playgroup – deep dive LLM day

Mor Consulting 2025-09

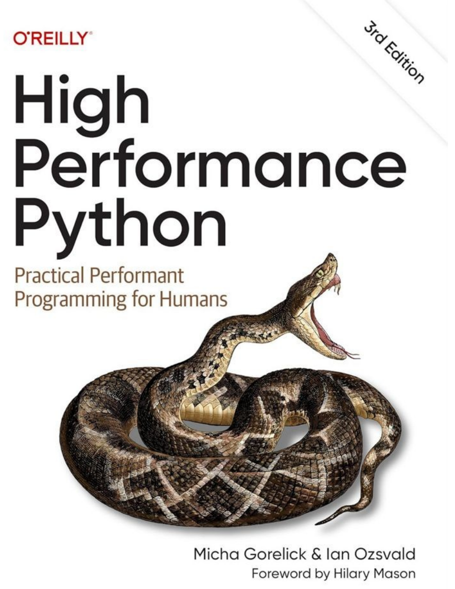
@IanOzsvald – ianozsvald.com

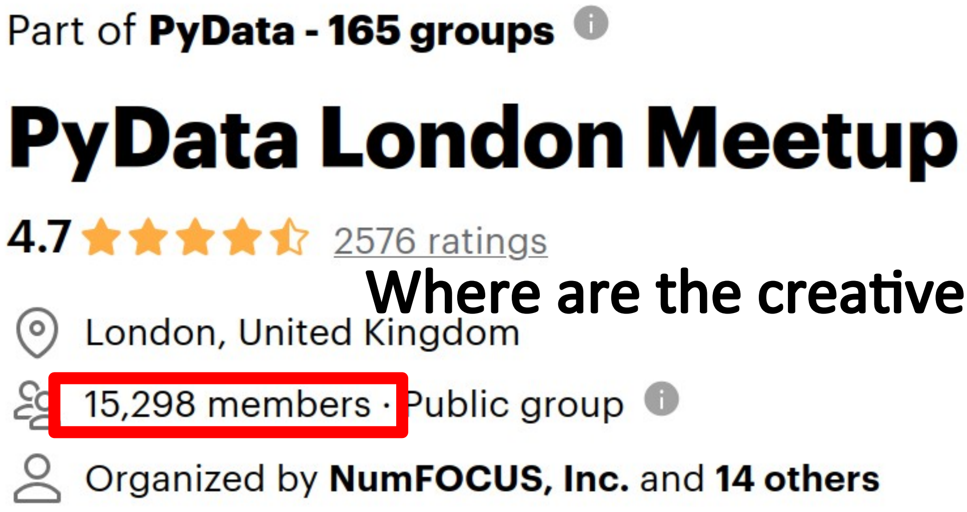
Rebel AI

Interim Chief Data Scientist

- Strategist/Trainer/Speaker/Author 25+ years
- Figuring where LLMs fit into DS

LANTERN


O'REILLY
3rd Edition
High Performance Python
Practical Performant Programming for Humans
Micha Gorelick & Ian Ozsvald
Foreword by Hilary Mason


Part of **PyData - 165 groups**

PyData London Meetup

4.7 ★★★★☆ 2576 ratings

London, United Kingdom

15,298 members · Public group

Organized by **NumFOCUS, Inc.** and **14 others**

Where are the creatives?

