When to USE Reasoning LLMs

Good fit for Reasoning LLMs

1. Generate single files for complex problems
2. Hallucinates less
3. Medical Diagnoses
4. Explanations
5. Evals from other LLMs

Poor Fit for Reasoning LLMs

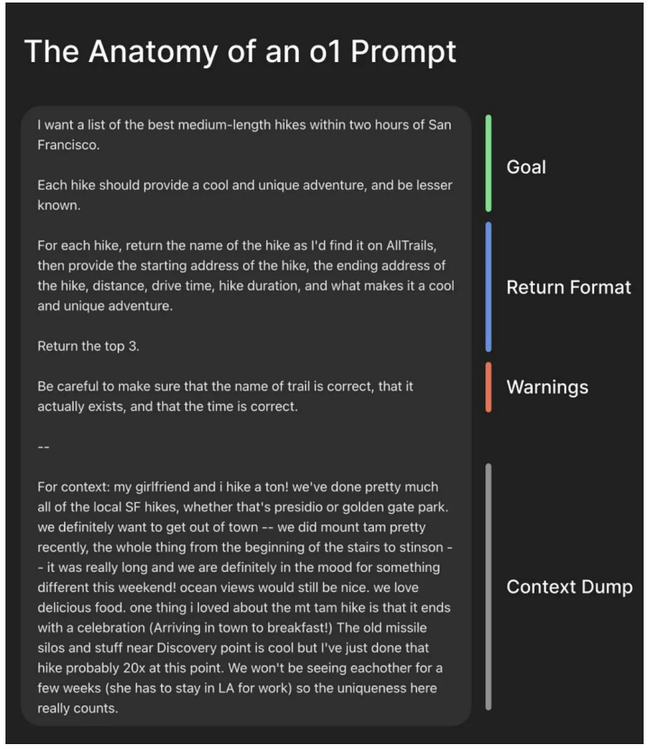
1. Writing in specific styles
2. Building full apps

1. Navigating ambiguous tasks
2. Finding a needle in a haystack
3. Finding relationships and nuance across a large dataset

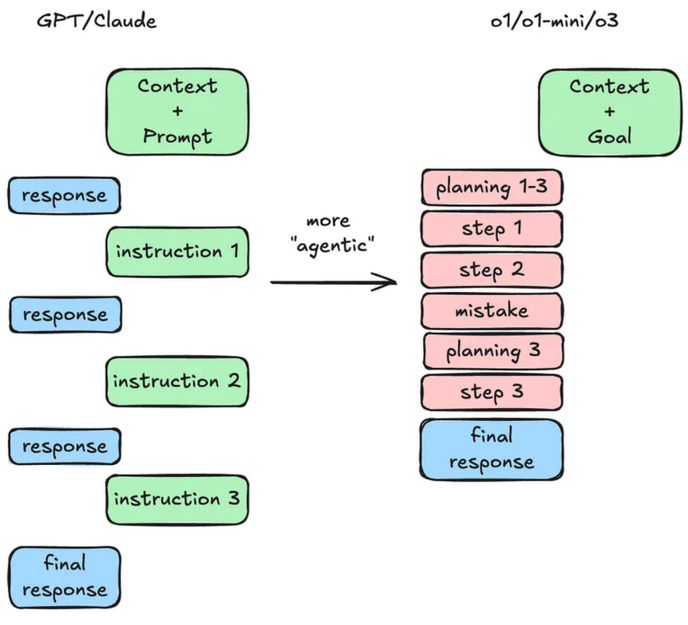
How to Prompt Reasoning LLMs

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 good x bad

https://x.com/benhylak/status/1878514144766480777



https://www.latent.space/p/o1-skill-issue

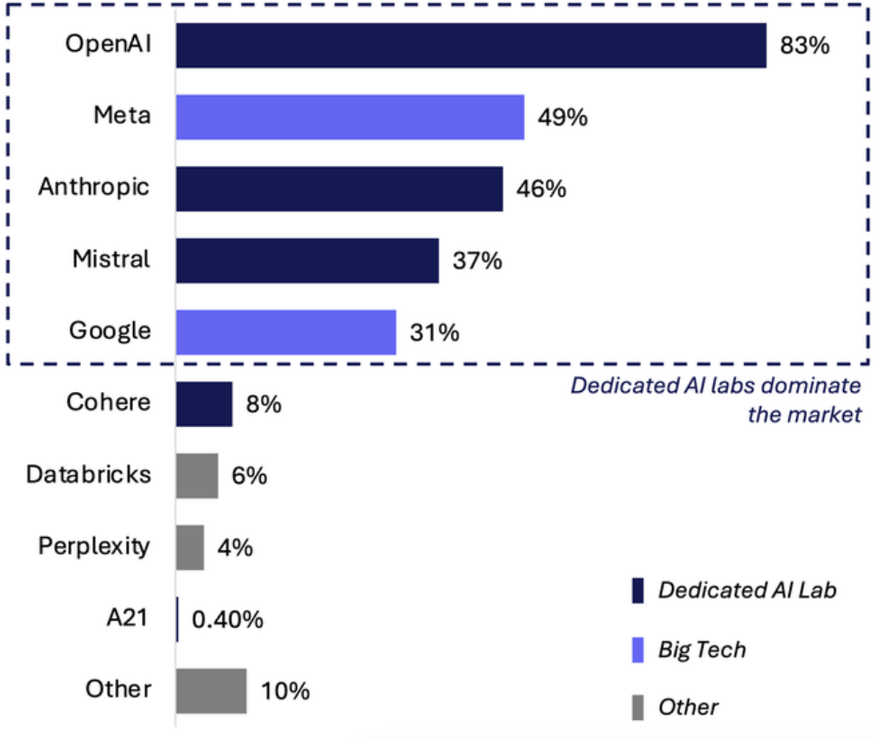


swyx's poor illustration attempt

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%

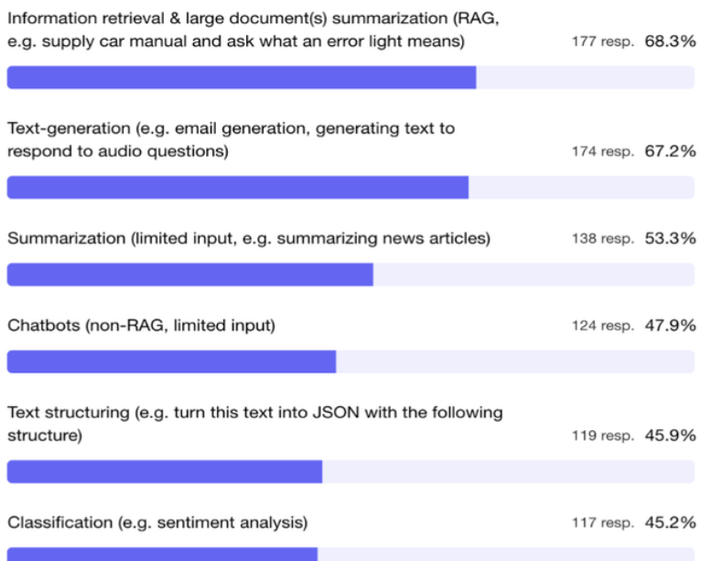
What Matters when Choosing LLMs?

