What is a Reasoning LLM?

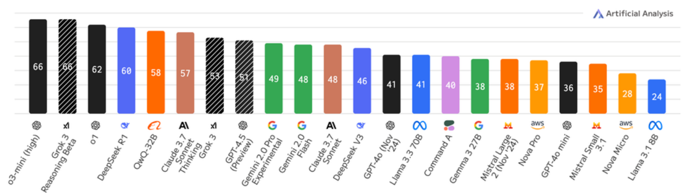
input → model → reasoning 1 | reasoning ... → output

o1 Gemini DeepSeekR1
Claude 3.7 Sonnet + Extended Thinking
Grok3? Qwen-32b

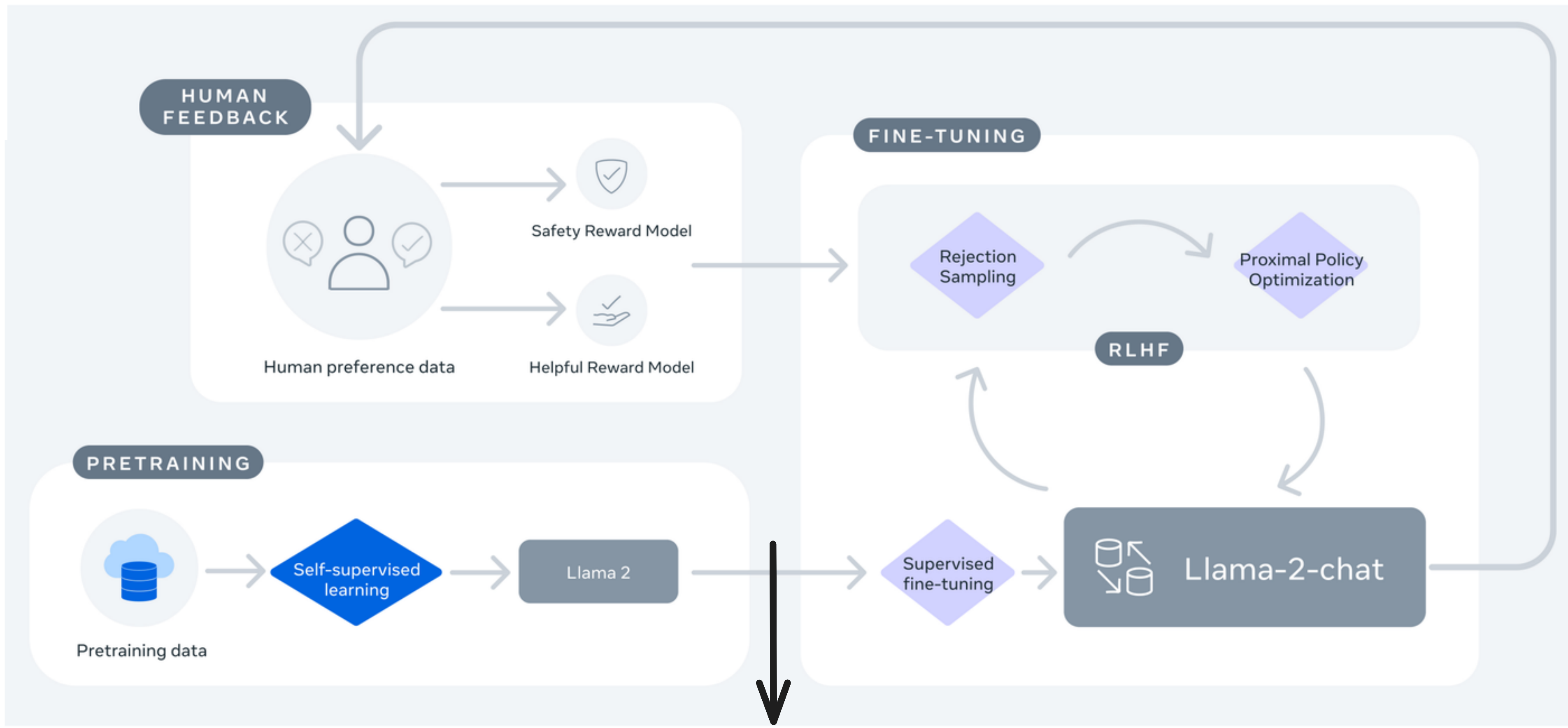
pre-trained LLM → fine tune it with thousands of chain-of-thought examples → Apply RL on top

Artificial Analysis Intelligence Index

Intelligence Index incorporates 7 evaluations spanning reasoning, knowledge, math & coding
Estimate (independent evaluation forthcoming)



Artificial Analysis Intelligence Index: Combination metric covering multiple dimensions of intelligence - the simplest way to compare how smart models are. Version 2 was released in Feb '25 and includes: MMLU-Pro, GPQA Diamond, Humanity's Last Exam, LiveCodeBench, SciCode, AIME, MATH-500. See Intelligence Index methodology for further details, including a breakdown of each evaluation and how we run them.