Man Group

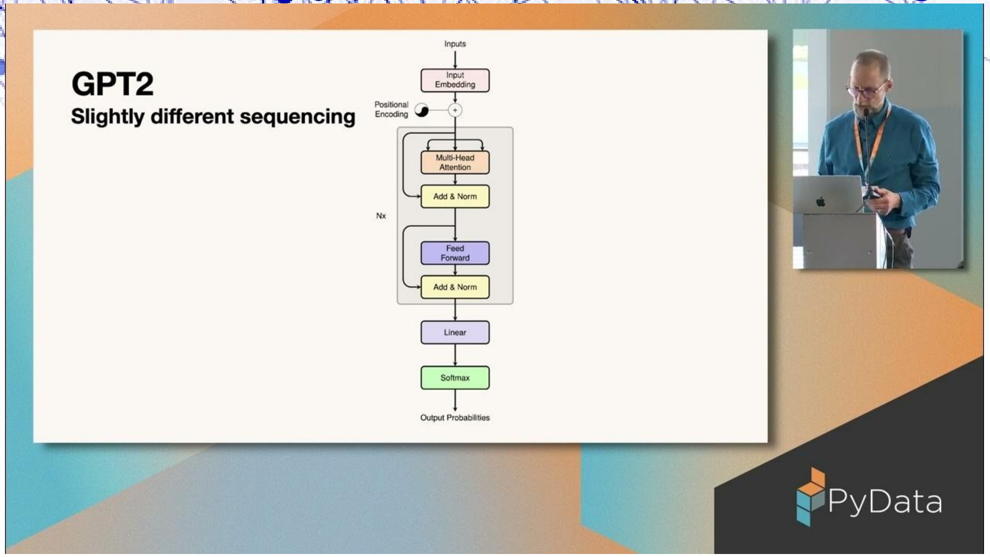
4

aspect capital
The Science of Investment

Hotels.com™

By [ian]@ianozsvald[.com]