1. reasoning quality
2. Price
3. Embedded knowledge
4. Speed!
5. context window

Companies are using LLMs in a wide range of technical approaches with no single approach dominant; most LLM users intend to use multimodal capabilities

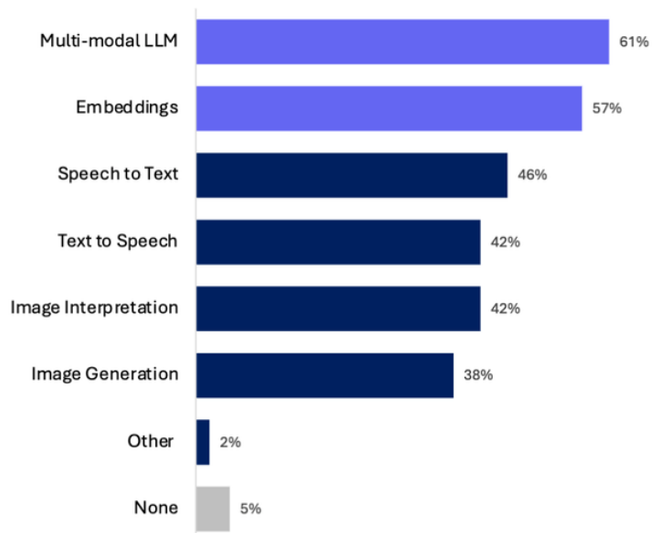
Adoption of Technical Approaches for Using LLMs

What technical uses do you intend to use LLMs for?, N=242



Demand for Multimodal Capabilities

What other AI capabilities do you use or intend to use?, N=252



Note: Results from the Artificial Analysis Developer Survey conducted from March to August 2024. Respondents represented a range of organization sizes and locations. Results should be considered indicative only and may be biased by Artificial Analysis's audience and impacted by survey timing.

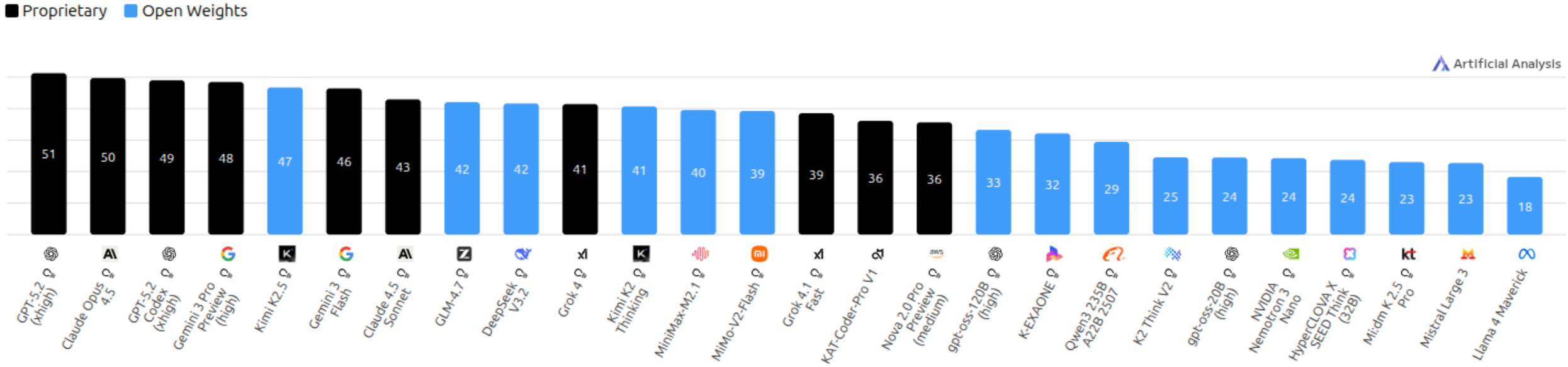
Artificial Analysis

Artificial Analysis Intelligence Index by Open Weights vs Proprietary

Artificial Analysis Intelligence Index v4.0 incorporates 10 evaluations: GDPval-AA, τ²-Bench Telecom, Terminal-Bench Hard, SciCode, AA-LCR, AA-Omniscience, IFBench, Humanity's Last Exam, GPQA Diamond, CritPt

26 of 392 models

+ Add model from specific provider



Artificial Analysis Intelligence Index

https://x.com/rohanpaul_ai/status/1860341557553168620

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

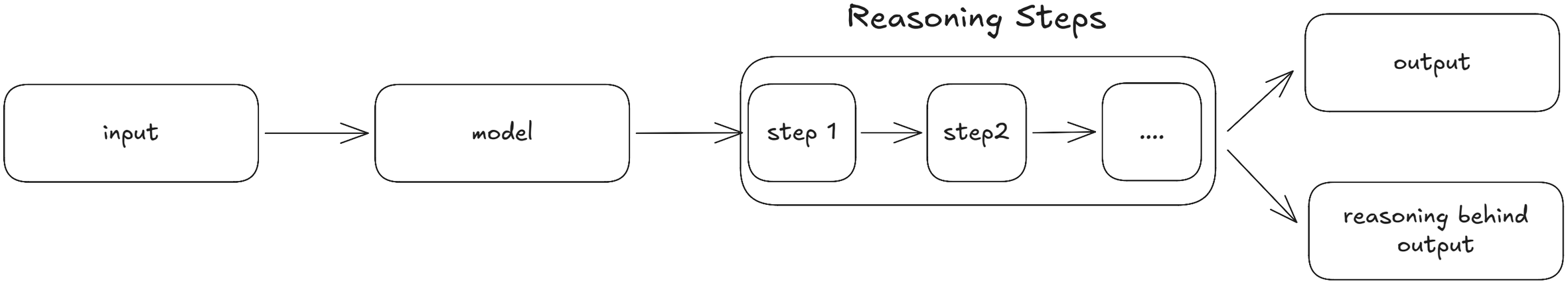
Transformer large language models (LLMs) have sparked admiration for their exceptional performance on tasks that demand intricate multi-step reasoning. Yet, these models simultaneously show failures on surprisingly trivial problems. This begs the question: Are these errors incidental, or do they signal 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.

