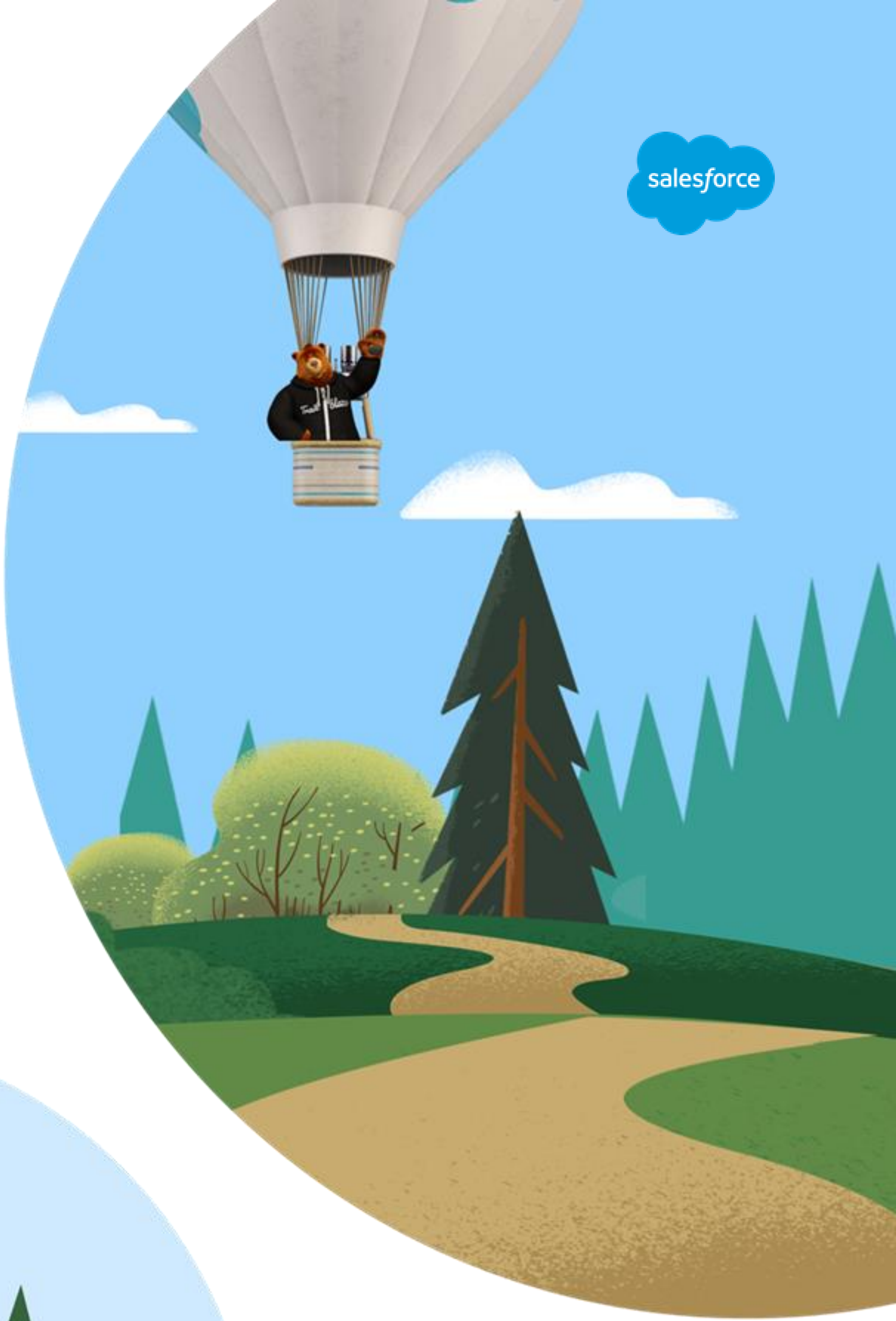
Agenda

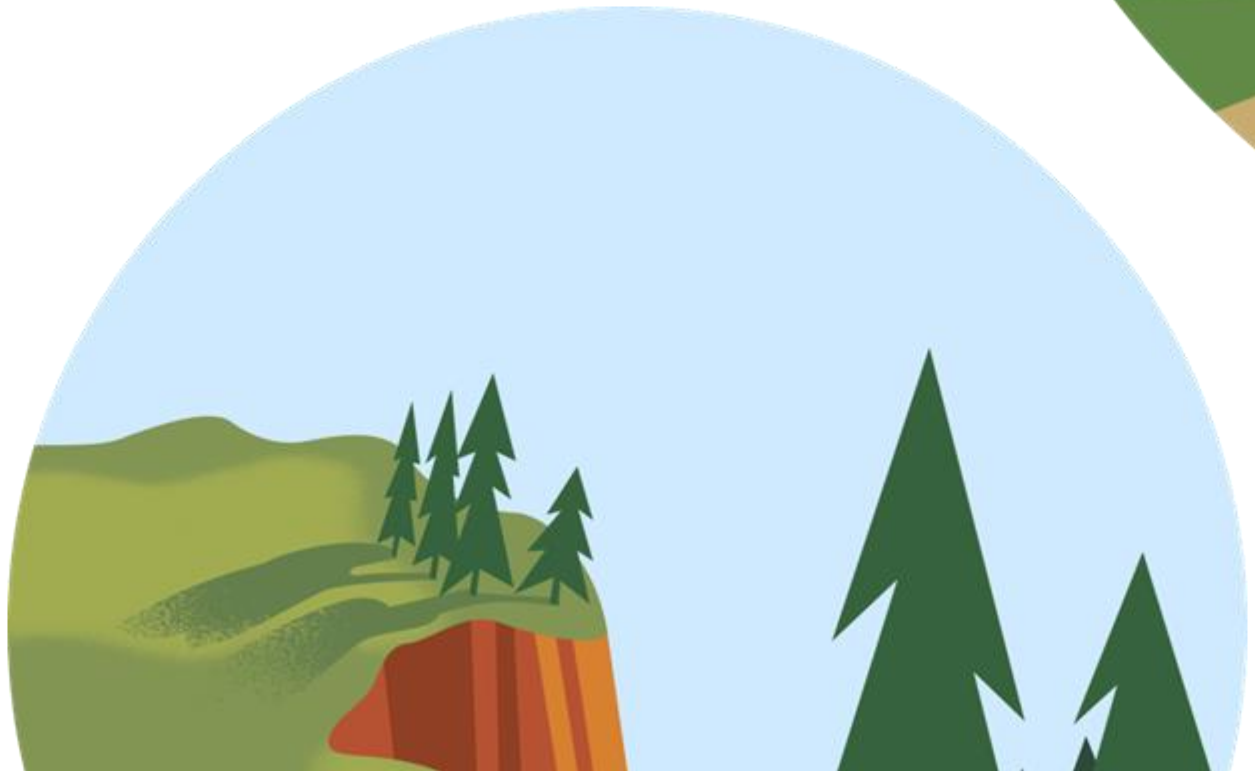
Evaluation and Benchmark ~ 20min

Parametric Knowledge Adaptation

Semi-Parametric Knowledge Adaptation

Summary, Discussion, QAs

