Local AI in Practice: Automating Assessment and Simulation Workflows

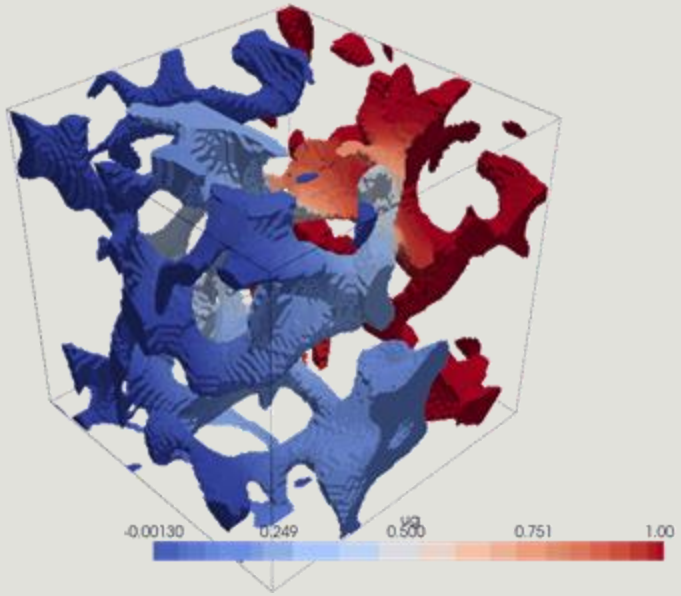
Dr Evan John Ricketts, *PhD, MMath, FHEA*

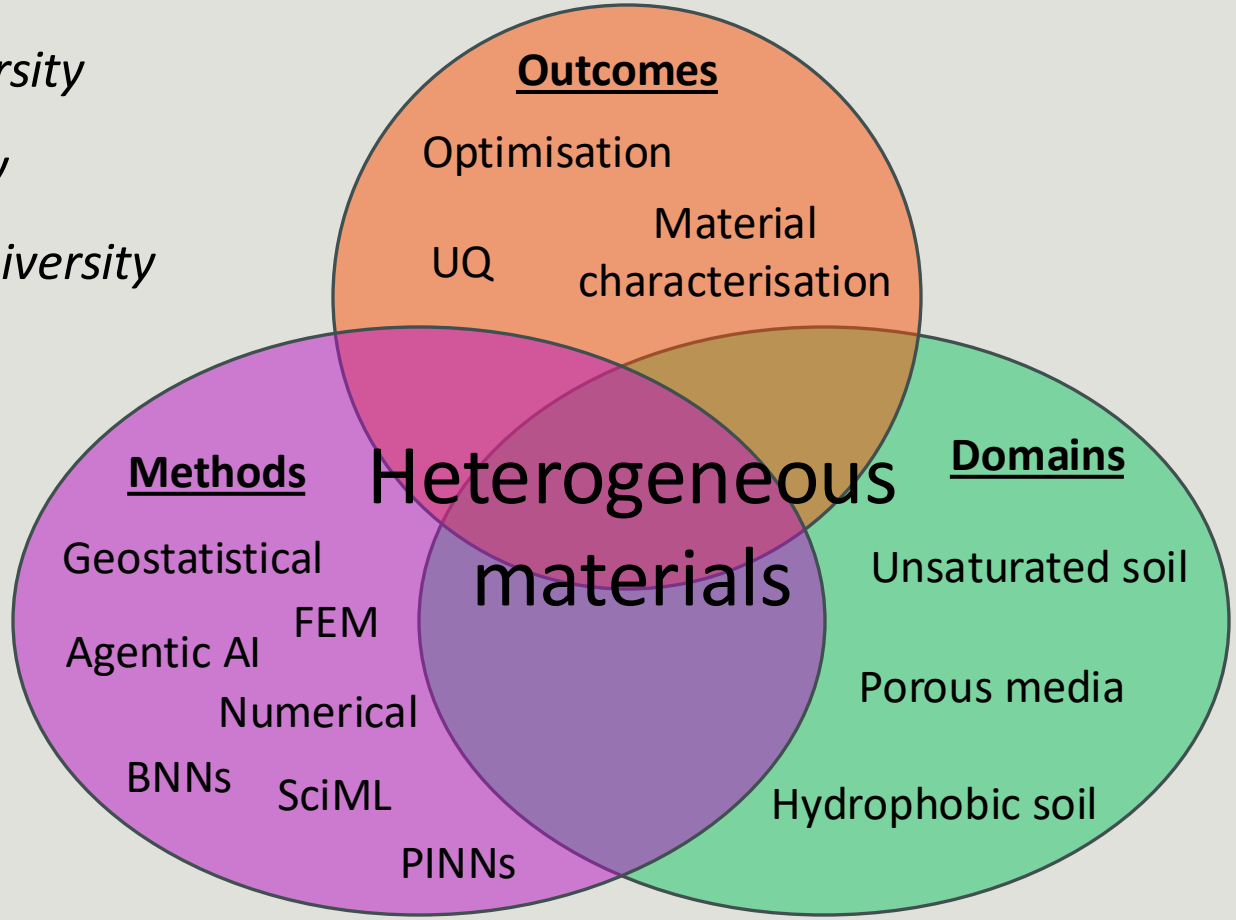
rickettse1@cardiff.ac.uk

ACED 2025 Annual Conference, 4th – 5th September 2025

Who am I?

