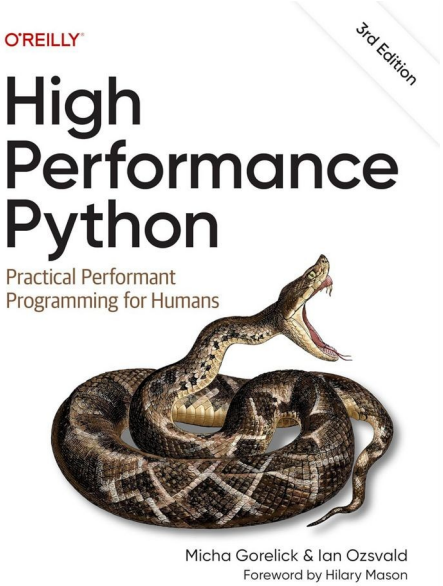
playgroup – deep dive LLM day

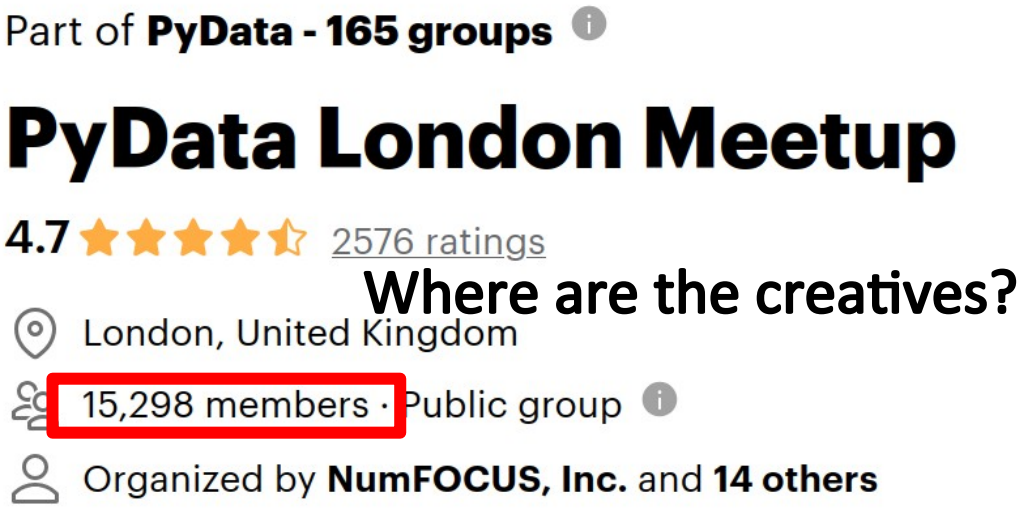
Mor Consulting 2025-06

@IanOzsvald – ianozsvald.com

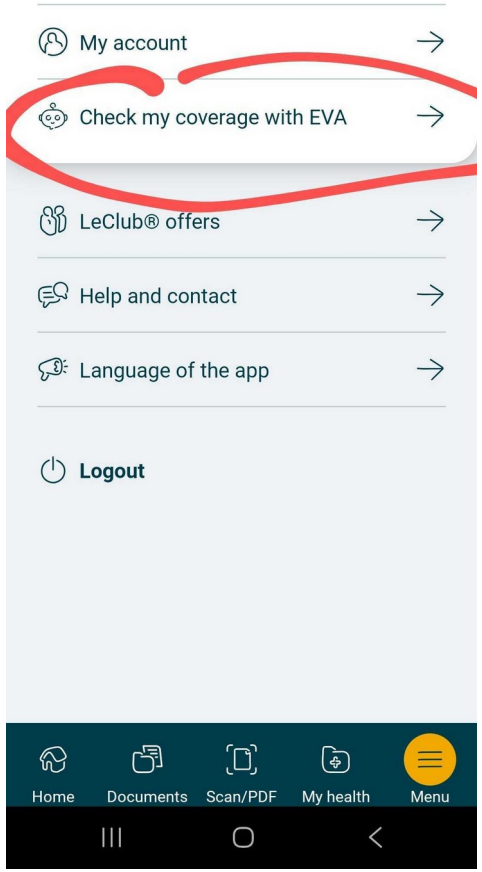
Interim Chief Data Scientist

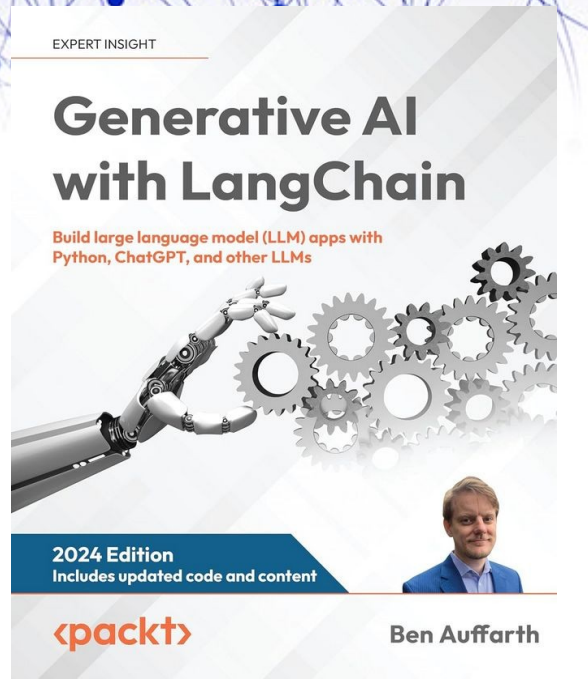
- Strategist/Trainer/Speaker/Author 25+ years
- Figuring where LLMs fit into DS





Where are the creatives?





Pydata London