- LLMs do not "think" or engage in systematic multi-step reasoning.
- Instead, they perform heuristic linear approximations by matching linearized subgraphs from their training data—essentially pattern matching—rather than applying genuine problem-solving algorithms.

What is a Reasoning LLM?

Traditional LLMs

Reasoning LLMs

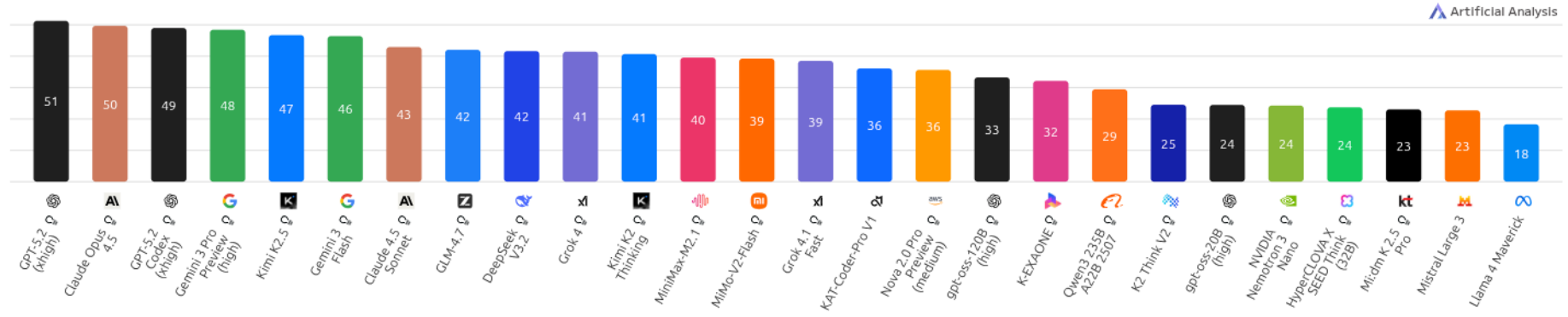


Artificial Analysis Intelligence Index

Artificial Analysis Intelligence Index v4.0 incorporates 10 evaluations: GDPval-AA, τ²-Bench Telecom, Terminal-Bench Hard, SciCode, AA-LCR, AA-Omniscience, IFBench, Humanity's Last Exam, GPQA Diamond, CritPt

26 of 392 models

+ Add model from specific provider



Artificial Analysis Intelligence Index

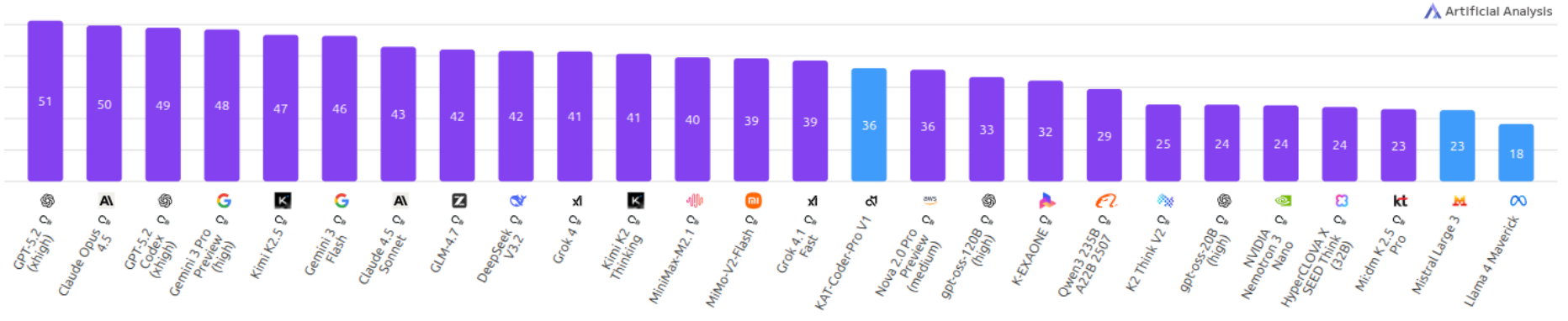
Artificial Analysis Intelligence Index by Model Type

Artificial Analysis Intelligence Index v4.0 incorporates 10 evaluations: GDPval-AA, τ²-Bench Telecom, Terminal-Bench Hard, SciCode, AA-LCR, AA-Omniscience, IFBench, Humanity's Last Exam, GPQA Diamond, CritPt

