Working with Reasoning Models

How and when to use LLMs for thinking and reasoning

Instructor: Lucas B. Nicolosi Soares

Agenda

1. What is a Reasoning LLM?

Agenda

1. What is a Reasoning LLM?
2. Reasoning Models vs Traditional LLMs

Agenda

1. What is a Reasoning LLM?
2. Reasoning Models vs Traditional LLMs
3. Why use a Reasoning LLM?

Agenda

1. What is a Reasoning LLM?
2. Reasoning Models vs Traditional LLMs
3. Why use a Reasoning LLM?
4. When to use a Reasoning LLM?

Agenda

1. What is a Reasoning LLM?
2. Reasoning Models vs Traditional LLMs
3. Why use a Reasoning LLM?
4. When to use a Reasoning LLM?
5. How to Use a Reasoning LLM?

Agenda

1. What is a Reasoning LLM?
2. Reasoning Models vs Traditional LLMs
3. Why use a Reasoning LLM?
4. When to use a Reasoning LLM?
5. How to Use a Reasoning LLM?
6. Limitations of Reasoning LLMs

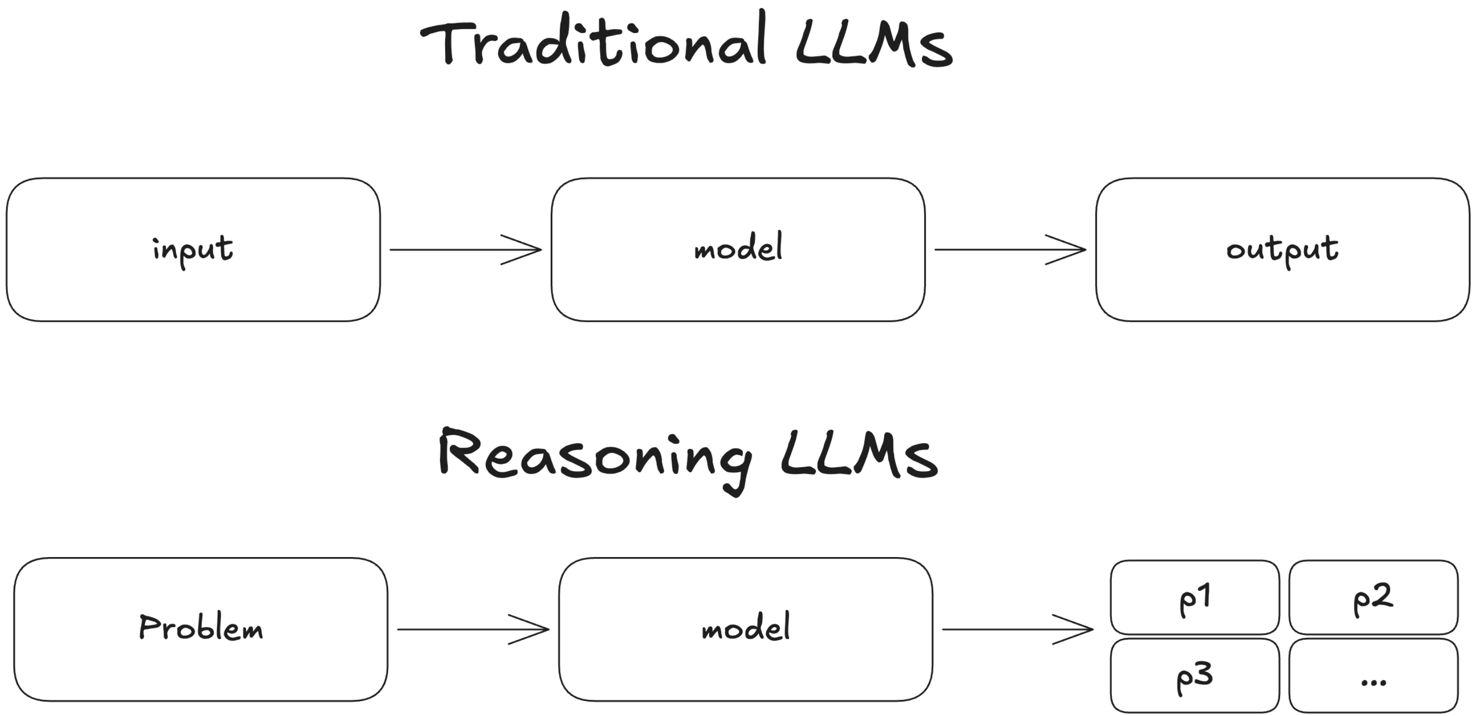
Agenda

1. What is a Reasoning LLM?
2. Reasoning Models vs Traditional LLMs
3. Why use a Reasoning LLM?
4. When to use a Reasoning LLM?
5. How to Use a Reasoning LLM?
6. Limitations of Reasoning LLMs
7. Choosing a Reasoning LLM