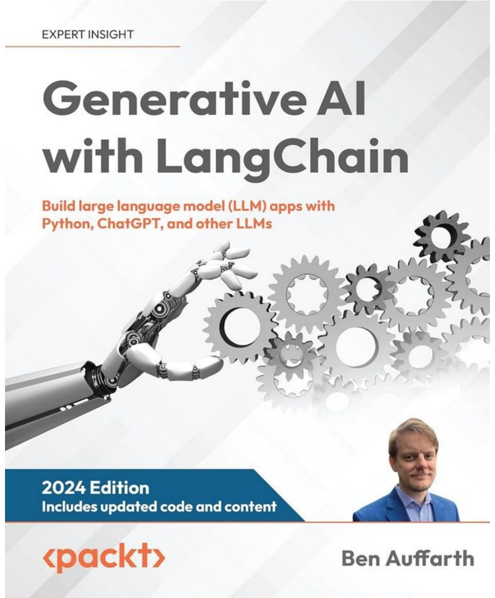
Ian Ozsvald



GPT2

Slightly different sequencing

Pydata London



EXPERT INSIGHT

Generative AI with LangChain

Build large language model (LLM) apps with Python, ChatGPT, and other LLMs

2024 Edition

Includes updated code and content

‹packt›

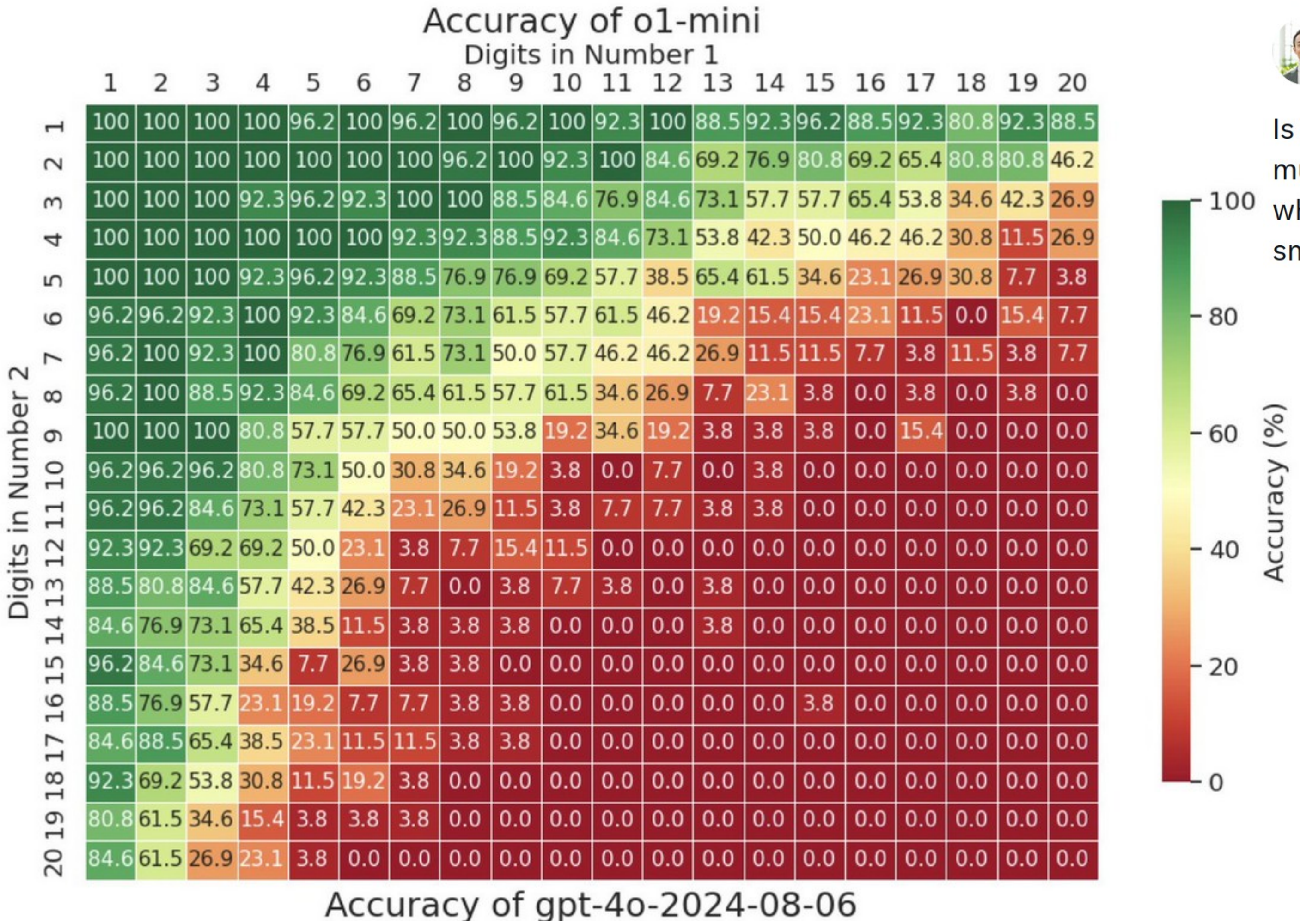
Ben Auffarth

Goal

- Will *agents take over the world* or are we living in a world of *approximate retrieval*? Is AGI nearly here?
- Can an LLM solve novel problems? See? Reflect?
- You – think on a novel problem, meet interesting folk, get your qs answered