26 of 392 models

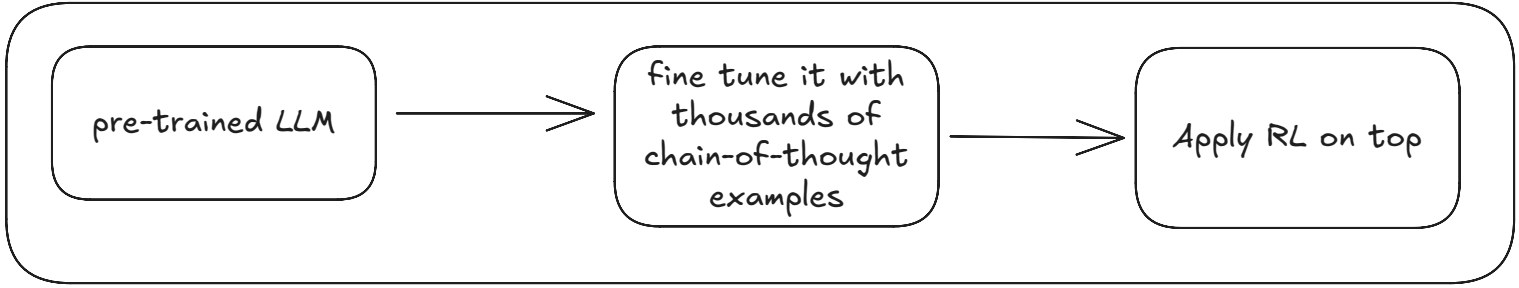
+ Add model from specific provider

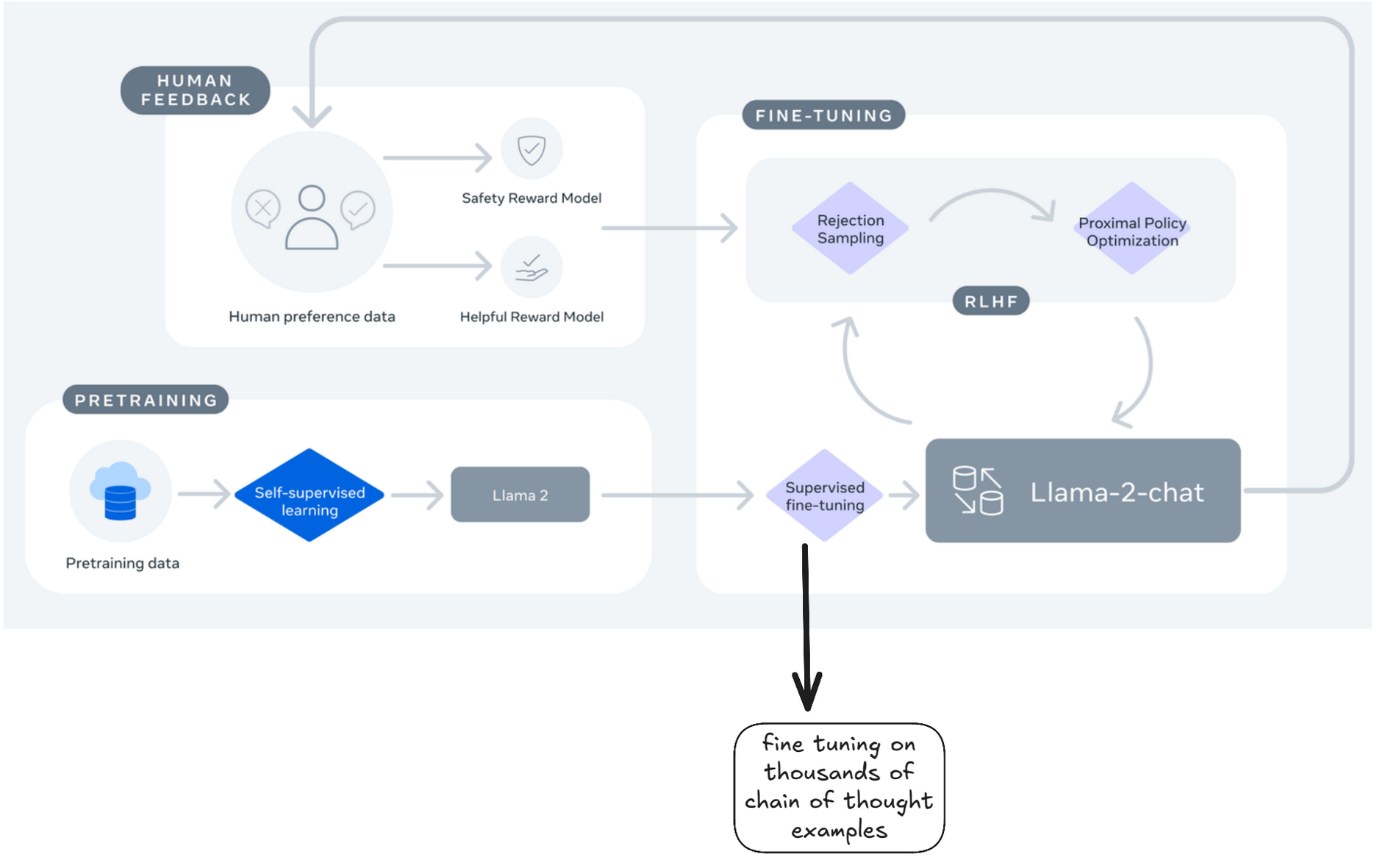
Reasoning Model Non-Reasoning Model



Artificial Analysis Intelligence Index

Llama 2 Paper


pre-trained LLM
fine tune it with thousands of chain-of-thought examples
Apply RL on top



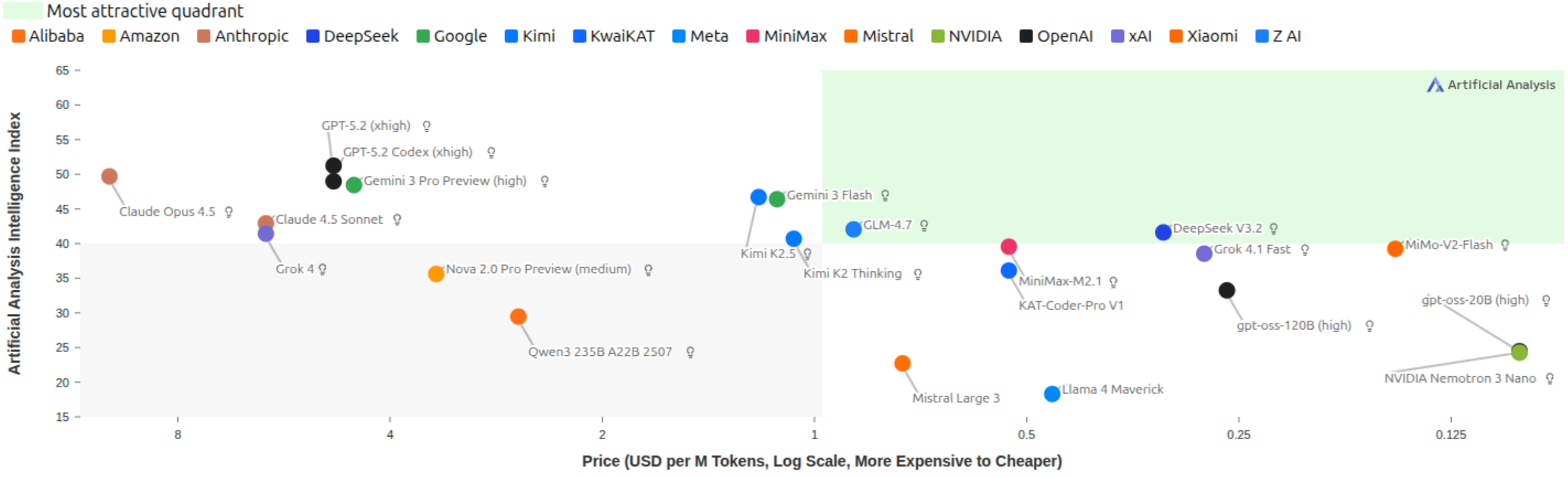
Stage	Traditional LLM	Thinking Model (DeepSeek-R1)
Pretrain	Internet-scale corpus	Same
SFT	Human-annotated QA	Chain-of-Thought (CoT) data
RL	Helpfulness/safety reward	Reasoning reward + Format reward + Language consistency
Emergent Behavior	Few CoTs	Self-verification, long CoTs, reflection