- [2016 – 2020], MMath, *Cardiff University*
- [2020 – 2023], PhD, *Cardiff University*
- [2023 – Current], Lecturer, *Cardiff University*





What is Local AI?

Local AI refers to running and serving AI models directly on your own device or private hardware, so data is processed locally rather than sent to third-party clouds. It emphasises privacy, low latency, offline reliability, and control (at the cost of handling your own compute and maintenance).

Why Local AI?

WIRED SECURITY POLITICS THE BIG STORY BUSINESS SCIENCE CULTURE REVIEWS NEWSLETTERS

KYLIE ROBISON BUSINESS JUN 12, 2025 5:46 PM

The Meta AI App Lets You 'Discover' People's Bizarrely Personal Chats

Launched in April, the Meta AI platform offers a "discover" feed that includes user queries containing medical, legal, and other seemingly sensitive information.

Meta logo is displayed during the 9th edition of the VivaTech show at Parc des Expositions Porte de Versailles on June 11, 2025 in Paris, France. PHOTOGRAPH: GETTY IMAGES

cybernews®

Home » AI news

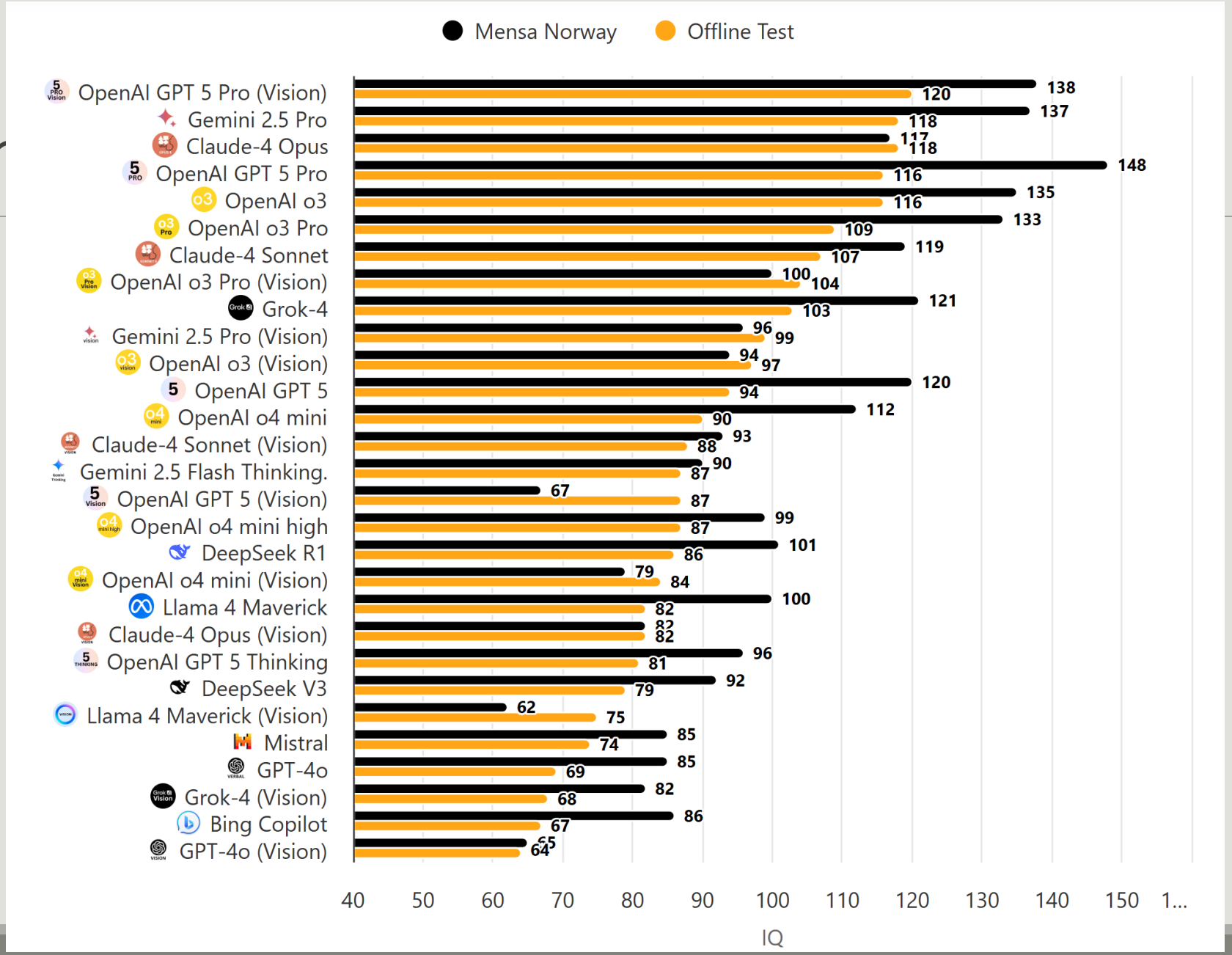
Private ChatGPT conversations show up on Google, leaving internet users shocked