https://x.com/yuntiandeng/status/1836114401213989366

Not so good at multiplication



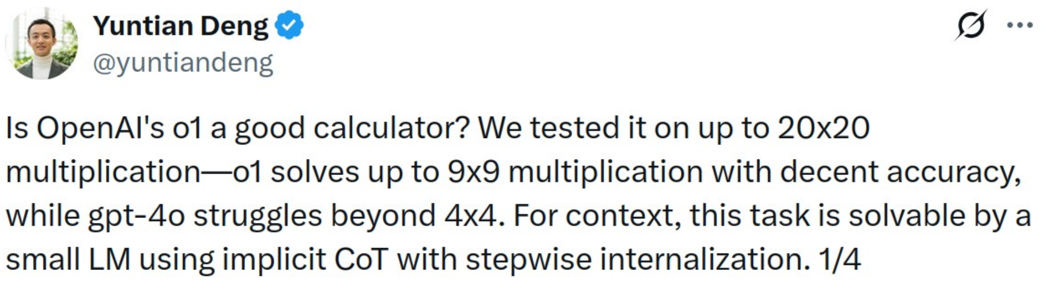
Accuracy of o1-mini

Digits in Number 1

Digits in Number 2

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	100	100	100	100	96.2	100	96.2	100	96.2	100	92.3	100	88.5	92.3	96.2	88.5	92.3	80.8	92.3	88.5
2	100	100	100	100	100	100	100	96.2	100	92.3	100	84.6	69.2	76.9	80.8	69.2	65.4	80.8	80.8	46.2
3	100	100	100	92.3	96.2	92.3	100	100	88.5	84.6	76.9	84.6	73.1	57.7	57.7	65.4	53.8	34.6	42.3	26.9
4	100	100	100	100	100	100	92.3	92.3	88.5	92.3	84.6	73.1	53.8	42.3	50.0	46.2	46.2	30.8	11.5	26.9
5	100	100	100	92.3	96.2	92.3	88.5	76.9	76.9	69.2	57.7	38.5	65.4	61.5	34.6	23.1	26.9	30.8	7.7	3.8
6	96.2	96.2	92.3	100	92.3	84.6	69.2	73.1	61.5	57.7	61.5	46.2	19.2	15.4	15.4	23.1	11.5	0.0	15.4	7.7
7	96.2	100	92.3	100	80.8	76.9	61.5	73.1	50.0	57.7	46.2	46.2	26.9	11.5	11.5	7.7	3.8	11.5	3.8	7.7
8	96.2	100	88.5	92.3	84.6	69.2	65.4	61.5	57.7	61.5	34.6	26.9	7.7	23.1	3.8	0.0	3.8	0.0	3.8	0.0
9	100	100	100	80.8	57.7	57.7	50.0	50.0	53.8	19.2	34.6	19.2	3.8	3.8	3.8	0.0	15.4	0.0	0.0	0.0
10	96.2	96.2	96.2	80.8	73.1	50.0	30.8	34.6	19.2	3.8	0.0	7.7	0.0	3.8	0.0	0.0	0.0	0.0	0.0	0.0
11	96.2	96.2	84.6	73.1	57.7	42.3	23.1	26.9	11.5	3.8	7.7	7.7	3.8	3.8	0.0	0.0	0.0	0.0	0.0	0.0
12	92.3	92.3	69.2	69.2	50.0	23.1	3.8	7.7	15.4	11.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
13	88.5	80.8	84.6	57.7	42.3	26.9	7.7	0.0	3.8	7.7	3.8	0.0	3.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0
14	84.6	76.9	73.1	65.4	38.5	11.5	3.8	3.8	3.8	0.0	0.0	0.0	3.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0
15	96.2	84.6	73.1	34.6	7.7	26.9	3.8	3.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
16	88.5	76.9	57.7	23.1	19.2	7.7	7.7	3.8	3.8	0.0	0.0	0.0	0.0	0.0	3.8	0.0	0.0	0.0	0.0	0.0
17	84.6	88.5	65.4	38.5	23.1	11.5	11.5	3.8	3.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
18	92.3	69.2	53.8	30.8	11.5	19.2	3.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
19	80.8	61.5	34.6	15.4	3.8	3.8	3.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
20	84.6	61.5	26.9	23.1	3.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Accuracy of gpt-4o-2024-08-06



Yuntian Deng @yuntiandeng

Is OpenAI's o1 a good calculator? We tested it on up to 20x20 multiplication—o1 solves up to 9x9 multiplication with decent accuracy, while gpt-4o struggles beyond 4x4. For context, this task is solvable by a small LM using implicit CoT with stepwise internalization. 1/4

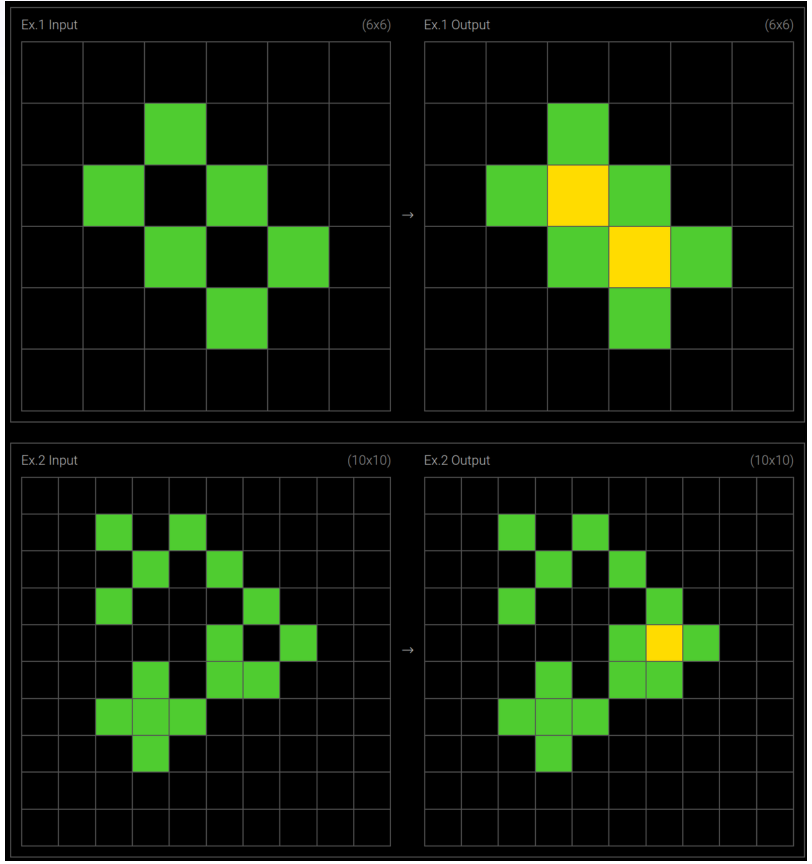
Maybe it lacks short term memory and iterative processing?