Llama 2 Paper

fine tuning on chain of thought example

https://github.com/deepseek-ai/DeepSeek-R1/blob/main/DeepSeek_R1.pdf

When to USE Reasoning LLMs

✅

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

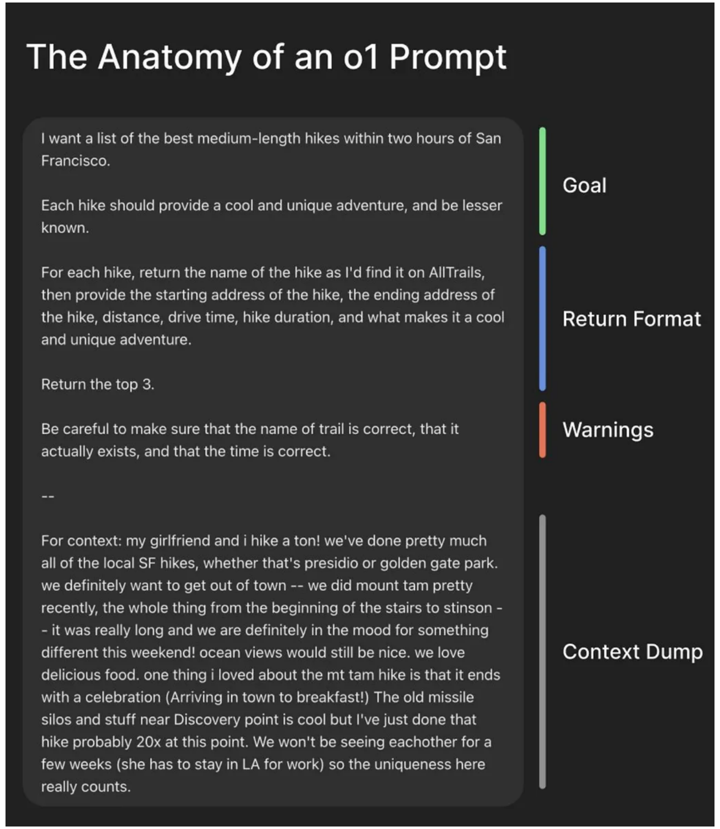
❌

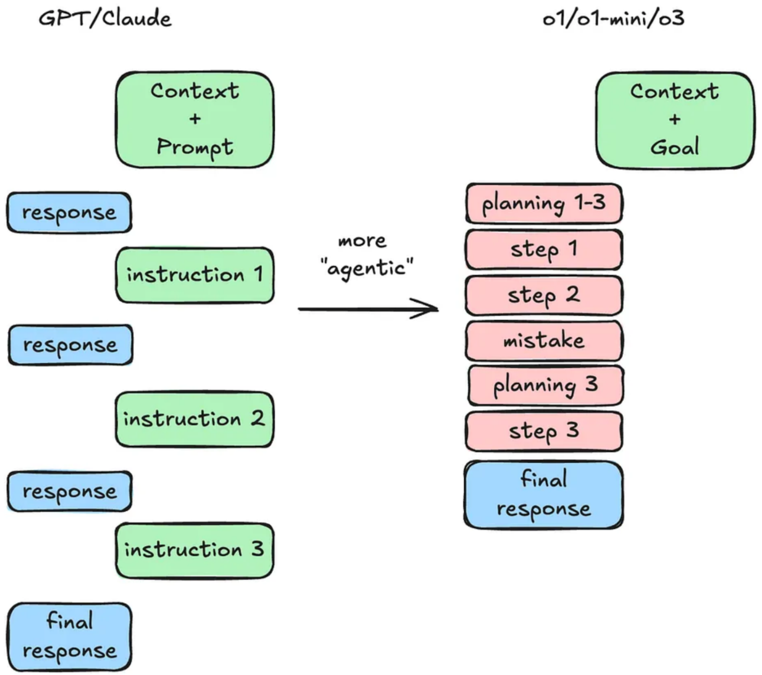
1. Writing in specific styles
2. Building full apps