Agenda

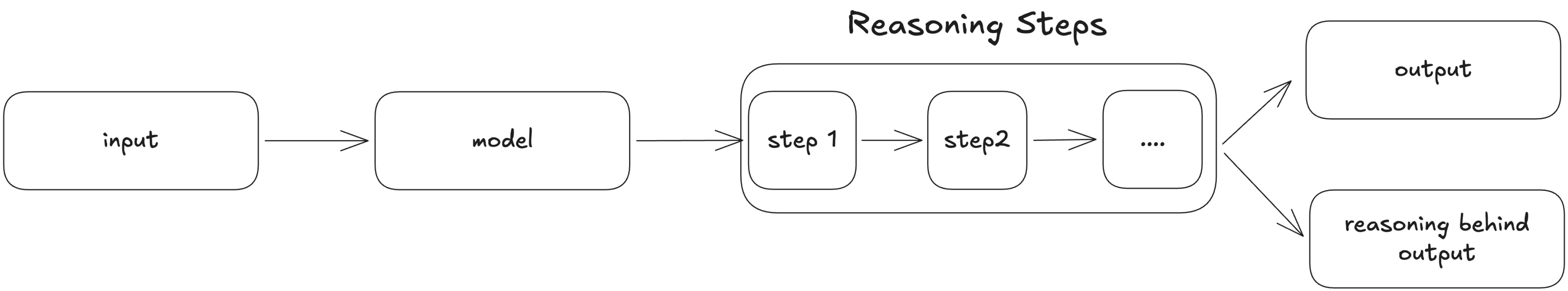
1. What is a Reasoning LLM?
2. Reasoning Models vs Traditional LLMs
3. Why use a Reasoning LLM?
4. When to use a Reasoning LLM?
5. How to Use a Reasoning LLM?
6. Limitations of Reasoning LLMs
7. Choosing a Reasoning LLM
8. Hands-on with Reasoning LLMs

What is a Reasoning LLM?

