playgroup – deep dive LLM day

Mor Consulting 2025-06

@IanOzsvald – ianozsvald.com

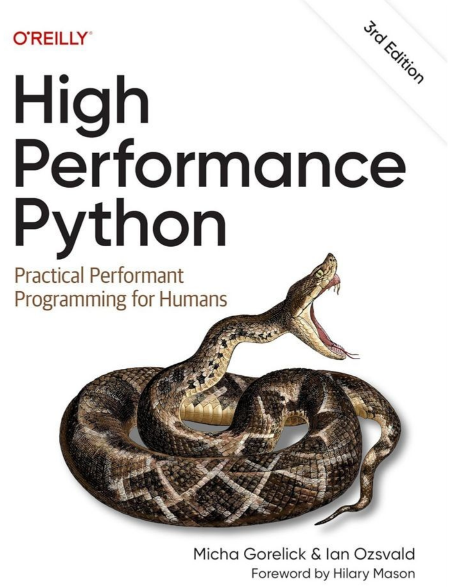
Rebel AI

Interim Chief Data Scientist

- Strategist/Trainer/Speaker/Author 25+ years

- Figuring where LLMs fit into DS

LANTERN

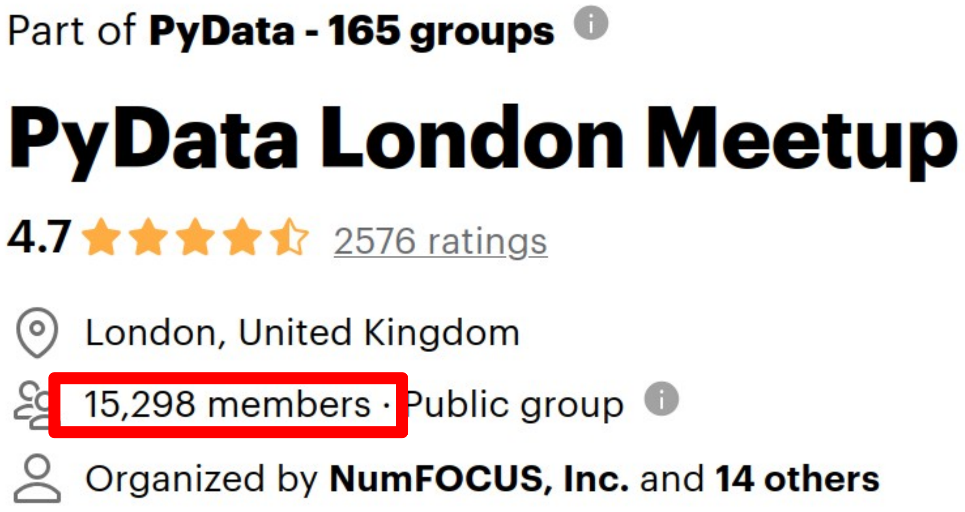


High Performance Python

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**

Man

Man Group

4

aspect capital

The Science of Investment

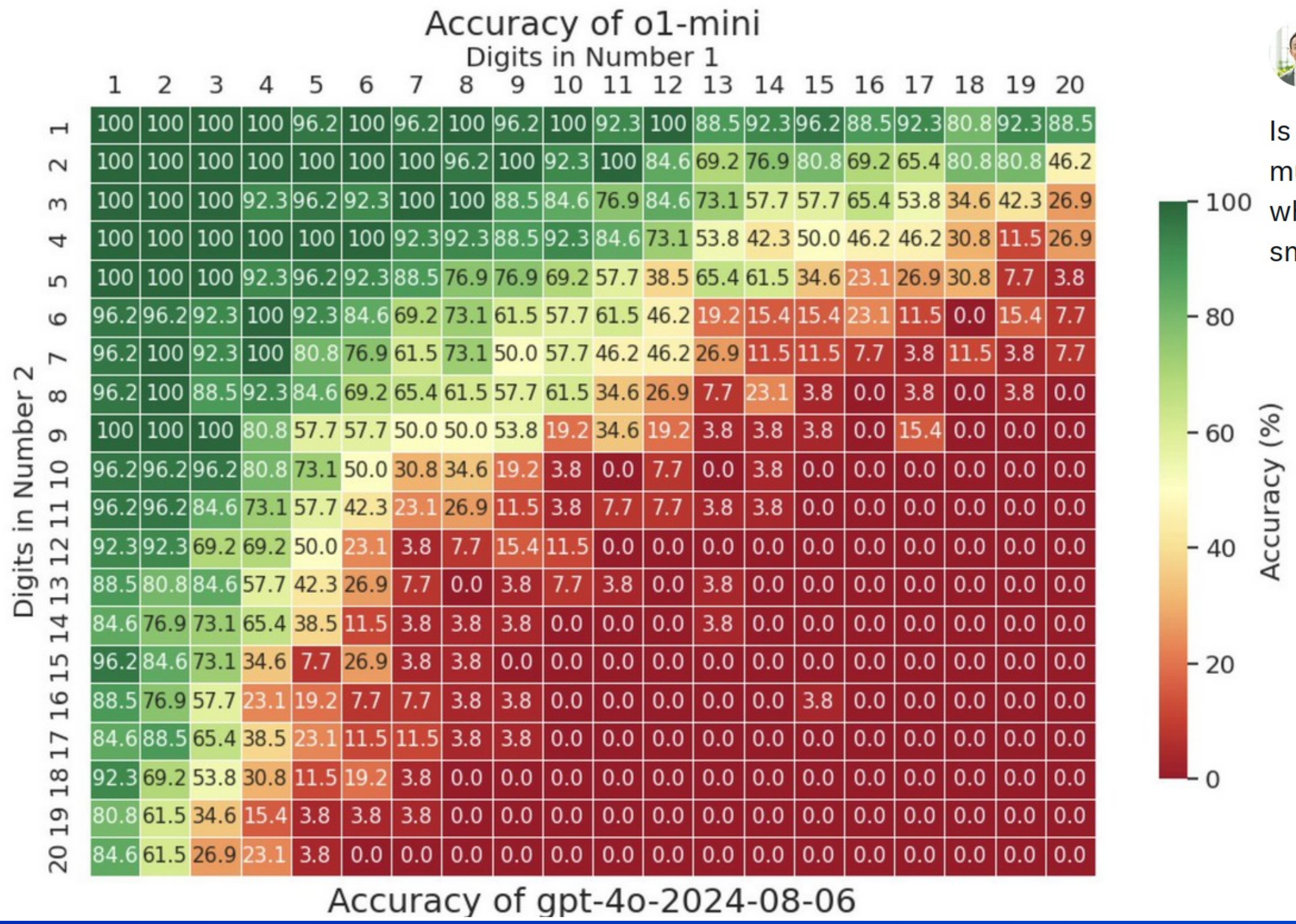
Hotels.com™

Goal

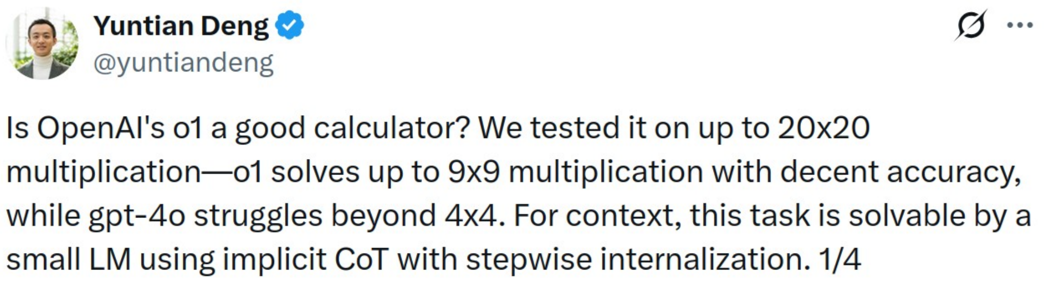
- Will *agents take over the world* or are we living in a world of *approximate retrieval*? Is AGI nearly here?
- Can an LLM solve novel problems? See? Reflect?
- You – think on a novel problem, meet interesting folk, get your qs answered

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

Not so good at multiplication



	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



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
- Get LLM to solve some (maybe)
- Can a vLLM describe an image? Can you make img?
- Can an agent(?) reflect and improve?