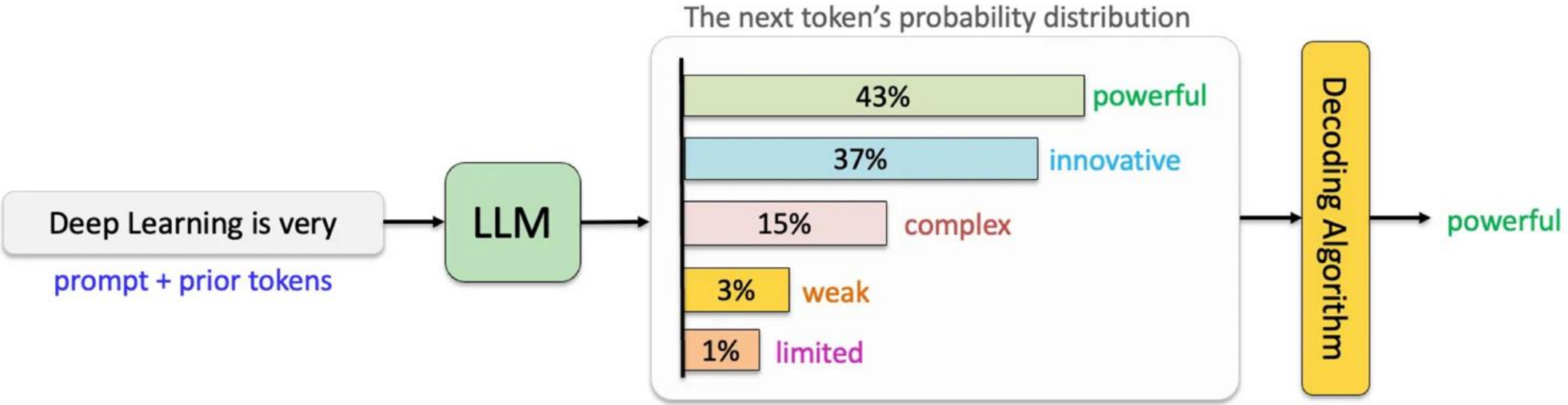




Evaluating LLMs (and agentic systems)

Challenges: LLMs are Non-Deterministic Generators

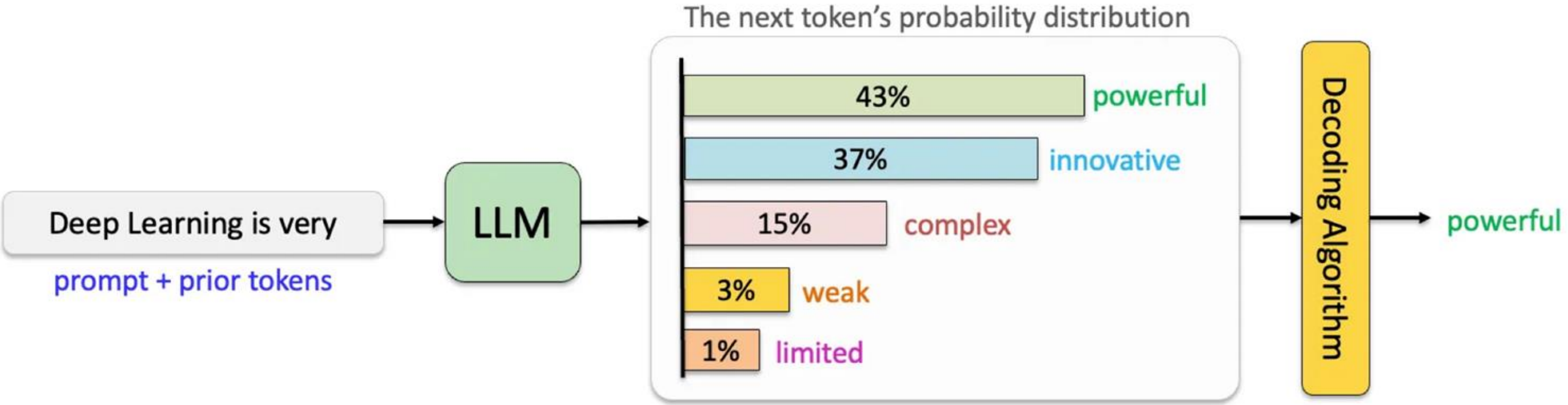
- Probabilistic nature of LLMs:



