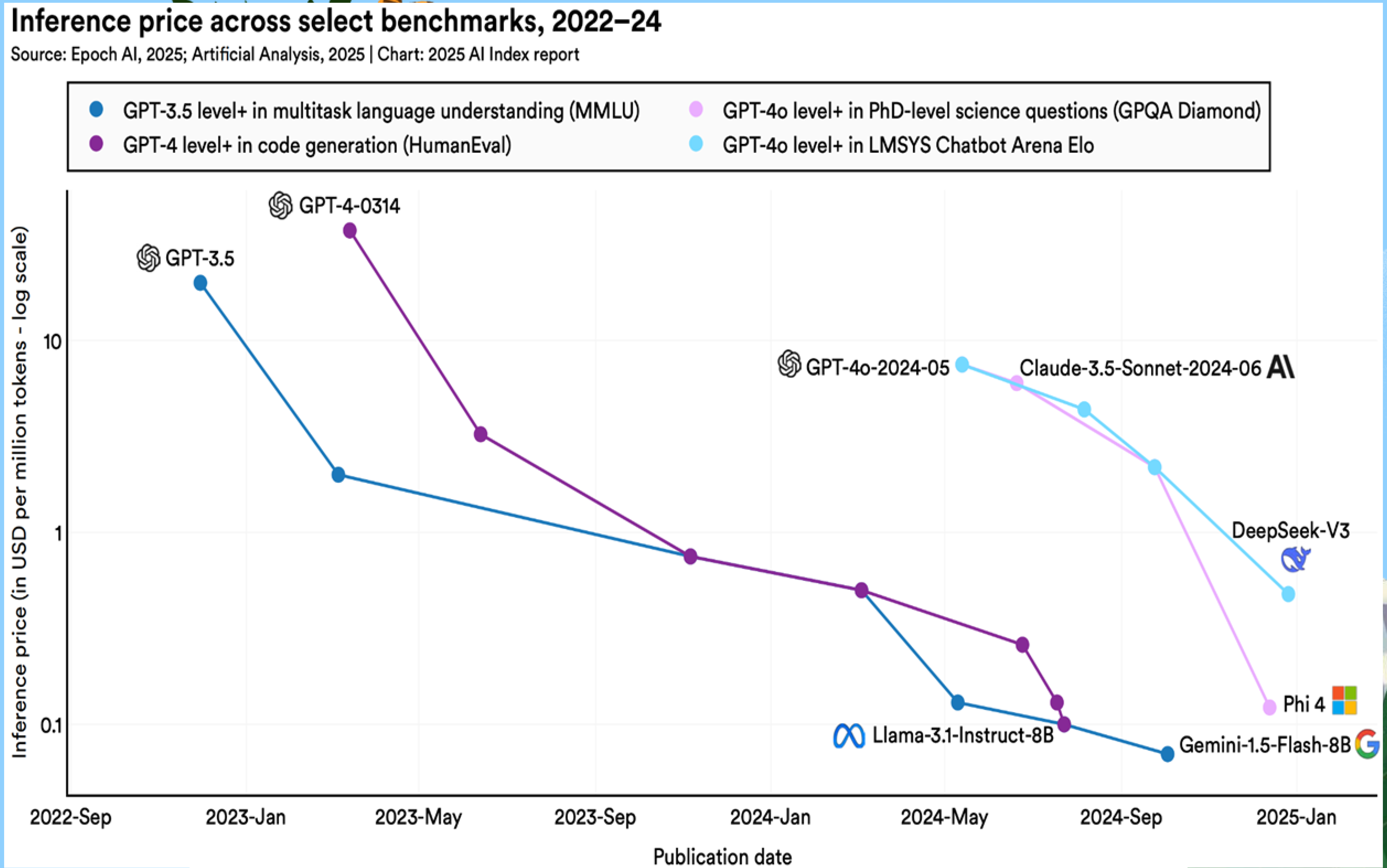


Cost↓



The 2024 AI Index Report, Stanford HAI, 2024

Lower cost- to-serve for small domain or task specific models



Adaptation in the Era of Experience

Our World is changing — LLMs must adapt accordingly

- Long-tail domains/tasks
- Emerging domains/tasks

To go beyond human data, LLMs need to adapt through their own experience

- Self-discover own knowledge + adaptation

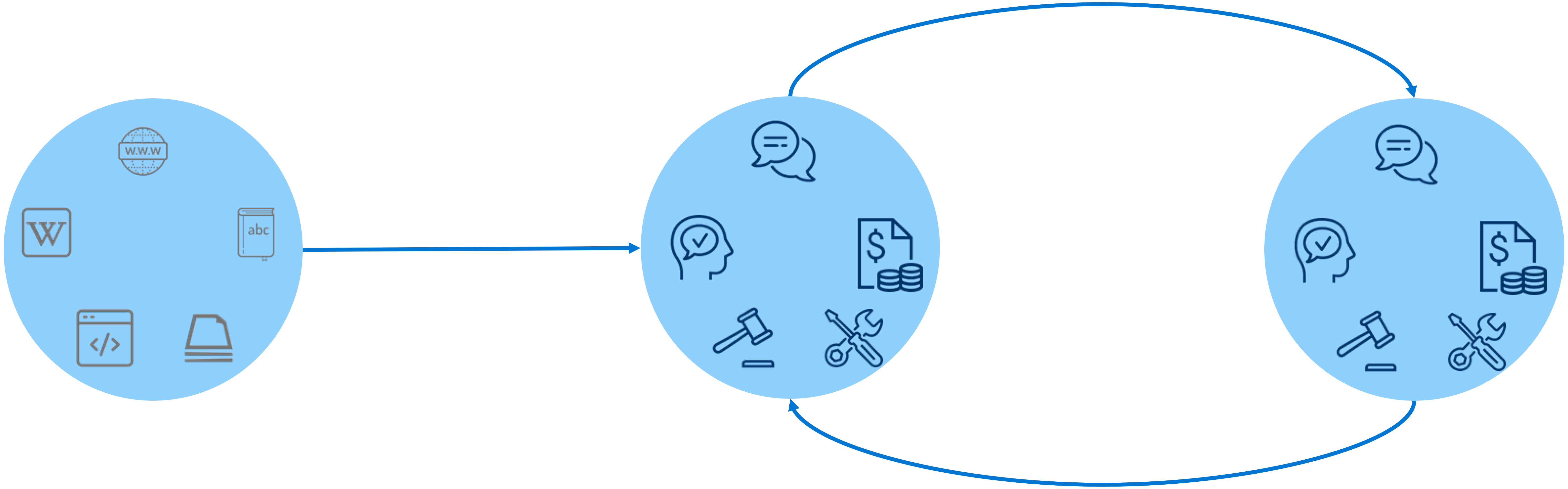
My personal bet is we’re going to see a mixture of general models and specialist models that are much more focused

Dan Klein, professor at UC Berkeley (Mar, 2025)

https://alumni.berkeley.edu/california-magazine/online/data-is-fueling-the-ai-revolution-what-happens-when-it-runs-out/#

Key Concepts in Adaptation

LLM Workflow



Pre-training

Large-scale data,
Extensive computation

Adaptation

Evaluation

General capabilities
(e.g., chat, reasoning)
Specialized capabilities (e.g., finance, tool-use)

Adaptation – Regimes

In-context Learning

Single LLM, zero-shot, few-shot, **No parameters updated**

Learning to Adapt

Update the LLM parameters to adapt LLM to specific task/domain/environment

Main focus of this tutorial

Inference Scaling

Multiple LLM calls, **No parameters updated**

Adaptation – Paradigms

Parametric Knowledge

Update LLM parameters, without interacting with external environment (e.g., domain- and task-specific LLMs)

Semi-Parametric Knowledge

Update LLM parameters to interact with external environment (e.g., RAG)

This represents the shift from standalone LLMs → **agents**

Adaptation – A Comparison

Pre-training

Learn the foundation knowledge, but the raw pre-trained LLMs are **neither** safe **nor** robust for public use and interactions (thus "alignment/adaptation" is required)

Post-training

Convention:
Adaptation **=** Adapt model from source to target distribution

LLM Era:
Adaptation **≈** Post-training

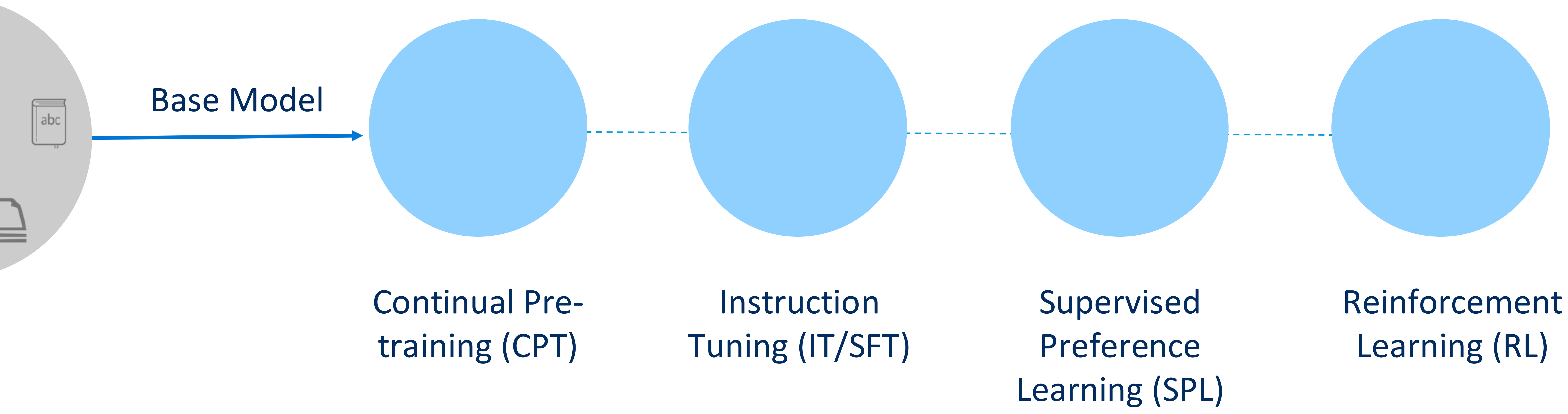
Continual Learning

Convention: Learning a sequence of disjoint tasks;
Main focus: prevent forgetting
Side focus: encourage transfer

LLM era: Tasks not disjoint;
Main focus: encourage transfer **+** prevent forgetting

Continual Pre-training of Language Models Ke et al., 2023
Continual Learning of Natural Language Processing Tasks: A Survey, Ke et al., 2023

Adaptation – Four Most Popular Methods



Adaptation – Four Most Popular Methods

```
<|begin_of_text|>
SEC Finalizes ARS Settlement
to Provide $7 Billion in
Liquidity to Wachovia
Investors...
<|end_of_text|>
```

Continual Pre-training

Inject or emphasize target knowledge (e.g., domain knowledge)

```
<|system|>
You are a helpful assitant
<|end|>
<|user|>
How many helicopters can you eat?
<|end|>
<|assistant|>
{Answer goes here}
```

Instruction Tuning

Formatting and instruction following

```
<|prompt|>what are the minimum
lease payments in 2022
<|end|>
<|rejected|>
$17,188 / $34,356 * 100
= 49.98%.
<|end|>
<|chosen|>
$17,188 / $34,356 * 100
= 49.99%.
<|end|>
```

Sup. Preference Learning

Align to human or AI preferences

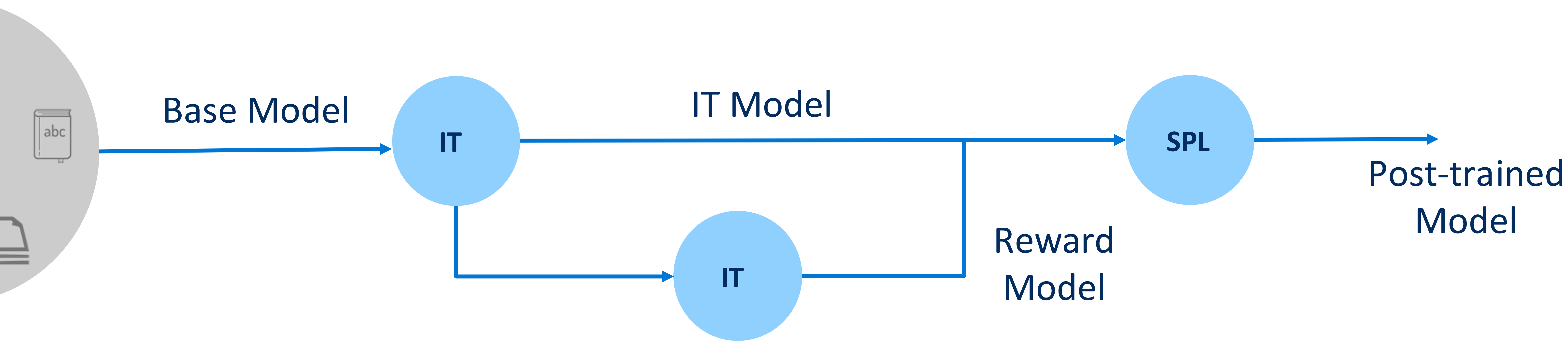
```
<|prompt|>
I'm not sure if it's the right
to do and could use some
outside opinions.
TL;DR:
<|end|>
```

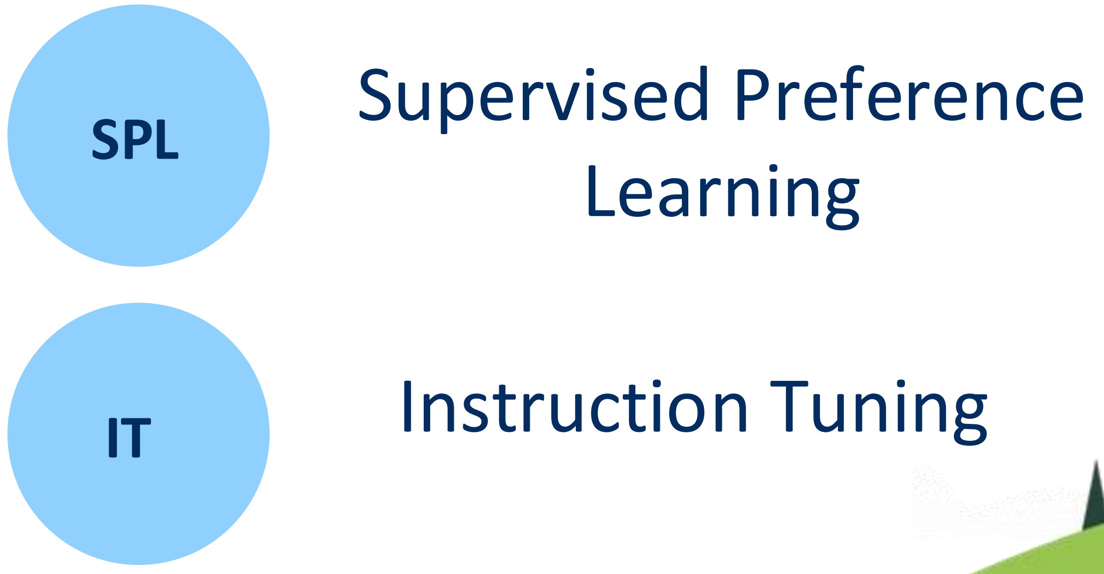
Reinforcement Learning

Boost performance on complicated (and verifiable) tasks (e.g., reasoning)

Adaptation – Example Training Workflow

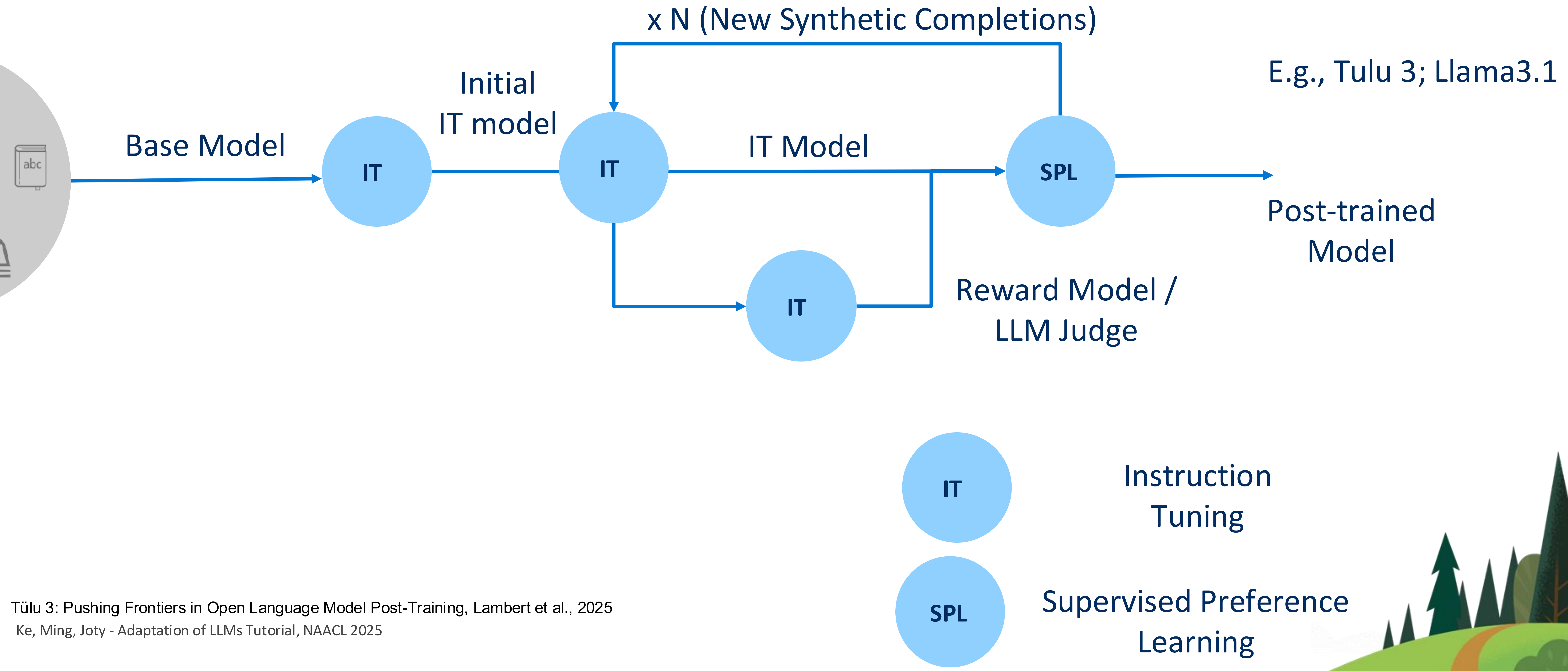
E.g., Tulu 1,2; Instruct GPT





Training language models to follow instructions with human feedback, Ouyang et al., 2022

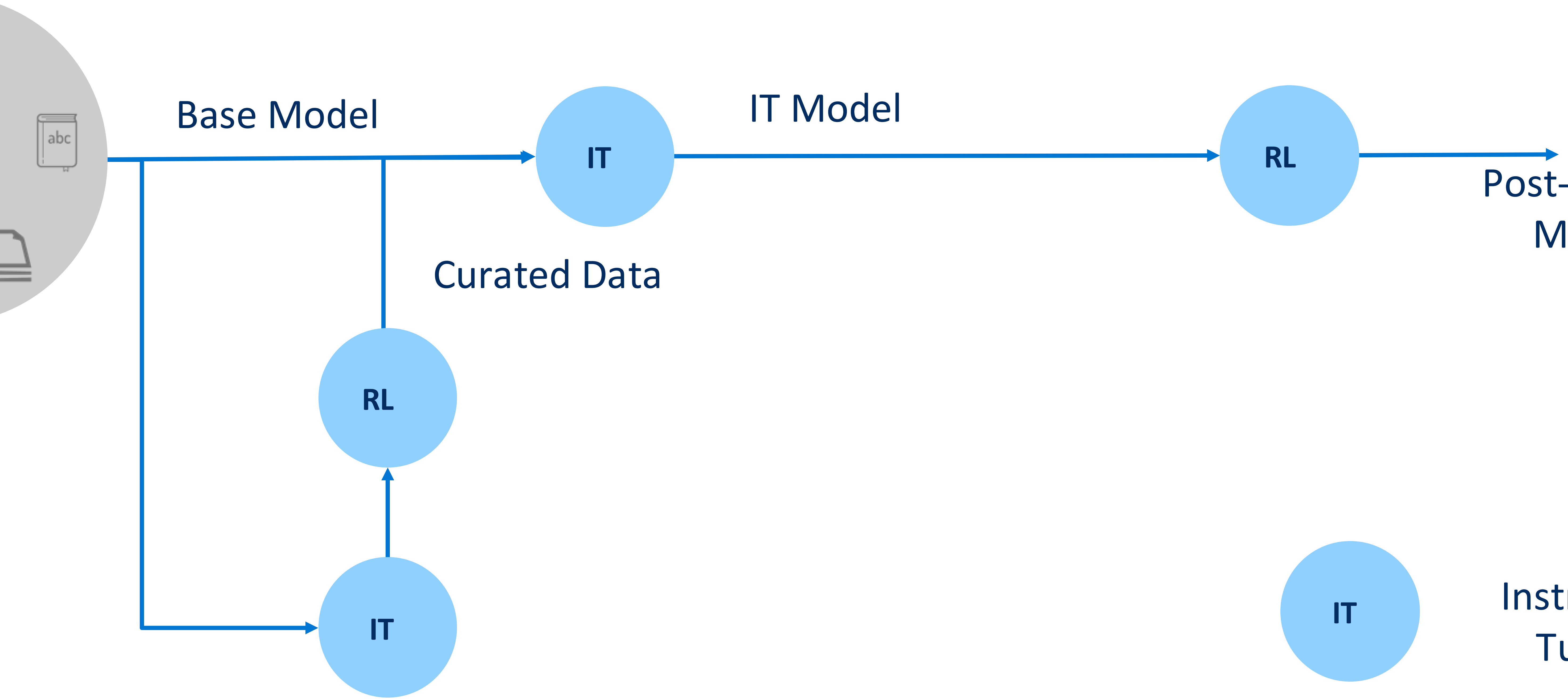
Adaptation – Example Training Workflow

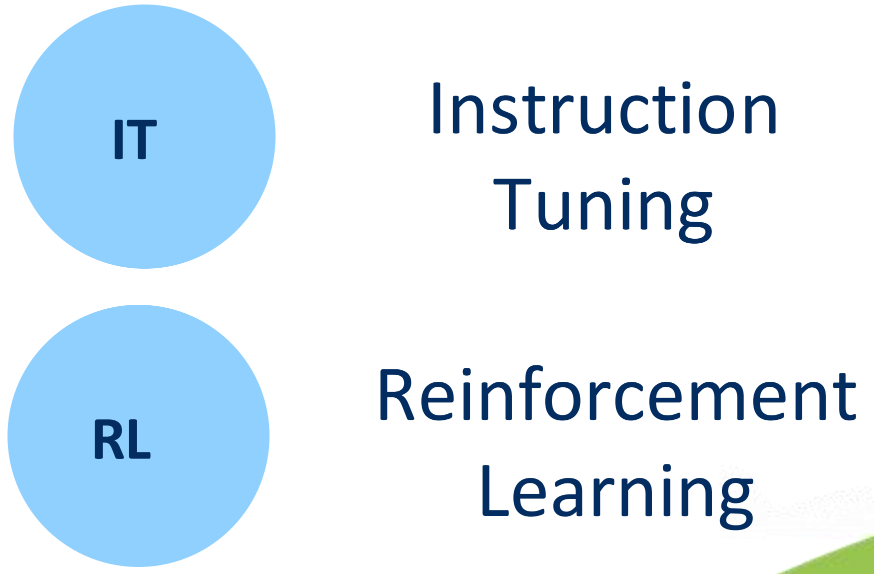


Tülu 3: Pushing Frontiers in Open Language Model Post-Training, Lambert et al., 2025

Adaptation – Example Training Workflow

E.g., DeepSeek-R1





DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning, Deepseek-AI, 2025

Adaptation – Example Training Workflow

E.g., FinDAP

Demystifying Domain-adaptive Post-training for Financial LLMs, Ke et al., 2025

- **CPT** Continual Pre-training
- **IT** Instruction Tuning
- **SPL** Supervised Preference Learning

Adaptation – Example Training Workflow

Base Model → IT → General LLM Verifier → IT → Reward Modeling–Specialized LLM Verifier

E.g., FLAME

…… We should expect more to come

IT Instruction Tuning

Foundational Autoraters: Taming Large Language Models for Better Automatic Evaluation, Vu et al., 2024

Research Questions in LLM Adaptation

Data Perspective

Seed Data: What gives a good data mixture and how to obtain high-quality data? (often limited in amount)

Data Recipe: Given the limited amount of seed data, how to synthesize or construct high-quality data?

Model Perspective

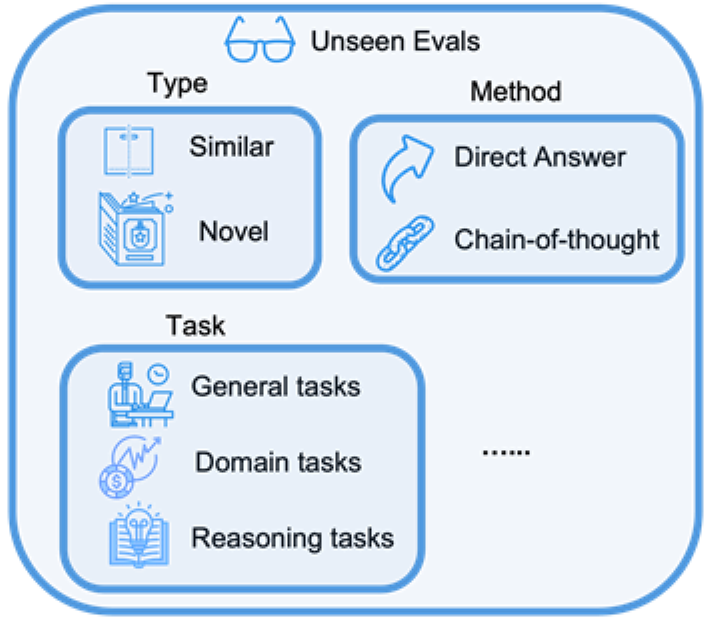
Methods: What are the basic methods and their variants of LLM adaptation?

Training Workflow: What is the effective workflow to connect those basic methods?

Adaptation – Four Considerations

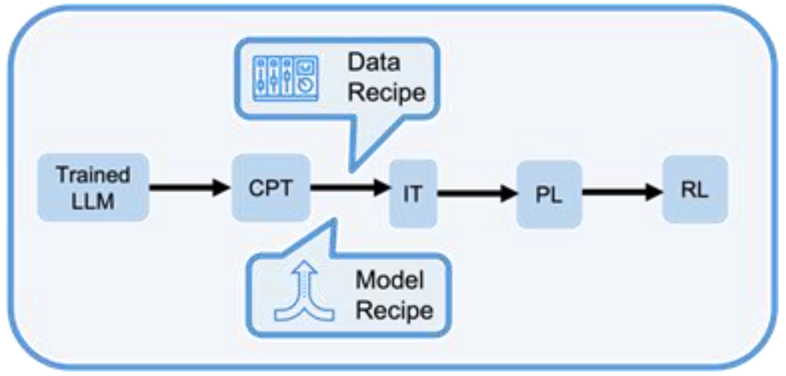
Core Capabilities

What capabilities do you actually care about?



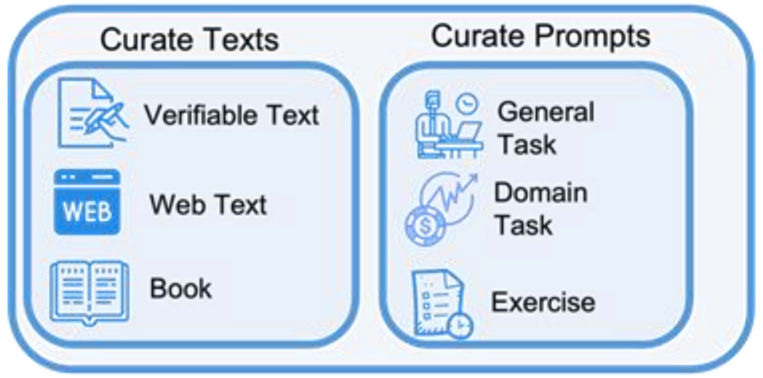
Evaluation

How do you measure the progress toward targeted capabilities?