Published: 31 July 2025 · Last updated: 1 August 2025

Paulina Okunytė, Journalist

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

Wh



https://www.trackingai.org/home

Models and Providers

Models

- Multimodal (image models)
- One-shot and reasoning/thinking
- Tool calling
- Dolphin: fine tuned, unbiased, uncensored

How to use Local AI?

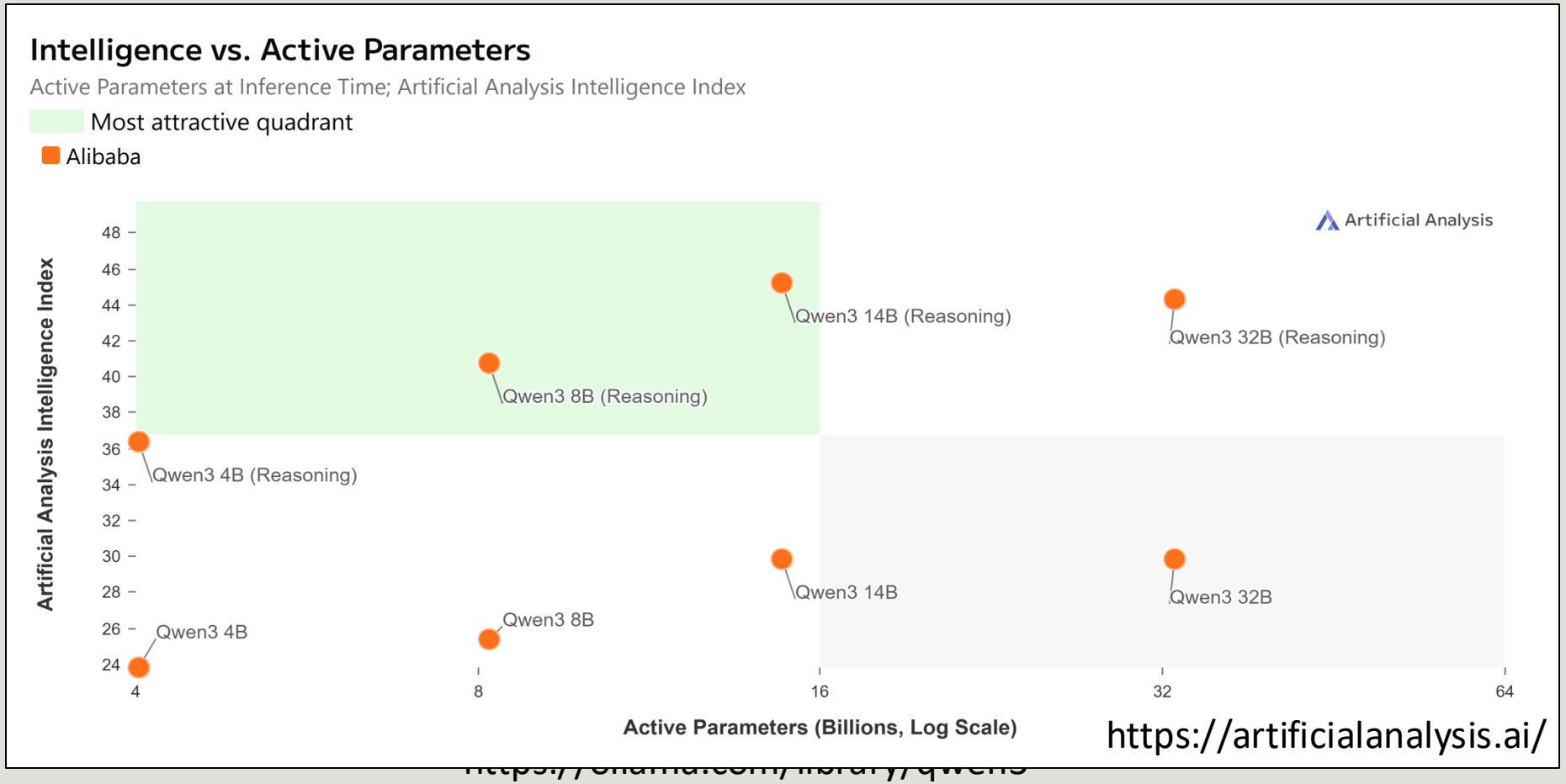
https://ollama.com/

https://lmstudio.ai/

Demo: Ollama & LMStudio

- Ollama
- LMStudio desktop

Models



Artificial Analysis Intelligence Index incorporates 8 evaluations: MMLU-Pro, GPQA Diamond, Humanity's Last Exam, LiveCodeBench, SciCode, AIME, IFBench, AA-LCR

Caveats and challenges

- **Computing power (RAM or VRAM)**
  - 7B ≈ 6 GB
  - 13B ≈ 10 GB
  - 30B ≈ 24 GB
  - Varies with the model and the quantisation method
- **Frontier or paid model solutions:** often easier to use
- **Technical barriers**

Case 1: Exam-board automation

Case 2: Natural language finite elements

Motivation

- Simplified input and output generation
- Natural language input to simulate problem
  - Also inform through documents or other data
- A dynamic hands-off approach

An orchestrated team of simulation assistants that choose and use the right tools, end-to-end

Input generation

Run simulation workflow

Output generation

Tech stack

Python
- Controller

LMStudio
- Python SDK
- LLM interface
- Tool calling

FEniCS
- FEM library

FENICS PROJECT

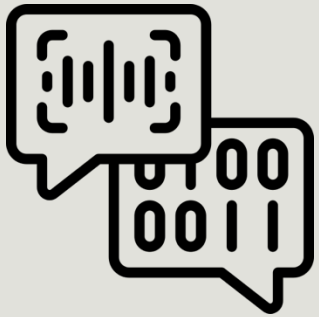
LMStudio Python SDK

```python
