#### playgroup – deep dive LLM day

Mor Consulting 2025-06

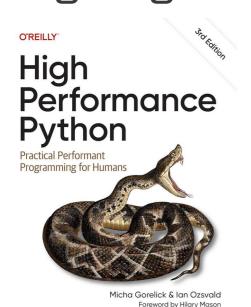
@lanOzsvald – ianozsvald.com





Strategist/Trainer/Speaker/Author 25+ years

Figuring where LLMs fit into DS



Part of PyData - 165 groups 10

#### **PyData London Meetup**

**4.7** ★★★★ 2576 ratings

Where are the creatives?

London, United Kingdom

2 15,298 members · Public group 🕕

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













# 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

#### Accuracy of o1-mini

Digits in Number 1 12 13 14 15 16 17 18 19 20 100 100 100 96.2 100 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 88.5 100 100 100 100 100 100 100 100 96.2 100 92.3 100 84.6 69.2 76.9 80.8 69.2 65.4 80.8 80.8 46.2 100 100 100 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 26.9 100 100 100 100 100 100 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 26.9 100 100 100 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 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 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 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 3.8 | 0.0 | 3.8 | 0.0 | 3.8 | 0.0 100 100 100 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 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 0.0 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 0.0 34.6 <mark>76.9 73.1 65.4 38.5</mark> 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 88.5 <mark>76.9 57.7</mark> 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 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 <mark>861.5 34.6</mark> 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

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

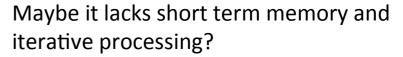
3.8 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0



80

- 20

Is OpenAl'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



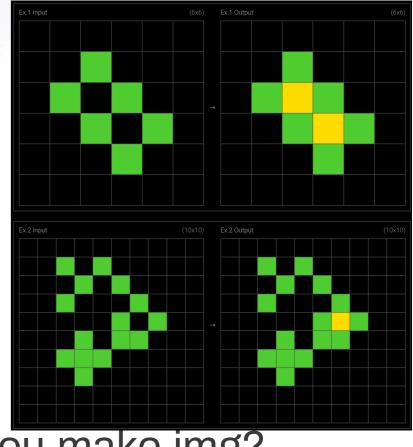
Tokens – representation issues?

Approximate retrieval at work?

Ø ...



- 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?



## 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

### Kick off ...

- Do you have the Gdoc? Do you have the code?
  - Add to the Gdoc with shared notes, branch code
- Tables when is GenAl 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

ms

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%

23.3%

21.2%

•ARC AGI 1 (few years), now ARC AGI 2025

Bryan Landers

- •400+ problems, public and private (offline) set
- •ARC AGI 1 "solved" by GPT o3 88% public \$70k (xmas)

o1-pro (Low)

Claude 3.7 (8K)

0.9%

0.9%

#### ARC AGI 2025 (today)

```
ARC-AGI-2 LEADERBOARD
                           Score
                                   $/Task
AI System
o3 (medium)
                           3.9X
                                   $2.53
                                   $0.55
o3-mini (high)
                           3.9%
ARChitects (2024)
                           2.5X
                                   $0.20
                                   $0.23
o4-mini (Medium)
                           2.4%
                                   $0.08
DeepSeek R1
                           1.3X
Gemini 2.0 Flash
                           1.3X
                                   $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



- 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?