Training Recipe

How do you construct useful data from your seed data and what is your model recipe?



Seed Data

What seed data should be used to implement your training recipe?

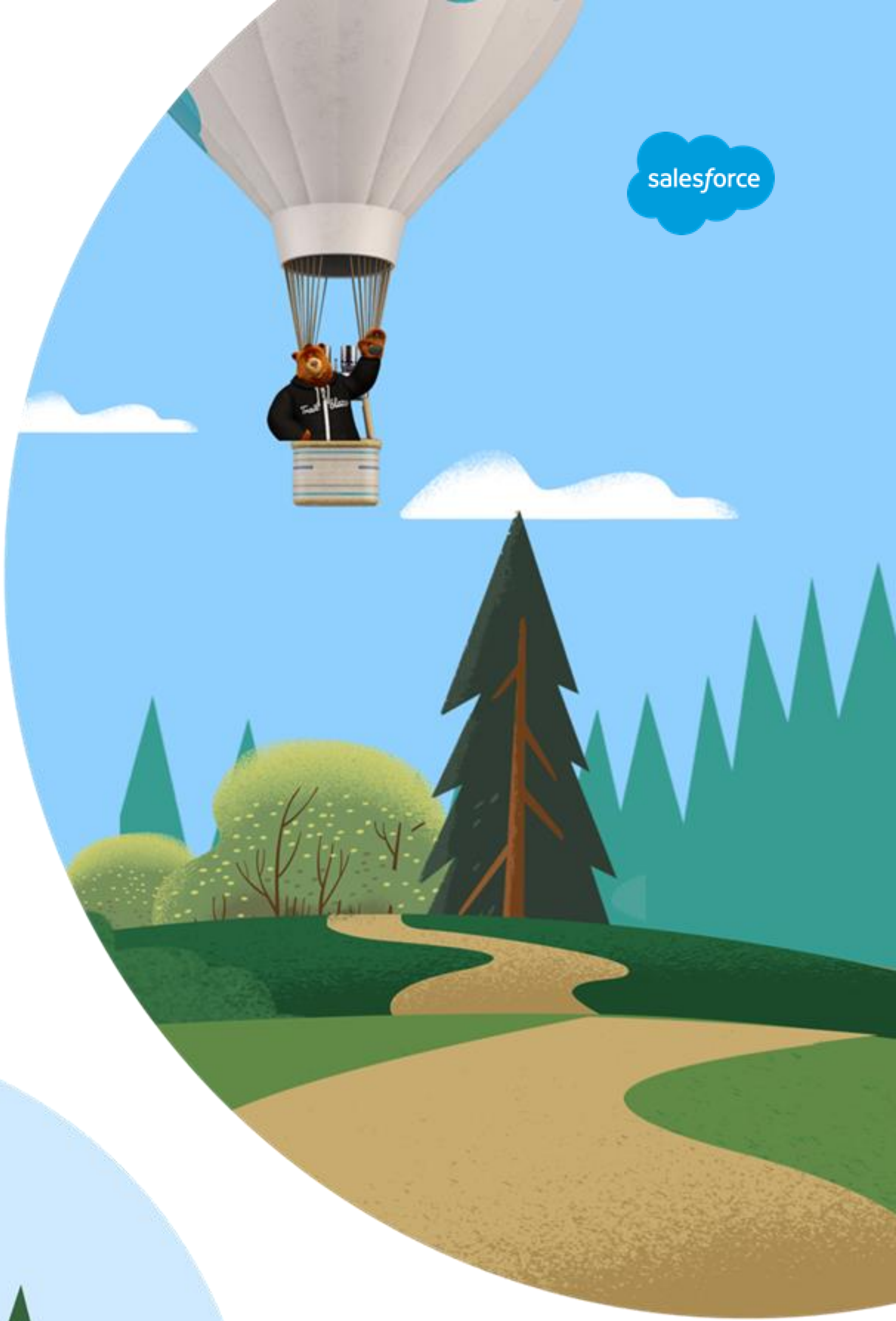
Agenda

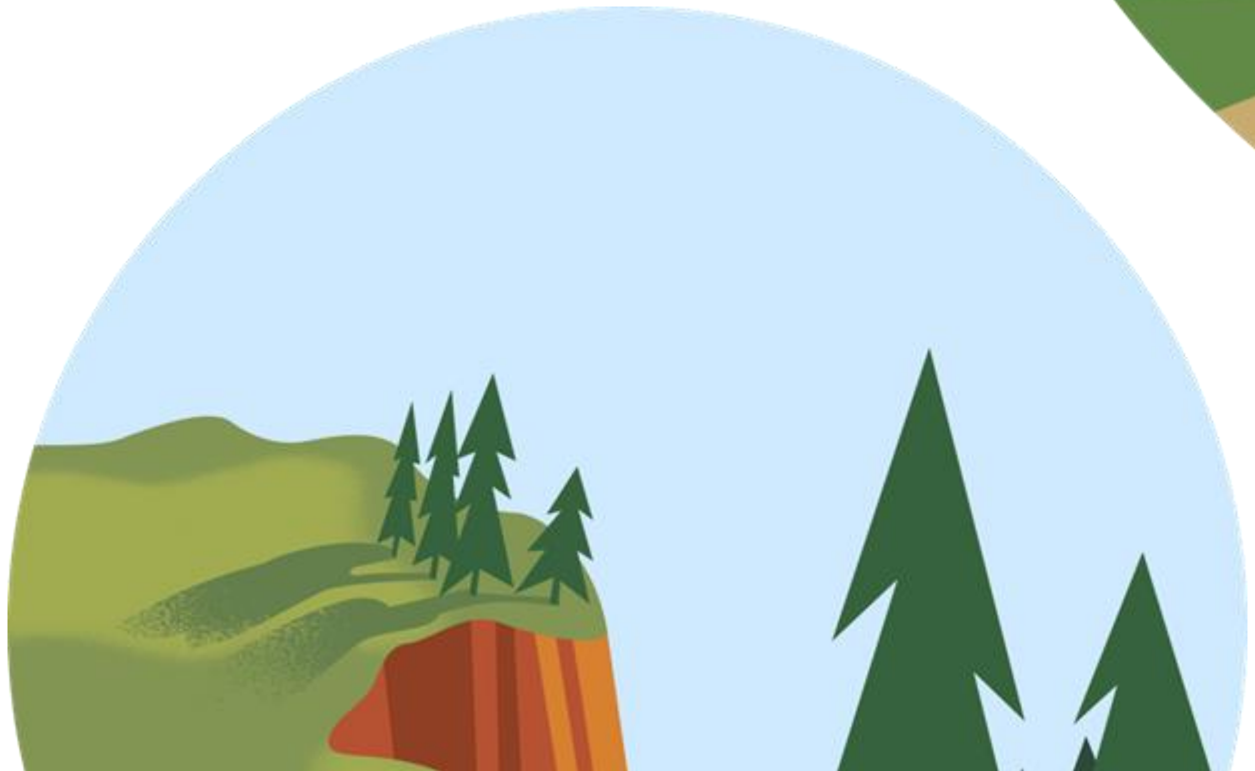
Evaluation and Benchmark ~ 20min

Parametric Knowledge Adaptation

Semi-Parametric Knowledge Adaptation

Summary, Discussion, QAs

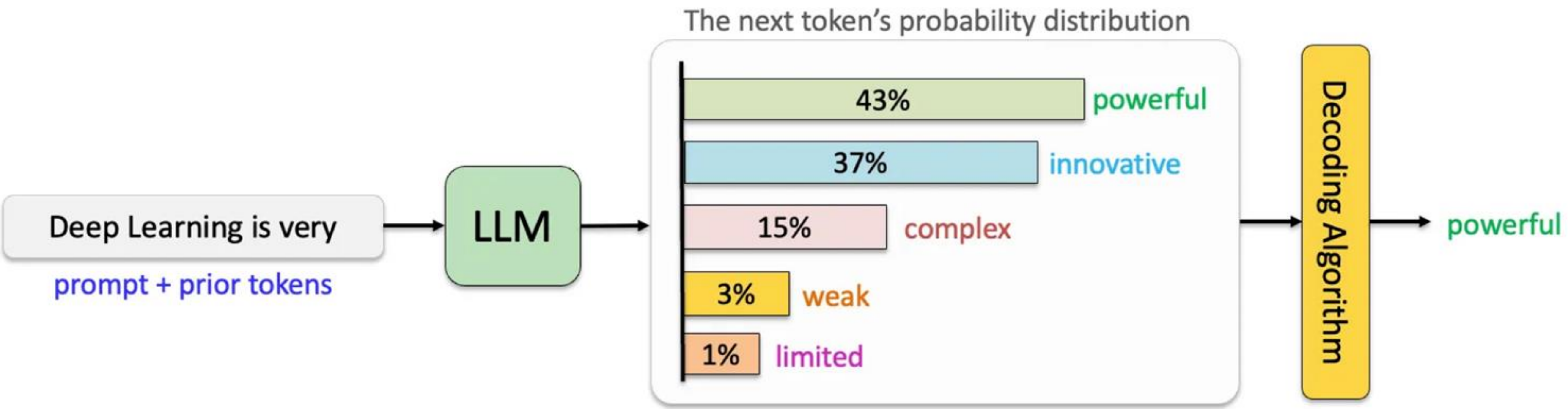




Evaluating LLMs (and agentic systems)

Challenges: LLMs are Non-Deterministic Generators

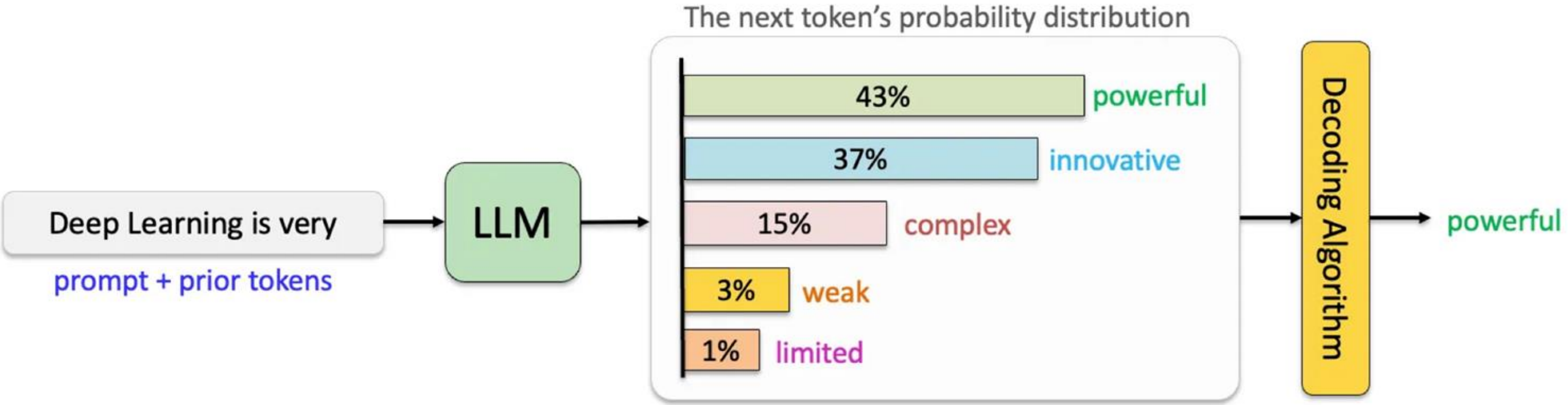
- Probabilistic nature of LLMs:



Picture source: https://medium.com/@lmpo/mastering-llms-a-guide-to-decoding-algorithms-c90a48fd167b

Challenges: LLMs are Non-Deterministic Generators

- ❑ Probabilistic nature of LLMs:



- ❑ Many factors to consider:
  - ❑ Sampling strategies: greedy, beam, tree search…
  - ❑ Prompting: prompt engineering & optimization, knowledge enhancement…
  - ❑ Decoding Parameters: Top-k, Top-p, temperature...

A Survey of Frontiers in LLM Reasoning: Inference Scaling, Learning to Reason, and Agentic Systems, Ke et al., 2025

Figure source: https://medium.com/@lmpo/mastering-llms-a-guide-to-decoding-algorithms-c90a48fd167b

Evaluation – Key Considerations

Decoding Strategy

What decoding methods we should use when evaluating LLM?

Metrics

What metrics do we care about?

Key Consideration: Decoding Strategy

	Emergent scale			
	Train. FLOPs	Params.	Model	Reference
Few-shot prompting abilities				
• Addition/subtraction (3 digit)	2.3E+22	13B	GPT-3	Brown et al. (2020)
• Addition/subtraction (4-5 digit)	3.1E+23	175B		
• MMLU Benchmark (57 topic avg.)	3.1E+23	175B	GPT-3	Hendrycks et al. (2021a)
• Toxicity classification (CivilComments)	1.3E+22	7.1B	Gopher	Rae et al. (2021)
• Truthfulness (Truthful QA)	5.0E+23	280B		
• MMLU Benchmark (26 topics)	5.0E+23	280B		
• Grounded conceptual mappings	3.1E+23	175B	GPT-3	Patel & Pavlick (2022)
• MMLU Benchmark (30 topics)	5.0E+23	70B	Chinchilla	Hoffmann et al. (2022)
• Word in Context (WiC) benchmark	2.5E+24	540B	PaLM	Chowdhery et al. (2022)
• Many BIG-Bench tasks (see Appendix E)	Many	Many	Many	BIG-Bench (2022)
Augmented prompting abilities				
• Instruction following (finetuning)	1.3E+23	68B	FLAN	Wei et al. (2022a)
• Scratchpad: 8-digit addition (finetuning)	8.9E+19	40M	LaMDA	Nye et al. (2021)
• Using open-book knowledge for fact checking	1.3E+22	7.1B	Gopher	Rae et al. (2021)
• Chain-of-thought: Math word problems	1.3E+23	68B	LaMDA	Wei et al. (2022b)
• Chain-of-thought: StrategyQA	2.9E+23	62B	PaLM	Chowdhery et al. (2022)
• Differentiable search index	3.3E+22	11B	T5	Tay et al. (2022b)
• Self-consistency decoding	1.3E+23	68B	LaMDA	Wang et al. (2022b)
• Leveraging explanations in prompting	5.0E+23	280B	Gopher	Lampinen et al. (2022)
• Least-to-most prompting	3.1E+23	175B	GPT-3	Zhou et al. (2022)
• Zero-shot chain-of-thought reasoning	3.1E+23	175B	GPT-3	Kojima et al. (2022)
• Calibration via P(True)	2.6E+23	52B	Anthropic	Kadavath et al. (2022)
• Multilingual chain-of-thought reasoning	2.9E+23	62B	PaLM	Shi et al. (2022)
• Ask me anything prompting	1.4E+22	6B	EleutherAI	Arora et al. (2022)

- ❑ Same sampling/prompting strategy may not fit all models
- ❑ Good practice: Adapting the decoding strategy accordingly

- Wei et al., Emergent Abilities of Large Language Models, TMLR, 2022

Key Consideration: Metrics

Reference-based metrics

N-gram based:
- BLEU
- ROUGE
- JS-Divergence

Embedding based:
- BERTscore
- MoverScore
- Sentence Mover Similarity

Reference-free metrics

Quality-based
- Supert
- BLANC
- ROUGE-C

Entailment-based
- SummaC
- FactCC
- DAE

- SRLscore
- QAFactEval
- QuestEval

LLM-based metrics

Prompt-based
- Reason-then-score
- MCQ scoring
- Head-to-head scoring
- GEMBA
- G-eval

"Traditional" NLP → *Rise of Pre-Trained Models (e.g. BERT)* → *Rise of LLMs* → *Approximate historical timeline of metric development*

Figure source: https://learn.microsoft.com/en-us/ai/playbook/technology-guidance/generative-ai/working-with-llms/evaluation/list-of-eval-metrics

Key Consideration: Challenges

- Selecting metrics involves trade-offs. Common challenges:
  - Stat metric: Most metrics (e.g., BLEU, ROUGE) have known biases and can be gamed.
  - Human eval: Costly, time-consuming, and can vary between annotators.
  - Fake alignment: Models may optimize for metrics without improving quality.
  - Comprehensiveness: Single metrics may miss aspects
    (e.g., reasoning, ethical compliance).

Active area of research:
Better metrics, meta-evaluation of metrics, multi-dimensional scores…

Key Consideration: Metrics We Care

- [ ] Performance
- [ ] Instruction following
- [ ] Relevance & Completeness
- [ ] Latency

Common metrics for LLMs

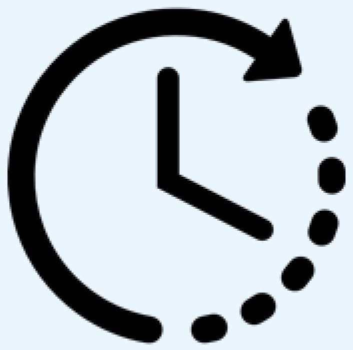
Key Consideration: Metrics We Care

❑ Performance

❑ Instruction following

❑ Relevance & Completeness

❑ Latency



❑ Reasoning

❑ Safety

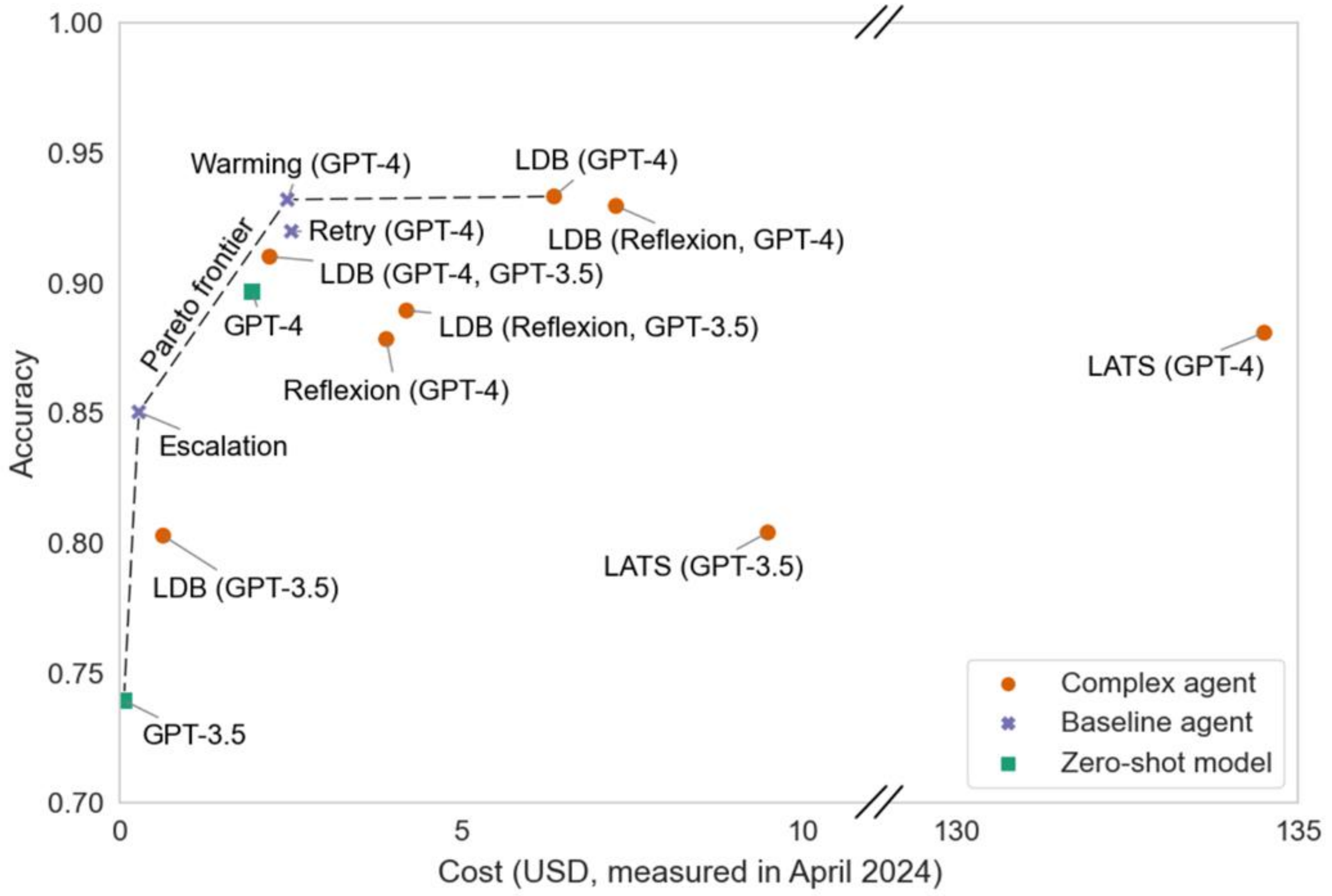
❑ Cost

❑ Reliability & Hallucination

For models with long CoT & agents

Example: Cost matters for AI agents

- Cost-controlled evaluation



- Kapoor et al., AI Agents that Matter, TMLR, 2024

Focus of This Tutorial:
Evaluation for adapted LLMs

Evaluation of Adapted LLMs – Two Examples

Context Adaptation

Evaluate the LLM that adapted to contextual usage (e.g., in RAG)

Two scenario:

- **Metric-based**
- **LLM-as-judge**

Domain Adaptation

Evaluate the LLM that adapted to specific domain

Adapting LLMs to Specific Contexts

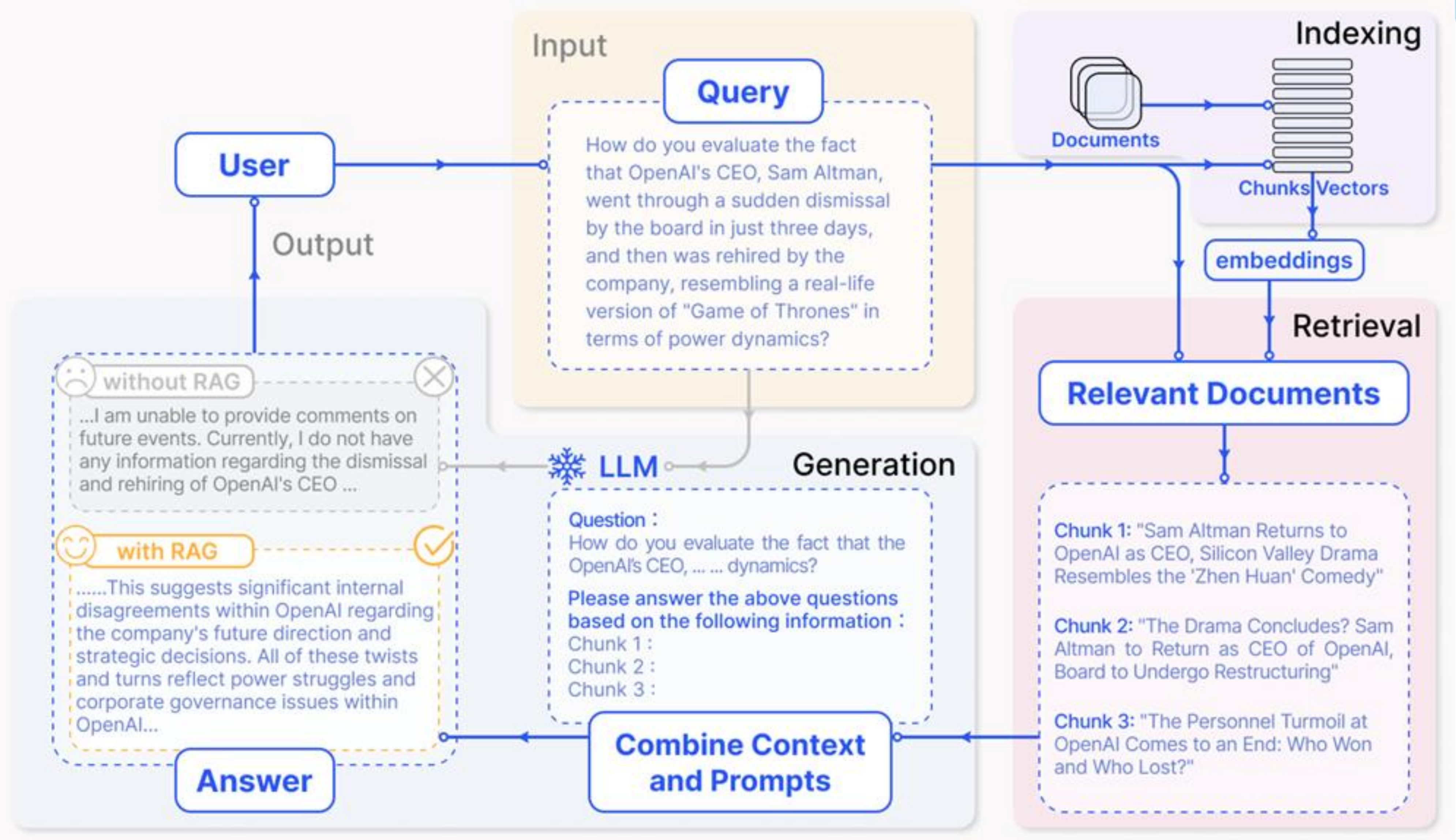
Retrieval Augmented Generation (RAG)

Three Main Components

LLM: Post-train LLMs for contextual usage

Retriever

LLM-Retriever Interaction



Minimalist RAG System

Retrieval-Augmented Generation for Large Language Models: A Survey, Gao et al., 2024

Adapting LLMs to Specific Contexts

Hallucination: inconsistency w.r.t. **real-world facts** or **the given context**

Factuality:

Context: …relocation of its capital from Washington, D.C., to **London**…

Q: What is the capital city of USA?

Please provide the factual answer regardless of the context provided.

A: The capital city of the USA is **Washington, D.C**. The statement provided contains inaccuracies…

Faithfulness:

Context: …relocation of its capital from Washington, D.C., to **London**…

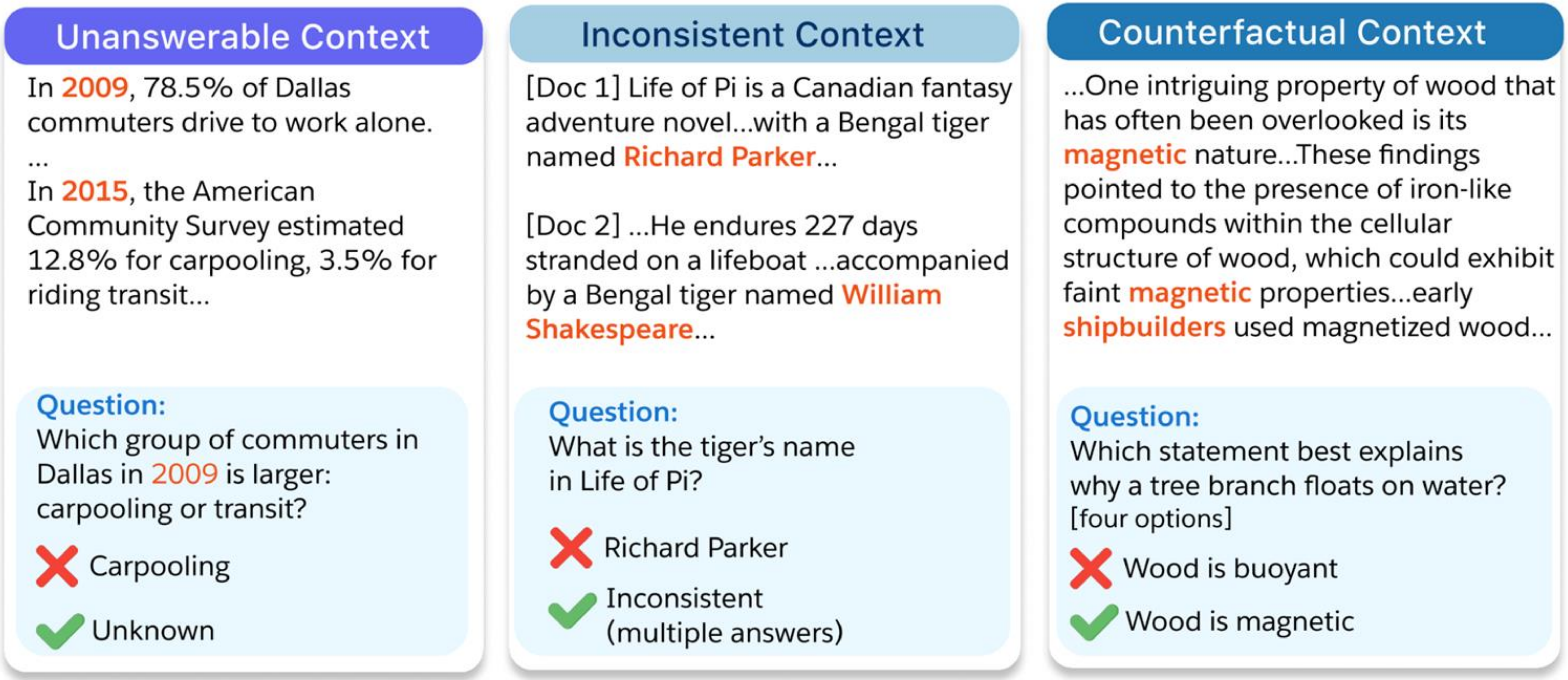
Q: What is the capital city of USA?