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

"Traditional" NLP

Reference-free metrics

Quality-based
- Supert
- BLANC
- ROUGE-C

Entailment-based
- SummaC
- FactCC
- DAE

- SRLscore
- QAFactEval
- QuestEval

Rise of Pre-Trained Models (e.g. BERT)

LLM-based metrics

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

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

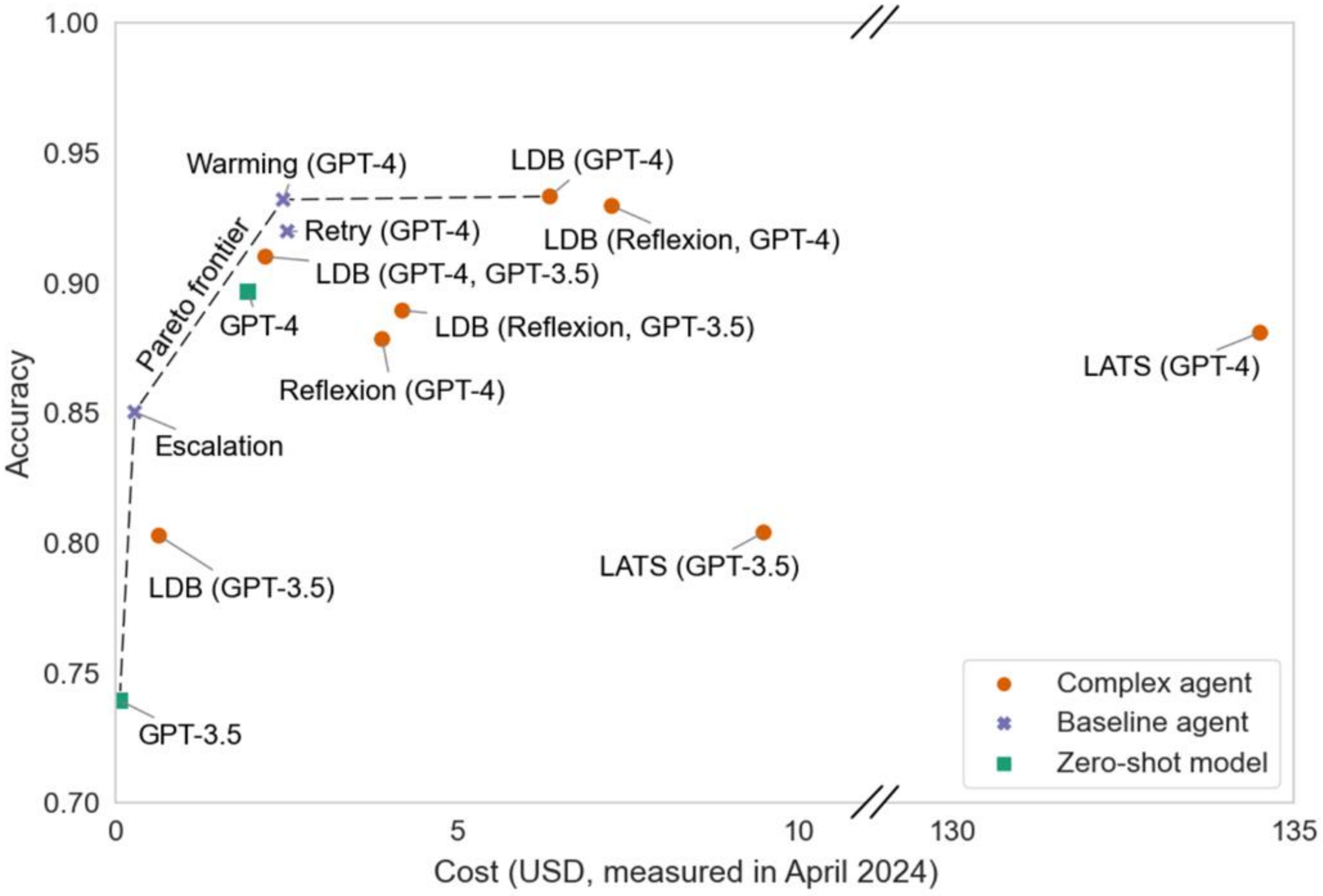
- ❑ Cost

- ❑ 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

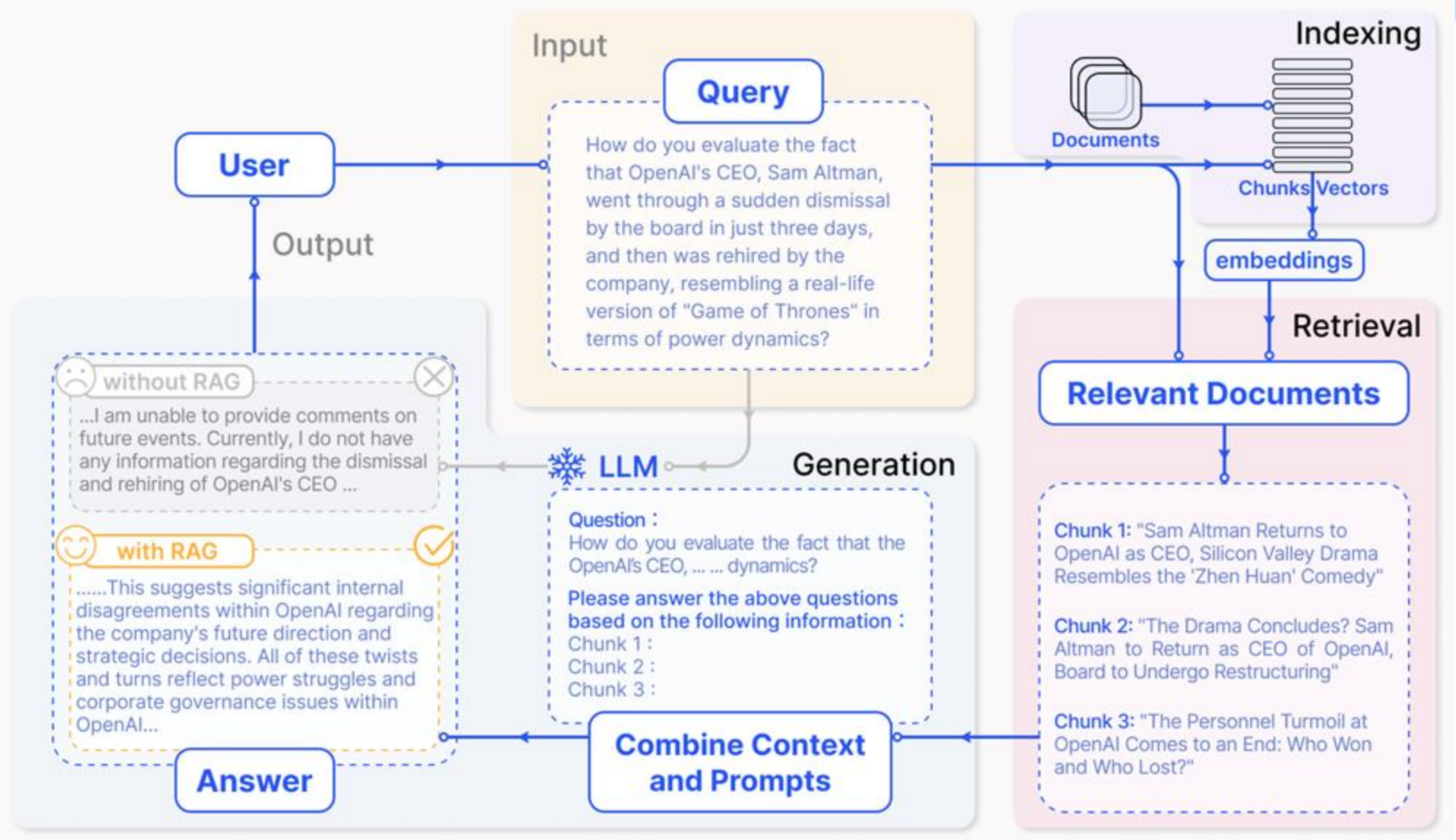
Retrieval Augmented Generation (RAG)

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**…

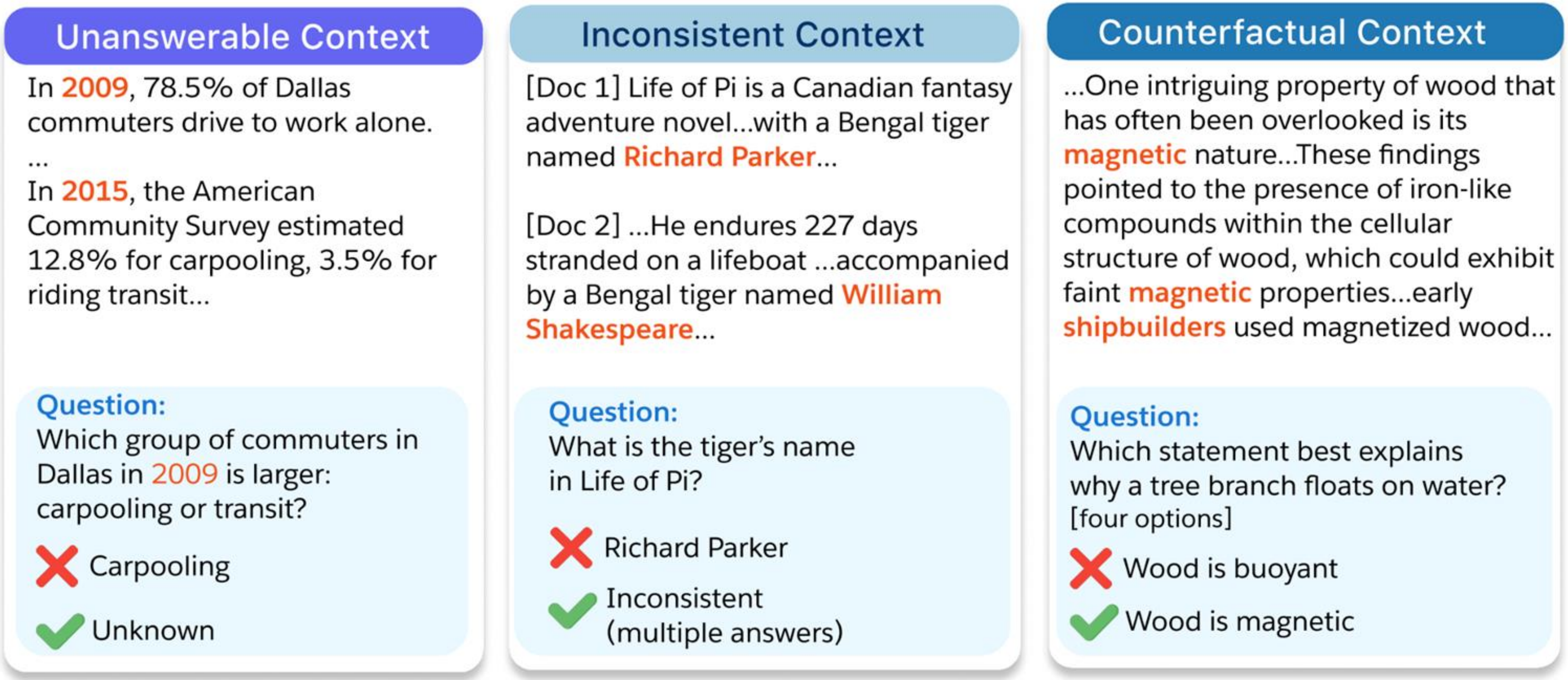
Q: What is the capital city of USA?