Tokens – representation issues?

Approximate retrieval at work?

Agenda

- Talk about ARC AGI, try manually
- Can an agentic(?) method reflect and improve?
- → office: prompting, testing, auto-code SQL, resilience



Business thoughts

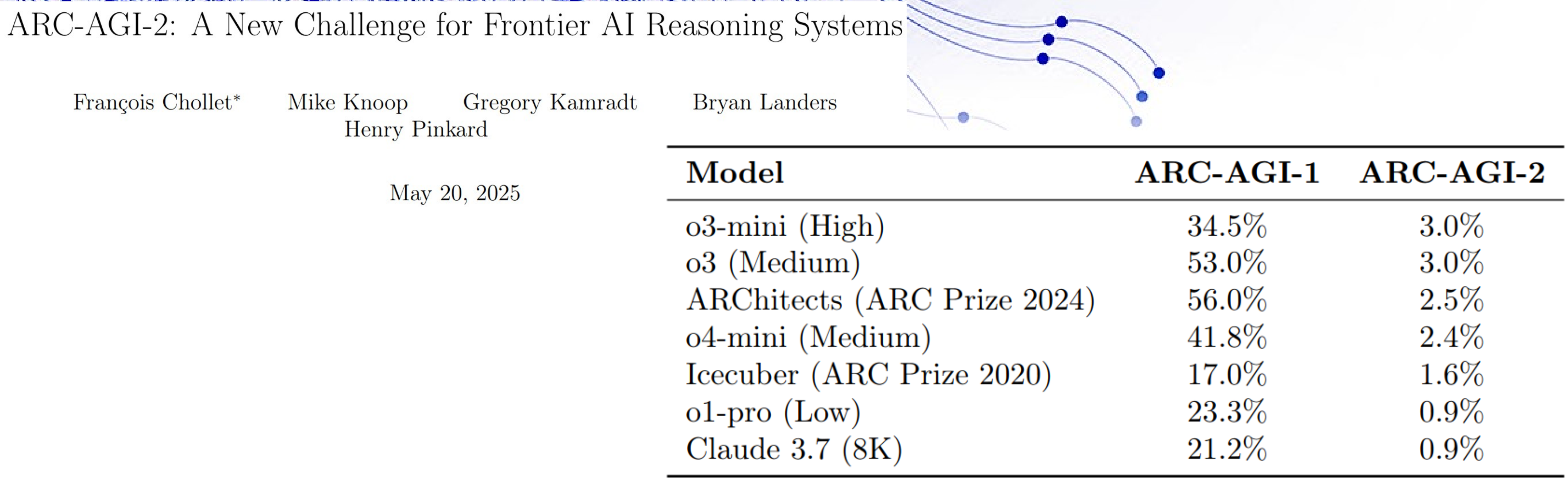
- VCs will want their cash back at some point
- Scaling is expensive – can we keep our solution?
- Keep IP in-house
- Maybe we don't need to burn the planet on LLMs

Am I asking the right question?