Atharva Lad • 1st
I write, design and code (sometimes).
4mo •

Meetup #58: **PyData London** 92st Meetup
Tuesday, 7th Jan 2025
(**#proofofnetwork**)

Serge Kozlov, from **Conundrum** shared insights on deploying optimal control systems in factories. His talk addressed the challenge of maintaining processor

Goal

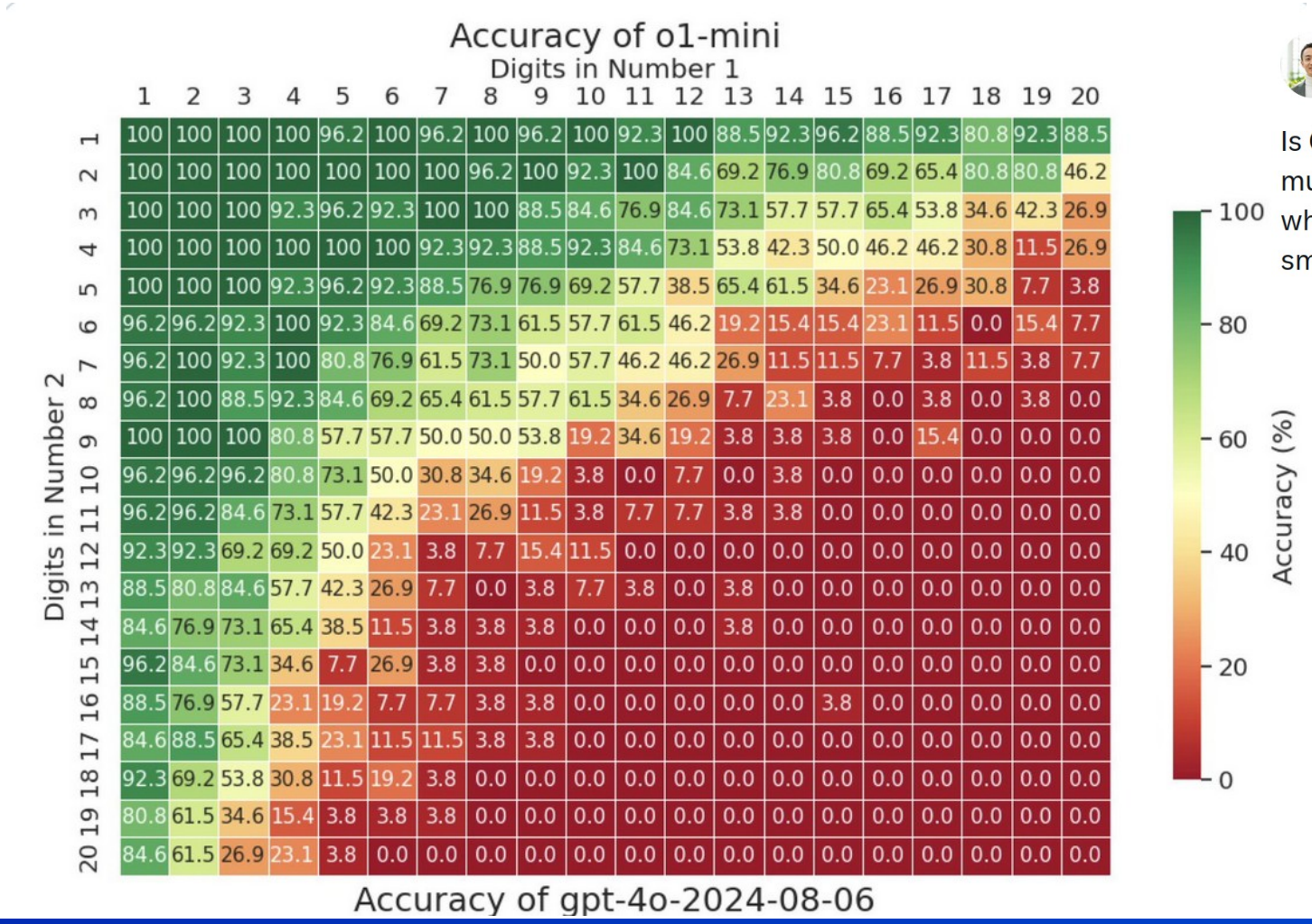
Valuable Lessons Learned on Kaggle's ARC AGI LLM challenge