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:

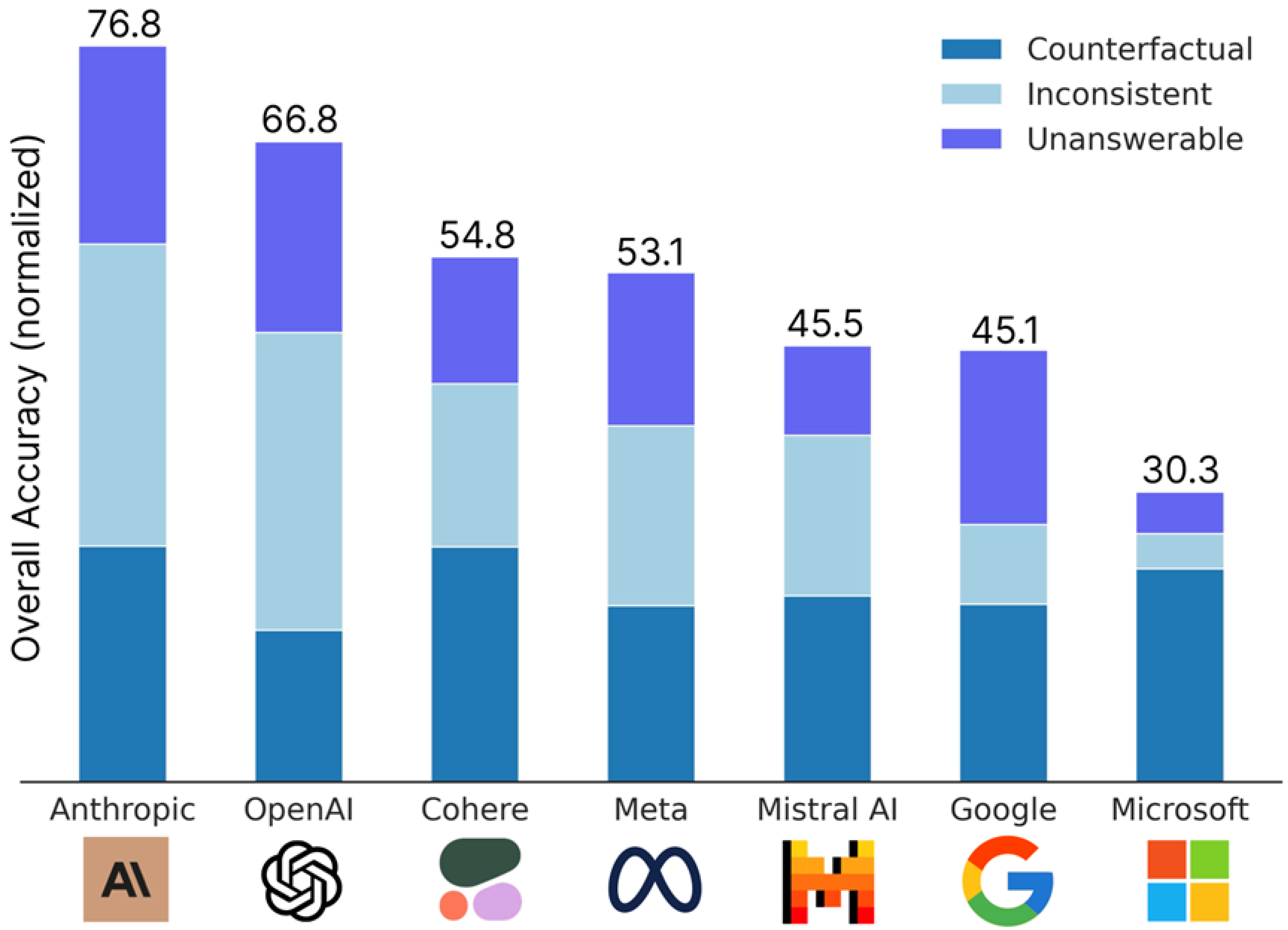


- Ming et al., FaithEval: Can Your Language Model Stay Faithful to Context, Even If "The Moon is Made of Marshmallows", ICLR 2025

Adapting LLMs to Specific Contexts

❑ How good are frontier LLMs against noisy contexts?