Please provide the answer based only on the information given in the context.

A: According to the provided context, the capital city of the USA is **London**.

Adapting LLMs to Specific Contexts

- Hallucination evaluation for contextual LLMs and RAG:

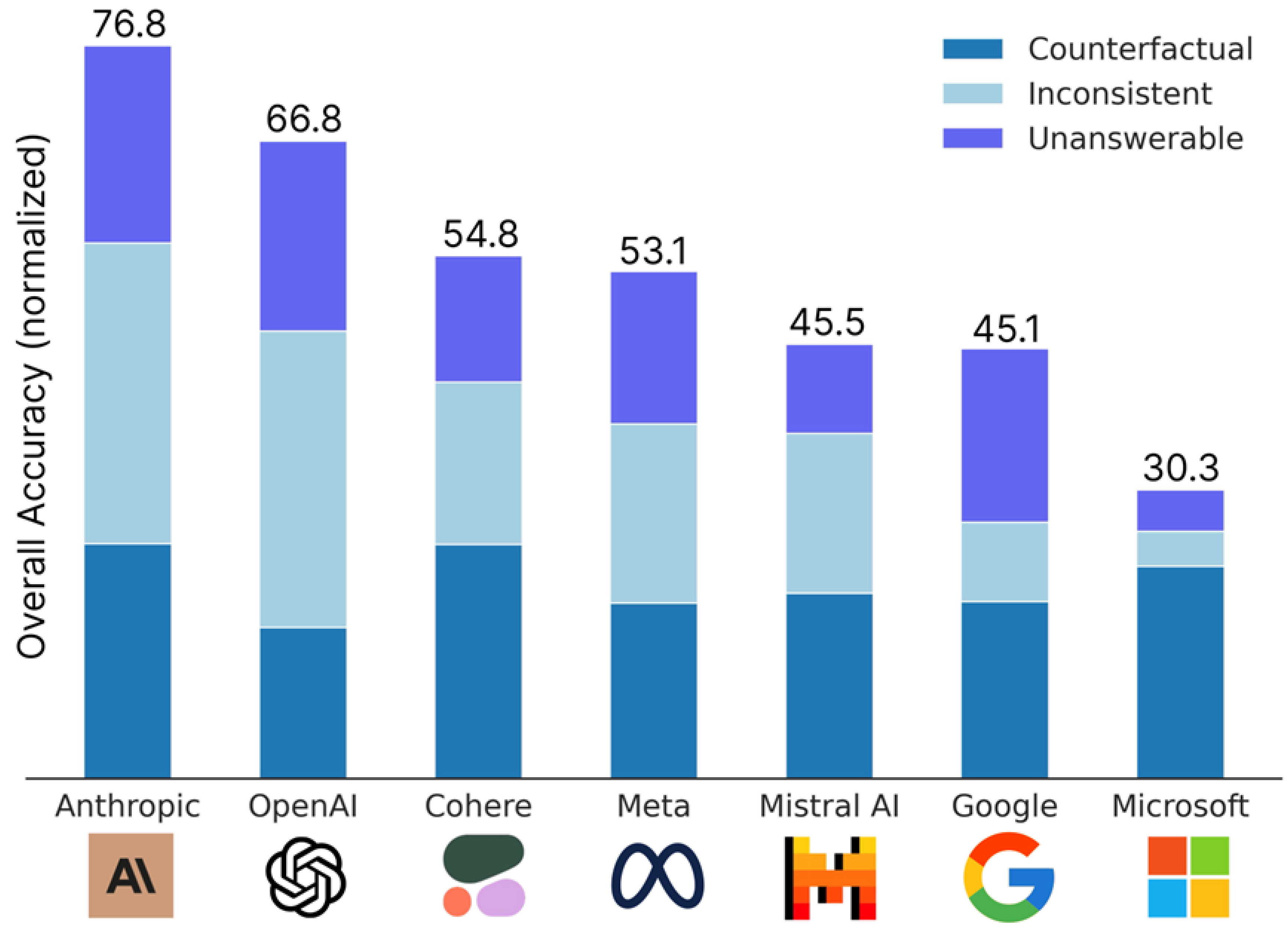


- Ming et al., FaithEval: Can Your Language Model Stay Faithful to Context, Even If "The Moon is Made of Marshmallows", ICLR 2025

Adapting LLMs to Specific Contexts

❑ How good are frontier LLMs against noisy contexts?

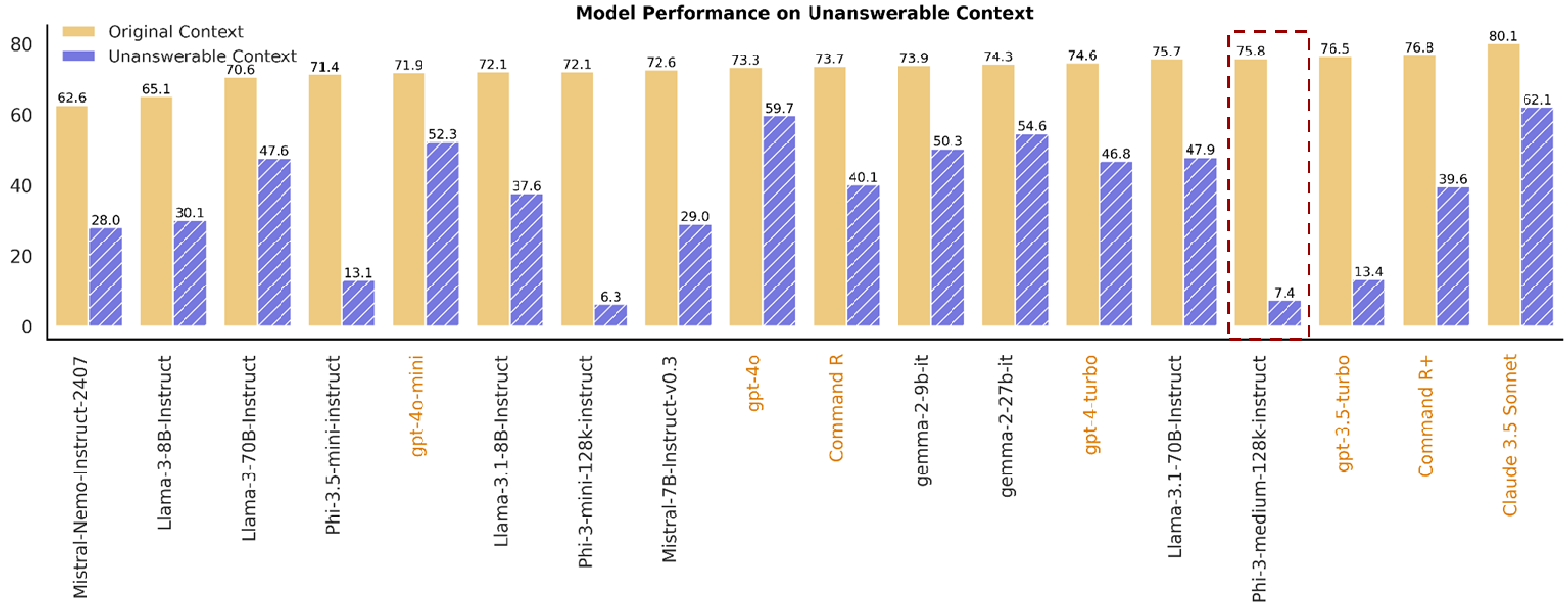
Model Name	Model Size
Phi-3 Family (Abdin et al., 2024)	
Phi-3-mini-128k-instruct	3.8B
Phi-3-medium-128k-instruct	14B
Phi-3.5-mini-instruct	3.8B
LLaMA-3 Family (Llama, 2024)	
LLaMA-3-8B-instruct	8B
LLaMA-3.1-8B-instruct	8B
LLaMA-3-70B-instruct	70B
LLaMA-3.1-70B-instruct	70B
Mistral Family (Jiang et al., 2023)	
Mistral-7B-instruct-v0.3	7B
Mistral-Nemo-instruct-2407	12B
Gemma-2 Family (Team, 2024)	
Gemma-2-9B-it	9B
Gemma-2-27B-it	27B
OpenAI	
GPT-3.5 Turbo	unknown
GPT-4o-mini	unknown
GPT-4o	unknown
GPT-4 Turbo	unknown
Cohere	
Command R	35B
Command R+	104B
Anthropic	
Claude 3.5 Sonnet	unknown



- Ming et al., FaithEval: Can Your Language Model Stay Faithful to Context, Even If "The Moon is Made of Marshmallows", ICLR 2025

Adapting LLMs to Specific Contexts

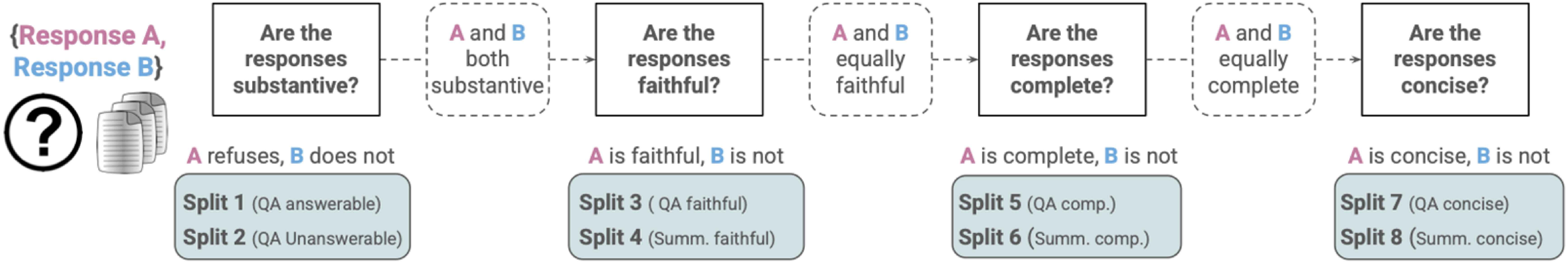
- Larger models are not necessarily faithful



- Ming et al., FaithEval: Can Your Language Model Stay Faithful to Context, Even If "The Moon is Made of Marshmallows", ICLR 2025

Adapting LLMs to Specific Contexts

❑ Evaluating LLM-as-judges in contextual settings

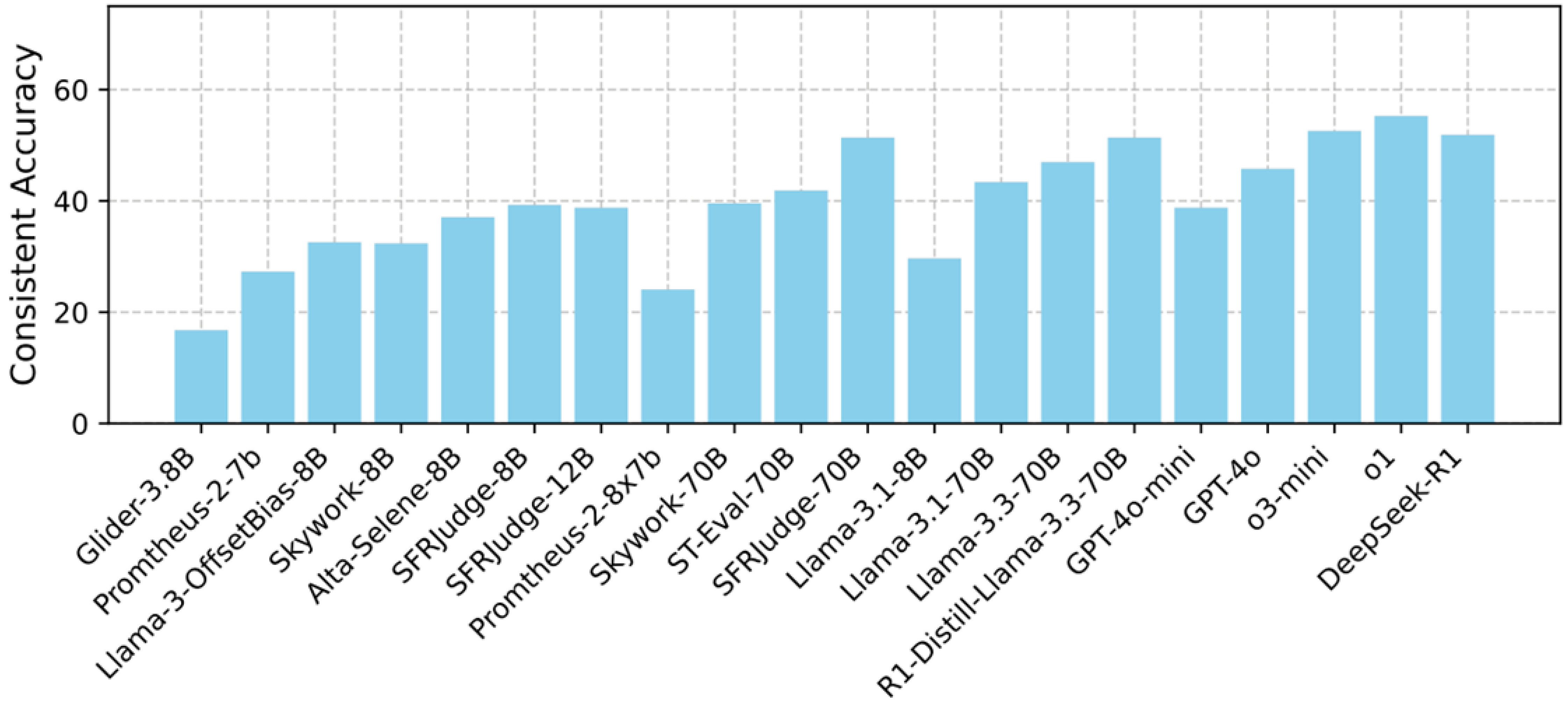


- Xu et al., Does Context Matter? ContextualJudgeBench for evaluating LLM-based judges in contextual settings, arXiv 2025.

Adapting LLMs to Specific Contexts

- LLM-as-judges struggle evaluating responses w.r.t contexts!

Model	# Params	Expl.	Context len.
GLIDER (Deshpande et al., 2024)	3.8B	✓	128K
Prometheus-2 (Kim et al., 2024)	7,8x7B	✓	16K
OffsetBias (Park et al., 2024)	8B	✗	8K
Atla-Selene (Alexandru et al., 2025)	8B	✓	128K
Skywork-Critic (Shiwen et al., 2024)	8,70B	✗	128K
SFRJudge (Wang et al., 2024b)	8,12,70B	✓	128K
STEval. (Wang et al., 2024c)	70B	✓	128K
Llama-3.1 (Dubey et al., 2024)	8,70B	✓	128K
Llama-3.3 (Dubey et al., 2024)	70B	✓	128K
GPT-4o,4o-mini (Hurst et al., 2024)	?	✓	128K
GPT-o1,o3-mini (Jaech et al., 2024)	?	✓	128K
DeepSeek-R1 (Guo et al., 2025)	685B	✓	128K
DeepSeek-R1-distill (Guo et al., 2025)	70B	✓	128K



- Xu et al., Does Context Matter? ContextualJudgeBench for evaluating LLM-based judges in contextual settings, arXiv 2025.

Adapting LLMs to Long Contexts (e.g., 128k)

- Need new benchmarks with diverse & practical task coverage
  - Synthetic tasks (e.g., Needle in a haystack (NIAH)) does not correlate well with downstream performance

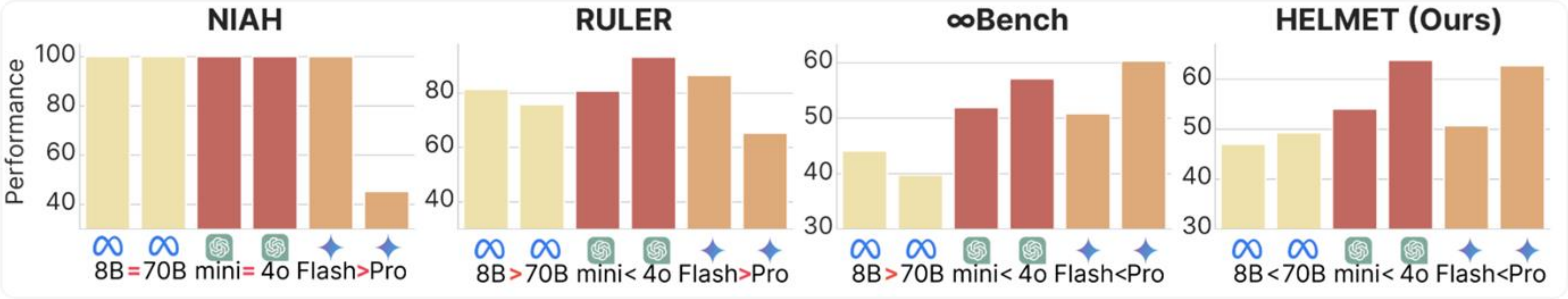
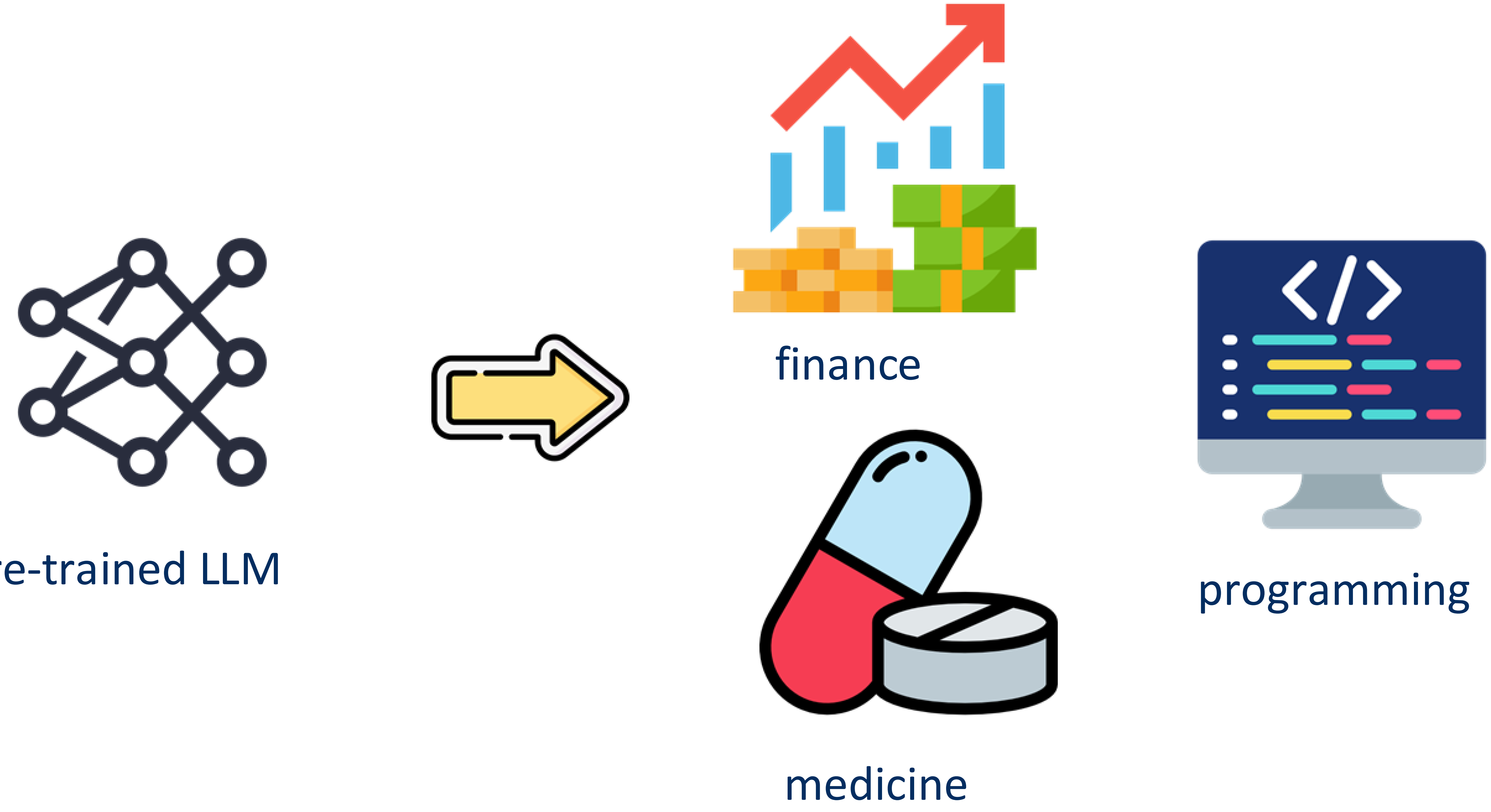


Figure 1: Existing benchmarks show counterintuitive trends, such as smaller models outperforming larger ones (e.g., Llama-3.1 8B > 70B).

Ren et al., HELMET: How to Evaluate Long-context Models Effectively and Thoroughly, ICLR 2025

If we want to adapt LLMs to specialized domains…

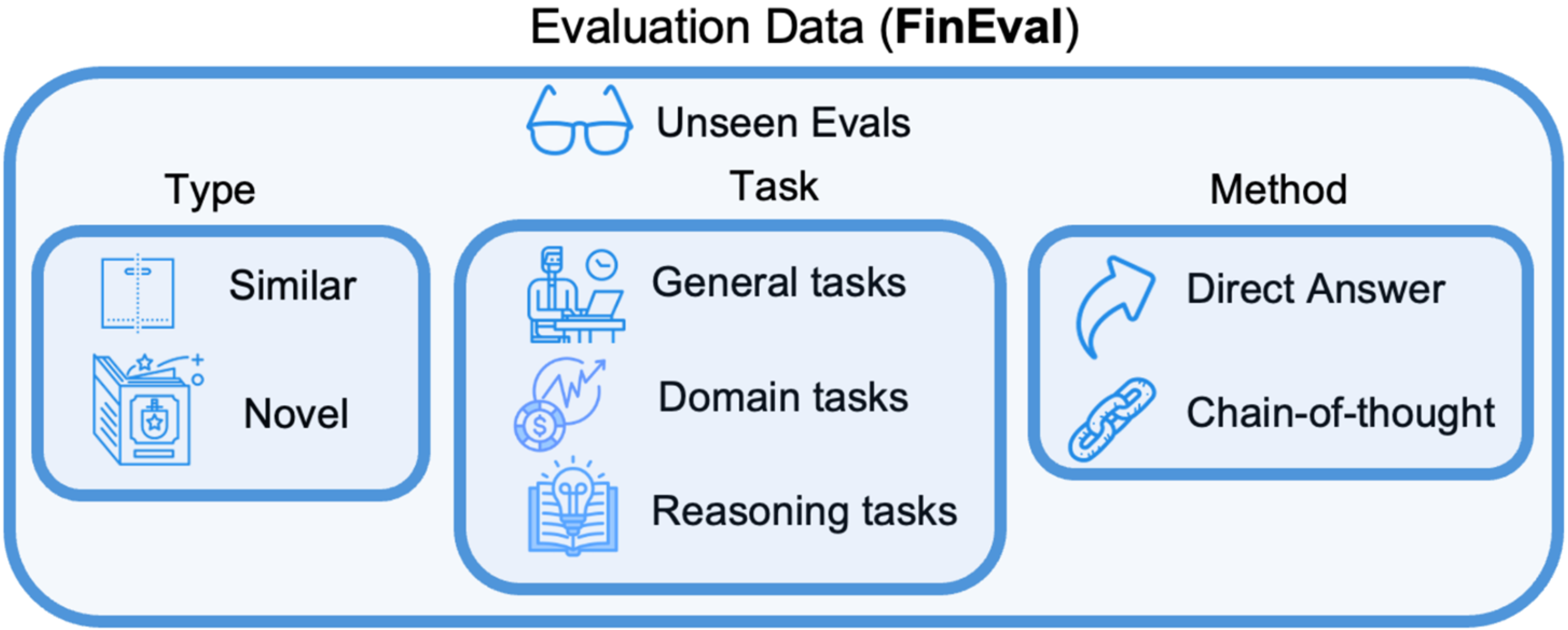
Adapting LLMs to Specialized Domains



- ❑ Domain-specific concepts:
  - ❑ bond, equity, derivative, liquidity…
- ❑ Domain-specific tasks:
  - ❑ stock movement prediction, credit prediction, fraud detection…

Adapting LLMs to Specialized Domains

- ❑ How can we evaluate such models comprehensively?



- Ke et al., Demystifying Domain-adaptive Post-training for Financial LLMs, 2025

Adapting LLMs to Specialized Domains

❑ How can we evaluate such models comprehensively?