PyDataGlobal 2024-12 talk

- 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



Yuntian Deng ✓
@yuntiandeng

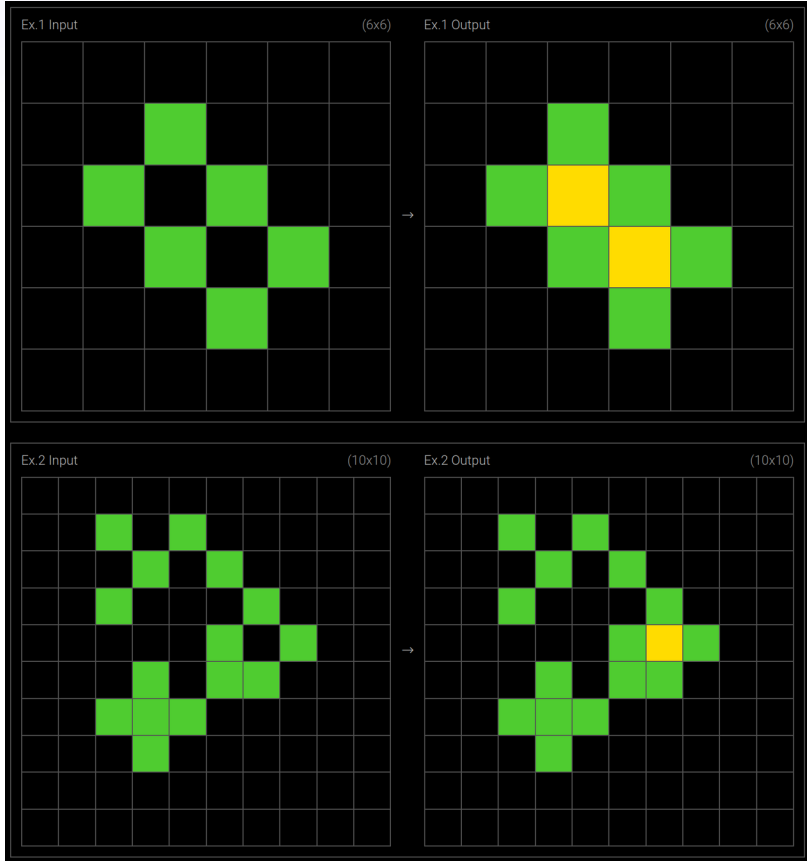
Is OpenAI's o1 a good calculator? We tested it on up to 20x20 multiplication—o1 solves up to 9x9 multiplication with decent accuracy, while gpt-4o struggles beyond 4x4. For context, this task is solvable by a small LM using implicit CoT with stepwise internalization. 1/4

Maybe it lacks short term memory and iterative processing?