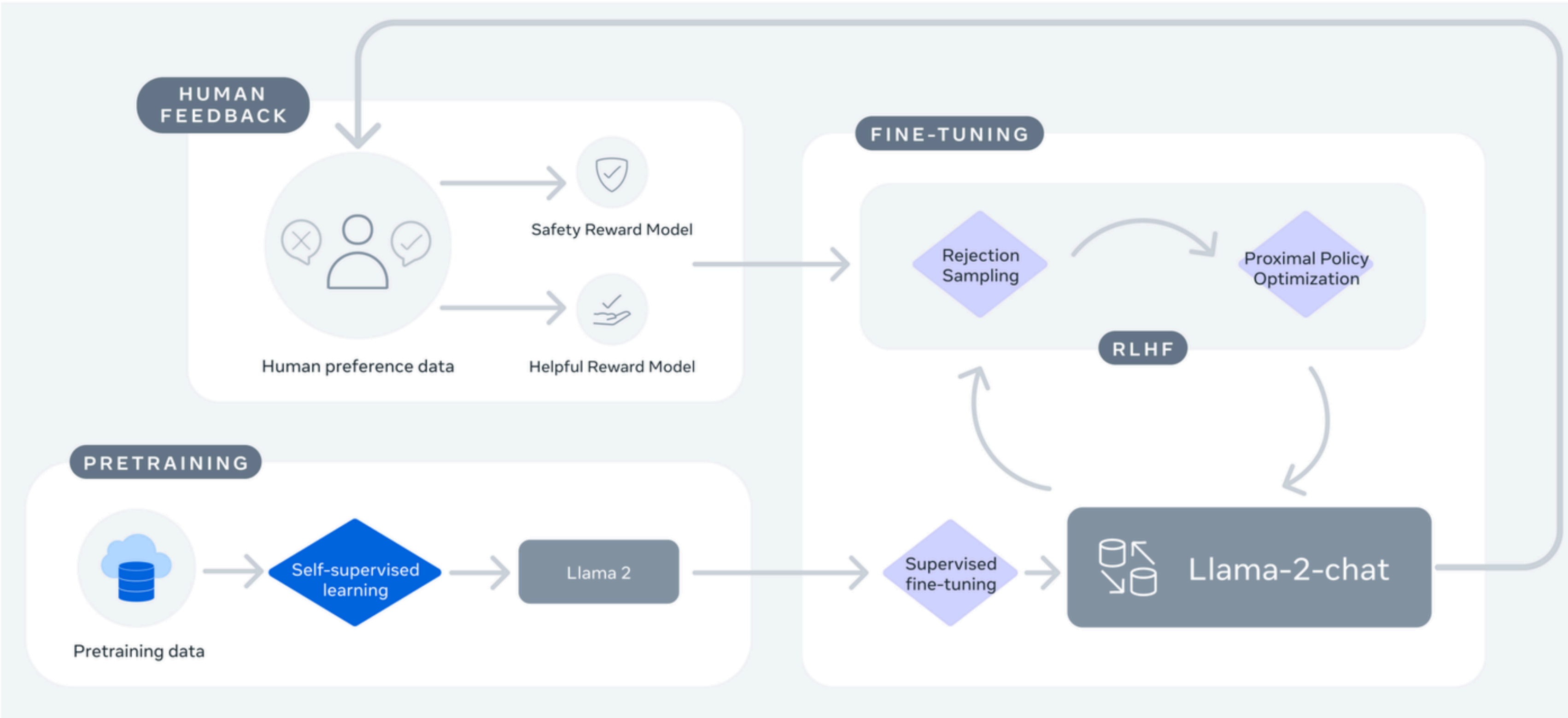


How?

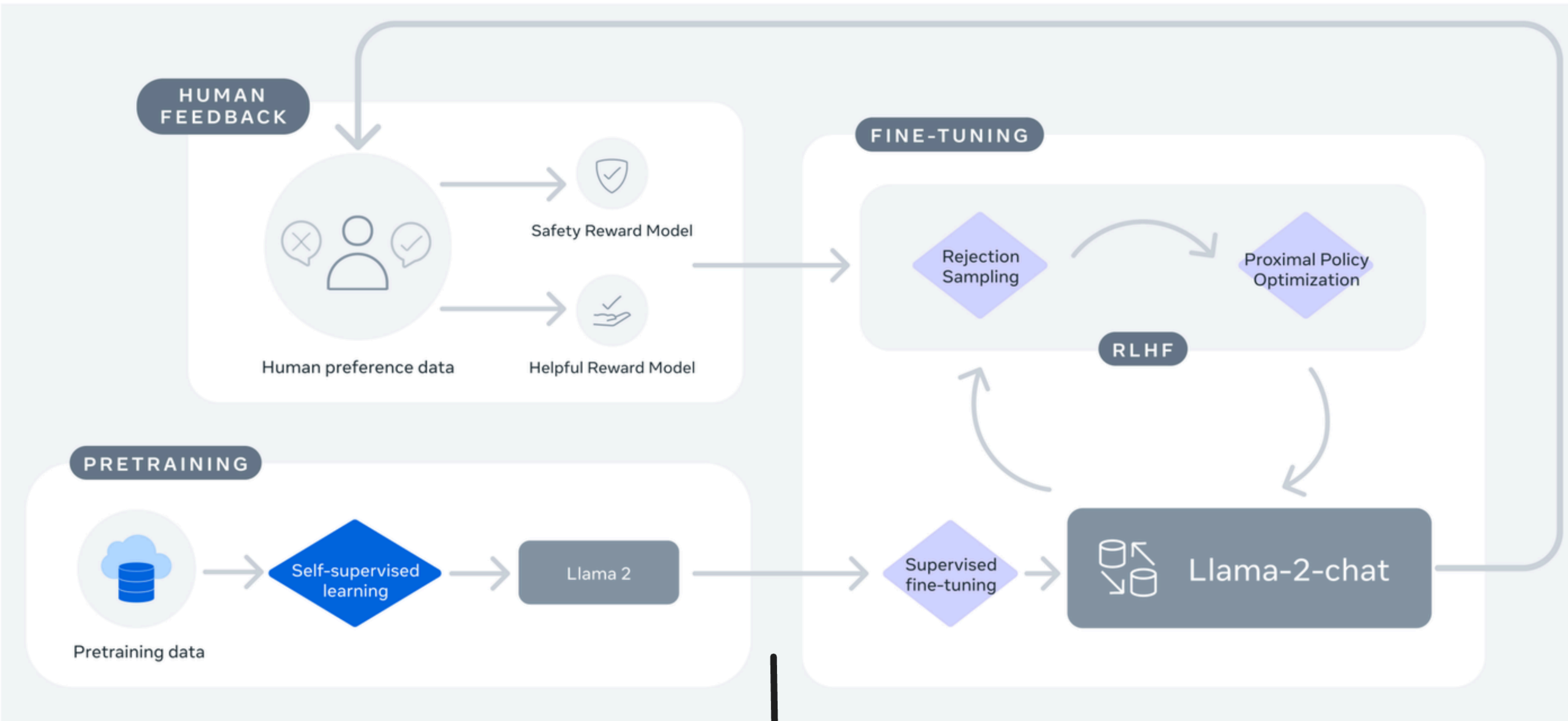
How?