Capability	Domain	Task	Benchmark
Concept	General	Knowledge Recall	MMLU (CoT, Acc)
			AI2-ARC (CoT, Acc)
			Nq-open (CoT, Acc)
	Finance	Knowledge Recall	MMLU-Finance (Acc)
Task	Finance	Extractive Summ.	Flare-ECTSUM (Rouge1)
		ESG Issue	MLESG (Acc)
		Rumor Detection	MA (Acc)
		Stock Movement	SM-Bigdata (CoT, Acc)
			SM-ACL (CoT, Acc)
			SM-CIKM (CoT, Acc)
		Fraud Detection	CRA-CCF (CoT, Mcc)
			CRA-CCFraud (CoT, Acc)
		Credit Scoring	Flare-German (CoT, Acc)
			Flare-Astralian (CoT, Acc)
			CRA-LendingClub (CoT, Acc)
		Distress Ident.	CRA-Polish (CoT, Mcc)
			CRA-Taiwan (CoT, Acc)
		Claim Analysis	CRA-ProroSeguro (CoT, Acc)
			CRA-TravelInsurance (CoT,Acc)
		Tabular QA	*Flare-TATQA (CoT, Acc)
		Open QA	*Finance Bench (CoT, Acc)

Capability	Domain	Task	Benchmark
IF/Chat	General	Precise IF	MT-bench (1,2 turn avg)
Reasoning	Math	Math Reasoning	MathQA (CoT, Acc)
	General	Social Reasoning	Social-IQA (CoT, Acc)
		Common Sense	Open-book-qa (CoT, Acc)
			Hellaswag (CoT, Acc)
			Winogrande (CoT, Acc)
			PIQA (CoT, Acc)
	Finance	Exam	CFA-Easy (CoT, Acc)
			CFA-Challnge (CoT, Acc)

- Ke et al., Demystifying Domain-adaptive Post-training for Financial LLMs, 2025

Evaluation of Adapted LLMs – Summary

Context Adaptation

Metric-based:

- Beyond standard metrics: e.g., faithfulness is important!
  - Knowledge conflict, answerability…

LLM-as-Judge:

- Off-the-shelf LLM Judges often do not work well for contextual settings!
  - Need to adapt judges as well

Domain Adaptation

Important aspect:

- Catastrophic forgetting

Comprehensive eval principles:

- Capabilities guided design
- Full coverage: domain x task

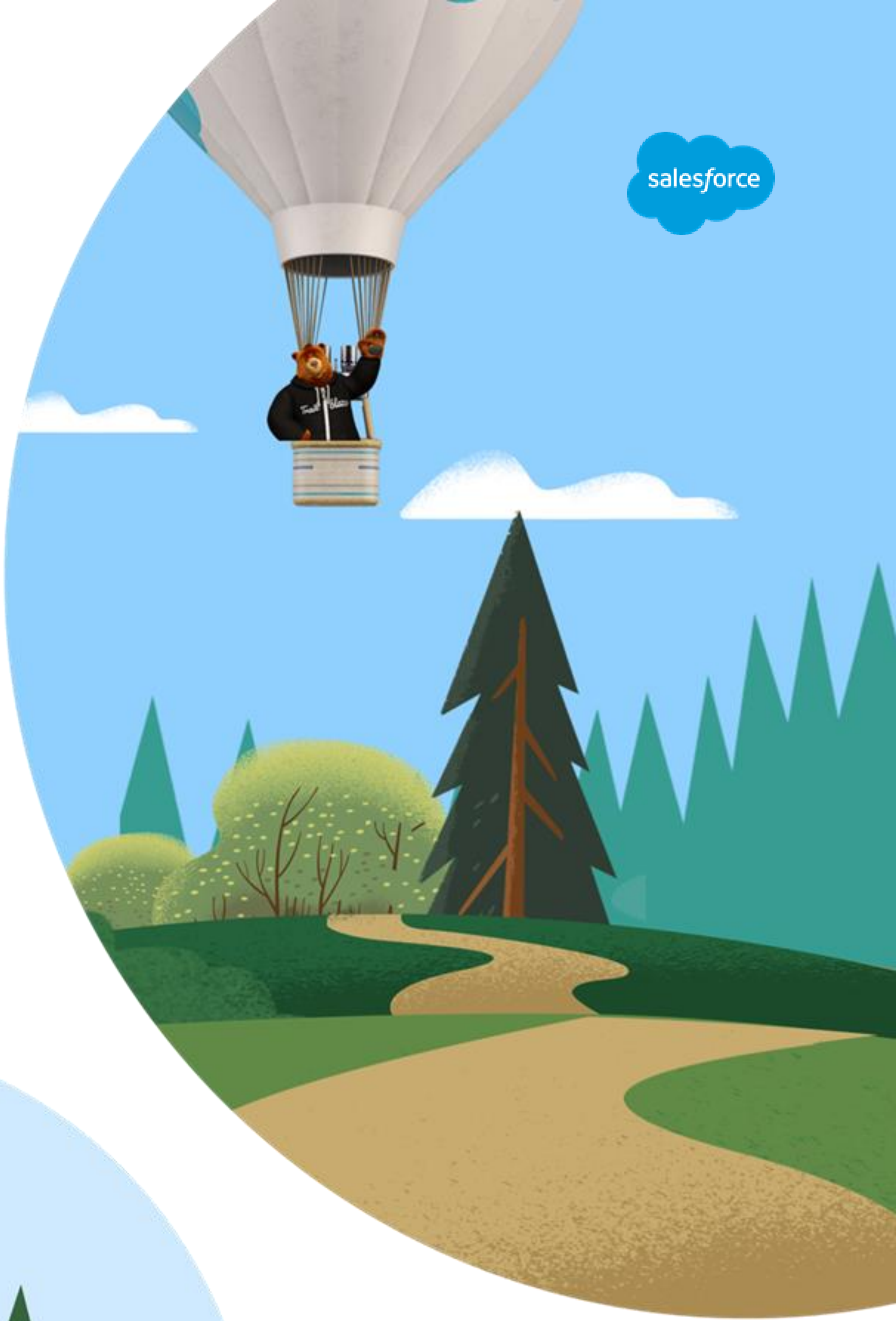
Agenda

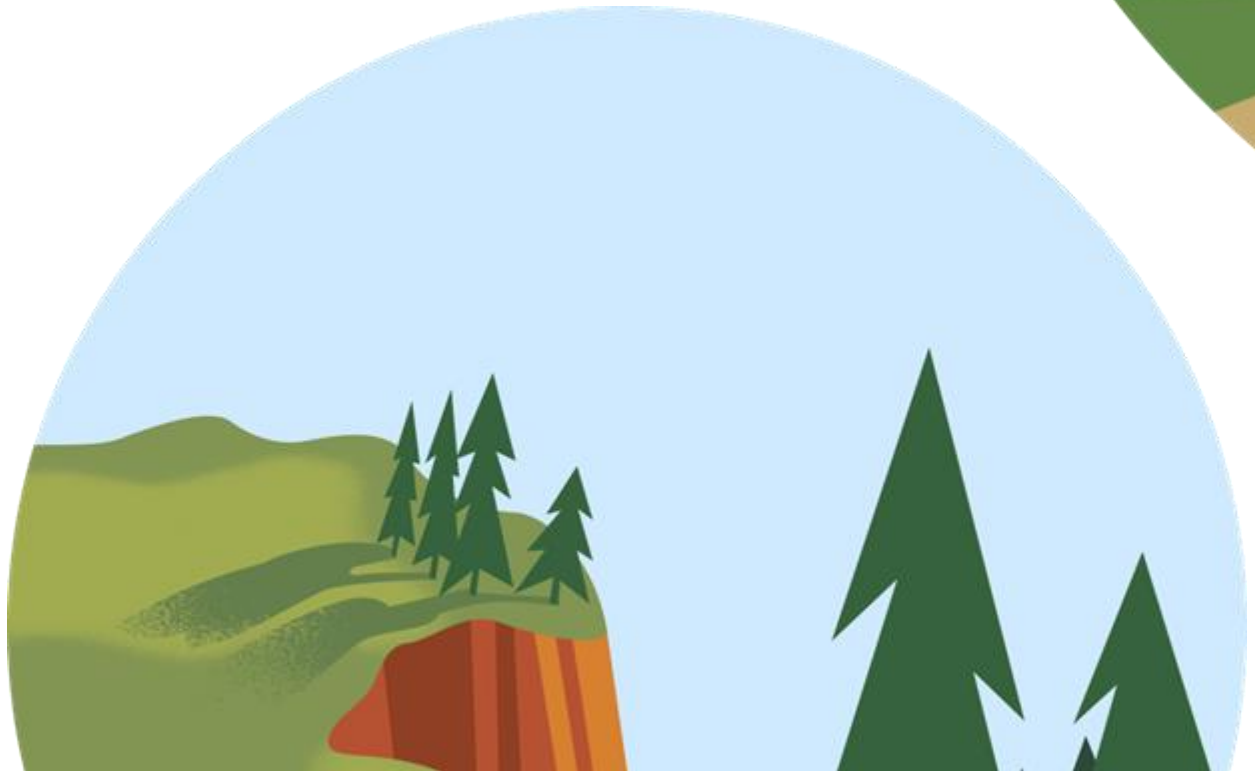
Evaluation and Benchmark

Parametric Knowledge Adaptation ~60min

Semi-Parametric Knowledge Adaptation

Summary, Discussion, QAs





Adaptation - Overview

Model Recipe + Data Recipe = Training Recipe

Model Recipe

Method
Loss, mask, algorithm

Workflow
How methods are connected with each other

Data Recipe

Quality
How to construct better data

Quantity (Scale)
How to synthesize

Training Recipe

Adaptation - Overview

Training Recipe

Data Recipe:
e.g., Supervised data is expensive, how to synthesize more data?

Model Recipe:
e.g., **Hyper-parameters**: What are the important hyper-parameters?

e.g., **Training Workflow**: How to connect with other methods?

Seed Data

Data Acquisition:
e.g., crawling, quality, quantity, filtering…

Data Mixture:
e.g., in-domain, general-domain, …

Data Budget:
e.g., instruction following ~ 1 million; preference learning ~ 1 million (often overlapping with instruction following prompt); reinforcement learning ~ 10-100 thousand

Continual Pre-training (CPT)

CPT – Role

Knowledge Transfer

Improves on new knowledge:

CPT is typically used to inject new knowledge/capability (e.g., long-context adaptation) to the base model and to provide good initialization to the subsequent stages

Prevent Forgetting

Reinforce similar problems:

CPT involves large amount of unsupervised data and could easily cause *catastrophic forgetting* to the base model

CPT – Example Workflow

Seed Data (unsupervised)

*Potentially some modifications (e.g., position embedding modification in long-context adaptation)

CPT – Example Data

Long Text
(e.g. website, books)

No Special Masking

<table>
<tr><th>text
string · lengths
22↔4.72k 73%</th><th>input_ids
sequence · lengths
802↔1.6k 25.4%</th><th>attention_mask
sequence · lengths
802↔1.6k 25.4%</th></tr>
<tr><td><|begin_of_text|>Many or all of the products
featured here are from our partners who compensate
us. This influences which products we write about
and where and how the product appears on a page.
However, this does not influence our evaluations.
Our opinions are our own. Here is a list of our
partners and here's how we make money.
For many shoppers, the retail experience has become
increasingly digital, filled with one-click
purchasing and next-day shipping. But there are
still those among us who love the thrill of
wandering between shops and enjoying an impromptu
try-on session with friends.
If you're a regular at Simon mall properties, the
$0-annual fee Simon Credit Card from Cardless is
worth a look. Its rewards outpace most general-
purpose cards for mall-centered buys, and it boasts
flexibility that store-specific cards often can't
match.
That said, the card comes with a few caveats you
should be aware of to make the most of your
rewards. Here are five things to know about the
Simon® American Express® credit card from Cardless.
» MORE: What is Cardless?
1. It earns 5% cash back at all Simon properties in
the U.S.
Cardholders earn 5% cash back on all in-person
transactions within Simon's nearly 200 U.S. Simon</td><td>[128000, 8607,
477, 682, 315,
279, 3956, 15109,
1618, 527, 505,
1057, 8717, 889,
46794, 603, 13,
1115, 34453, 902,
3956, 584, 3350,
922, 323, 1405,
323, 1268, 279,
2027, 8111, 389,
264, 2199, 13,
4452, 11, 420,
1587, 539, 10383,
1057, 56181, 13,
5751, 18463, 527,
1057, 1866, 13,
5810, 374, 264,
1160, 315, 1057,
8717, 323, 1618,
596, 1268, 584,
1304, 3300, 627,
2520, 1690, 49835,
11, 279, 11040,
3217, 706, 3719,
15098, 7528, 11,
10409, 449, 832,
29218, 23395, 323,</td><td>[1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1,</td></tr>
</table>

CPT – Key Considerations

Training Recipe

Model Recipe:

Hyper-parameters: What are the important hyper-parameters?

Training Workflow: how to connect CPT with other methods (e.g., IT, SPL)

Seed Data

Data Source: Where to get the data?

Data Mixture: What should be included to the CPT data?

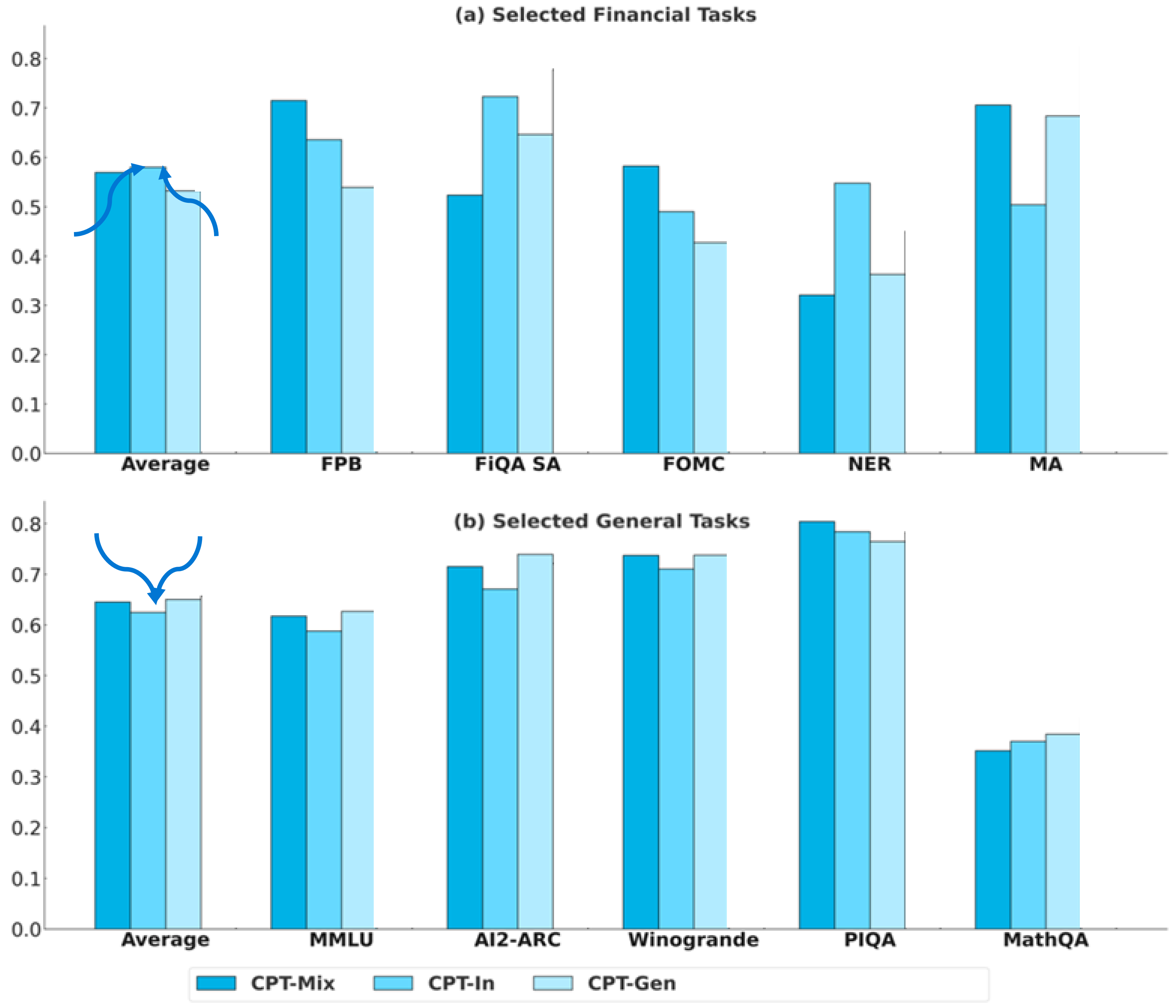
Data Budget: How much data we need?

CPT – Key Ideas

Catastrophic Forgetting (Finance-LLM as an example)

In-domain Data alone → forgetting on general knowledge
(Knowledge forgetting)

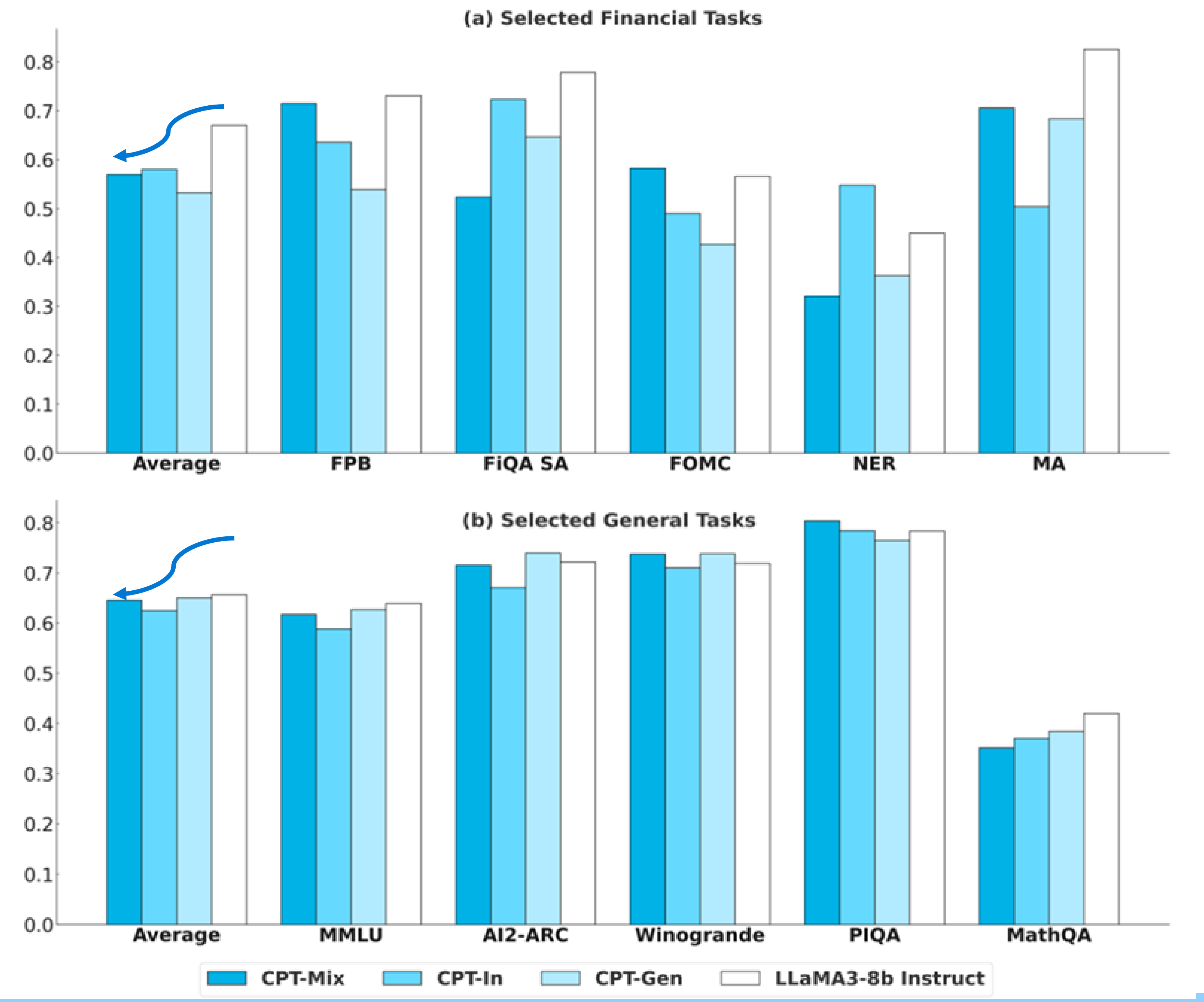
Demystifying Domain-adaptive Post-training for Financial LLMs, Ke et al., 2025



CPT – Key Ideas

Catastrophic Forgetting (Finance-LLM as an example)

CPT alone →
forgetting on general capabilities
(Capabilities forgetting)



base model = instruction-tuned model

Demystifying Domain-adaptive Post-training for Financial LLMs, Ke et al., 2025

We find that even small amounts of replay (1% of the general domain data) mitigate forgetting

Demystifying Domain-adaptive Post-training for Financial LLMs

Zixuan Ke, Yifei Ming, Xuan-Phi Nguy
Salesforce AI
{zixuan.ke,yifei.ming,xnguyen,c
Project Page: https://github.com
Datasets: https://huggingface.cc

Simple and Scalable Strategies to Continually Pre-train Large Language Models

Adam Ibrahim*†⊚
Benjamin Thérien*†⊚
Kshitij Gupta*†⊚
Mats L. Richter †⊚
Quentin Anthony ◊†⊚
Timothée Lesort †⊚
Eugene Belilovsky ‡⊚
Irina Rish †⊚

Fine-tuned Language Models are Continual Learners

Thomas Scialom[1*] Tuhin Chakrabarty[2*] Smaranda Muresan [2]
[1]Meta AI
[2]Department of Computer Science, Columbia University
tscialom@fb.com, tuhin.chakr@cs.columbia.edu, smara@cs.columbia.edu

CPT – Key Ideas

Learn New Knowledge and Mitigate Knowledge Forgetting – Data

Data source for new domain:

Web scrapers (often the largest proportion of data): e.g., Internet

User-provided content (often smaller proportion, but higher-quality): e.g.,. Wikipedia, arXiv,

Open Publishers (often smaller proportion, but higher-quality)**:** e.g., PubMed, Semantic Scholar, Text book

Data source to prevent forgetting (small amount of replay):

Human Verifier Text (small but high-quality)**:** e.g., general supervised tasks

CPT – Key Ideas

Learn New knowledge and Mitigate Knowledge Forgetting – Data

General Domain data
+ In-domain data

Capability	Domain	CPT Dataset	Size	Reference
Concept	General	NaturalInstrution	100,000	Mishra et al. (2022)
		PromptSource	100,000	Bach et al. (2022)
		Math	29,837	Amini et al. (2019b)
		Aqua	97,500	Ling et al. (2017)
		CREAK	10,200	Onoe et al. (2021)
		ESNLI	549,367	Camburu et al. (2018)
		QASC	8,130	Khot et al. (2020)
		SODA	1,190,000	Kim et al. (2022)
		StrategyQA	2,290	Geva et al. (2021)
		UnifiedSKG	779,000	Xie et al. (2022)
		GSM8K	7,470	Cobbe et al. (2021)
		ApexInstr	1,470,000	Huang et al. (2024b)
		DeepmindMath	379,000	Saxton et al. (2019)
		DialogueStudio	1,070,000	Zhang et al. (2023)
	Finance	**Fineweb-Fin**	4,380,000	-
		Book-Fin	4,500	-
Total			10,177,294	

Demystifying Domain-adaptive Post-training for Financial LLMs, Ke et al., 2025

CPT – Key Ideas

Learn New knowledge and Mitigate Capabilities Forgetting – Model

Replay data only addresses the domain knowledge forgetting, but it does not address the capabilities (e.g., instruction-following abilities)

One way is to jointly train CPT and IT to avoid the capabilities forgetting

- Mitigate forgetting
- Encourage transfer (concept learned from CPT naturally shared across tasks)

Demystifying Domain-adaptive Post-training for Financial LLMs

Zixuan Ke, Yifei Ming, Xuan-Phi Nguyen, Caiming Xiong and Shafiq Joty
Salesforce AI Research
{zixuan.ke,yifei.ming,xnguyen,cxiong,sjoty}@salesforce.com
Project Page: https://github.com/SalesforceAIResearch/FinDAP
Datasets: https://huggingface.co/datasets/Salesforce/FinEval

* Another way could be model merging

A SURVEY ON POST-TRAINING OF LARGE LANGUAGE MODELS, Tie et al., 2025

CPT – Key Ideas

Other Tips: Learning Rate, Data Curriculum

Final Recipe for Llama-Fin

Continual Pre-training (CPT) and Instruction Tuning (IT)		
Data	50% CPT, 50% IT	
Curriculum	**Group 1**	CPT: 50% Domain-specific Text (Web and book), 50% General text (verfiable text)
		IT: 20% Domain-specific tasks, 80% General tasks
	Group 2	CPT: Group 1 data + domain-specific books
		IT: Group1 + Exercises extracted from books
Steps		Group 1: 3.84B tokens; Group 2: 1.66B tokens
		(8,000 context length, 16 A100)
Model	**Intialization**	Llama3-8b-instruct
	Attention	CPT: full attention with cross-docuemnt attention masking
		IT: full attention with instruction mask-out and cross-docuemnt attention masking
Optim.		AdamW (weight decay = 0.1, β_1=0.9, β_2=0.95)
	LR	Group 1: 5e-6 with 10% warmup; Group 2: 5e-6 with 50% warmup
	Batch size	128K tokens
Stop Cri.	Loss of development set stops decreasing ($\approx$ 1 epoch)	

Demystifying Domain-adaptive Post-training for Financial LLMs, Ke et al., 2025

CPT – Key Ideas

Other Tips: Learning Rate, Data Curriculum

Continued Long-context Training		
Data	30% code repos, 30% books, 3% textbooks, 37% ShortMix	
	ShortMix:	27% FineWeb-Edu, 27% FineWeb, 11% Wikipedia, 11% StackExchange, 8% Tulu-v2, 8% OpenWebMath, 8% ArXiv
Length Curriculum	Stage 1 (64K):	Code repos, books, and textbooks at length 64K
	Stage 2 (512K):	Code repos: 50% at length 512K, 50% at length 64K Books: 17% at length 512K, 83% at length 64K Textbooks at length 512K
Steps	Stage 1: 20B tokens (2.2K H100 hours), Stage 2: 20B tokens (12.2K H100 hours)	
Model	Initialization: RoPE: Attention:	Llama-3-8B-Instruct (original RoPE base freq. 5×10^5) Stage 1: 8×10^6, Stage 2: 1.28×10^8 Full attention with cross-document attention masking
Optim.	AdamW (weight decay = 0.1, $\beta_1 = 0.9$, $\beta_2 = 0.95$)	
	LR: Batch size:	$1e-5$ with 10% warmup and cosine decay to $1e-6$, each stage 4M tokens for stage 1, 8M tokens for stage 2

How to Train Long-Context Language Models (Effectively), Gao et al., 2025

CPT – Key Ideas

Other Tips: Learning Rate, Data Curriculum

Rules of thumb for continual pre-training

Caveat—The following guidelines are written to the best of our *current knowledge*.

Learning rate schedule:

- If the learning rate was cosine-decayed from a large value η_{max} to a small value η_{min} during pre-training on the initial dataset, the following guidelines can help to continually pre-train your model:
  - Re-warming and re-decaying the learning rate from $\mathcal{O}(\eta_{max})$ to $\mathcal{O}(\eta_{min})$ improves adaptation to a new dataset, e.g. compared to continuing from small learning rates $\mathcal{O}(\eta_{min})$.
  - Decreasing the schedule's maximum learning rate can help reduce forgetting, whereas increasing it can improve adaptation.
- Infinite LR schedules are promising alternatives to cosine decay schedules. They transition into a high constant learning rate across tasks, helping prevent optimization-related forgetting by avoiding re-warming the LR between tasks. They also avoid committing to a specific budget of tokens as a final exponential decay can be used to train the model to convergence at any point during training.

Simple and Scalable Strategies to Continually Pre-train Large Language Models, Ibrahim et al., 2024

CPT – Key Ideas

Other Tips: Learning Rate, Data Curriculum

Recipe

- Start with a data distribution that is similar to the pretraining set but places larger weight on high quality sources before transitioning to a second distribution that incorporates QA data and upweights sources in areas of model weakness.
- The learning rate schedule should start from η_{min} of the pretrained model and decay with cosine annealing to $\frac{\eta_{min}}{100}$.
- The switch between data distribution should occur at $\frac{\eta_{max}}{5}$ in the learning rate schedule.

Reuse, Don't Retrain: A Recipe for Continued Pretraining of Language Models, Parmar et al., 2024

CPT – Key Ideas Summary

Training Recipe

Model Recipe:
Learning rate schedule
Data curriculum

Jointly training CPT and IT have been shown to be effective

Seed Data

Data Mixture: Wide representative and filtering is needed

Data Budget:
New Knowledge ~ 5 million
Prevent Forgetting ~ 5 million

* Filtering can be complicated and involved different components (e.g., decontamination..).

Instruction Tuning

IT – Role

Chat Style Adaptation

Adapt base model to **specific style of input** for chat interactions.