Intelligence vs. Price (Log Scale)

Artificial Analysis Intelligence Index; Price: USD per 1M Tokens; Inspired by prior analysis by Swyx

27 of 392 models

+ Add model from specific provider

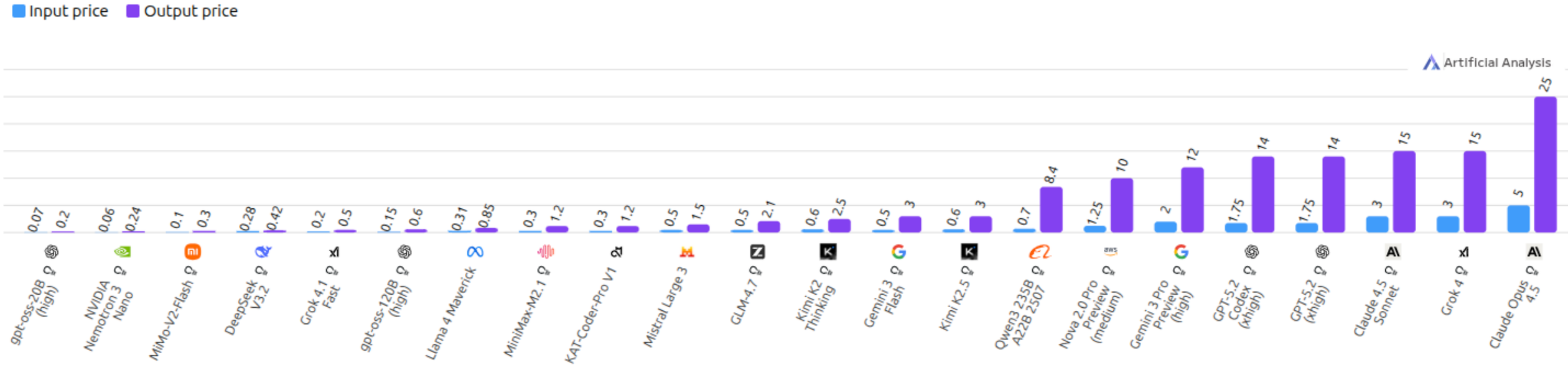


Pricing: Input and Output Prices

Price: USD per 1M Tokens

27 of 392 models

+ Add model from specific provider



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

DeepSeek-AI

research@deepseek.com

Abstract

We introduce our first-generation reasoning models, DeepSeek-R1-Zero and DeepSeek-R1. DeepSeek-R1-Zero, a model trained via large-scale reinforcement learning (RL) without supervised fine-tuning (SFT) as a preliminary step, demonstrates remarkable reasoning capabilities. Through RL, DeepSeek-R1-Zero naturally emerges with numerous powerful and intriguing reasoning behaviors. However, it encounters challenges such as poor readability, and language mixing. To address these issues and further enhance reasoning performance, we introduce DeepSeek-R1, which incorporates multi-stage training and cold-start data before RL. DeepSeek-R1 achieves performance comparable to OpenAI-o1-1217 on reasoning tasks. To support the research community, we open-source DeepSeek-R1-Zero, DeepSeek-R1, and six dense models (1.5B, 7B, 8B, 14B, 32B, 70B) distilled from DeepSeek-R1 based on Qwen and Llama.