Reasoning Steps
input
model
step 1
step2
....
output
reasoning behind output

How an LLM is Trained - Busy Person Guide



How an LLM is Trained - Busy Person Guide



fine tuning on thousands of chain of thought examples

What is a Reasoning LLM?

- Language models designed for complex, multi-step problems

What is a Reasoning LLM?

- Language models designed for complex, multi-step problems
- Generate intermediate reasoning steps or "thought processes"

What is a Reasoning LLM?

- Language models designed for complex, multi-step problems
- Generate intermediate reasoning steps or "thought processes"
- Break down tasks into logical sub-steps

What is a Reasoning LLM?

- Language models designed for complex, multi-step problems
- Generate intermediate reasoning steps or "thought processes"
- Break down tasks into logical sub-steps
- Similar to human step-by-step problem solving

What is a Reasoning LLM?

- Language models designed for complex, multi-step problems
- Generate intermediate reasoning steps or "thought processes"
- Break down tasks into logical sub-steps
- Similar to human step-by-step problem solving
- More accurate and explainable results on challenging tasks:
  - Mathematical problems
  - Logic puzzles
  - Code debugging

Q&A & Break

Reasoning Models vs Traditional LLMs

Traditional LLMs

- Direct pattern-based prediction

Traditional LLMs

- Direct pattern-based prediction
- One-shot "I'm feeling lucky" approach

Traditional LLMs

- Direct pattern-based prediction
- One-shot "I'm feeling lucky" approach
- Quick but less reliable on complex tasks

Traditional LLMs

- Direct pattern-based prediction
- One-shot "I'm feeling lucky" approach
- Quick but less reliable on complex tasks
- No explicit reasoning steps

Reasoning LLMs

Traditional LLMs

- Direct pattern-based prediction
- One-shot "I'm feeling lucky" approach
- Quick but less reliable on complex tasks
- No explicit reasoning steps

Reasoning LLMs

- Step-by-step problem solving

Traditional LLMs

- Direct pattern-based prediction
- One-shot "I'm feeling lucky" approach
- Quick but less reliable on complex tasks
- No explicit reasoning steps

Reasoning LLMs

- Step-by-step problem solving
- Chain-of-thought (CoT) approach

Traditional LLMs

- Direct pattern-based prediction
- One-shot "I'm feeling lucky" approach
- Quick but less reliable on complex tasks
- No explicit reasoning steps

Reasoning LLMs

- Step-by-step problem solving
- Chain-of-thought (CoT) approach
- More methodical but slower

Traditional LLMs

- Direct pattern-based prediction
- One-shot "I'm feeling lucky" approach
- Quick but less reliable on complex tasks
- No explicit reasoning steps

Reasoning LLMs

- Step-by-step problem solving
- Chain-of-thought (CoT) approach
- More methodical but slower
- Shows intermediate steps

Q&A & Break

Key Capabilities of Reasoning LLMs

1. **Chain-of-Thought Reasoning**

Key Capabilities of Reasoning LLMs

1. **Chain-of-Thought Reasoning**
    - Internal dialogue approach

Key Capabilities of Reasoning LLMs

1. **Chain-of-Thought Reasoning**
    - Internal dialogue approach
    - Step-by-step problem solving

Key Capabilities of Reasoning LLMs

1. **Chain-of-Thought Reasoning**
    - Internal dialogue approach
    - Step-by-step problem solving
    - "Thinking aloud" in text

Key Capabilities of Reasoning LLMs

1. **Chain-of-Thought Reasoning**
    - Internal dialogue approach
    - Step-by-step problem solving
    - "Thinking aloud" in text
2. **Self-Consistency**