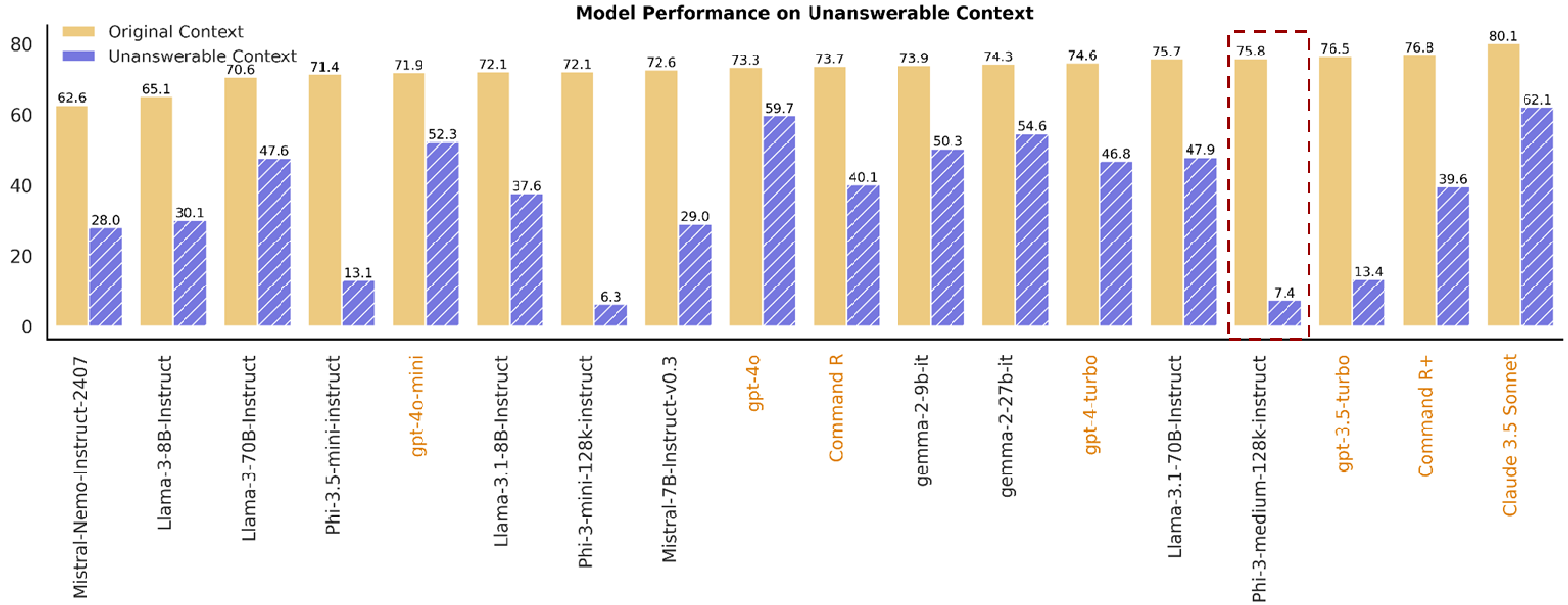
Model Name	Model Size
Phi-3 Family (Abdin et al., 2024)	
Phi-3-mini-128k-instruct	3.8B
Phi-3-medium-128k-instruct	14B
Phi-3.5-mini-instruct	3.8B
LLaMA-3 Family (Llama, 2024)	
LLaMA-3-8B-instruct	8B
LLaMA-3.1-8B-instruct	8B
LLaMA-3-70B-instruct	70B
LLaMA-3.1-70B-instruct	70B
Mistral Family (Jiang et al., 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

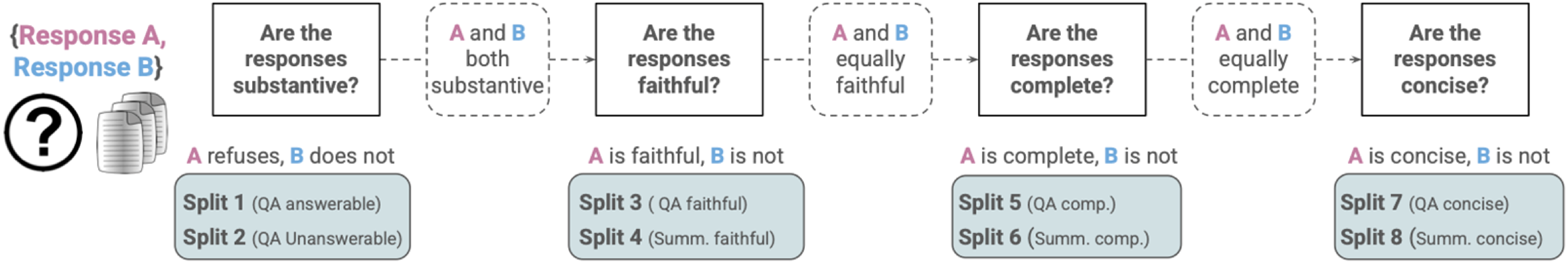
- Larger models are not necessarily faithful