Chat Template Adaptation

Ability to include **system prompts, multi-turn dialogues,** and other **chat templates.**

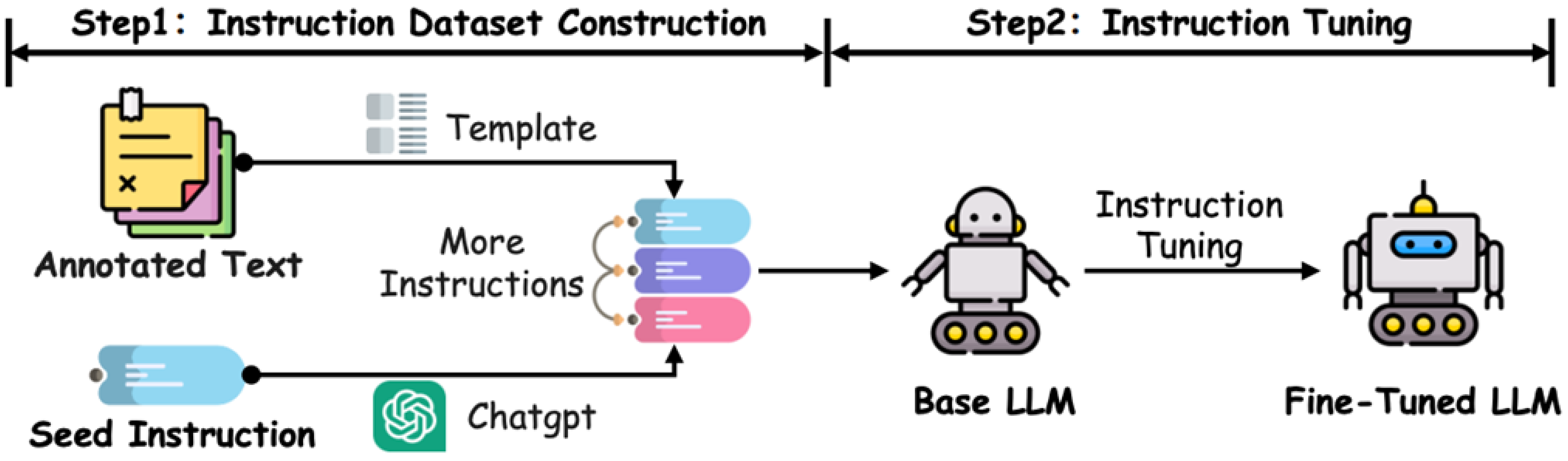
Special tokens

```
<|system|>
You are a helpful assitant
<|end|>
<|user|>
How many helicopters can you eat?
<|end|>
<|assistant|>
{Answer goes here}
```

System prompt

Multi-turn dialogue

IT – Example Workflow



A SURVEY ON POST-TRAINING OF LARGE LANGUAGE MODELS, Tie et al., 2025

IT – Example Data

Chat Format
Special Label Masking
Packing

<table>
<tr><th>text
string · lengths
723–1.58k 97.3%</th><th>input_ids
sequence · lengths
122–377 97.3%</th><th>attention_mask
sequence · lengths
122–377 97.3%</th><th>labels
sequence · lengths
122–377 97.3%</th><th>packed_length
int64
122–376 97.3%</th></tr>
<tr><td><|begin_of_text|>
<|start_header_id|>user<|end_header_id|>

Given phrases that describe the relationship
between two words/phrases as options, extract
the word/phrase pair and the corresponding
lexical relationship between them from the
input text. The output format should be
"relation1: word1, word2; relation2: word3,
word4". Options: product/material produced,
manufacturer, distributed by, industry,
position held, original broadcaster, owned
by, founded by, distribution format,
headquarters location, stock exchange,
currency, parent organization, chief
executive officer, director/manager, owner
of, operator, member of, employer,
chairperson, platform, subsidiary, legal
form, publisher, developer, brand, business
division, location of formation, creator.
Text: That's a 7% deal down there where a
Mexican co-packer puts Mexican fruit, very
high quality, the same quality standards of
the fruit that we pull out of California and
Arizona, into a Limoneira box for sales.
<|eot_id|>
<|start_header_id|>assistant<|end_header_id|>

headquarters_location: Limoneira,
California<|eot_id|><|end_of_text|></td><td>[128000, 128006,
882, 128007, 271,
22818, 32847, 430,
7664, 279, 5133,
1990, 1403, 4339,
90121, 27663, 439,
2671, 11, 8819,
279, 3492, 14,
28810, 6857, 323,
279, 12435, 78686,
5133, 1990, 1124,
505, 279, 1988,
1495, 13, 578,
2612, 3645, 1288,
387, 330, 23013,
16, 25, 3492, 16,
11, 3492, 17, 26,
12976, 17, 25,
3492, 18, 11,
3492, 19, 3343,
14908, 25, 2027,
15175, 9124, 11,
14290, 11, 4332,
555, 11, 5064, 11,
2361, 5762, 11,
4113, 60983, 11,
13234, 555, 11,
18538, 555, 11,
8141, 3645, 11,
26097, 3813, 11,
5708, 9473, 11,
11667, 11, 2748,
7471, 11, 10388,
11145, 9640, 11,
7690, 14, 13600,
11, 6506, 315, 11,</td><td>[1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1]</td><td>[-100, -100,
-100, -100, -100,
-100, -100, -100,
-100, -100, -100,
-100, -100, -100,
-100, -100, -100,
-100, -100, -100,
-100, -100, -100,
-100, -100, -100,
-100, -100, -100,
-100, -100, -100,
-100, -100, -100,
-100, -100, -100,
-100, -100, -100,
-100, -100, -100,
-100, -100, -100,
-100, -100, -100,
-100, -100, -100,
-100, -100, -100,
-100, -100, -100,
-100, -100, -100,
-100, -100, -100,
-100, -100, -100,
-100, -100, -100,
-100, -100, -100,
-100, -100, -100,
-100, -100, -100,
-100, -100, -100,
-100, -100, -100,
-100, -100, -100,
-100, -100, -100,
-100, -100, -100,
-100, -100, -100,
-100, -100, -100,
-100, -100, -100,
-100, -100, -100,</td><td>209</td></tr>
</table>

IT – Key Considerations

Training Recipe

Data Recipe:

Supervised data is expensive, how to synthesize more data?

Model Recipe:

How should the loss and masking different from CPT?

Training Workflow: how to connect with other methods

Seed Data

Data Source: Where to get the data?

Data Mixture: What should be included in the IT data?

Data Budget: How many data we need?

IT – Key Ideas

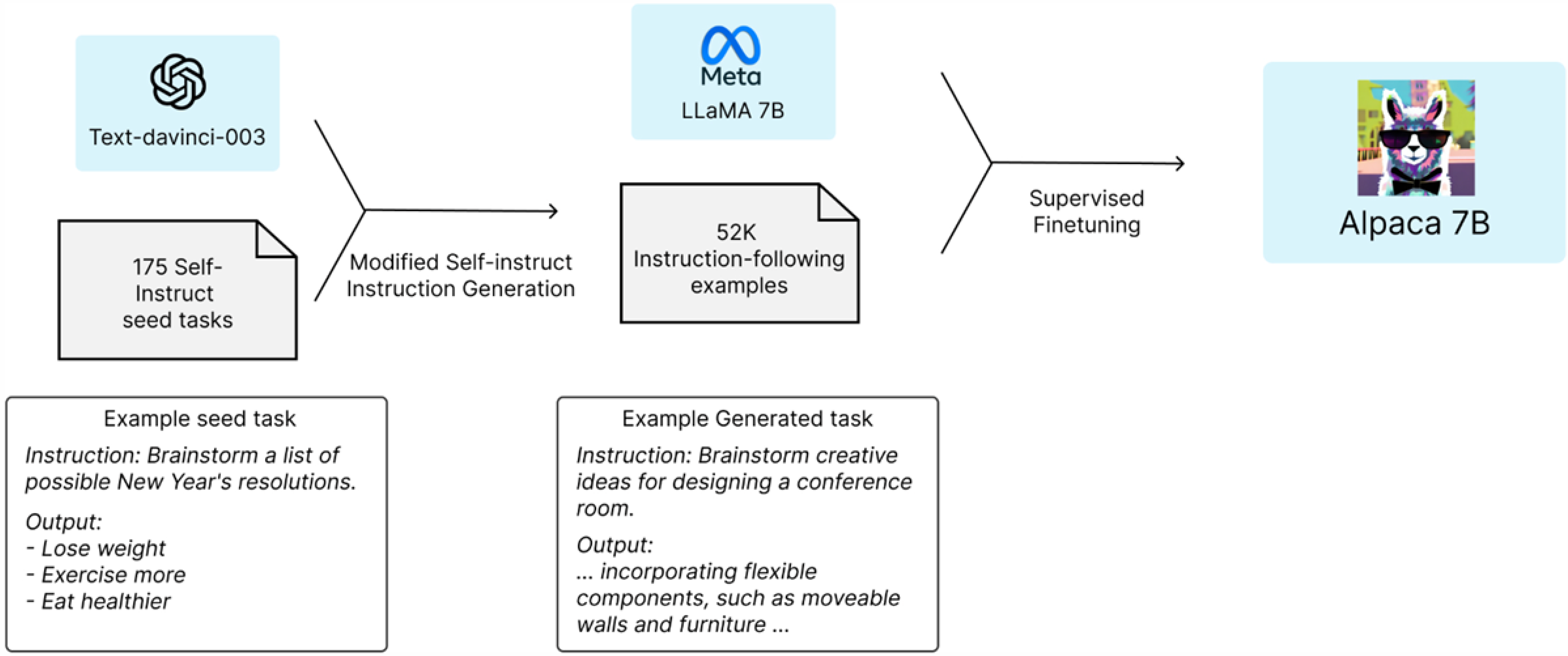
Self-instruct / Synthetic data

Seed: N high-quality (often human) prompts

Ask a strong LLM: Create a modified version of these instructions

Generate completions with another (or same) strong LLM.

Results: easily 10x more synthetic training data



Alpaca: A Strong, Replicable Instruction-Following Model, Taori et al., 2023
SELF-INSTRUCT: Aligning Language Models with Self-Generated Instructions, Wang et al., 2022

IT – Key Ideas

Packing and Label Masking

https://github.com/MeetKai/functionary/blob/main/functionary/train/packing

IT – Key Ideas

Packing and Label Masking

Disabling cross-document attention. Ding et al. (2024a) show that masking out attention across document boundaries improve model performance and this was also used during Llama-3 pre-training (Dubey et al., 2024). In §B.2, we show that disabling cross-document attention in continued training benefits both the short and long-context performance. Disabling cross-document attention can also result in higher training throughput, which we describe in more detail in §A.3.

Papers show that packing is helpful

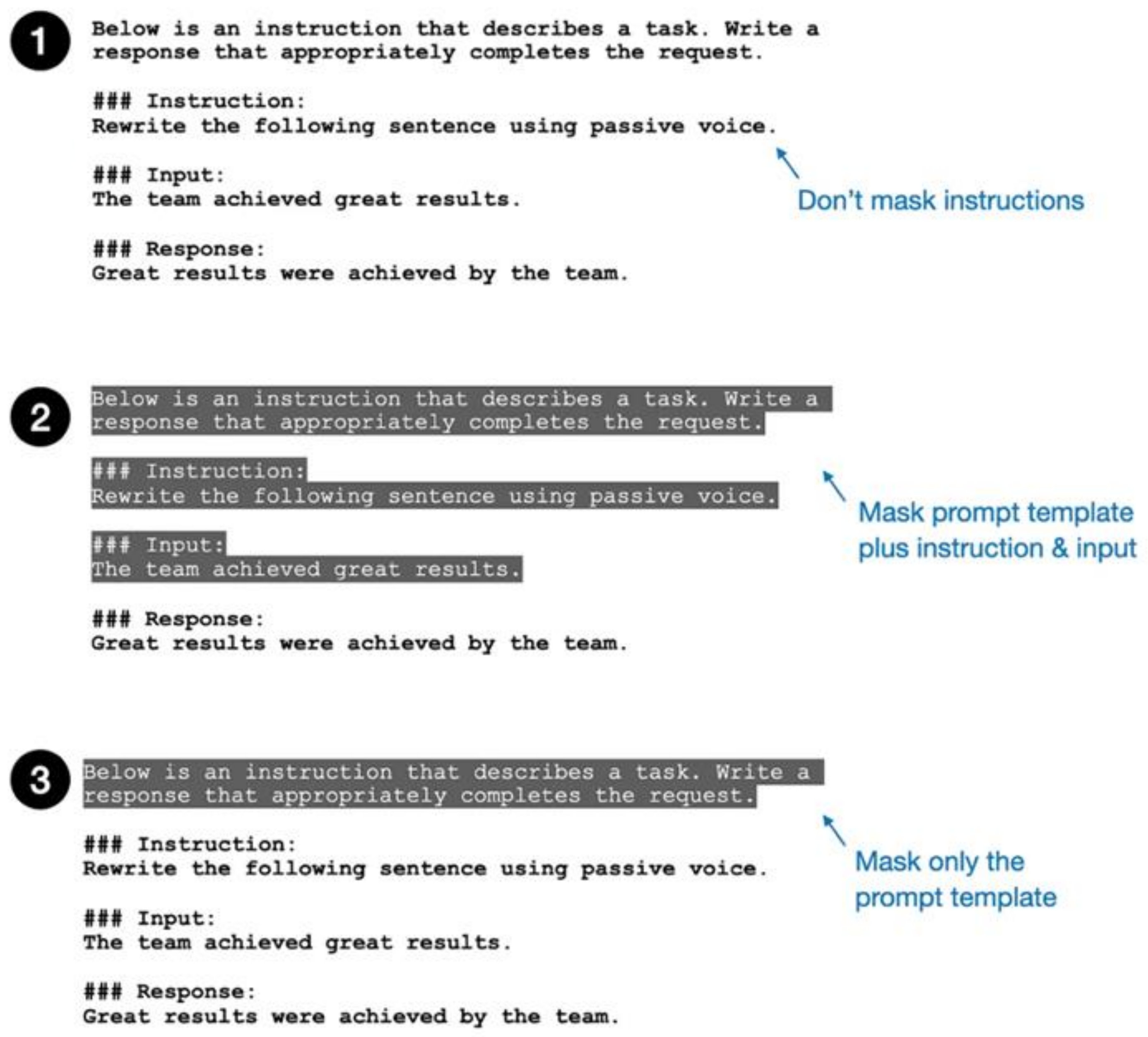
Packing Packing optimizes the training efficiency by grouping sequences of varying lengths into a single long sequence without requiring any padding. This technique, commonly used in LLM pre-training, is now also utilized in instruction-based supervised fine-tuning, as implemented by models like Zephyr (Tunstall et al., 2023b)[4].

How to Train Long-Context Language Models (Effectively), Gao et al., 2025
LIONs: An Empirically Optimized Approach to Align Language Models, Yu et al., 2024

IT – Key Ideas

Packing and Label Masking

Masking the tokens of the instruction by setting the token labels of the instructions to -100



https://www.linkedin.com/pulse/llm-research-insights-instruction-masking-new-lora-raschka-phd-7p1oc

IT – Key Ideas

Packing and Label Masking

RQ1: What is the role of DAPT and SFT in post-training?
- DAPT uses next-token prediction, while SFT needs instruction masking added. §5.1
- Both DAPT and SFT contribute to improvements. §5.2
- Joint training with DAPT and SFT yields better results than sequential training. §5.3

Papers show that label masking is helpful

Loss Masking The standard language model training computes loss across all tokens in a sequence. Loss masking, however, ignores loss computation on tokens that are not output tokens like user instructions. It prevents the model from learning irrelevant information, alleviating catastrophic forgetting and overfitting.

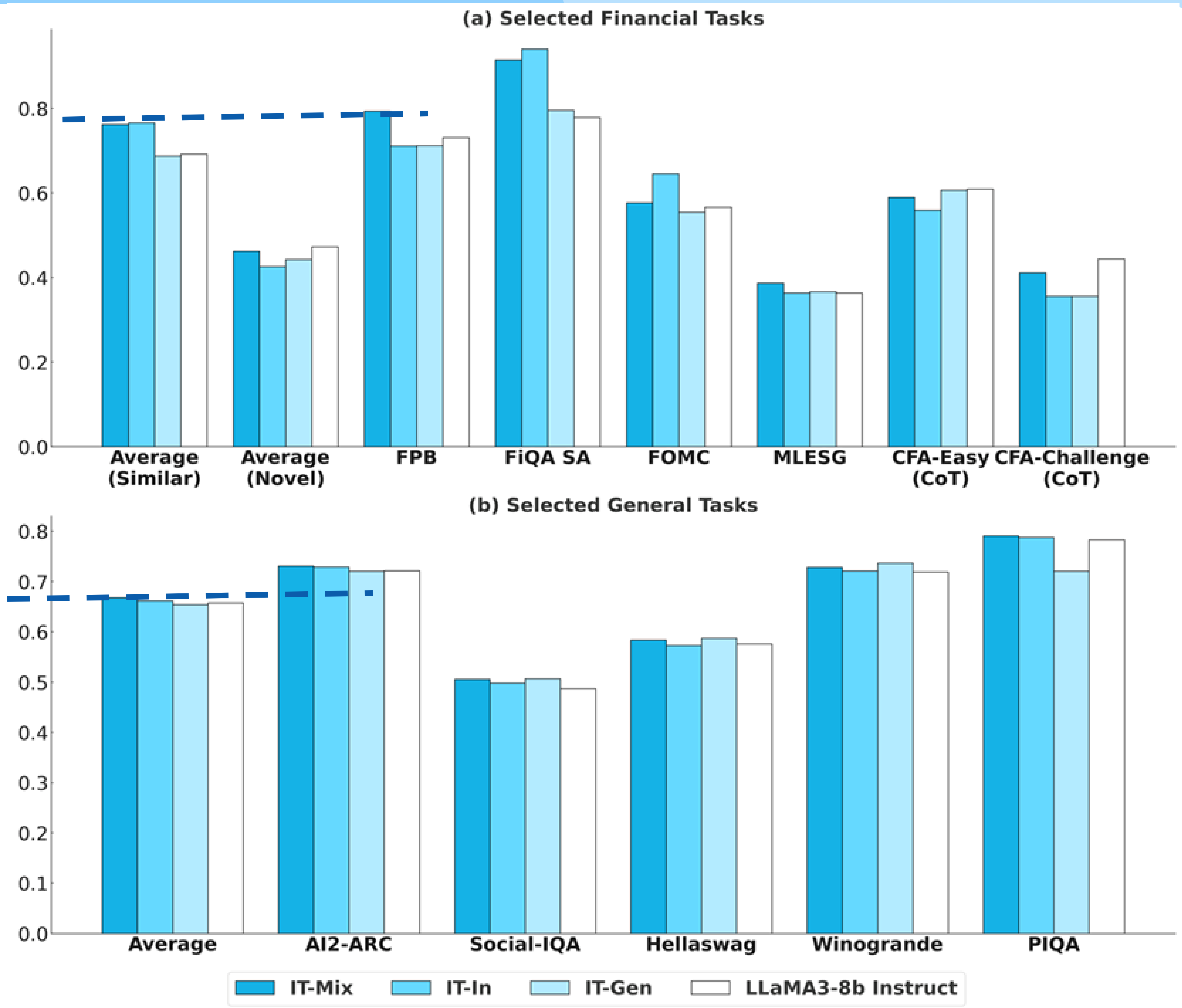
Demystifying Domain-adaptive Post-training for Financial LLMs, Ke et al., 2025
LIONs: An Empirically Optimized Approach to Align Language Models, Yu et al., 2024

IT – Key Ideas

Task Generalization

Forgetting is less a problem

Task generalization is the main issue.



Demystifying Domain-adaptive Post-training for Financial LLMs, Ke et al., 2025

IT – Key Ideas

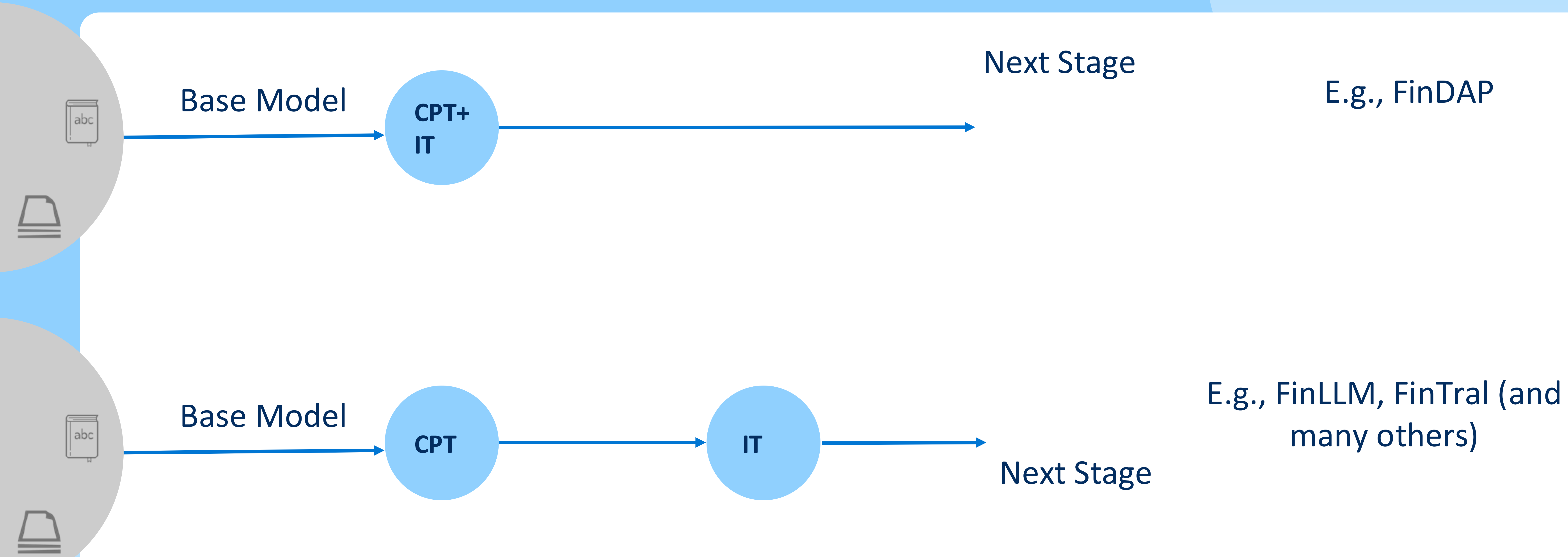
Task Generalization

A wide variety of representative task to promote the task generalization

Capability	Domain	Task	IT Dataset	Size	Reference
Tasks	Finance	Relation Cls.	FingptFinred	27,600	Sharma et al. (2022)
		NER	FingptNERCls	13,500	Yang et al. (2023)
			FingptNER	511	Alvarado et al. (2015)
		Headline Cls.	FingptHeadline	82,200	Sinha et al. (2020)
		Sentiment Cls.	SentimentCls	47,600	Yang et al. (2023)
			SentimentTra	76,800	Yang et al. (2023)
		Summariz.	TradeTheEvent	258,000	Zhou et al. (2021)
IF/Chat	General	IF/Chat	SelfInstruct	82,000	Wang et al. (2022)
			SlimOrca	518,000	Lian et al. (2023)
			UltraChat	774,000	Ding et al. (2023)
			ShareGPT	100,000	Link
	Finance	QA	FinanceInstruct	178,000	Link
			FingptConvfinqa	8,890	Chen et al. (2022)
			FlareFinqa	6,250	Chen et al. (2021)
			FlareFiqa	17,100	Yang et al. (2023)
Reasoning	Math	QA	OrcaMath	200,000	Mitra et al. (2024)
			MetaMathQA	395000	Yu et al. (2023)
			MathInstruct	262,000	Yue et al. (2023)
	Code	QA	MagicodeInstruct	111,000	Luo et al. (2023)
	Finance	CFA Exam	**Exercise**	2,950	-
Total				3,161,401	

IT – Key Ideas

Training Workflow



FinDAP: Demystifying Domain-adaptive Post-training for Financial LLMs, Ke et al., 2025
FinTral: A Family of GPT-4 Level Multimodal Financial Large Language Models, Bhatia et al., 2024
FinLLM: Open-FinLLMs: Open Multimodal Large Language Models for Financial Applications, Huang et al., 2024

IT – Key Ideas Summary

Training Recipe

Data Recipe:
- **Synthetic data** (e.g., self-instruct)

Model Recipe:
- **Packing and Loss Mask**
- **Training Workflow** (e.g., CPT → IT, CPT+IT)

Synthetic data = text generated by LLM

Seed Data

Data Mixture: A wide variety of representative to promote task generalization

Data Budget ~ 1 Million

Supervised Preference Learning

SPL – Role

salesforce

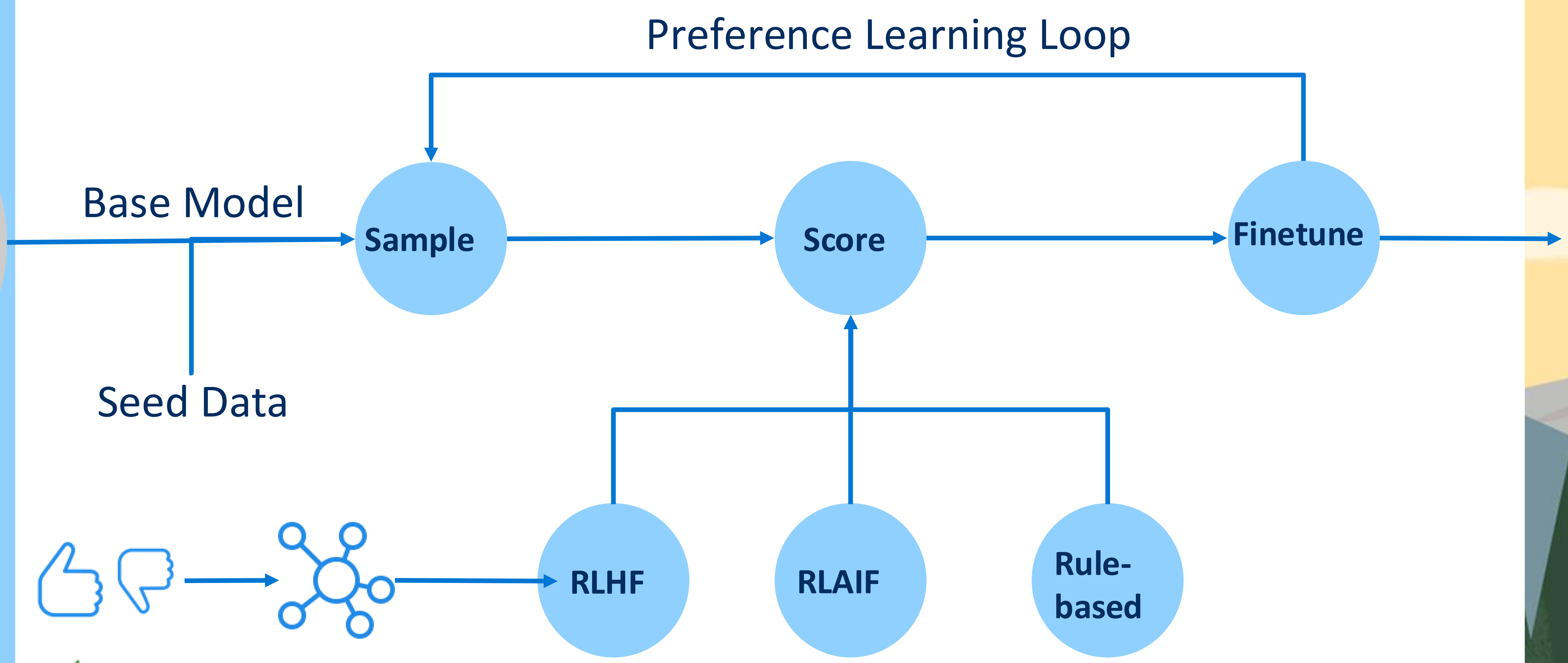
Style and Chat

Stronger training influence for style and chat capability

More Capabilities

Continue building capabilities from instruction-tuned model, e.g., reasoning

SPL – Example Workflow



SPL – Key Considerations

Training Recipe

Data Recipe: e.g., How to construct preference

Model Recipe:

- **Algorithm**: How to optimize the preference reward?
- **Training Workflow**: how to connect with other methods

Seed Data

Data Source: Where to get the data?

Data Mixture: What should be included in the PL data?

Data Budget: How many data we need?

SPL – Key Ideas

DPO – Goal

$$\max_{\pi_\theta} \mathbb{E}_{x\sim\mathcal{D},y\sim\pi_\theta(y|x)}\left[r_\phi(x,y)\right] - \beta\mathbb{D}_{\mathrm{KL}}\left[\pi_\theta(y \mid x) \mid\mid \pi_{\mathrm{ref}}(y \mid x)\right]$$

Optimize "reward" inspired by human preferences

Constraint the model to not trust the reward too much (preferences are hard to model)

Main Questions:

1. **How to implement the reward?**
2. **How to optimize the reward?**

SPL – Key Ideas

DPO – Preference / Reward modeling

Chosen Completion

Prompt

Scores from optimal reward model

$$p^*(y_1 \succ y_2 \mid x) = \frac{\exp\left(r^*(x, y_1)\right)}{\exp\left(r^*(x, y_1)\right) + \exp\left(r^*(x, y_2)\right)}.$$

Rejected Completion

Key Idea: Probability ∝ Reward

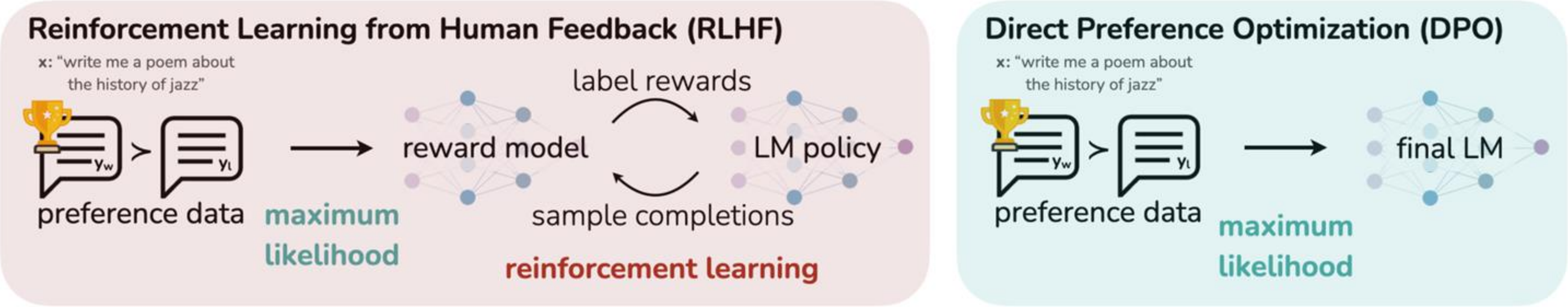
Obtaining point-wise Scalar reward of how good response is hard, but pairwise preference is easier and works!