- Representation
- Prompt
- Process
- What am I missing? What's a **big question** to ask?

Kick off

- Do you have **slack**? Do you have the code?
  - Add to the slack with shared notes, branch code
- **Tables – what’s a funny/useful GenAI story?** Share back, start in pairs, decide on someone’s example to share – 15 mins



ARC-AGI-2: A New Challenge for Frontier AI Reasoning Systems

François Chollet* Mike Knoop Gregory Kamradt Bryan Landers
Henry Pinkard

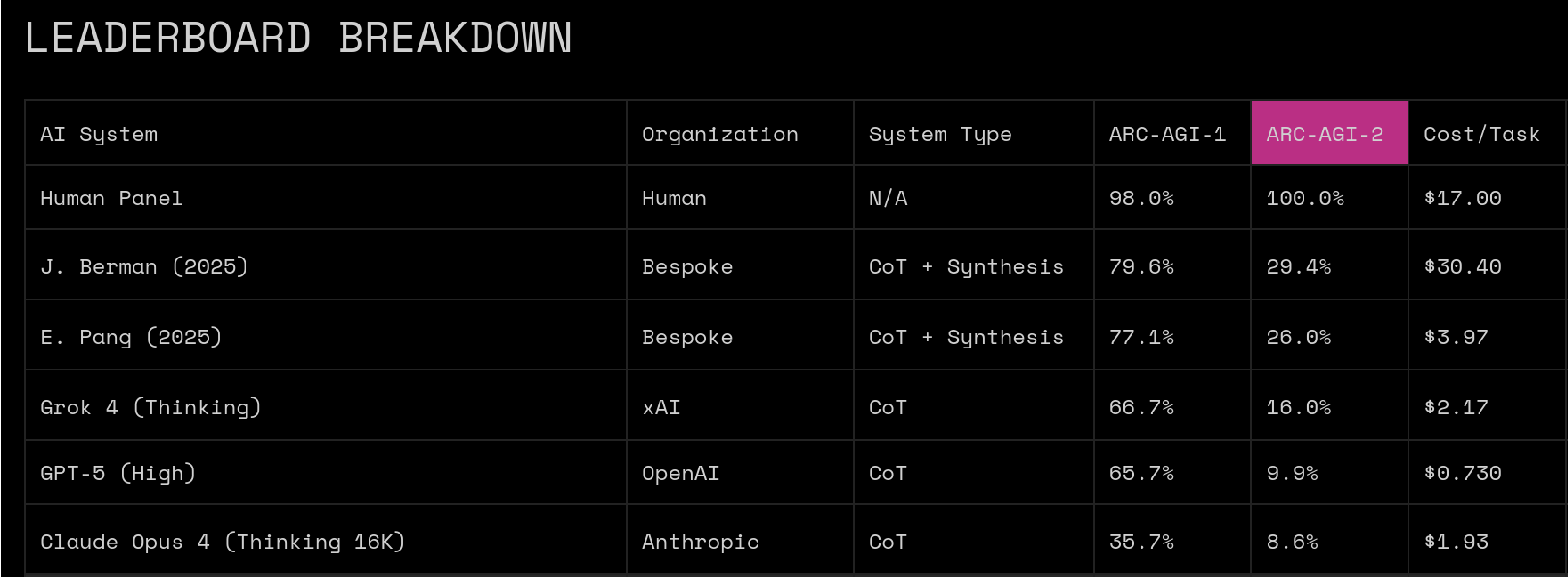
May 20, 2025

Model	ARC-AGI-1	ARC-AGI-2
o3-mini (High)	34.5%	3.0%
o3 (Medium)	53.0%	3.0%
ARChitects (ARC Prize 2024)	56.0%	2.5%
o4-mini (Medium)	41.8%	2.4%
Icecuber (ARC Prize 2020)	17.0%	1.6%
o1-pro (Low)	23.3%	0.9%
Claude 3.7 (8K)	21.2%	0.9%

- ARC AGI 1 (few years), now ARC AGI 2025
- 400+ problems, public and *private* (offline) set
- ARC AGI 1 “solved” by **GPT o3 88%** public $70k (xmas)