- 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

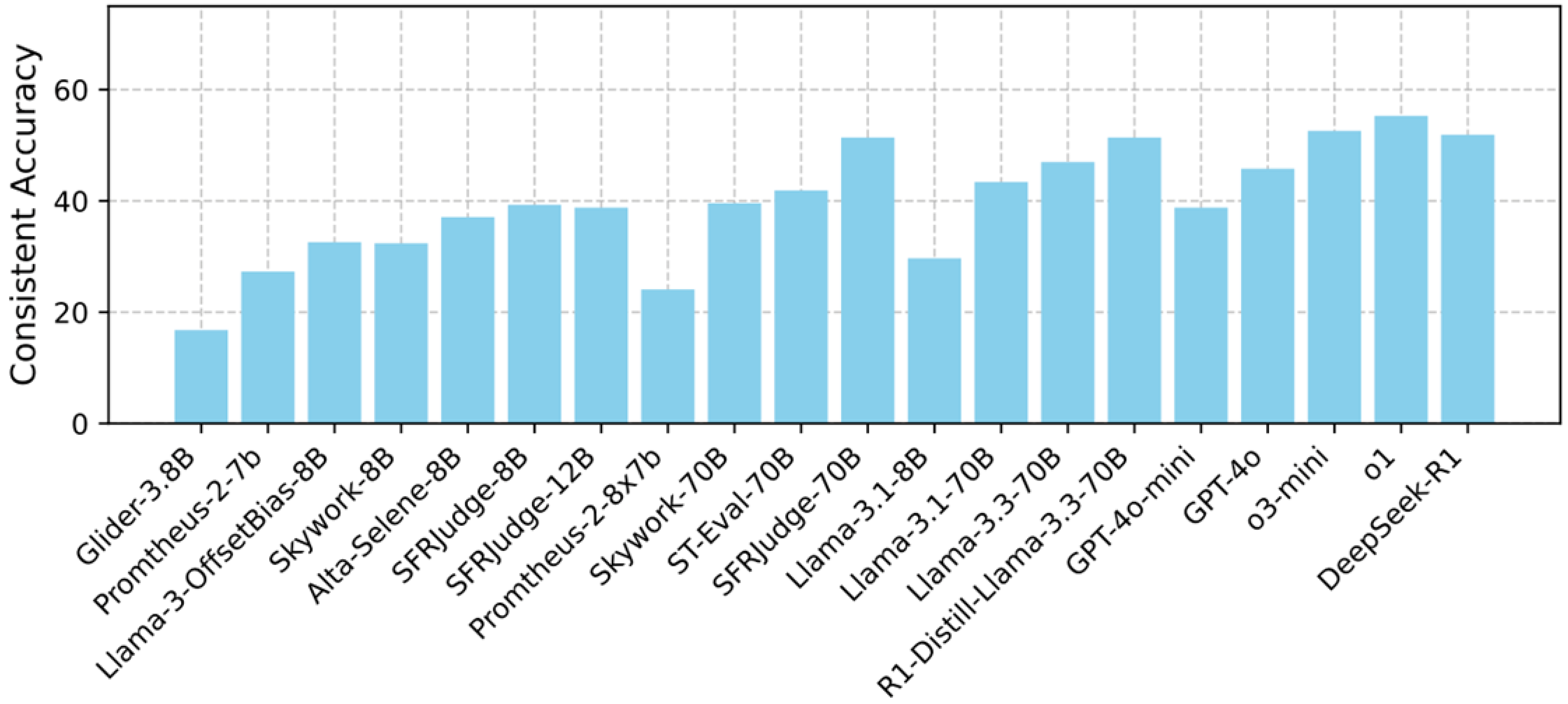


- Xu et al., Does Context Matter? ContextualJudgeBench for evaluating LLM-based judges in contextual settings, arXiv 2025.

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

	ICL	Cite	Re-rank	LongQA	Summ	Avg.
NIAH	0.44	0.71	0.75	0.76	0.72	0.68
RULER MK	0.48	0.73	0.84	0.79	0.87	0.74
RULER MV	0.61	0.71	0.77	0.83	0.79	0.74
RULER All	0.51	0.77	0.85	0.79	0.83	0.75
Recall	0.61	0.74	0.85	0.82	0.85	0.77
RAG	0.5	0.72	0.85	0.92	0.89	0.78

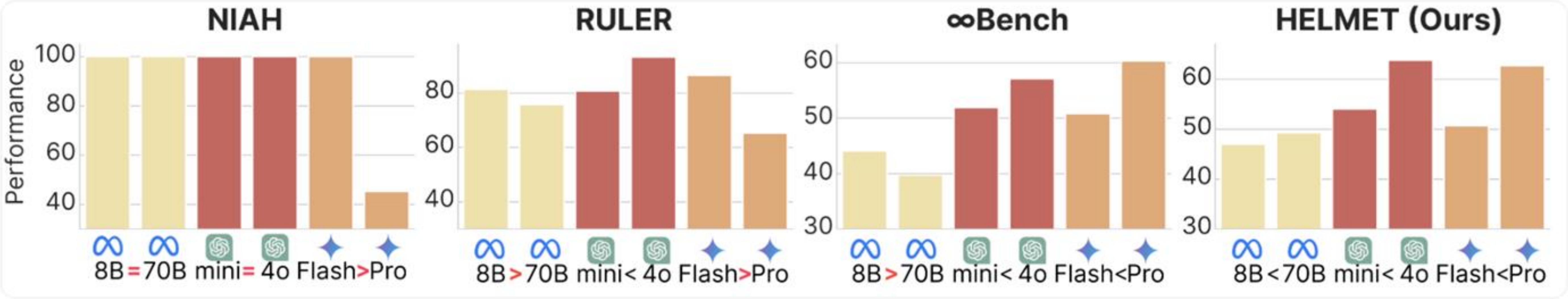
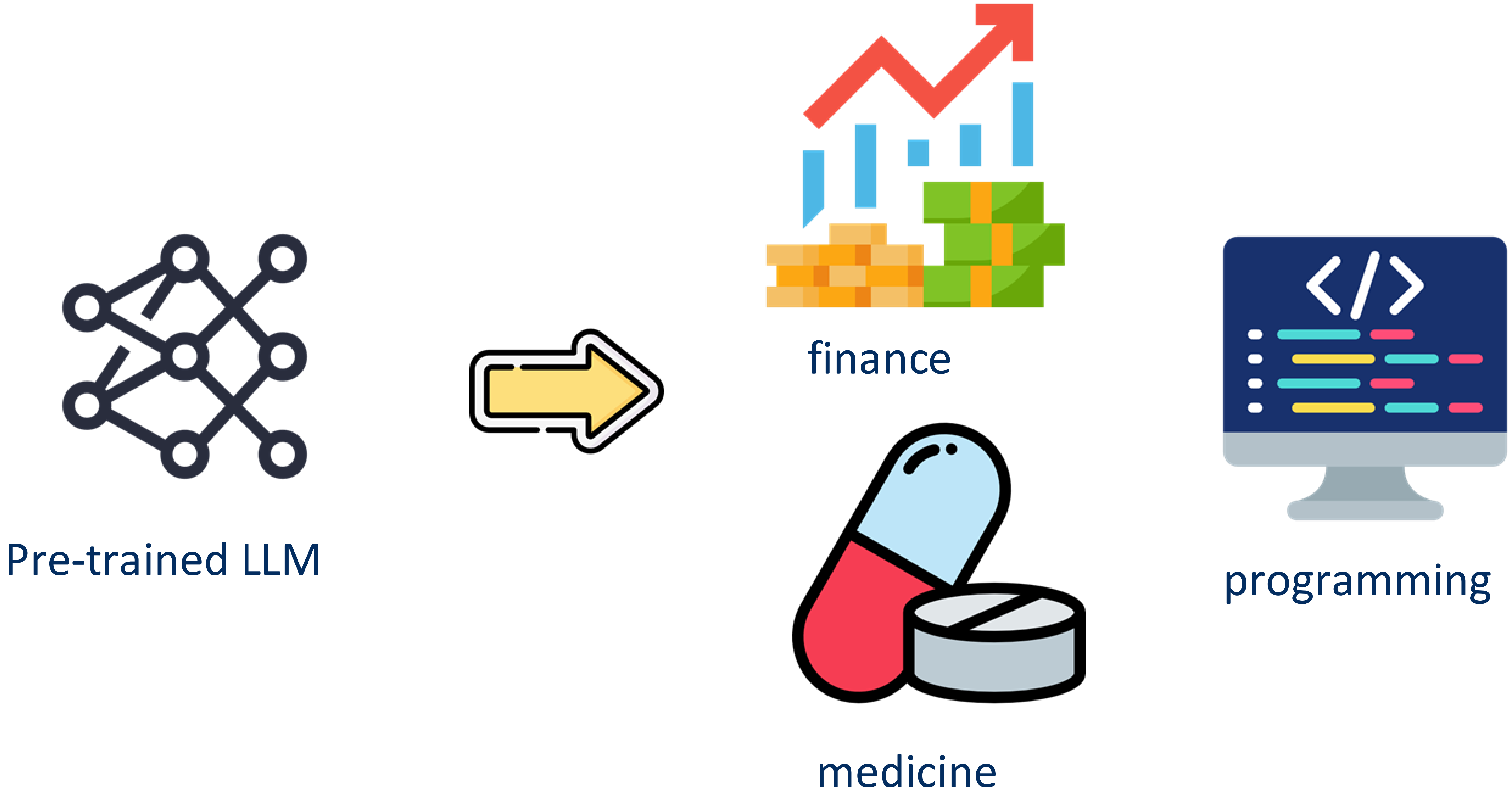