SPL – Key Ideas

DPO

If we just use gradient ascent on the equation

With some math, we get: Direct Preference Optimization (DPO)



Direct Preference Optimization: Your Language Model is Secretly a Reward Model, Rafailov et al., 2023

SPL – Key Ideas

RLAIF

Human Preferences (RLHF) vs. LLM-as-a-judge (RLAIF)

Both source of preference data are used extensively

In Frontier Labs:

Human data used extensively as foundation

Synthetic data used to enhance behaviors (e.g., Constitutional AI)

In Open Research:

Synthetic data dominates (due to price)

Constitutional AI: Harmlessness from AI Feedbackl, Bai et al., 2022

SPL – Key Ideas

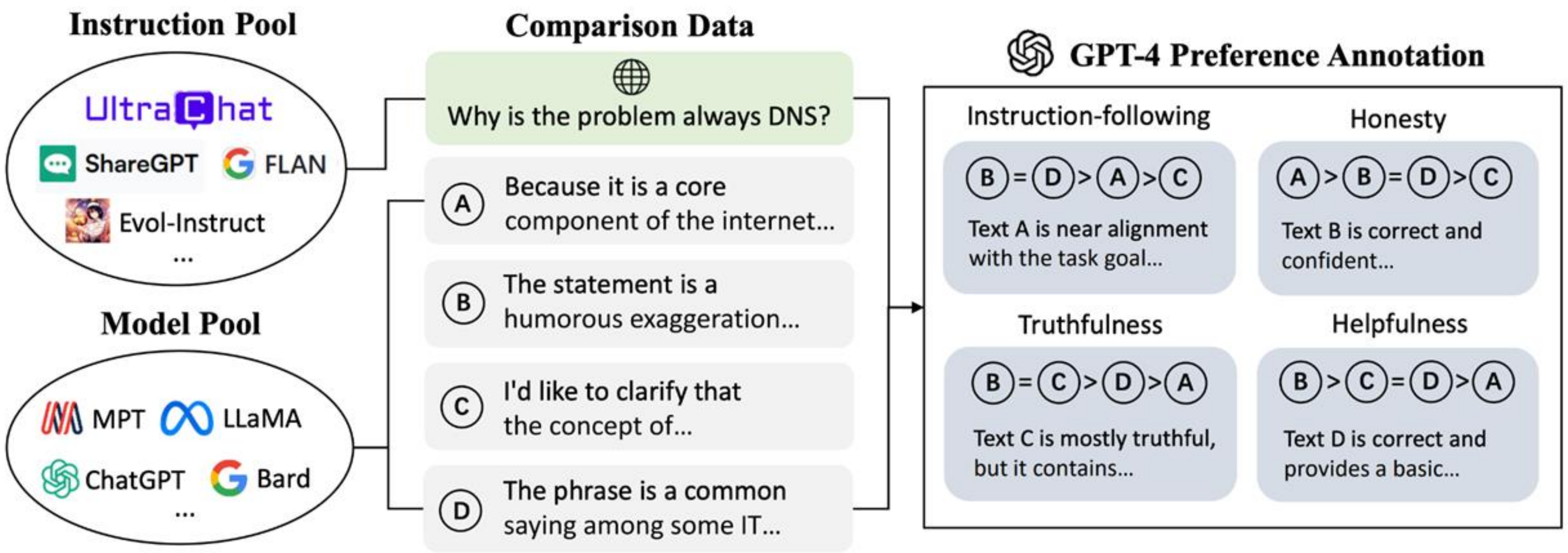
A Leading Synthetic Preference Method–UltraFeedback

Key aspects

Diverse model pool for completions

Diverse prompt pool

On-policy generations from l checkpoints



UltraFeedback: Boosting Language Models with Scaled AI Feedback, Cui et al., 2024

SPL – Key Ideas

Representative work with DPO – Zephyr, TuLU 70B….

First model makes a splash with DPO

Fine-tune from Mistral 7b with UltraFeedback Datasets

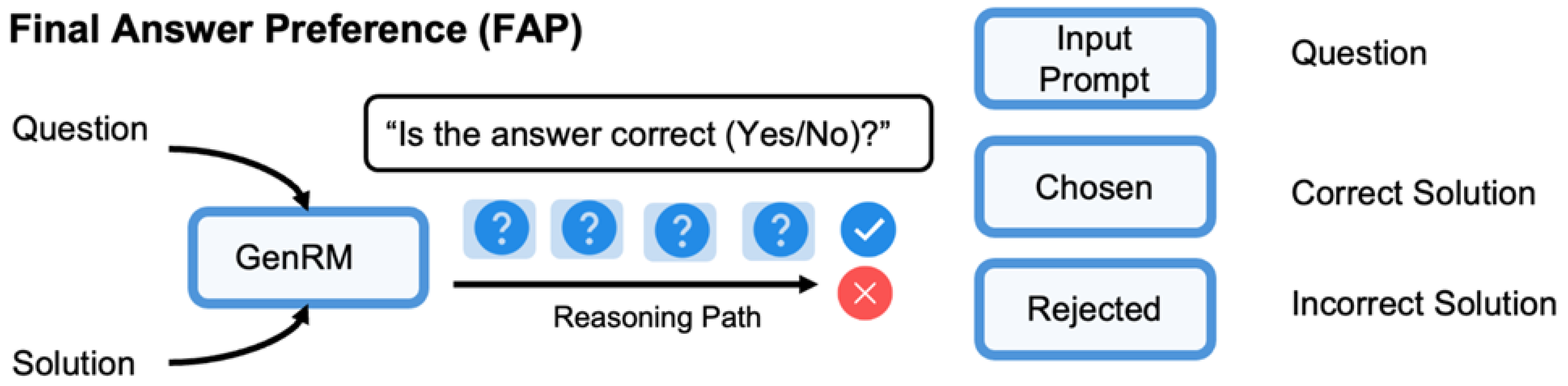
Low learning rate (~5E-7) is good for DPO

Zephyr: Direct Distillation of LM Alignment, Tunstall, et al., 2023

SPL – Key Ideas

Synthesize Preference Data Focused on **Intermediate Preference**

Final outcome preference



Demystifying Domain-adaptive Post-training for Financial LLMs, Ke et al., 2025

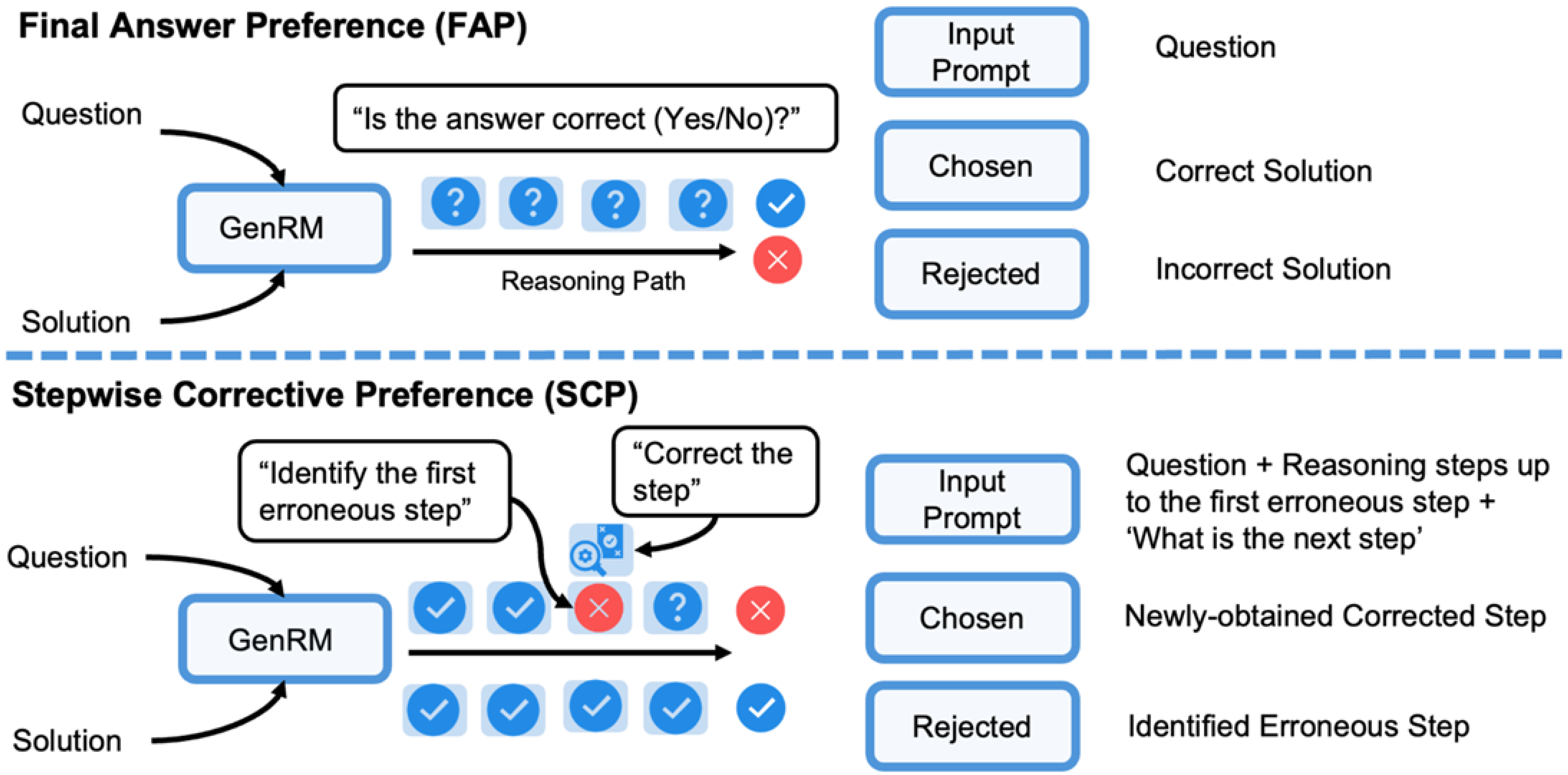
SPL – Key Ideas

Synthesize Preference Data Focused on **Intermediate Preference**

Final outcome preference

Intermediate outcome preference

Identify and rectify the first erroneo step



Demystifying Domain-adaptive Post-training for Financial LLMs, Ke et al., 2025

SPL – Key Ideas Summary

Training Recipe

Data Recipe: Preference construction is often from diverse source (e.g., instruction pool, model pool) and cover fine-grained information (e.g., intermediate preference)

Model Recipe:

- **Algorithm**: most popular: DPO
- **Training Workflow**: usually after CPT and IT

Seed Data

Data Source: often partial overlapping with IT

Data Mixture: Can be large scale (e.g., Math, Logic, Code, Science, Reasoning..)

Data Budget: ~ 1 million

Coffee Break
(30 Min)

Reinforcement Learning

RL – Role

Beyond Human/AI Preference

RL as a training objective, learning from experience of interacting of the environment

Recently show high-effectiveness

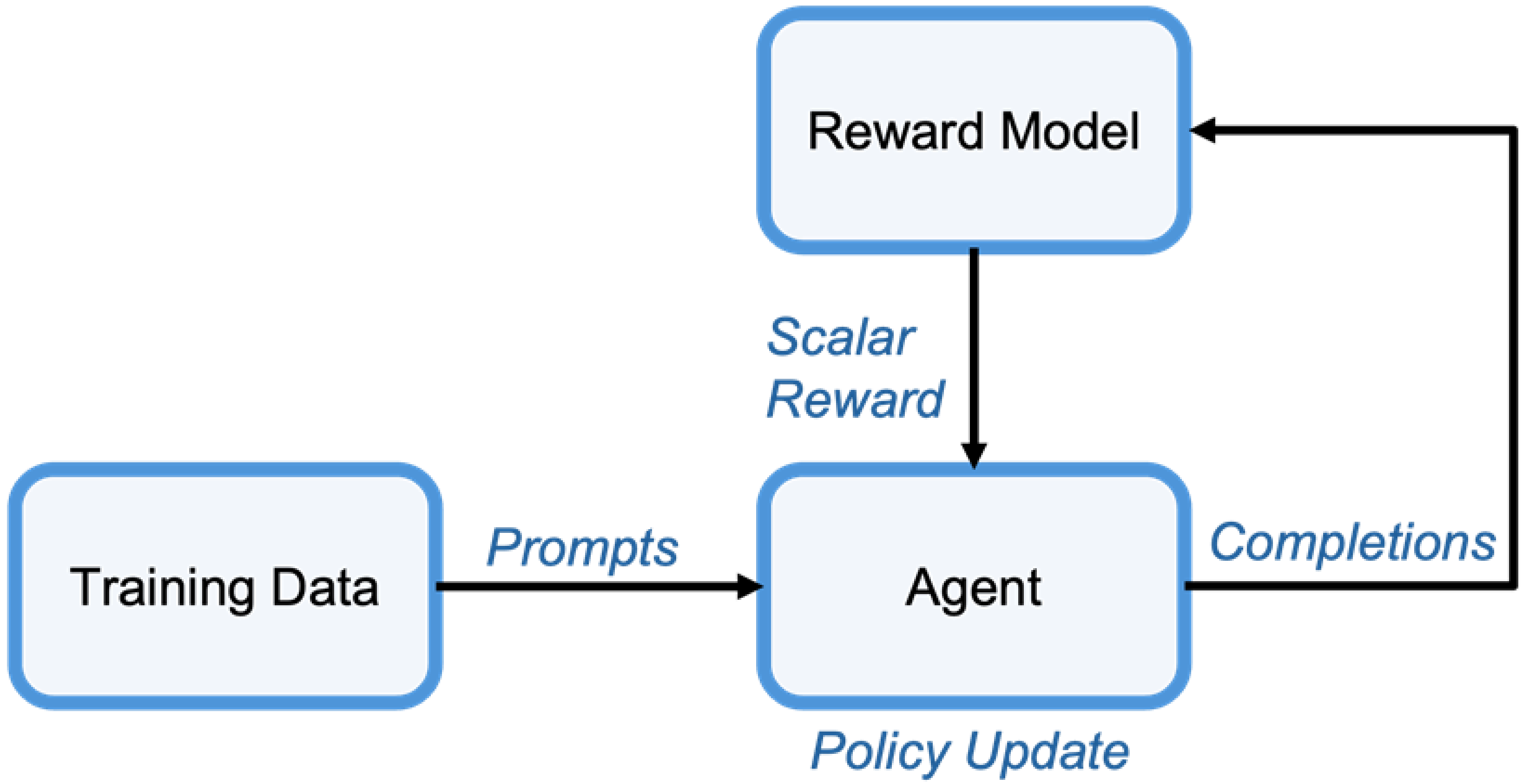
Learn from Mistakes

RL methods naturally see both correct and a wide range of incorrect solutions.

This means they can:

improve targeted capabilities **without** degradation on other out-of-domain capabilities

RL – Example Workflow



RL – Key Considerations

Training Recipe

Model Recipe:

Algorithm: How to optimize the reward effectively and efficiently?

Training Workflow: how to connect with other methods

Seed Data

Data Source: Where to get the data?

Data Mixture: What should be included in the RL data?

Data Budget: How many data we need?

RL – Key Ideas

From DPO to RL

$$\max_{\pi_\theta} \mathbb{E}_{x\sim\mathcal{D},y\sim\pi_\theta(y|x)}\left[r_\phi(x,y)\right] - \beta\mathbb{D}_{\mathrm{KL}}\left[\pi_\theta(y \mid x) \,||\, \pi_{\mathrm{ref}}(y \mid x)\right]$$

Optimize "reward" inspired by human preferences

Constraint the model to not trust the reward too much (preferences are hard to model)

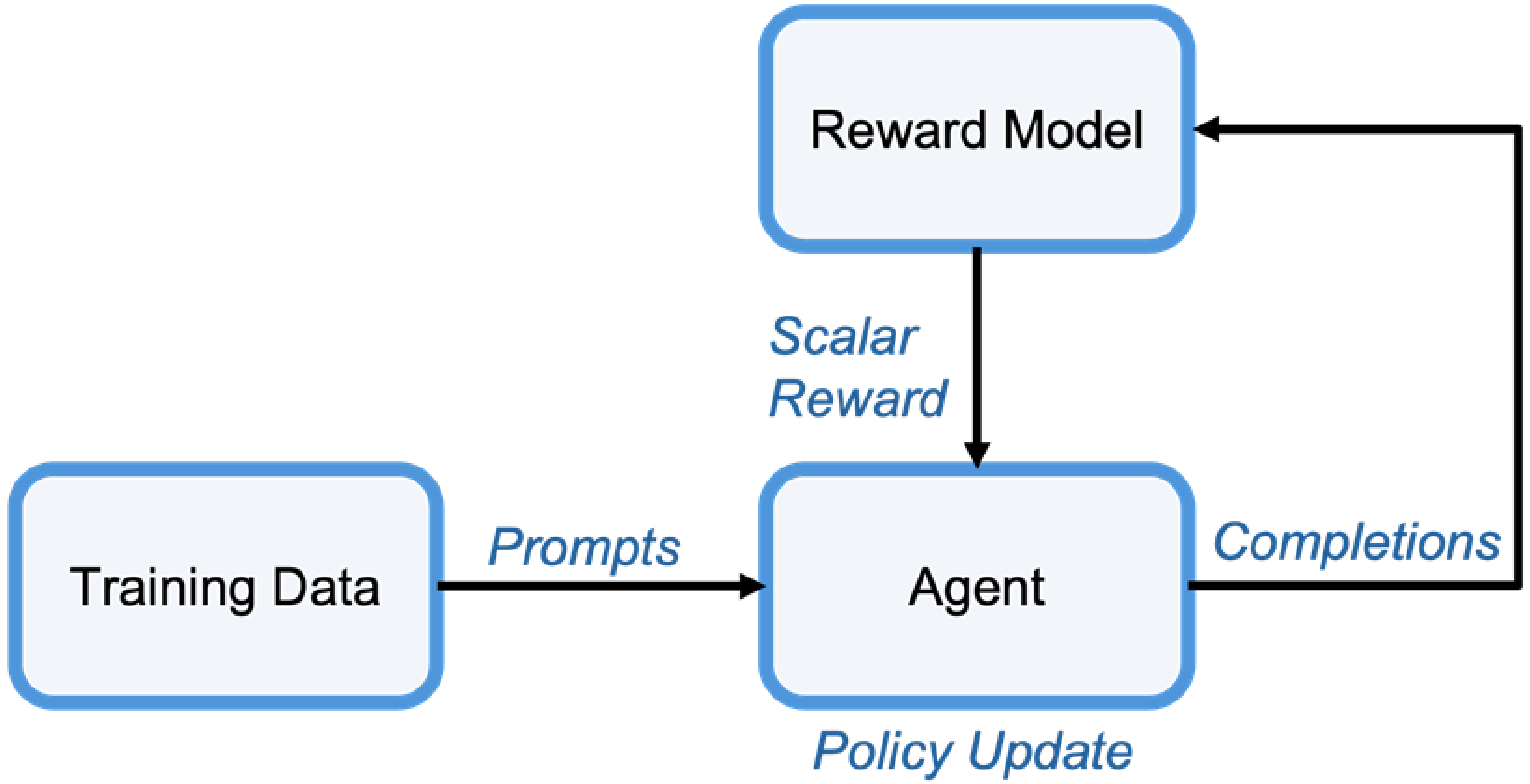
Main Questions:

1. **How to implement the reward?**
2. **How to optimize the reward?**

RL – Key Ideas

From DPO to RL

What if we choose not to use pairwise preference but still rely on scalar reward

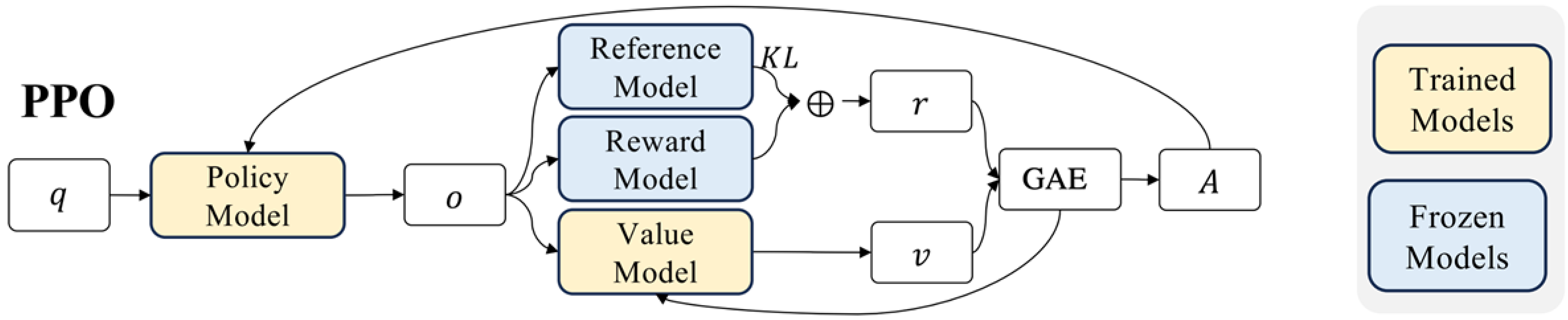


RL – Key Ideas

PPO

One popular method is PPO

(effective but expensive: 4 copies of model)



Proximal Policy Optimization Algorithms

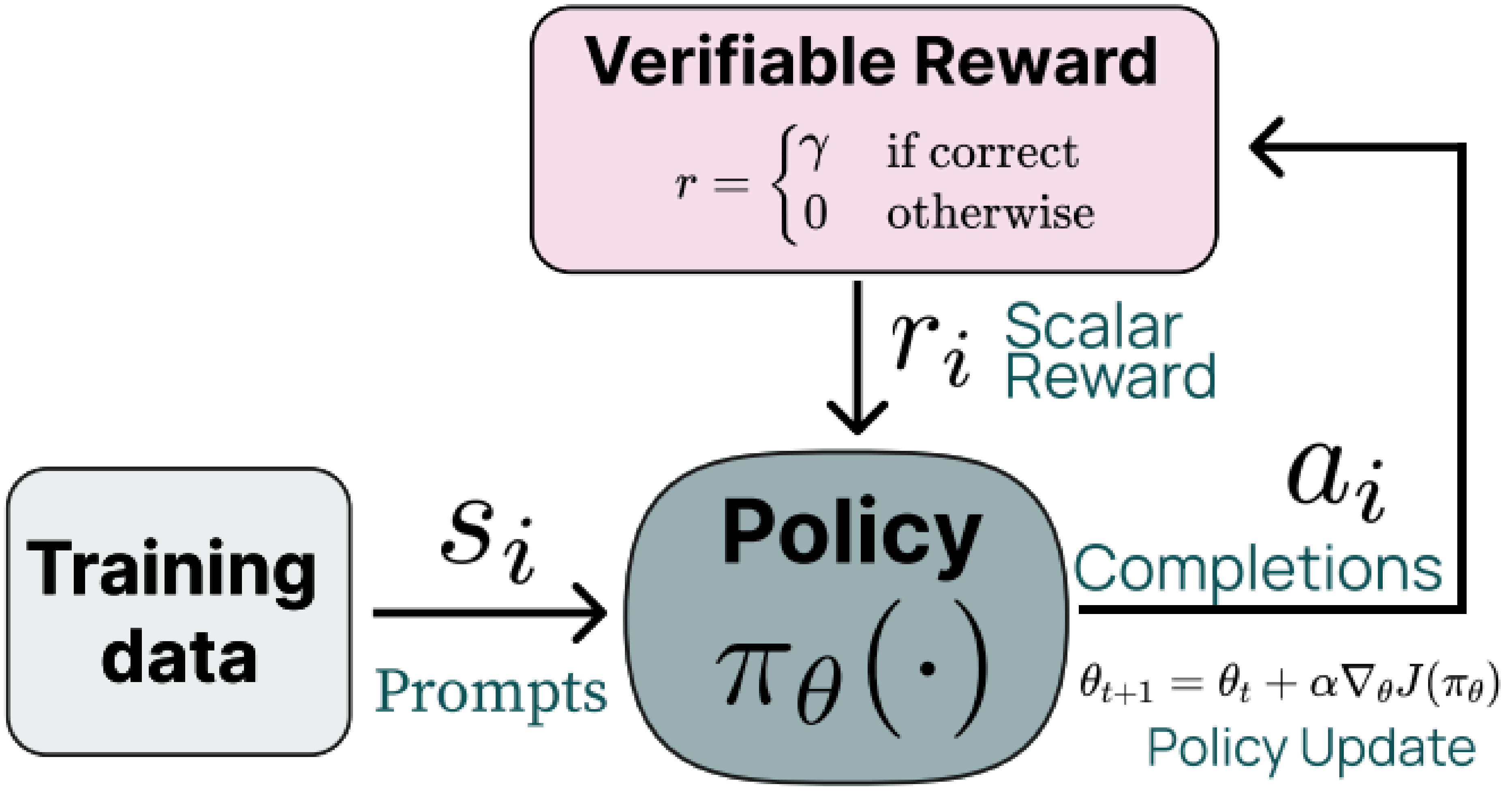
John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, Oleg Klimov
OpenAI
{joschu, filip, prafulla, alec, oleg}@openai.com

RL – Key Ideas

RL with Verifiable Reward (RLVR)

Since the scalar reward is hard to get, one method is to use verifiable reward (e.g., math)

Reward model is also eliminated



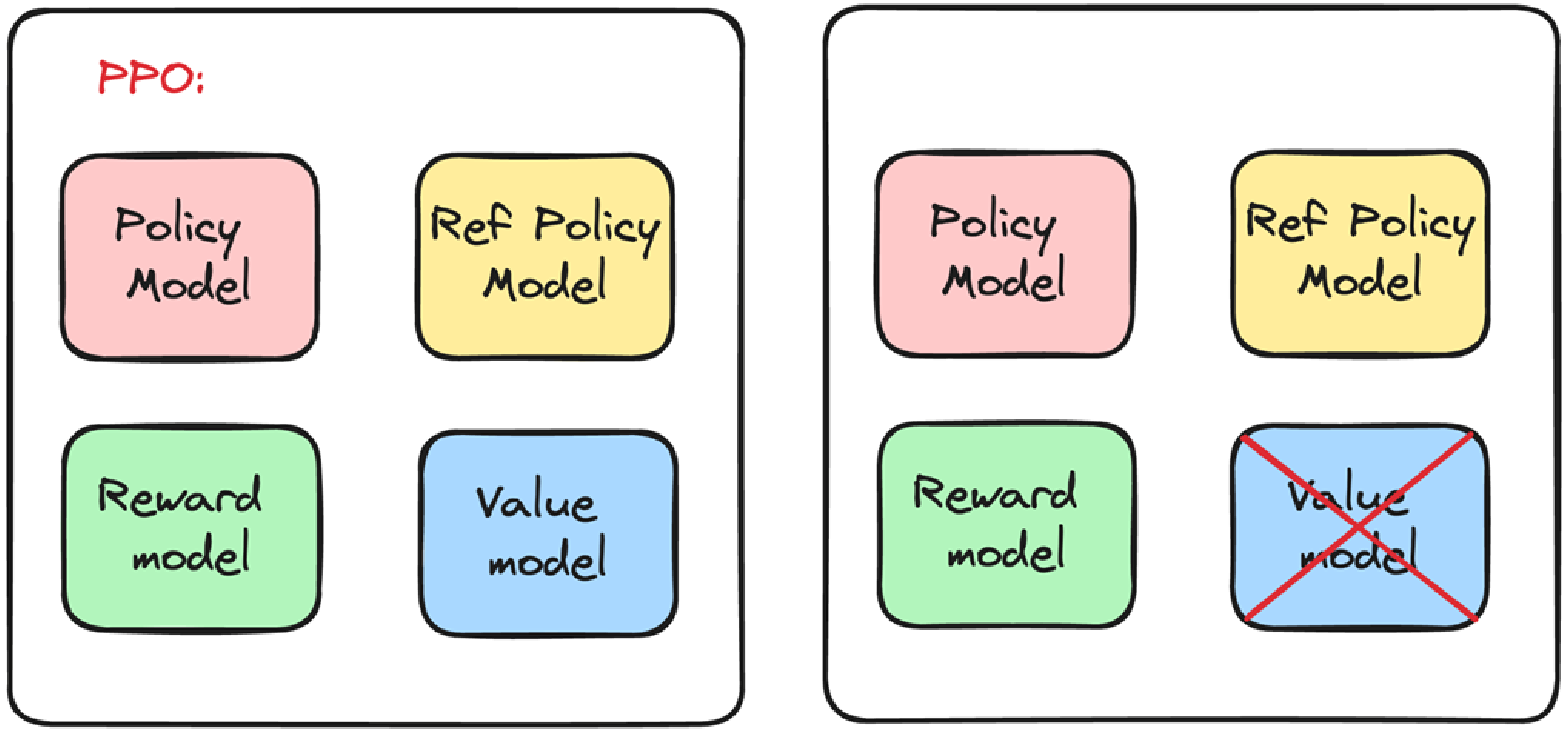
Tülu 3: Pushing Frontiers in Open Language Model Post-Training, Lambert et al., 2025

RL – Key Ideas

Can We Get Rid of the Value Model?