Key Capabilities of Reasoning LLMs

1. **Chain-of-Thought Reasoning**
   - Internal dialogue approach
   - Step-by-step problem solving
   - "Thinking aloud" in text
2. **Self-Consistency**
   - Verifies own answers

Key Capabilities of Reasoning LLMs

1. **Chain-of-Thought Reasoning**
    - Internal dialogue approach
    - Step-by-step problem solving
    - "Thinking aloud" in text
2. **Self-Consistency**
    - Verifies own answers
    - Revisits problematic solutions

Key Capabilities of Reasoning LLMs

1. **Chain-of-Thought Reasoning**
   - Internal dialogue approach
   - Step-by-step problem solving
   - "Thinking aloud" in text
2. **Self-Consistency**
   - Verifies own answers
   - Revisits problematic solutions
3. **Structured Outputs**

Key Capabilities of Reasoning LLMs

1. **Chain-of-Thought Reasoning**
   - Internal dialogue approach
   - Step-by-step problem solving
   - "Thinking aloud" in text
2. **Self-Consistency**
   - Verifies own answers
   - Revisits problematic solutions
3. **Structured Outputs**
   - Organized reasoning steps

Key Capabilities of Reasoning LLMs

1. **Chain-of-Thought Reasoning**
   - Internal dialogue approach
   - Step-by-step problem solving
   - "Thinking aloud" in text
2. **Self-Consistency**
   - Verifies own answers
   - Revisits problematic solutions
3. **Structured Outputs**
   - Organized reasoning steps
   - Numbered thoughts

Key Capabilities of Reasoning LLMs

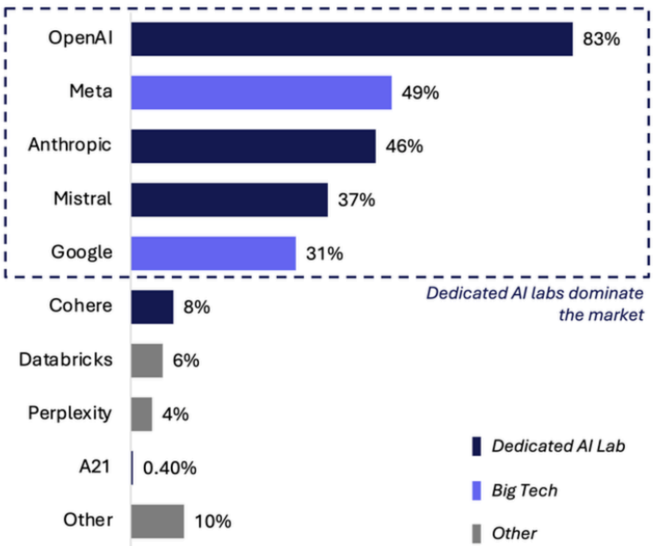
1. **Chain-of-Thought Reasoning**
   - Internal dialogue approach
   - Step-by-step problem solving
   - "Thinking aloud" in text
2. **Self-Consistency**
   - Verifies own answers
   - Revisits problematic solutions
3. **Structured Outputs**
   - Organized reasoning steps
   - Numbered thoughts
     - Traceable calculations

Why use a Reasoning LLM?

Demand for AI models is concentrated on releases from top AI labs; model reasoning quality and price are the primary decision drivers for choosing models

Model Demand by Provider

Which LLMs are you using or considering using?, N=270



Importance of Model Decision Criteria

How important are these criteria to you when choosing a model?, N=250

	Not important	Less Important	Important	Very important
Reasoning quality	0%	2.8%	32.5%	64.7%
Embedded knowledge	3.7%	21.5%	37.2%	37.6%
Context window	3.6%	20.6%	47.4%	28.3%
Speed / Throughput (Tokens ...	3.2%	21%	39.9%	35.9%
Latency (Time to First Token)	4.9%	27.6%	38.3%	29.2%
Price	2.4%	14.4%	33.2%	50%
Open-source	19.4%	31.2%	26.7%	22.7%
Function calling	8.5%	27.2%	38.6%	25.6%
JSON mode	13.4%	27.6%	34.6%	24.4%

[Artificial Analysis AI Review]

When to use a Reasoning LLM?

When to use a Reasoning LLM?

- Background tasks where latency isn't critical

When to use a Reasoning LLM?

- Background tasks where latency isn't critical
- Complex problems requiring deeper thinking

When to use a Reasoning LLM?