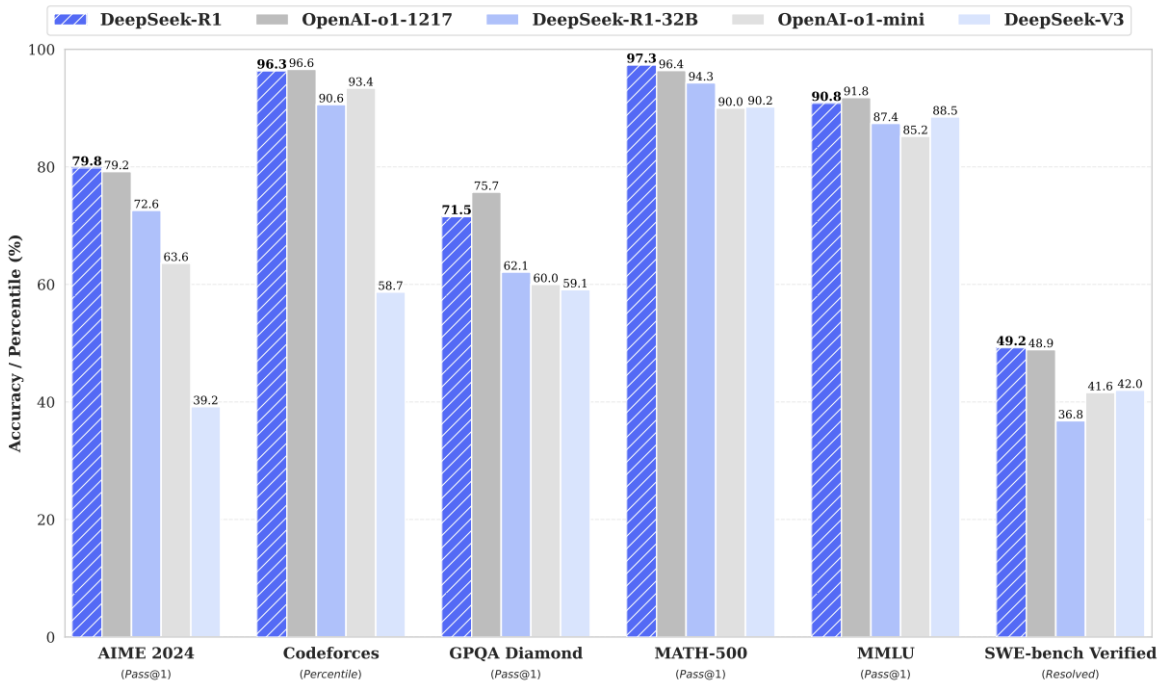
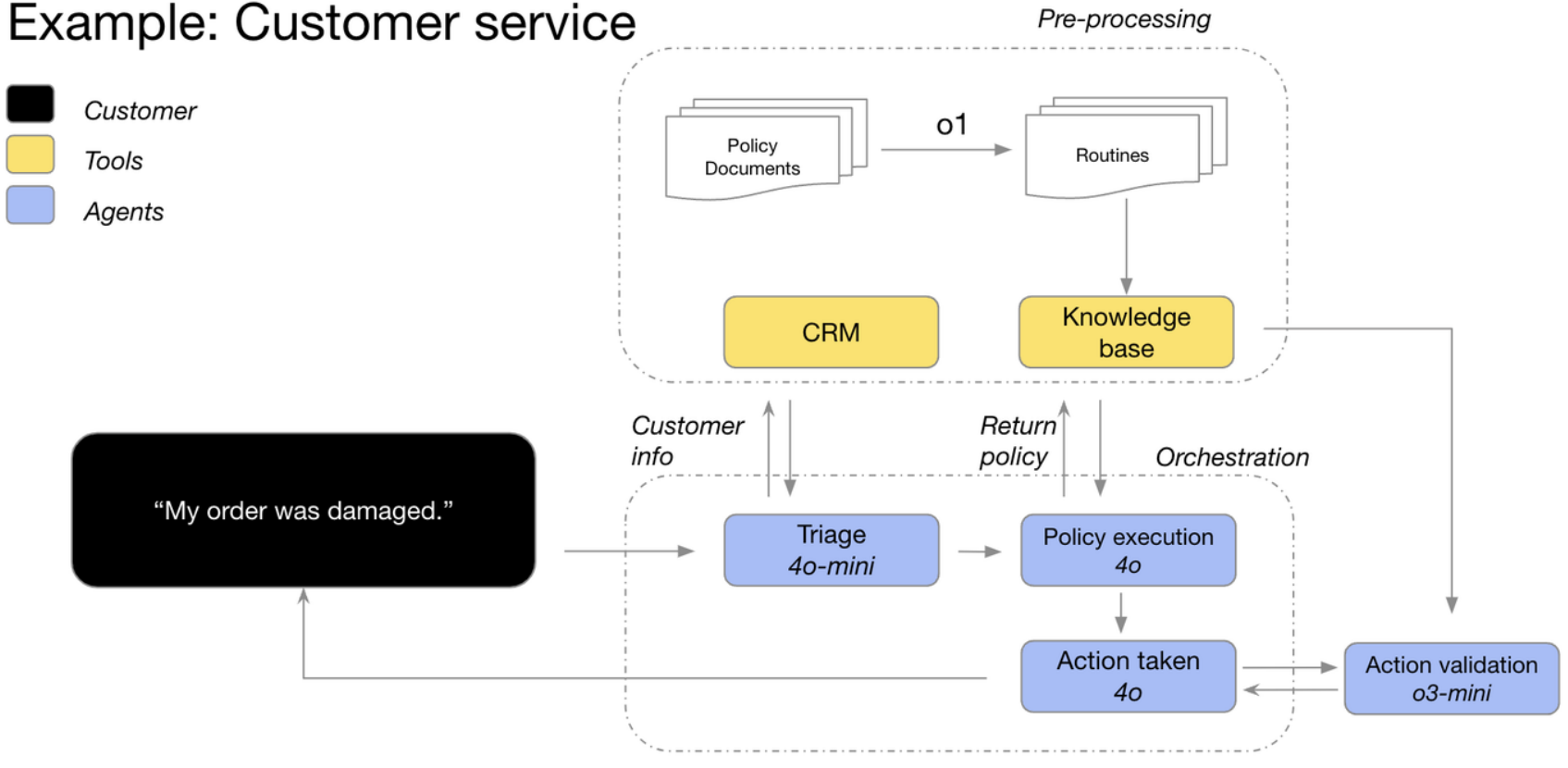


Figure 1 | Benchmark performance of DeepSeek-R1.

Example workflow blending reasoning and non reasoning LLMs



Chain-of-Thought Prompting Elicits Reasoning in Large Language Models

Jason Wei **Xuezhi Wang** **Dale Schuurmans** **Maarten Bosma**

Brian Ichter **Fei Xia** **Ed H. Chi** **Quoc V. Le** **Denny Zhou**

Google Research, Brain Team
{jasonwei,dennyzhou}@google.com

Abstract

We explore how generating a *chain of thought*—a series of intermediate reasoning steps—significantly improves the ability of large language models to perform complex reasoning. In particular, we show how such reasoning abilities emerge naturally in sufficiently large language models via a simple method called *chain-of-thought prompting*, where a few chain of thought demonstrations are provided as exemplars in prompting.

Experiments on three large language models show that chain-of-thought prompting improves performance on a range of arithmetic, commonsense, and symbolic reasoning tasks. The empirical gains can be striking. For instance, prompting a PaLM 540B with just eight chain-of-thought exemplars achieves state-of-the-art accuracy on the GSM8K benchmark of math word problems, surpassing even finetuned GPT-3 with a verifier.