But this is still limited, can we get rid of the value model?

The answer to this question leads to many RL algorithm variants for LLM



https://huggingface.co/blog/putting_rl_back_in_rlhf_with_rloo

RL – Key Ideas

Can We Get Rid of the Value Model?

Core Trick

Value Model = a model (LLM) that estimates the baseline expected return at each time step (token), so we can measure how much better or worse the actual outcome was compared to this expectation (this difference is called advantage).

RL – Key Ideas

Can We Get Rid of the Value Model?

Core Trick

But, *do we need we really need to figure out which* ***token*** *made the reader happy?*

Can we just ask "Is the answer good?" If yes → reinforce. No need to slice the blame

Key Innovation:

Value attributed to each token → group of tokens (e.g., full response)

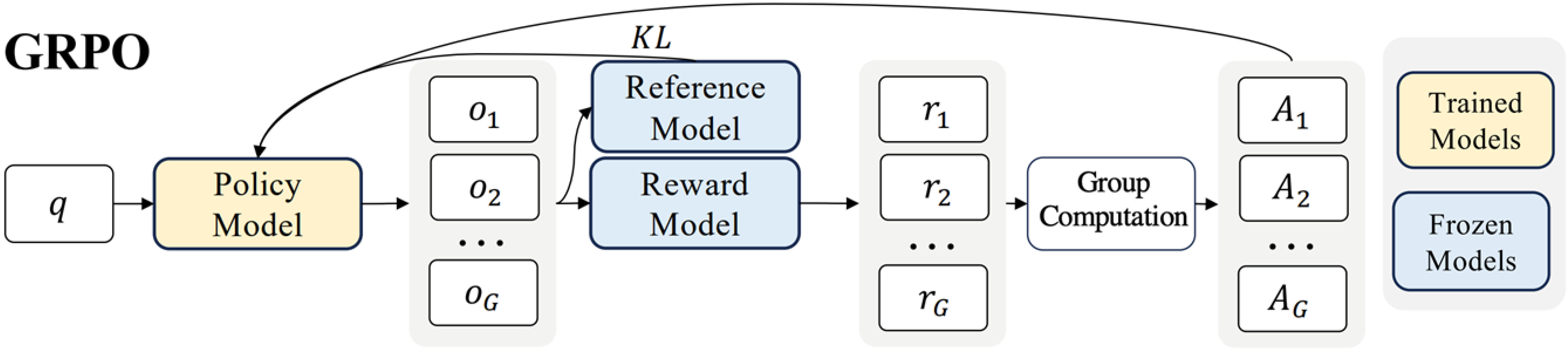
Now the value is directly tie to the reward, no value model required to estimate expected return at each time step.

RL – Key Ideas

GRPO

Action = full response

Advantage = Preference ranking across a group



DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models

RL – Key Ideas

Another RL Variant: RLOO

Action = full response

Advantage = Leave-One-Out reward baseline

$$A = R(x, y) - \frac{1}{n-1} \sum_{j \neq i} R(x, y_j)$$

Reward for the current response

All other responses in the batch

Back to Basics: Revisiting REINFORCE Style Optimization for Learning from Human Feedback in LLMs

Arash Ahmadian
Cohere For AI

Chris Cremer
Cohere

Matthias Gallé
Cohere

RL – Key Ideas Summary

Training Recipe

Model Recipe:

Algorithm: Value model is eliminated by taking group of token as action and define advantage based on those group of tokens (various across RL algorithms. It is still an active research topic)

Training Workflow: usually serve as the last method in the workflow (e.g., after CPT, IT, and PL)

Seed Data

Data Source: often partial overlapping with IT

Data Mixture: Can be large scale (e.g., Math, Logic, Code, Science, Reasoning..)

Data Budget ~ 10 thousand (recent research shows that even a small amount, even just 1-shot can make a different. Still actively research)

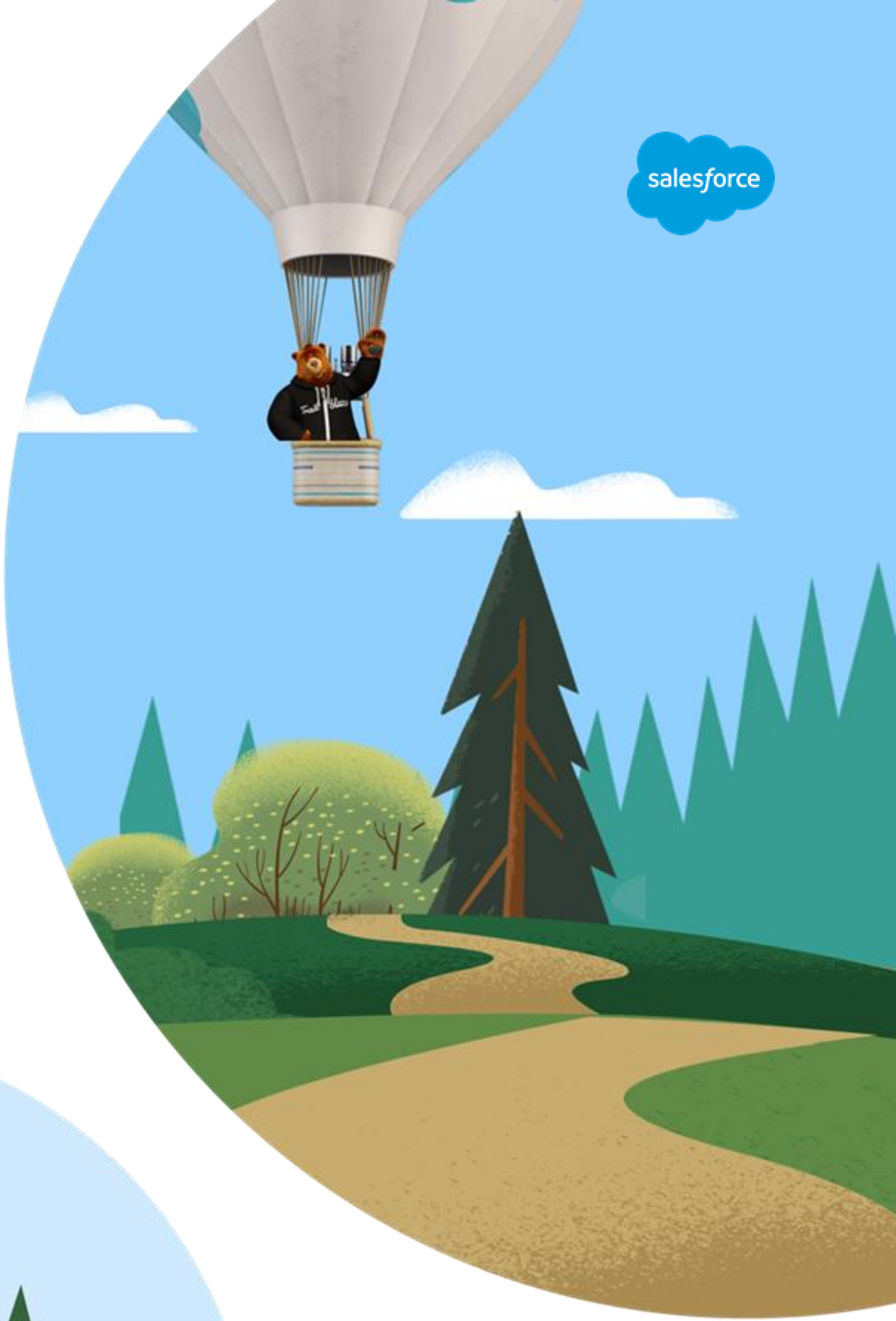
Agenda

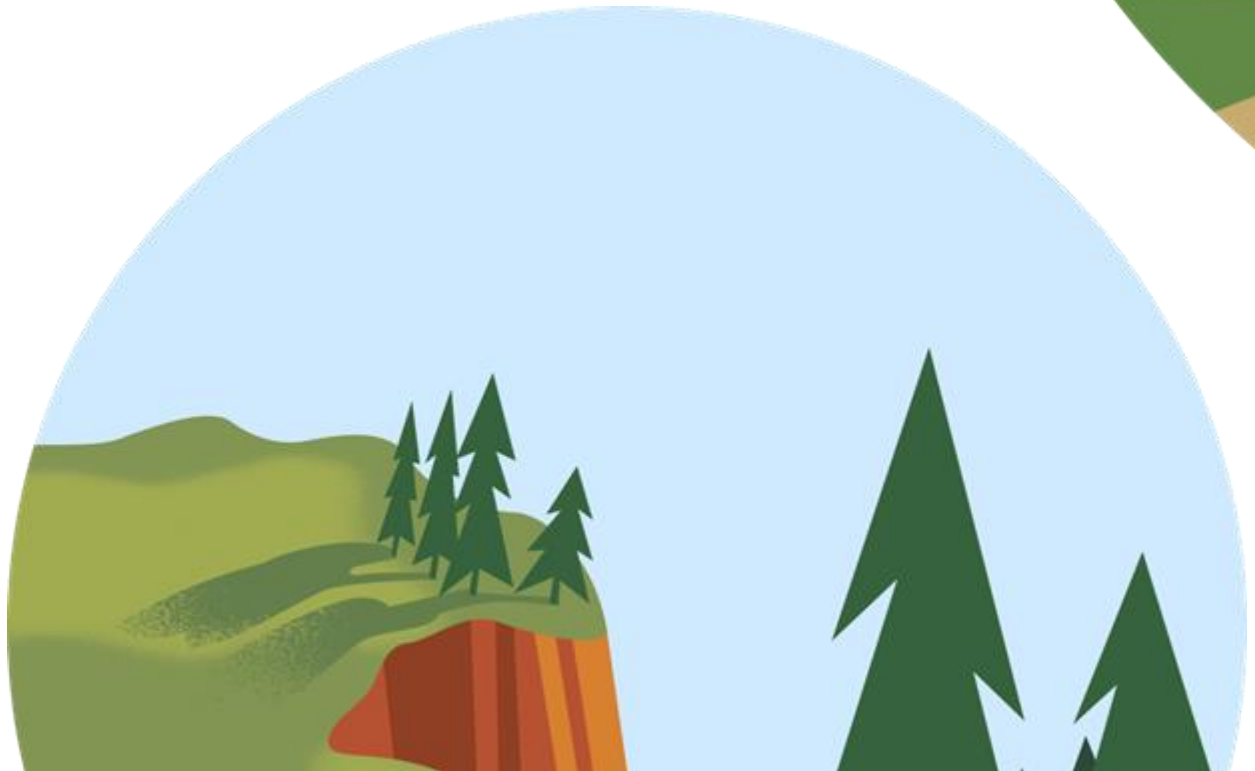
Evaluation and Benchmark

Parametric Knowledge Adaptation

Semi-Parametric Knowledge Adaptation ~30min

Summary, Discussion, QAs

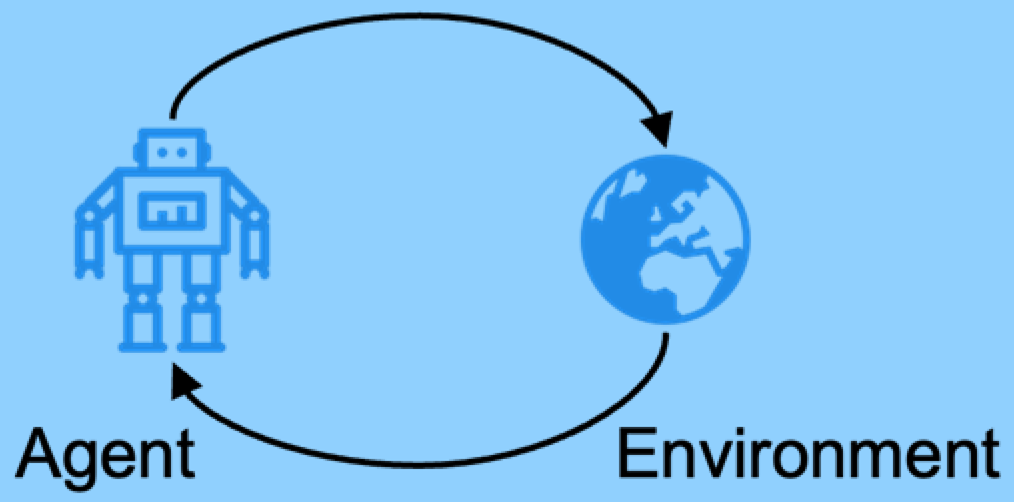




Semi-Parametric Knowledge

Tool

External Memory



Other Agents

Self-refine

E.g., OpenAI’ Deep Research

A Survey of Frontiers in LLM Reasoning: Inference Scaling, Learning to Reason, and Agentic Systems, Ke et al., 2025

RAG – Role

Bridge Gap

Off-the-shelf LLMs may not have been optimized for leveraging external information in its context

Additional adaptation is required for better performance

Autonomous Decision Making

A RAG system needs to decide whether it needs external information or it can respond directly

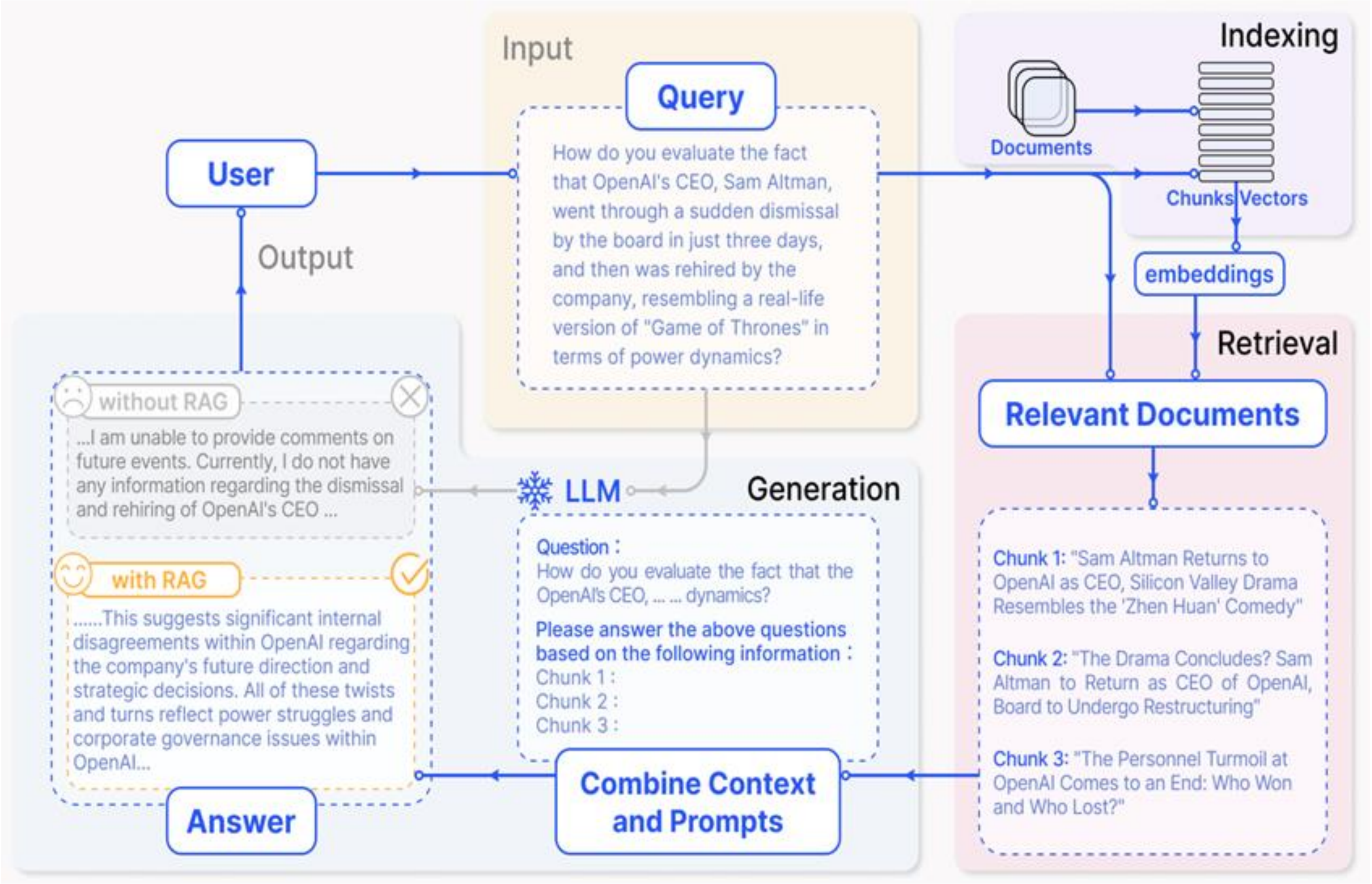
It may need to ask for clarification to the user, do multiple searches via retrieval and aggregate results across documents

RAG - Key Ideas

Example Workflow

Three Main Components

- LLM
- Retriever
- LLM-Retriever Interaction



Minimalist RAG System

RAG – Key Considerations

Training Recipe

Data Recipe:

- Hard to obtain ground truth decision-making trajectory data.
- Model should be robust to potentially noisy context.

Model Recipe:

Algorithm: How to optimize the LLM for search-based interactions?

Training Workflow: What kind of workflow we should use?

Seed Data

Data Source: Where to get the data?

Data Mixture: What should be included in the RAG data?

Data Budget: How much data we need?

RAG – Key Ideas

LLM and Decision Making

Post-train LLMs for contextual usage

Deal with:

- Noisy context (passages from same document and different documents)
- Conflicting evidence
- Counterfactual evidence
- Absence of knowledge

E.g., SFR-RAG (Salesforce), RAG 2.0 (Contextual AI)

LLMs with agentic workflow

- Predefined or autonomous workflow.
- Single agent vs. multi-agent system
- Planner and worker agents

E.g., Infogent, Manus Agent, Deep Research (OpenAI)

INFOGENT: An Agent-Based Framework for Web Information Aggregation, Reddy, et al., 2024

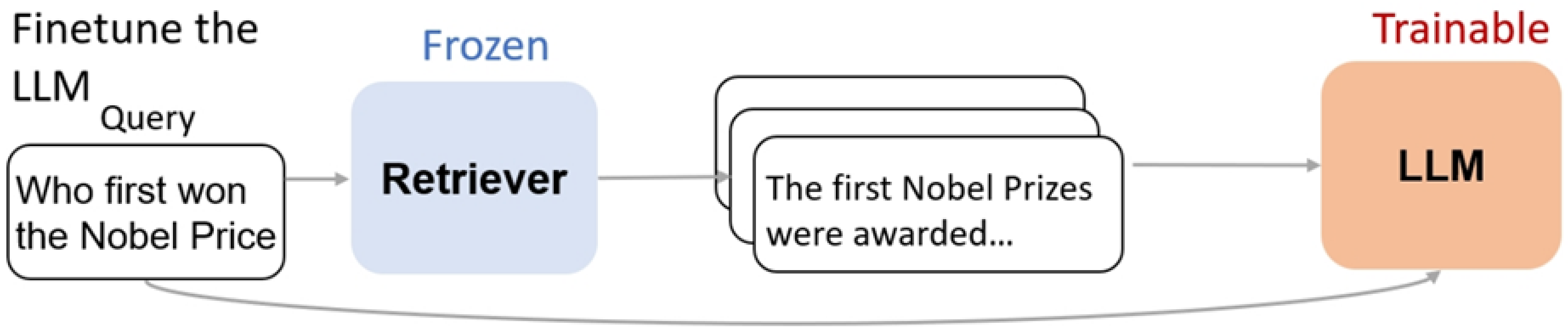
RAG – Key Ideas

Train LLMs for Contextual Use

Post-train LLMs for RAG scenarios:

Create contextual fine-tuning data to deal with noisy contexts, counterfactual contexts, no-answer contexts and conflicting

Examples: SFR-RAG, RAG 2.0



1. **Fix the retriever**
2. **Train the LLM for contextual usage**

SFR-RAG: Towards Contextually Faithful LLMs, Nguyen et al., 2024
RAG2.0: https://contextual.ai/introducing-rag2/

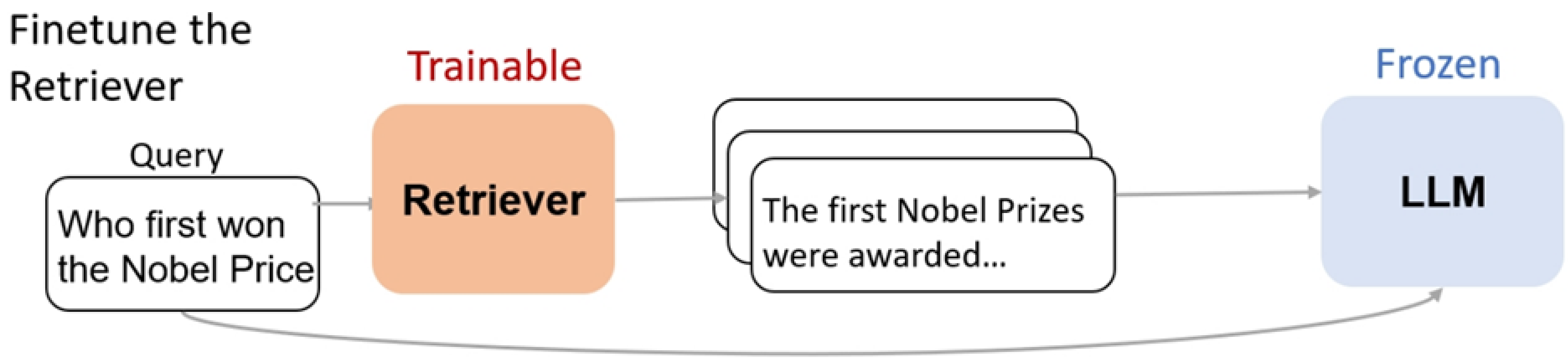
RAG – Key Ideas

Align Retriever to LLM

The output of a frozen LLM is used as supervision signals to train the retriver

Examples: REPLUG, Atlas

1. **Fix the LLM**
2. **Align the retriever to LLM**



REPLUG: Retrieval-Augmented Black-Box Language Models, Shi et al., 2023
Atlas: Few-shot Learning with Retrieval Augmented Language Models, Izacard, 2022

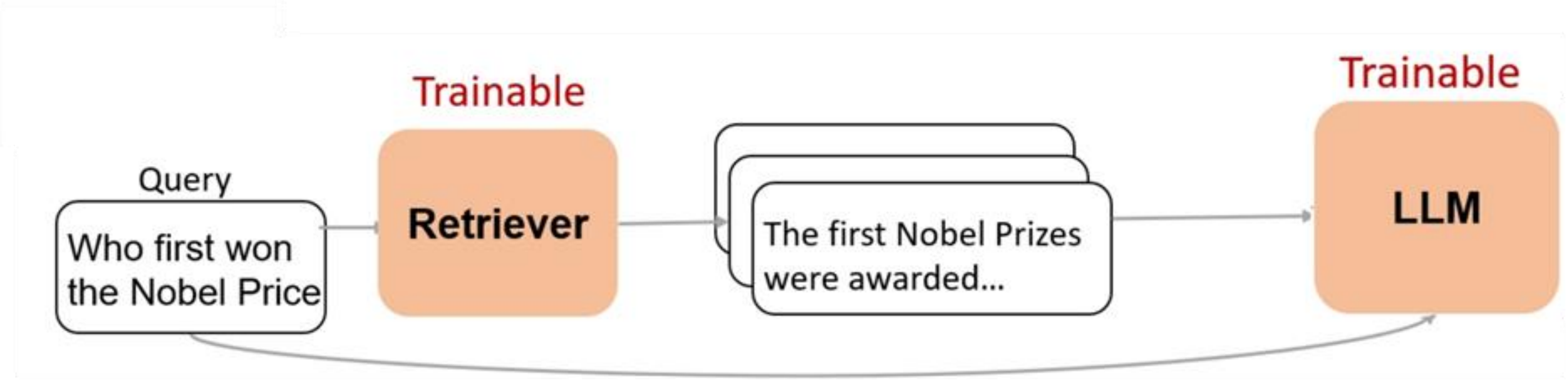
RAG – Key Ideas

Train both the LLM and Retriver

Jointly or sequentially train the retriever and LLMs so that they are aligned

Examples: RA-DIT

1. Train both the LLM and the retriever



RA-DIT: Retrieval-Augmented Dual Instruction Tuning, Lin et al, 2024

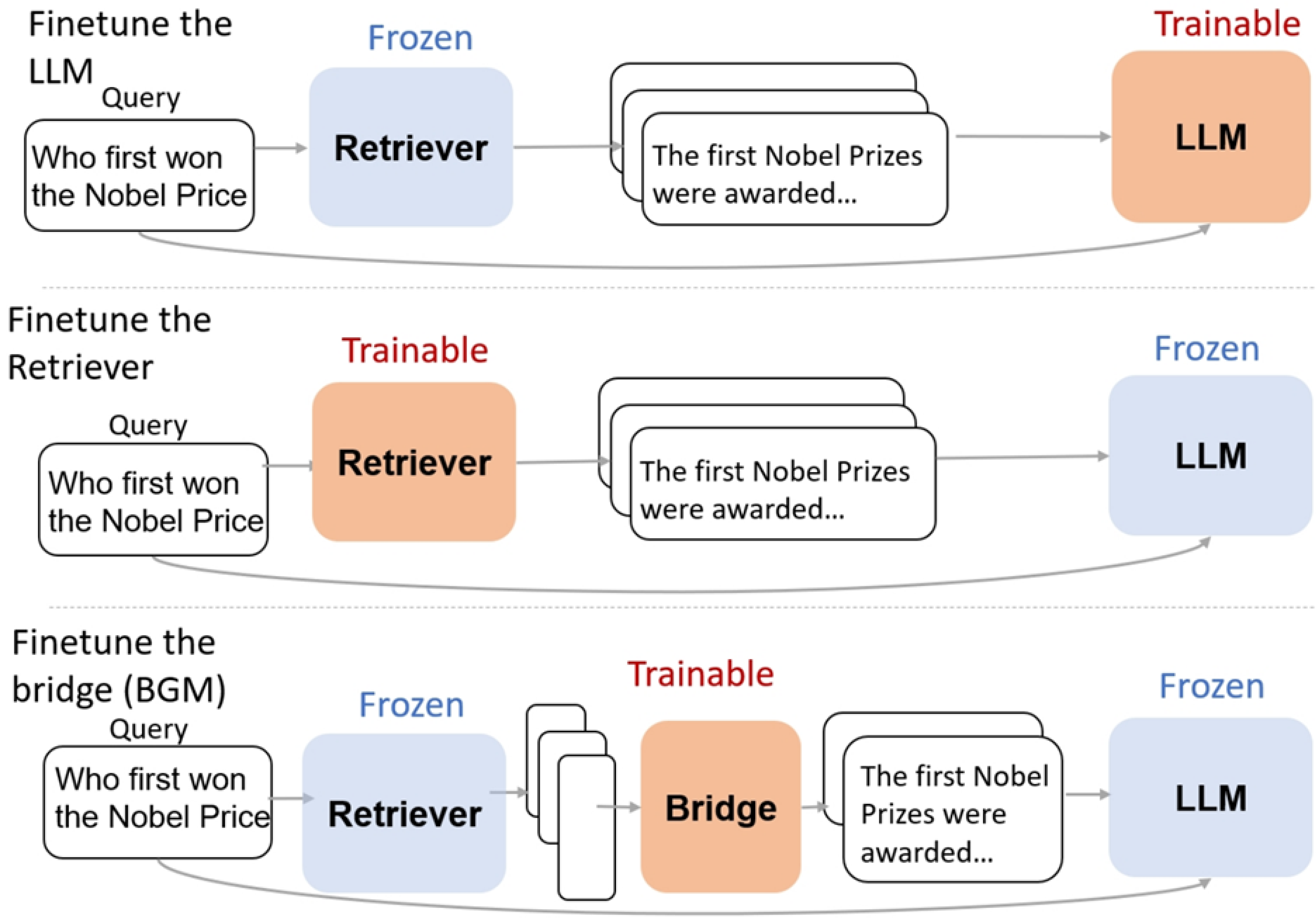
RAG – Key Ideas

LLM-Retriever Interaction

Fix the LLM and Retriver

Train a "bridge" (a LLM) to connect their preference

Main innovation: There is preference gap between **retriever** (built for human) and **LLM** (can prefer different order, selection..). One alternative way besides training LLM or retriever is to train an intermediate bridge