Tokens – representation issues?

Approximate retrieval at work?

Agenda



- Talk about ARC AGI, try manually
- Get (v)LLM to solve some (maybe)
- Can an agent(?) 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 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-2: A New Challenge for Frontier AI Reasoning Systems

François Chollet* Mike Knoop Gregory Kamradt Bryan Landers
Henry Pinkard

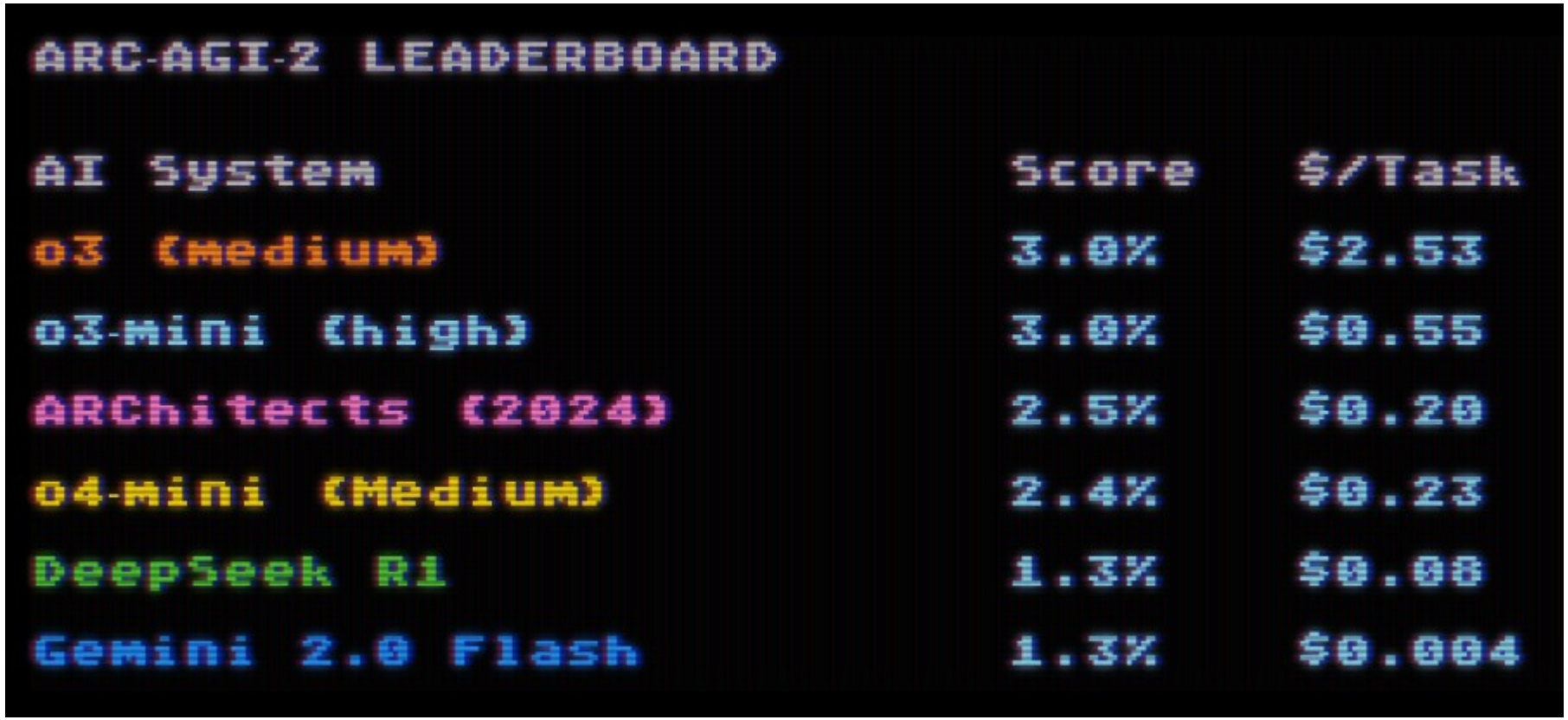
May 20, 2025

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/



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?

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