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

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





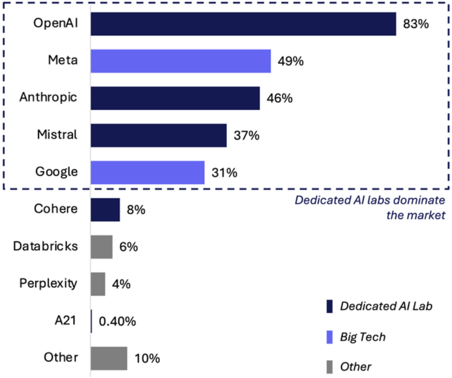
swyx's poor illustration attempt

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

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%

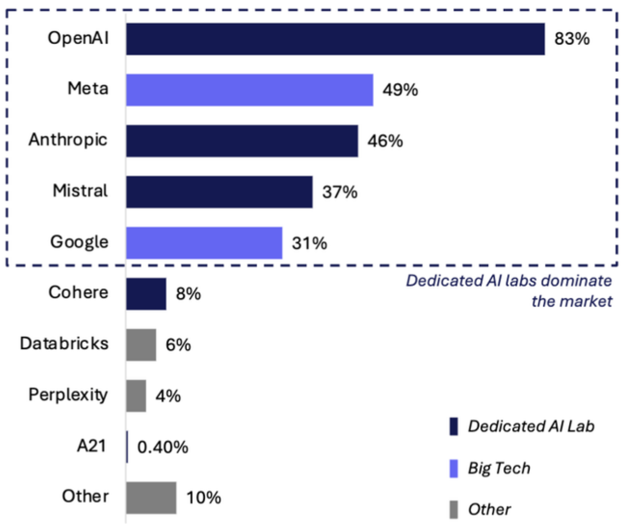
What Matters when Choosing LLMs?

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

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%

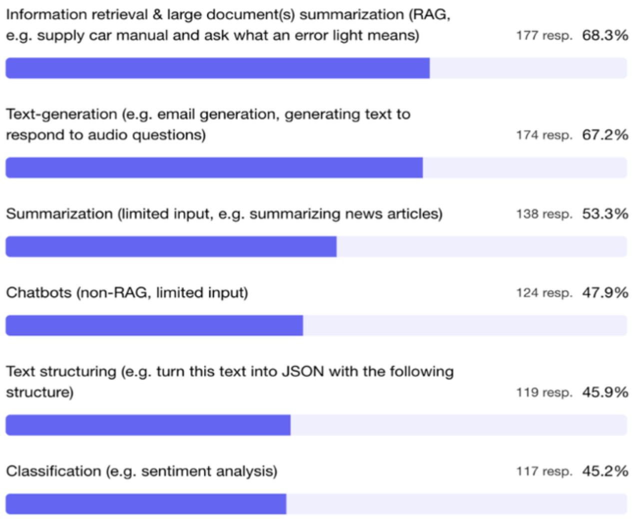
Note: Results from the Artificial Analysis Developer Survey conducted from March to August 2024. Respondents represented a range of organization sizes and locations. Results should be considered indicative only and may be biased by Artificial Analysis's audience. Results may also be affected by survey timing and when new models were added to the survey (typically within days of their release).

Artificial Analysis

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

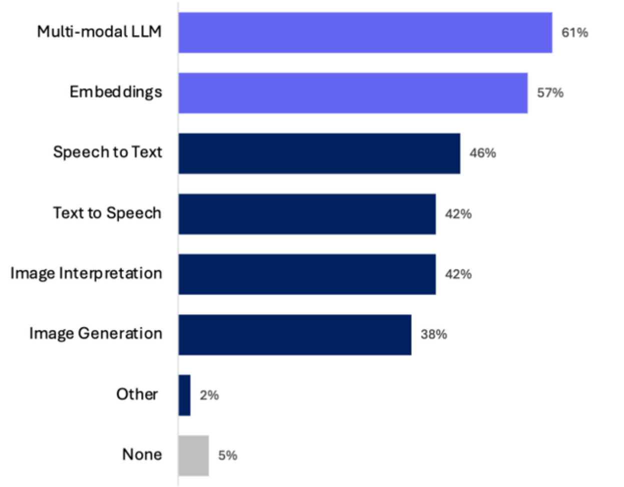
Adoption of Technical Approaches for Using LLMs

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.