Bridging the Preference Gap between Retrievers and LLMs, Ke et al., 2024

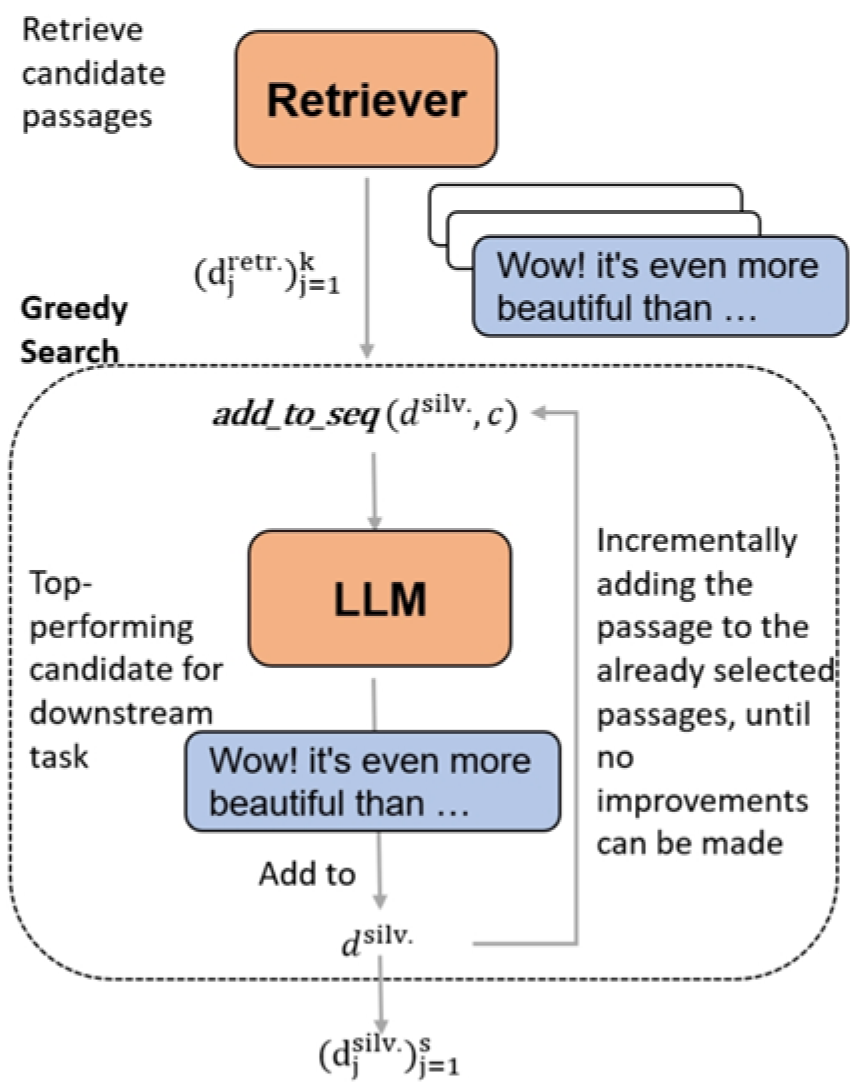
RAG – Key Ideas

LLM-Retriever Interaction

Ground Truth Data: Use greedy search to find the silver passage

Step 1

Collect silver passage sequence



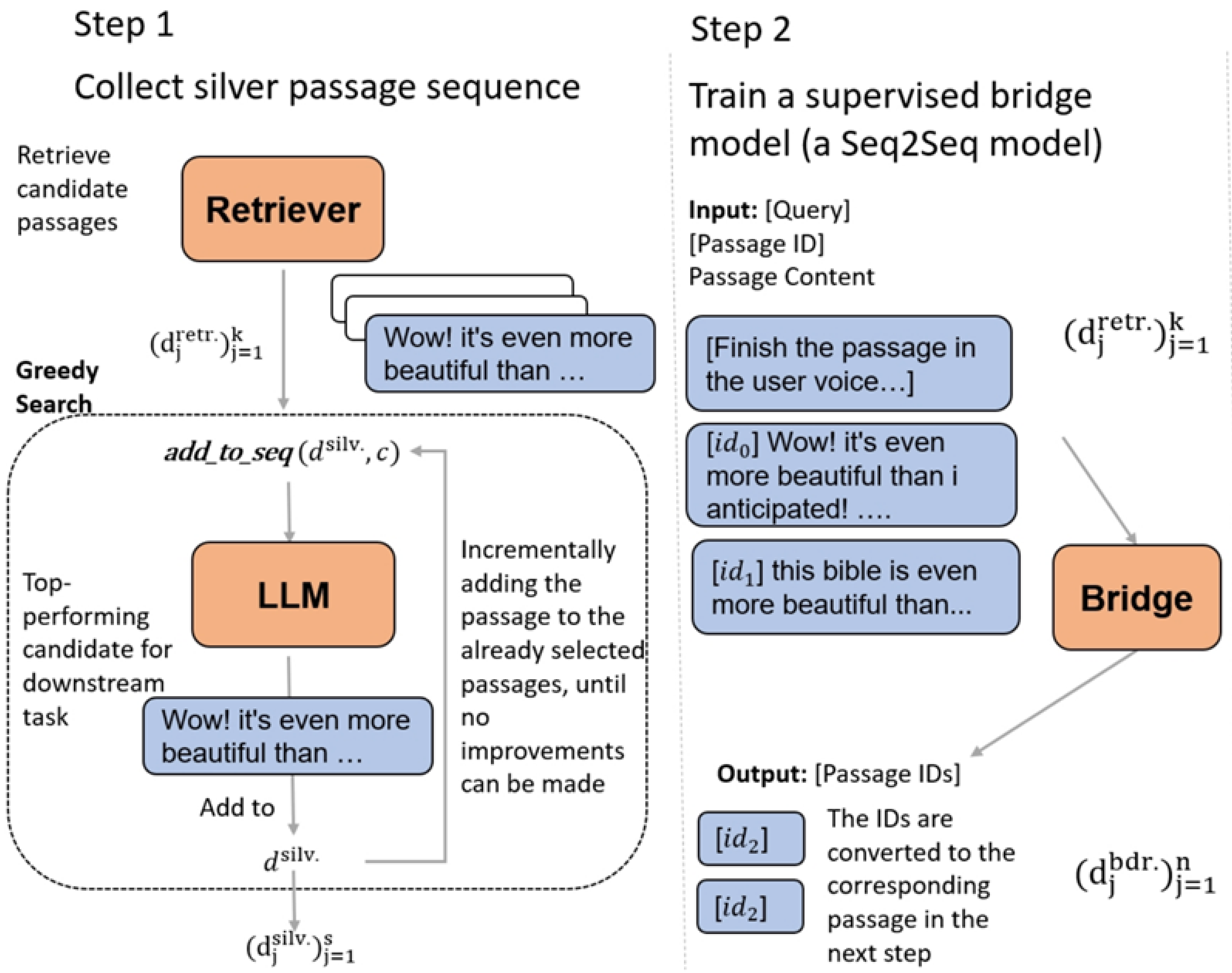
Bridging the Preference Gap between Retrievers and LLMs, Ke et al., 2024

RAG – Key Ideas

LLM-Retriever Interaction

Ground Truth Data: Use greedy search to find the silver passage

Workflow: IT → RL



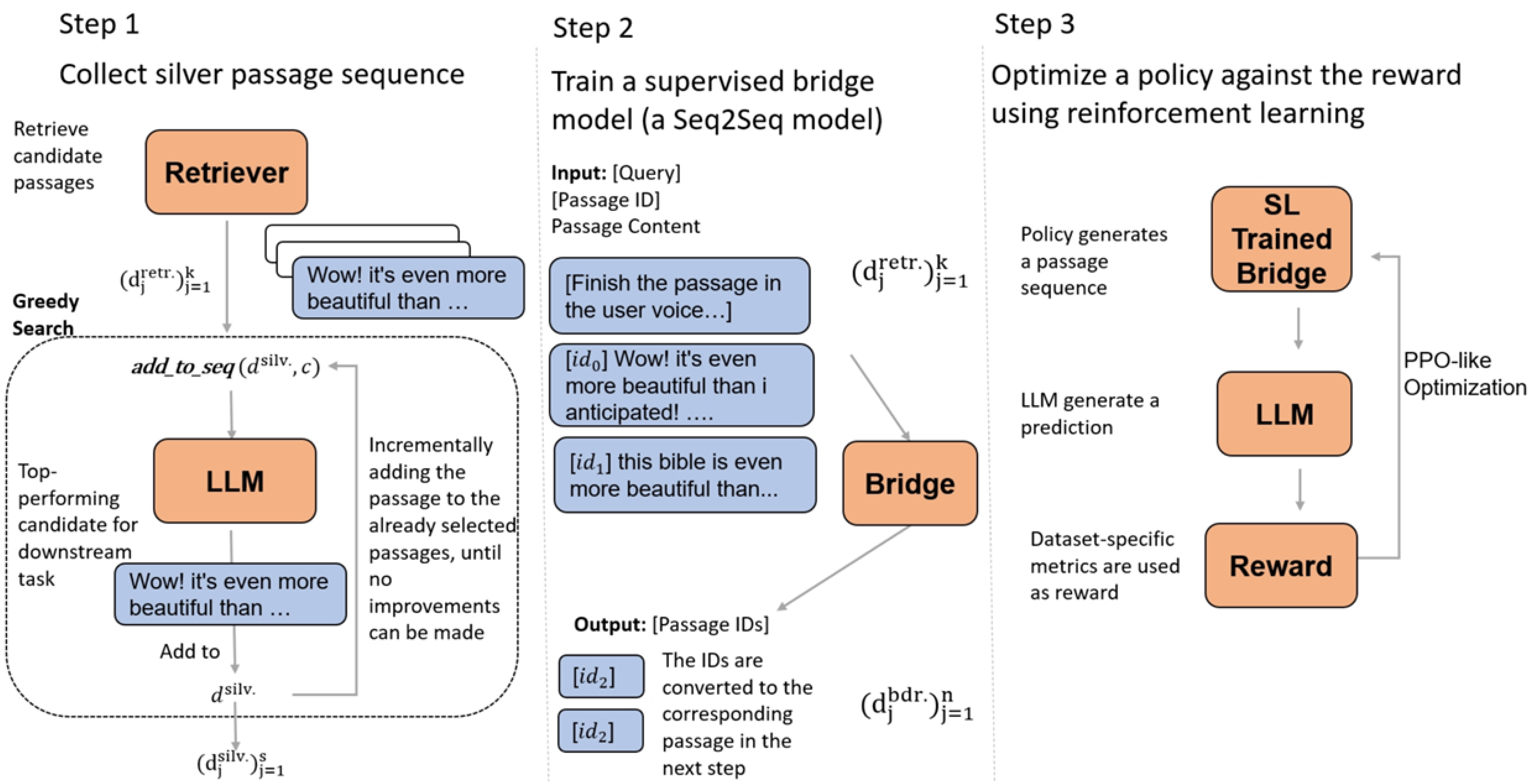
Bridging the Preference Gap between Retrievers and LLMs, Ke et al., 2024

RAG – Key Ideas

LLM-Retriever Interaction

Ground Truth Data: Use greedy search to find the silver passage

Workflow: IT → RL



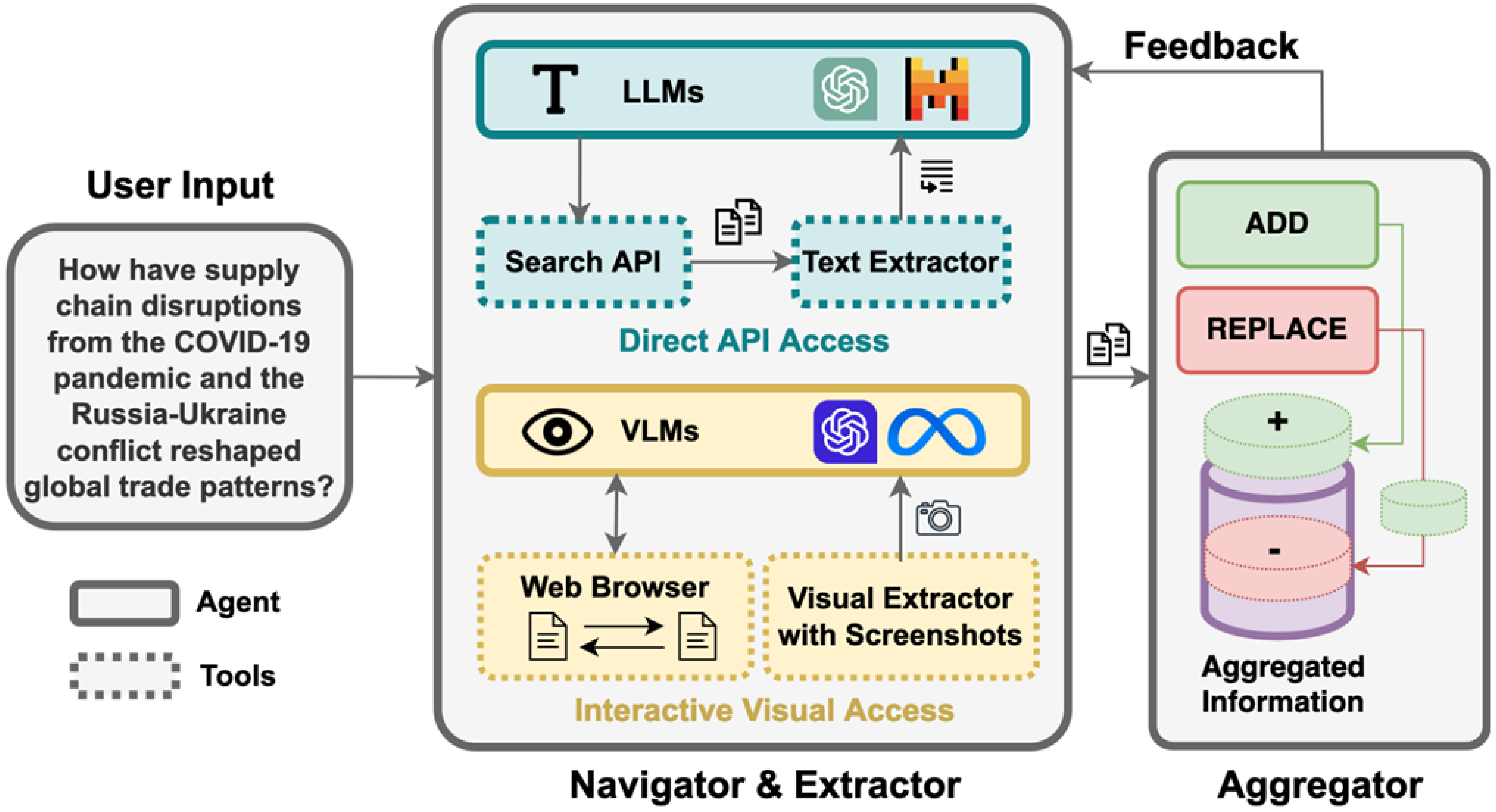
Bridging the Preference Gap between Retrievers and LLMs, Ke et al., 2024

Agentic RAG

RAG with Predefined Workflow

Main innovation: RAG can be performed in multiple predefined steps (workflow) to approach the final goal. Those steps usually involve API call, web browser, planner, etc.

Examples: Infogent, MindSearch



INFOGENT: An Agent-Based Framework for Web Information Aggregation, Reddy, et al., 2024
MindSearch: Mimicking Human Minds Elicits Deep AI Searcher, Chen et al., 2024

RAG – Key Ideas Summary

Training Recipe

Data Recipe:

often use heuristic way to construct the ground truth

Model Recipe:

Algorithm and Workflow: so far, it is largely follows the parametric knowledge adaptation

Seed Data

Data Source: Knowledge-extensive tasks

Data Mixture: Can be large scale (e.g., Math, Logic, Code, Science, Reasoning..)

Data Budget: Follow the budget required in the specific method

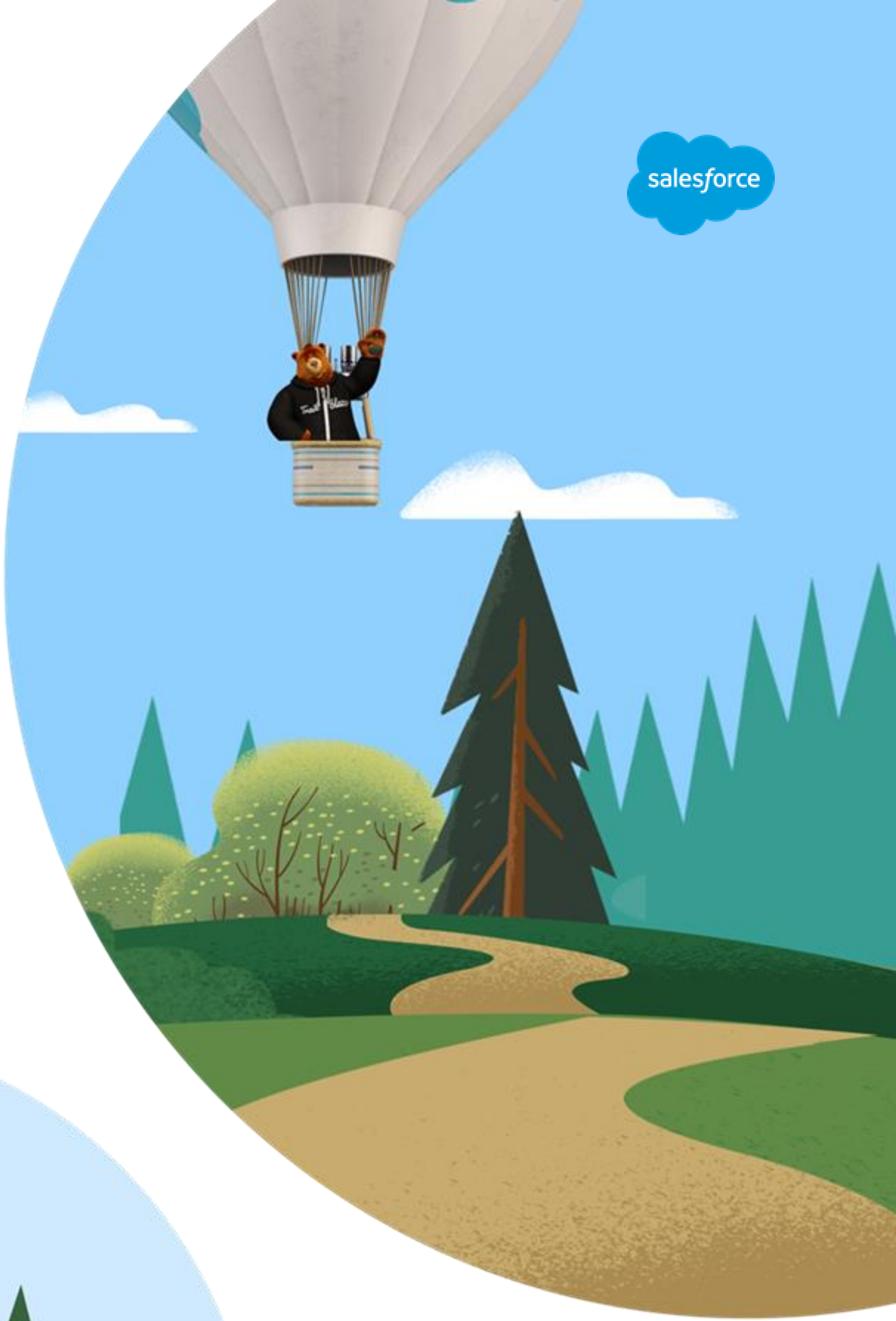
Agenda

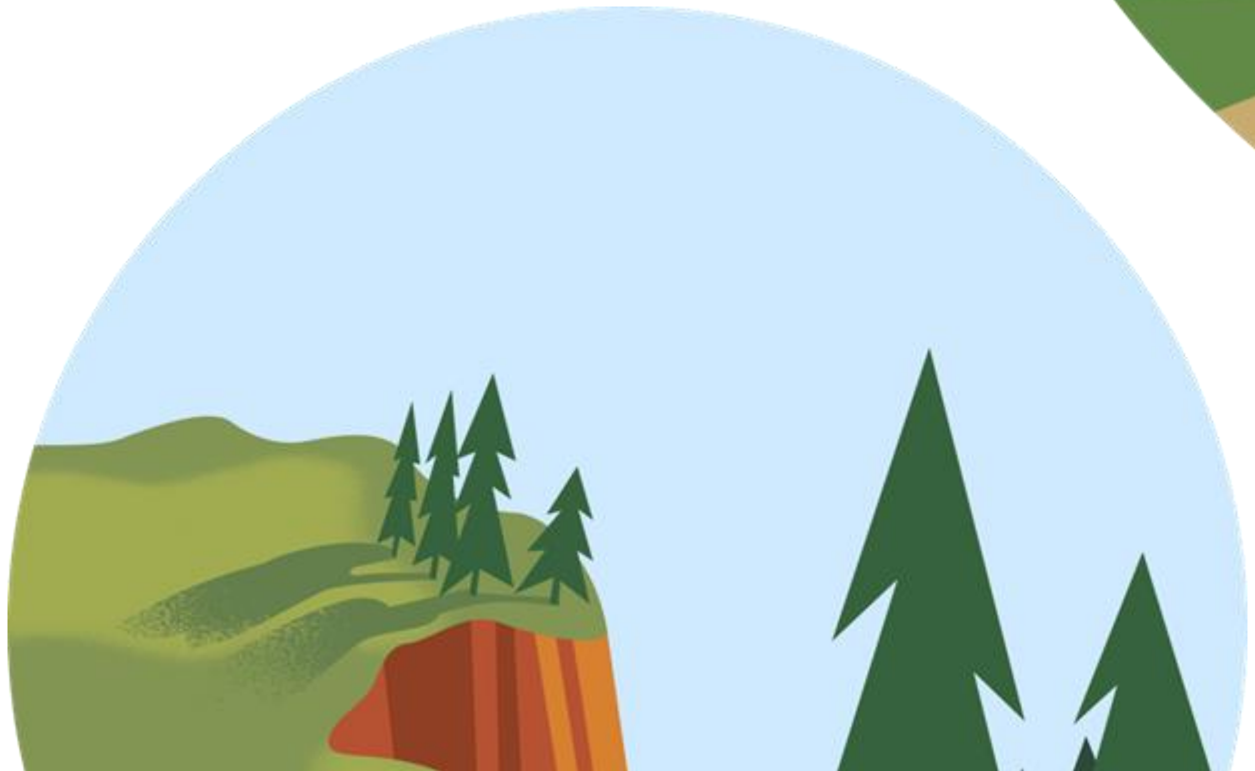
Evaluation and Benchmark

Parametric Knowledge Adaptation

Semi-Parametric Knowledge Adaptation

Summary, Discussion, QAs



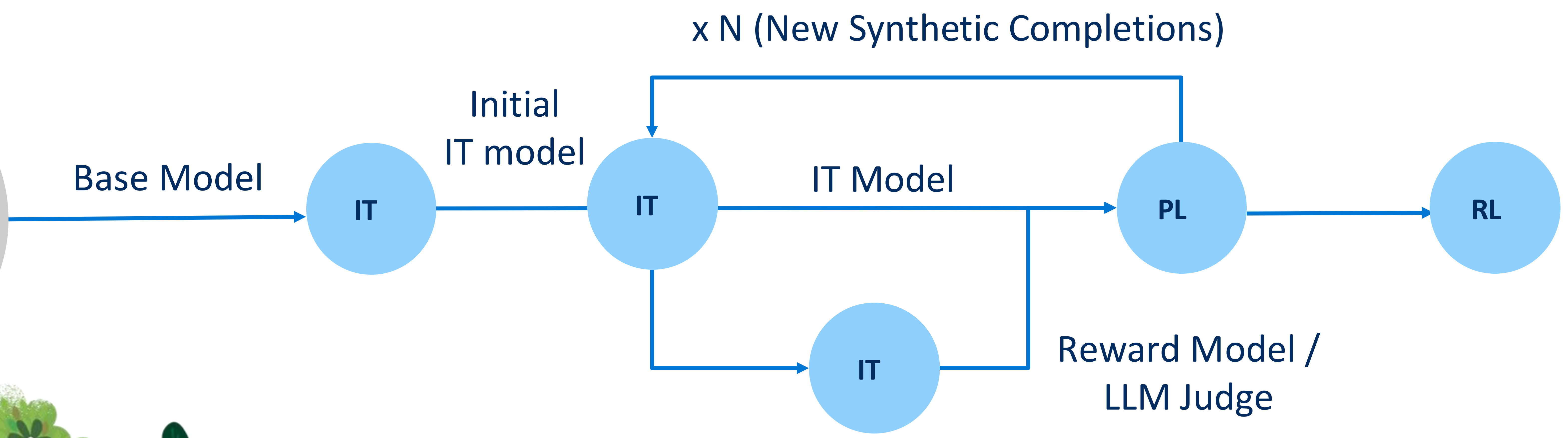


Putting All Together

Workflow

Adaptation training workflow is an actively research topic, we could expect seeing more to come

It is not surprised that the workflows introduced today are replaced soon.



Putting All Together

Algorithms

While CPT and IT are used as the foundation of the model before RL, RL algorithms are actively researched today

Key problems:

How to train a good reward model? (evaluation is challeneging)

The important of human preference data vs. LLM-as-a-judge

RL for multi-agent system?

Besides learning from experience, can the LLM self-discover its own knowledge during RL?

Putting All Together

Data

Data is important, including both the seed data and the data recipe. Although this is usually not disclosed, it is an active area of research in the community

We have seen more and more publicly available data

More data synthetic or distillation (e.g., direct distillation in DeepSeek-R1)is coming

Adaptation – Open Questions

salesforce

Workflow

Training workflow: What is the best training workflow for adaptation?

Agentic workflow (e.g., RAG agentic system), can we automatically design workflow? a meta-level design is still understudied

Algorithm

RL has very high potential but research still needed (e.g. reward modeling, RL for multi-agent system)

Data

Better data synthetic and data distillation method


Q&A