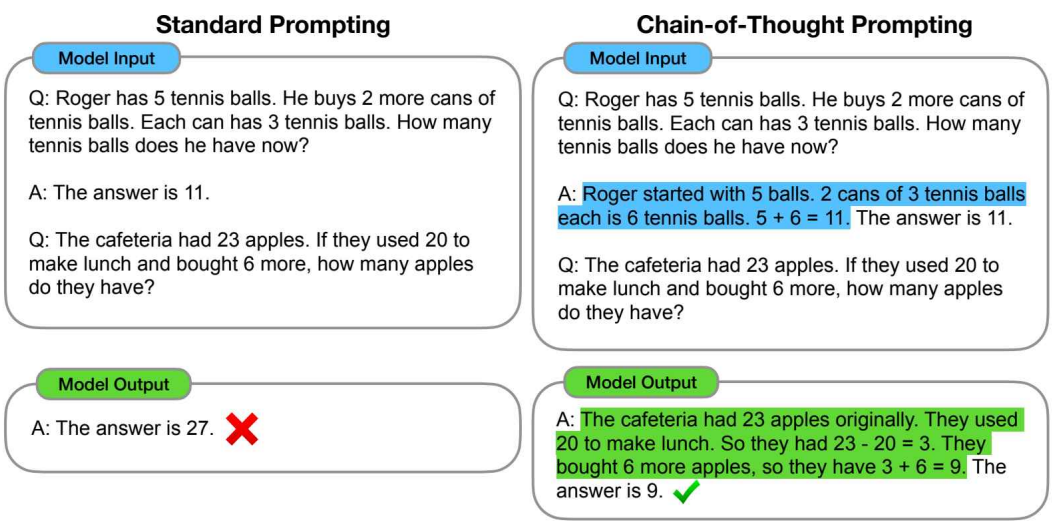


Figure 1: Chain-of-thought prompting enables large language models to tackle complex arithmetic, commonsense, and symbolic reasoning tasks. Chain-of-thought reasoning processes are highlighted.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

When to Use Reasoning LLMs Complement

Use Cases

- **Coding:** One-shotting entire files or sets of files
- **Planning & Agency**: Upfront planning for agentic workflows
- **Deep Reflection:** Analysis of meeting notes, documents, papers
- **Data Analysis**: Medical diagnostics, data interpretation
- **Research & Report Generation**: Deep research on complex topics
- **LLM as Judge**: Evaluation steps in workflows

When to Use Reasoning Models

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

The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity

Parshin Shojaee*† Iman Mirzadeh* Keivan Alizadeh
Maxwell Horton Samy Bengio Mehrdad Farajtabar

Apple

Abstract

Recent generations of frontier language models have introduced Large Reasoning Models (LRMs) that generate detailed thinking processes before providing answers. While these models demonstrate improved performance on reasoning benchmarks, their fundamental capabilities, scaling properties, and limitations remain insufficiently understood. Current evaluations primarily focus on established mathematical and coding benchmarks, emphasizing final answer accuracy. However, this evaluation paradigm often suffers from data contamination and does not provide insights into the reasoning traces' structure and quality. In this work, we systematically investigate these gaps with the help of controllable puzzle environments that allow precise manipulation of compositional complexity while maintaining consistent logical structures. This setup enables the analysis of not only final answers but also the internal reasoning traces, offering insights into how LRMs "think". Through extensive experimentation across diverse puzzles, we show that frontier LRMs face a complete accuracy collapse beyond certain complexities. Moreover, they exhibit a counterintuitive scaling limit: their reasoning effort increases with problem complexity up to a point, then declines despite having an adequate token budget. By comparing LRMs with their standard LLM counterparts under equivalent inference compute, we identify three performance regimes: (1) low-complexity tasks where standard models surprisingly outperform LRMs, (2) medium-complexity tasks where additional thinking in LRMs demonstrates advantage, and (3) high-complexity tasks where both models experience complete collapse. We found that LRMs have limitations in exact computation: they fail to use explicit algorithms and reason inconsistently across puzzles. We also investigate the reasoning traces in more depth, studying the patterns of explored solutions and analyzing the models' computational behavior, shedding light on their strengths, limitations, and ultimately raising crucial questions about their true reasoning capabilities.

The Illusion of the Illusion of Thinking A Comment on Shojaee et al. (2025)

C. Opus

Anthropic

A. Lawsen

Open Philanthropy

(June 10, 2025)

Abstract

Shojaee et al. (2025) report that Large Reasoning Models (LRMs) exhibit ”accuracy collapse” on planning puzzles beyond certain complexity thresholds. We demonstrate that their findings primarily reflect experimental design limitations rather than fundamental reasoning failures. Our analysis reveals three critical issues: (1) Tower of Hanoi experiments systematically exceed model output token limits at reported failure points, with models explicitly acknowledging these constraints in their outputs; (2) The authors’ automated evaluation framework fails to distinguish between reasoning failures and practical constraints, leading to misclassification of model capabilities; (3) Most concerningly, their River Crossing benchmarks include mathematically impossible instances for $N \geq 6$ due to insufficient boat capacity, yet models are scored as failures for not solving these unsolvable problems. When we control for these experimental artifacts, by requesting generating functions instead of exhaustive move lists, preliminary experiments across multiple models indicate high accuracy on Tower of Hanoi instances previously reported as complete failures. These findings highlight the importance of careful experimental design when evaluating AI reasoning capabilities.

Quantizing = reducing the size to fit into hardware
(loosing a bit on performance quality)