ARC AGI 2025 (today)

https://arcprize.org/



LEADERBOARD BREAKDOWN

AI System	Organization	System Type	ARC-AGI-1	ARC-AGI-2	Cost/Task
Human Panel	Human	N/A	98.0%	100.0%	$17.00
J. Berman (2025)	Bespoke	CoT + Synthesis	79.6%	29.4%	$30.40
E. Pang (2025)	Bespoke	CoT + Synthesis	77.1%	26.0%	$3.97
Grok 4 (Thinking)	xAI	CoT	66.7%	16.0%	$2.17
GPT-5 (High)	OpenAI	CoT	65.7%	9.9%	$0.730
Claude Opus 4 (Thinking 16K)	Anthropic	CoT	35.7%	8.6%	$1.93

Stages

- Limited GPU, Llama Scout (mm) about right – how should we represent the problem? Might vision help?
- We can try DeepSeek v3 0324 (quick, big context), circa 700GB VRAM at fp8 vs 90GB VRAM competition limit
- Does giving feedback help?
- Could ‘agent framework’ help? Open q

Over to you

- Run the code, notes are in the README
- I’ll tell you about our stages
- Try to talk to everyone in the room (cheatsheet)

How could it do better?

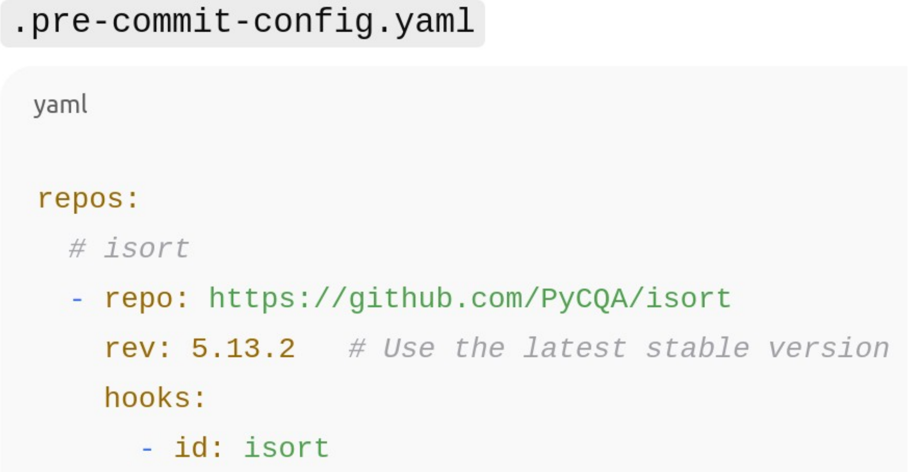
- Make hypotheses, critique, rank
- Implement, get graded feedback, iterate
- Extract library of useful fns
- Writing code – solved?

How did others solve it?

- GA on human-designed solver components (no LLM)
- Library of human-solved clues, synthetic dataset
- Test-time fine tuning on 3 examples
  - Restricted representation fine tune
- GA to evolve prompts

Problems with LLMs?

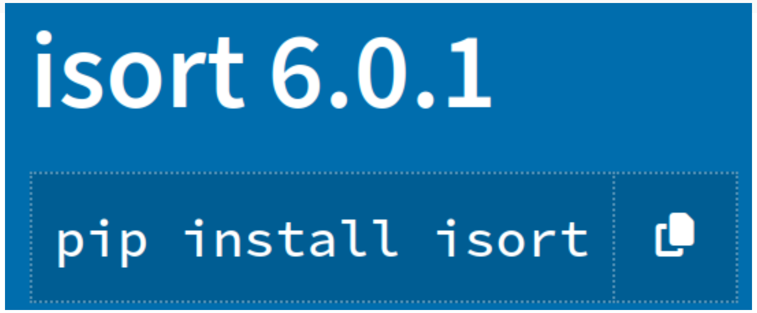
- Outdated knowledge



`.pre-commit-config.yaml`

yaml

```yaml
repos:
  # isort
  - repo: https://github.com/PyCQA/isort
    rev: 5.13.2   # Use the latest stable version
    hooks:
      - id: isort
```



isort 6.0.1

```
pip install isort
```