Figure 1: Existing benchmarks show counterintuitive trends, such as smaller models outperforming larger ones (e.g., Llama-3.1 8B > 70B).

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

If we want to adapt LLMs to specialized domains…

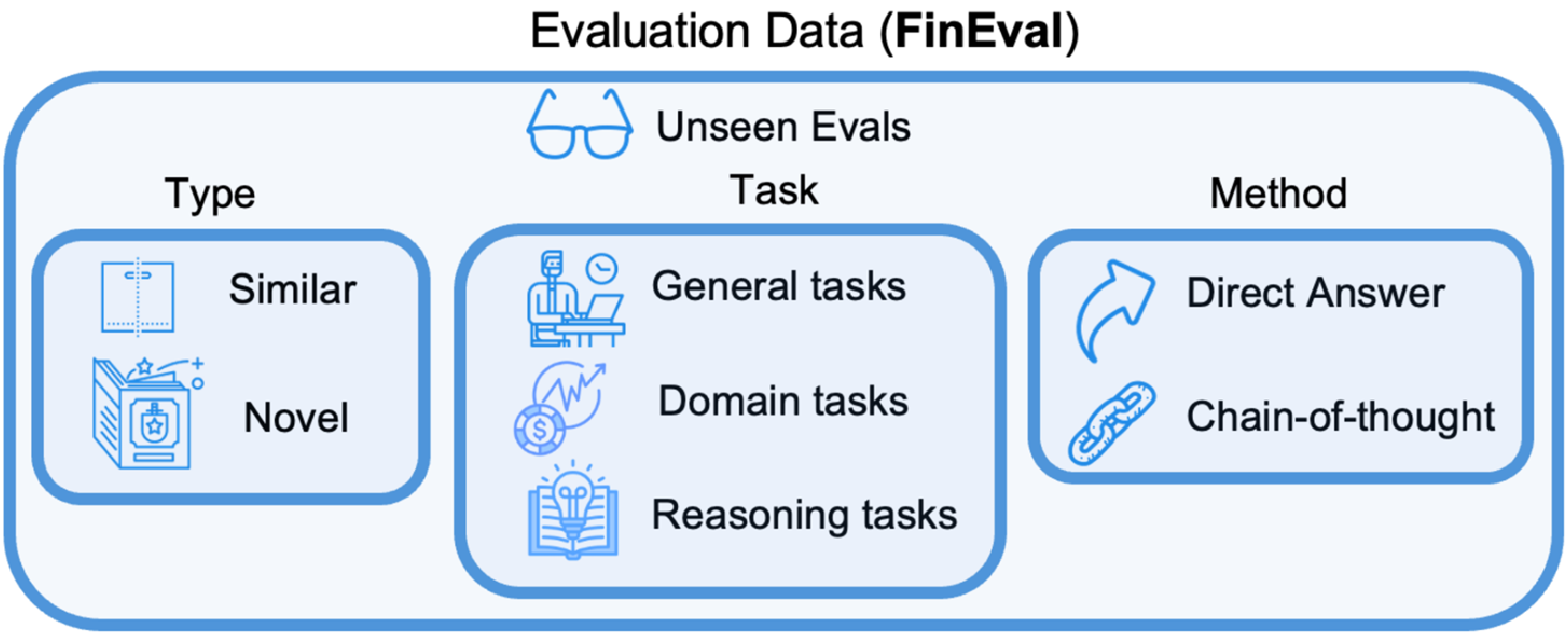
Adapting LLMs to Specialized Domains



- ❑ 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