Decision Making chart for Reasoning LLMs

0-100 0-500T/s 0-100$/1MT

0.2x 0.4x 0.4x

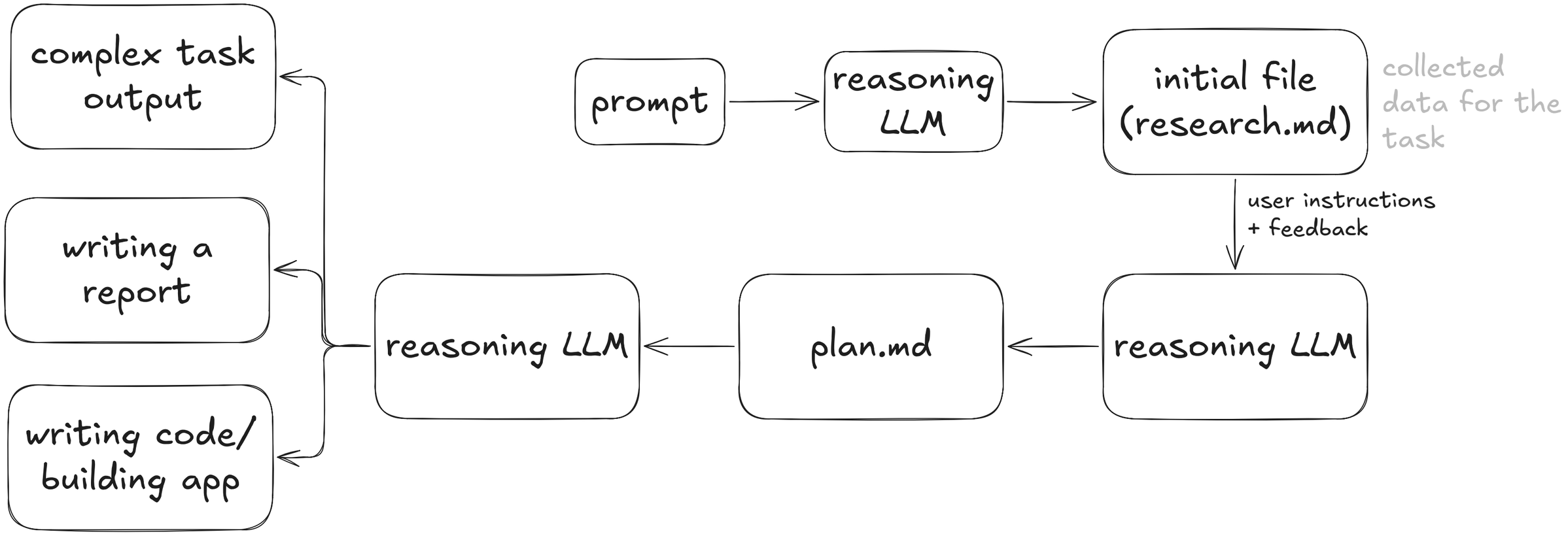
Model	Intelligence Score	Speed/Latency	Cost	Score (0-1)
DeepSeek R1	42/100	31/500	0.3$/100	0.508
Gemin 3 Pro	48/100	198/500	4.5$/100	0.636
GPT-5.2	51/100	311/500	4.8$/100	0.732
Claude Opus 4.5	50/100	69/500	10$/100	0.515

f(x)=normalization function (range 0-1?)

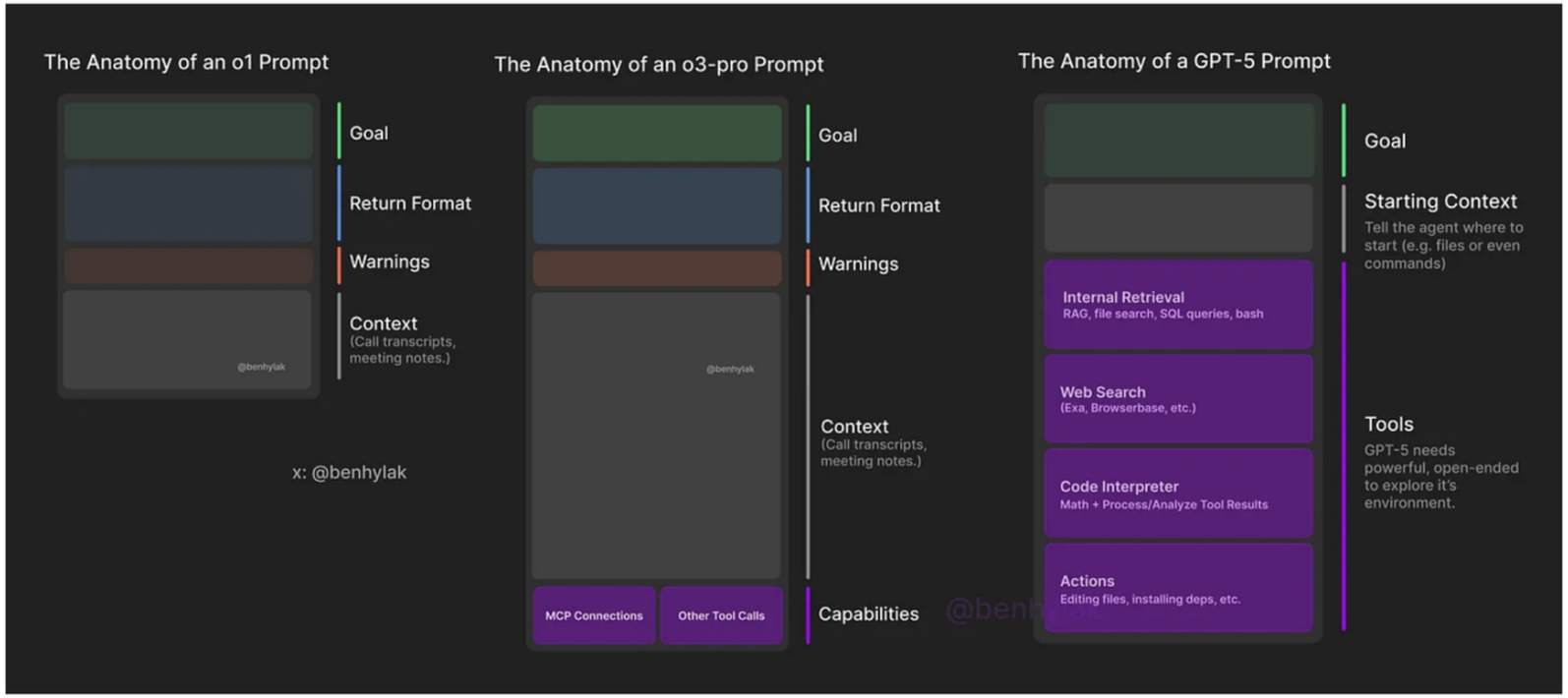
$$\text{Score} = 0.2 \cdot f_{\text{intel}} + 0.4 \cdot f_{\text{speed}} + 0.4 \cdot f_{\text{cost}}$$

Factors	Factor 1	Factor 2	Factor 3	Factor 4	Score
Weights					
Option 1					
Option 2					
Option 3					

Usual Workflow with Reasoning LLMs



Anatomy of a GPT-5 Prompt



You can't think of it like prompting a "model" anymore. **You have to think of it like prompting an agent.**

GPT-5 is a unified system with a **smart, efficient model** that answers most questions, a **deeper reasoning model** (GPT-5 thinking) for harder problems, and a **real-time router** that quickly decides which to use based on conversation type, complexity, tool needs, and your explicit intent (for example, if you say "think hard about this" in the prompt). The router is continuously trained on real signals, including when users switch models, preference rates for responses,