- Background tasks where latency isn't critical
- Complex problems requiring deeper thinking
- Tasks benefiting from extensive reasoning

When to use a Reasoning LLM?

- Background tasks where latency isn't critical
- Complex problems requiring deeper thinking
- Tasks benefiting from extensive reasoning
- Research and planning-heavy workflows

Ideal Use Cases

Ideal Use Cases

- **Complex Problem Solving**
    - Mathematical proofs
    - Logic puzzles
    - Multi-step reasoning

Ideal Use Cases

- **Complex Problem Solving**
  - Mathematical proofs
  - Logic puzzles
  - Multi-step reasoning
- **Deep Analysis**
  - Research papers
  - Document analysis
  - Meeting notes interpretation

Ideal Use Cases

- **Complex Problem Solving**
  - Mathematical proofs
  - Logic puzzles
  - Multi-step reasoning
- **Deep Analysis**
  - Research papers
  - Document analysis
  - Meeting notes interpretation
- **Planning & Agency**
  - Workflow planning
  - Agentic systems
  - Strategic decision-making

Practical Applications

Practical Applications

- **Data Analysis**
  - Medical diagnostics
  - Complex data interpretation
  - Anomaly detection

Practical Applications

- **Data Analysis**
  - Medical diagnostics
  - Complex data interpretation
  - Anomaly detection
- **Background Processing**
  - Batch processing workflows
  - Overnight analysis jobs

Practical Applications

- **Data Analysis**
  - Medical diagnostics
  - Complex data interpretation
  - Anomaly detection
- **Background Processing**
  - Batch processing workflows
  - Overnight analysis jobs
- **Evaluation Tasks**
  - LLM as judge
  - Quality assessment
  - Verification workflows

Q&A & Break

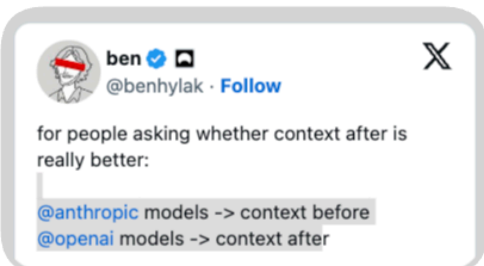
How to Use a Reasoning LLM?

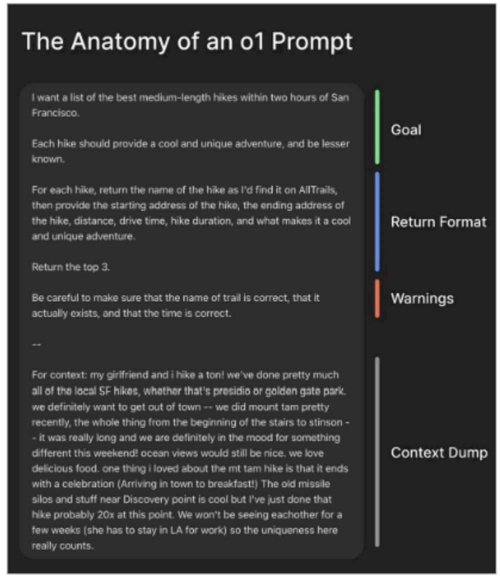
Whiteboard - How to Use a Reasoning LLM?

Hands-on - How to Use a Reasoning LLM?

How to Prompt Reasoning LLMs

1. prompts x briefs
2. "report generator."
3. push as much context as you can into o1.
4. use superwhisper for 1 minute then paste that into the model
5. develop good criteria for what is goodxbad





The Anatomy of an o1 Prompt

I want a list of the best medium-length hikes within two hours of San Francisco.

Each hike should provide a cool and unique adventure, and be lesser known.

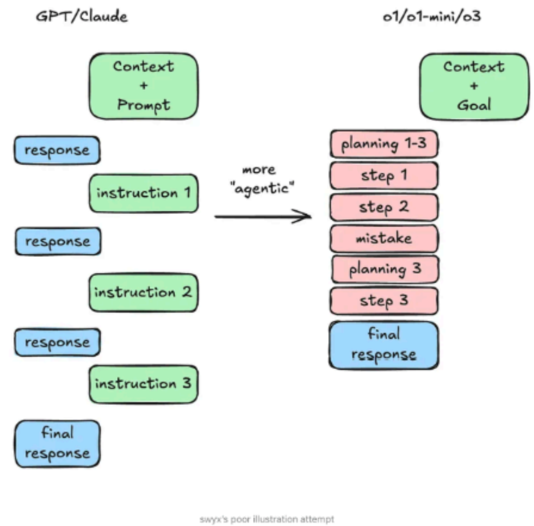
For each hike, return the name of the hike as I'd find it on AllTrails, then provide the starting address of the hike, the ending address of the hike, distance, drive time, hike duration, and what makes it a cool and unique adventure.