Kick off

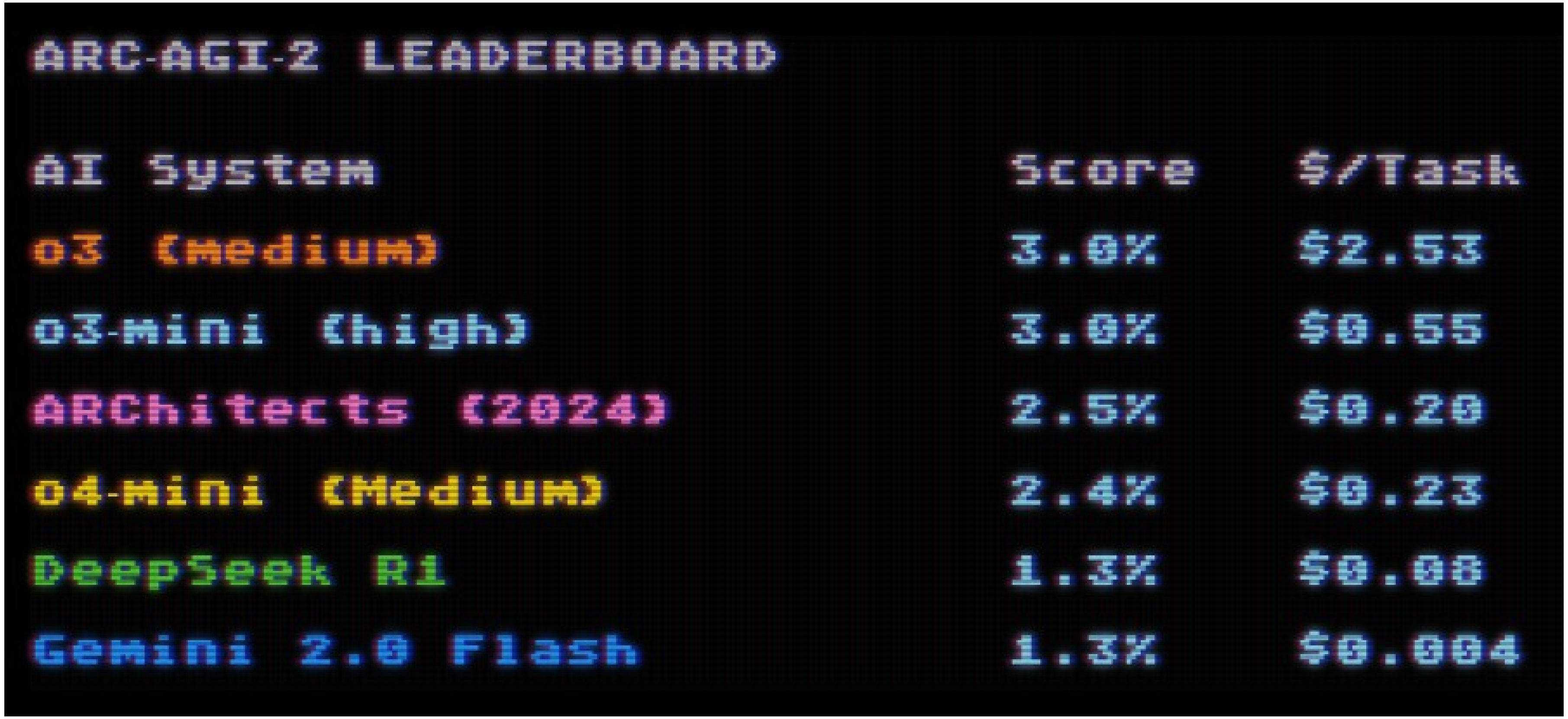
- Do you have the Gdoc? Do you have the code?
  - Add to the Gdoc with shared notes, branch code
- **Tables – when is GenAI useful?** Share back, start in pairs, decide on someone’s example to share – 15 mins

ARC AGI

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

ARC AGI 2025 (today)

https://arcprize.org/



ARC-AGI-2 LEADERBOARD

AI System	Score	$/Task
o3 (medium)	3.0%	$2.53
o3-mini (high)	3.0%	$0.55
ARChitects (2024)	2.5%	$0.20
o4-mini (Medium)	2.4%	$0.23
DeepSeek R1	1.3%	$0.08
Gemini 2.0 Flash	1.3%	$0.004

Stages

- Limited GPU, Llama Scout (mm) about right – how should we represent the problem? Might vision help?
- We can try DeepSeek, Opus 4 ($$$!)
- 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?