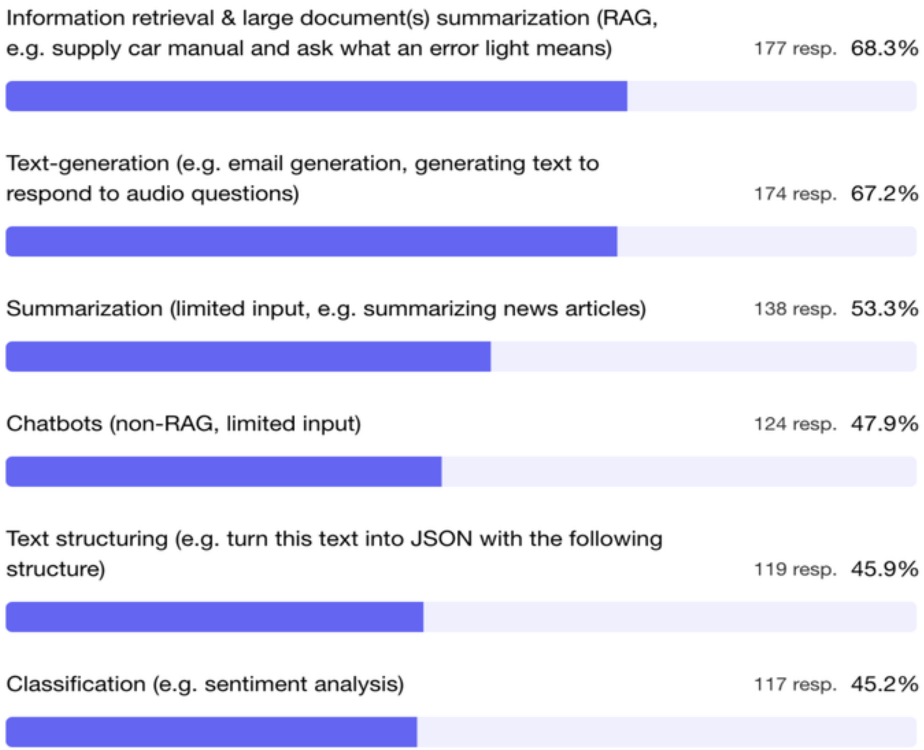
Artificial Analysis

What Are these Models used for?

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

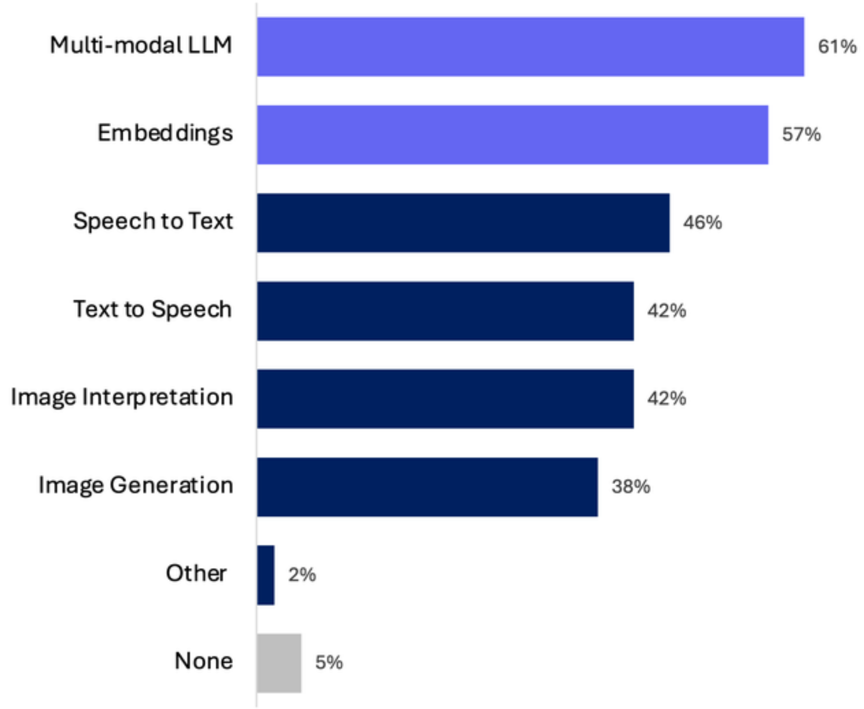
Adoption of Technical Approaches for Using LLMs

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



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.

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

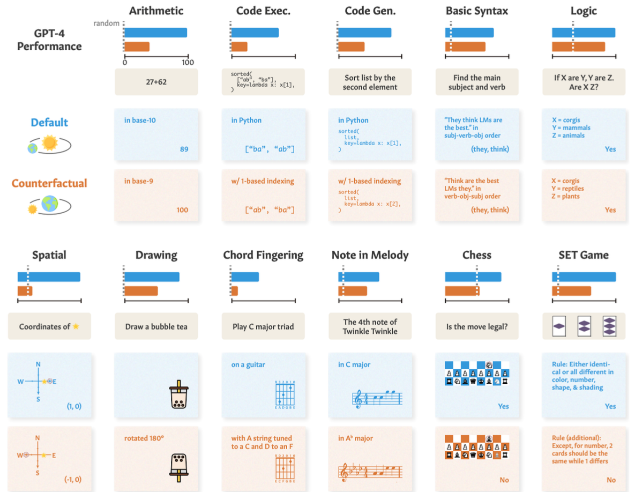


Figure 1: GPT-4's performance on the default version of various tasks (blue) and counterfactual counterparts (orange). The shown results use 0-shot chain-of-thought prompting (§4; Kojima et al., 2023). GPT-4 consistently and substantially underperforms on counterfactual variants compared to default task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