Return the top 3.

Be careful to make sure that the name of trail is correct, that it actually exists, and that the time is correct.

--

For context: my girlfriend and i hike a ton! we've done pretty much all of the local SF hikes, whether that's presidio or golden gate park. we definitely want to get out of town -- we did mount tam pretty recently, the whole thing from the beginning of the stairs to stinson -- it was really long and we are definitely in the mood for something different this weekend! ocean views would still be nice. we love delicious food. one thing i loved about the mt tam hike is that it ends with a celebration (Arriving in town to breakfast!) The old missile silos and stuff near Discovery point is cool but I've just done that hike probably 20x at this point. We won't be seeing eachother for a few weeks (she has to stay in LA for work) so the uniqueness here really counts.



Q&A & Break

Limitations of Reasoning LLMs

1. Performance Trade-offs

- **Increased latency**: Extended thinking process leads to significantly longer response times

1. Performance Trade-offs

- **Increased latency**: Extended thinking process leads to significantly longer response times
- **Higher resource requirements**: Often require more computational resources

1. Performance Trade-offs

- **Increased latency**: Extended thinking process leads to significantly longer response times
- **Higher resource requirements**: Often require more computational resources
- **Cost implications**: More tokens and processing time translate to higher operational costs

1. Performance Trade-offs

- **Increased latency**: Extended thinking process leads to significantly longer response times
- **Higher resource requirements**: Often require more computational resources
- **Cost implications**: More tokens and processing time translate to higher operational costs
- **Scalability challenges**: Multiple concurrent requests become more resource-intensive

2. Reasoning Quality Constraints

- **Domain-specific limitations**: Performance varies considerably across different subject areas

2. Reasoning Quality Constraints

- **Domain-specific limitations**: Performance varies considerably across different subject areas
- **Brittleness with novel problems**: May struggle with problem types not encountered during training

2. Reasoning Quality Constraints

- **Domain-specific limitations**: Performance varies considerably across different subject areas
- **Brittleness with novel problems**: May struggle with problem types not encountered during training
- **Inconsistent depth of reasoning**: Quality of reasoning can vary within the same model

2. Reasoning Quality Constraints

- **Domain-specific limitations**: Performance varies considerably across different subject areas
- **Brittleness with novel problems**: May struggle with problem types not encountered during training
- **Inconsistent depth of reasoning**: Quality of reasoning can vary within the same model
- **Over-confidence in incorrect reasoning**: May present flawed reasoning with high confidence

3. Knowledge and Context Limitations

- **Bounded knowledge**: Limited to training data available at model development time

3. Knowledge and Context Limitations

- **Bounded knowledge**: Limited to training data available at model development time
- **Context window constraints**: Struggle with very lengthy problems despite large context windows

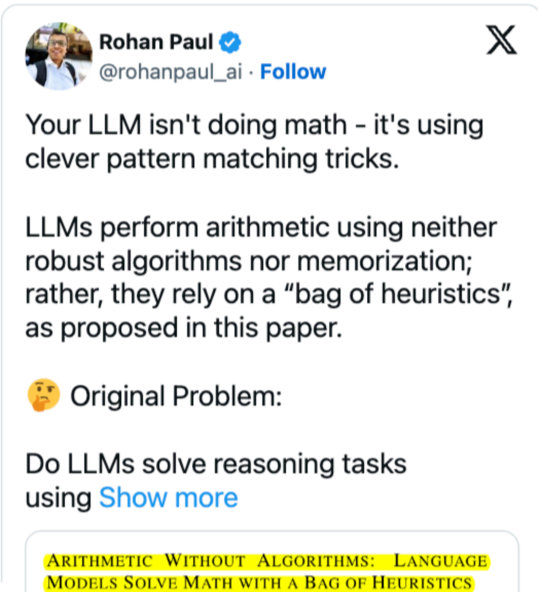
3. Knowledge and Context Limitations

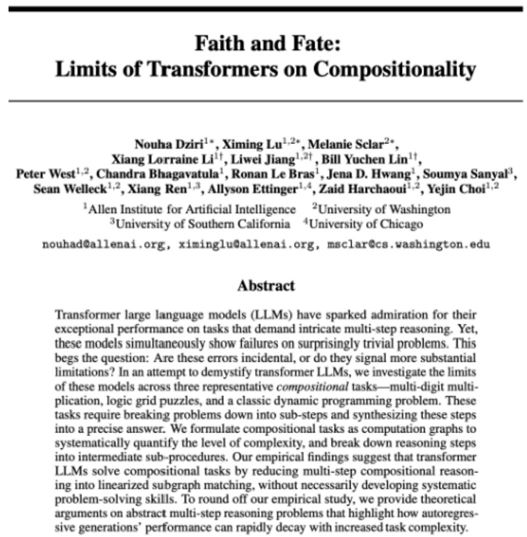
- **Bounded knowledge**: Limited to training data available at model development time
- **Context window constraints**: Struggle with very lengthy problems despite large context windows
- **Information integration challenges**: Difficulty maintaining coherence across extensive reasoning chains

3. Knowledge and Context Limitations

- **Bounded knowledge**: Limited to training data available at model development time
- **Context window constraints**: Struggle with very lengthy problems despite large context windows
- **Information integration challenges**: Difficulty maintaining coherence across extensive reasoning chains
- **Limited transfer learning**: Reasoning in one domain doesn't always transfer to other domains

Limitations of Reasoning LLMs





Faith and Fate:
Limits of Transformers on Compositionality

Nouha Dziri[1*], Ximing Lu[1,2*], Melanie Sclar[2*], Xiang Lorraine Li[1†], Liwei Jiang[1,2†], Bill Yuchen Lin[1†], Peter West[1,2], Chandra Bhagavatula[1], Ronan Le Bras[1], Jena D. Hwang[1], Soumya Sanyal[3], Sean Welleck[1,2], Xiang Ren[1,3], Allyson Ettinger[1,4], Zaid Harchaoui[1,2], Yejin Choi[1,2]

[1]Allen Institute for Artificial Intelligence [2]University of Washington
[3]University of Southern California [4]University of Chicago

nouhad@allenai.org, ximinglu@allenai.org, msclar@cs.washington.edu

Abstract

Transformer large language models (LLMs) have sparked admiration for their exceptional performance on tasks that demand intricate multi-step reasoning. Yet, these models simultaneously show failures on surprisingly trivial problems. This begs the question: Are these errors incidental, or do they signal more substantial limitations? In an attempt to demystify transformer LLMs, we investigate the limits of these models across three representative *compositional* tasks—multi-digit multiplication, logic grid puzzles, and a classic dynamic programming problem. These tasks require breaking problems down into sub-steps and synthesizing these steps into a precise answer. We formulate compositional tasks as computation graphs to systematically quantify the level of complexity, and break down reasoning steps into intermediate sub-procedures. Our empirical findings suggest that transformer LLMs solve compositional tasks by reducing multi-step compositional reasoning into linearized subgraph matching, without necessarily developing systematic problem-solving skills. To round off our empirical study, we provide theoretical arguments on abstract multi-step reasoning problems that highlight how autoregressive generations' performance can rapidly decay with increased task complexity.

https://arxiv.org/pdf/2305.18654

arXiv:2307.02477v3 [cs.CL] 28 Mar 2024

Reasoning or Reciting? Exploring the Capabilities and Limitations of Language Models Through Counterfactual Tasks

Zhaofeng Wu[@] Linlu Qiu[@] Alexis Ross[@] Ekin Akyürek[@] Boyuan Chen[@] Bailin Wang[@] Najoung Kim[Ω] Jacob Andreas[@] Yoon Kim[@]
[@]MIT [Ω]Boston University
zfw@csail.mit.edu

Figure 1: GPT-4's performance on the default version of various tasks (blue) and counterfactual counterparts (orange). The shown results use 0-shot chain-of-thought prompting (§4; Kojima et al., 2023). GPT-4 consistently and substantially underperforms on counterfactual variants compared to default task instantiations.

[Dziri et al. (2023)](#)

[Wu et al. (2024)](#)

Q&A & Break

Choosing a Reasoning LLM

Whiteboard - Choosing a Reasoning LLM

Hands-on with Reasoning LLMs

Connect With Me

📚 Course materials

🔗 LinkedIn

🐦 Twitter/X

📺 YouTube

📧 Email: lucasenkrateia@gmail.com