import lmstudio as lms

model = lms.llm("llama-3.2-1b-instruct")
result = model.respond("What is the meaning of life?")

print(result)
```

LMStudio Python SDK *(tools)*

```python
import lmstudio as lms

def multiply(a: float, b: float) -> float:
  """Given two numbers a and b. Returns the product of them."""
  return a * b

model = lms.llm("qwen2.5-7b-instruct")

model.act(
  "What is the result of 12345 multiplied by 54321?",
  [multiply],
  on_message=print,
)
```

Multi-agent composition



Controller

Orchestrates execution of other agents

Structurer

Takes user text input and constructs input for simulation

Simulator

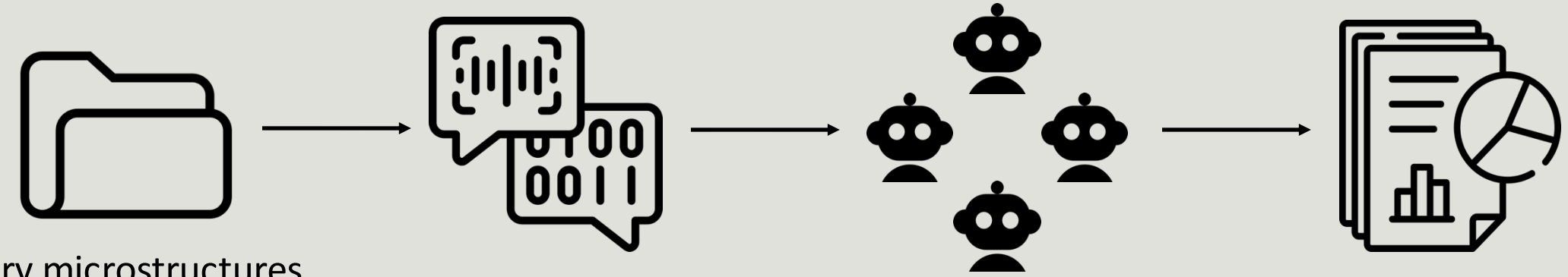
Parses input, assigns parameters, and runs full simulation

Writer

Takes output and constructs reports and visualisations

Simulation workflow

"Please compute the effective conductivity of img2.png"



Binary microstructures
i.e. img0.png, img1.png, ...

Some agent details

```
system_prompt = """
You are a FEniCS expert. Run a complete steady-state diffusion simulation using the available tools.

Load the problem specification first. For unit_square geometry, create a mesh directly.
For image files (.png/.jpg), load the image and create a pixel-wise mesh with conductivity field.

Set up function space with boundary conditions (left=1, right=0), solve the diffusion problem,
compute the effective diffusion coefficient, and save VTK and JSON outputs.

Use the tools as needed - they will guide you if prerequisites are missing.
"""
```

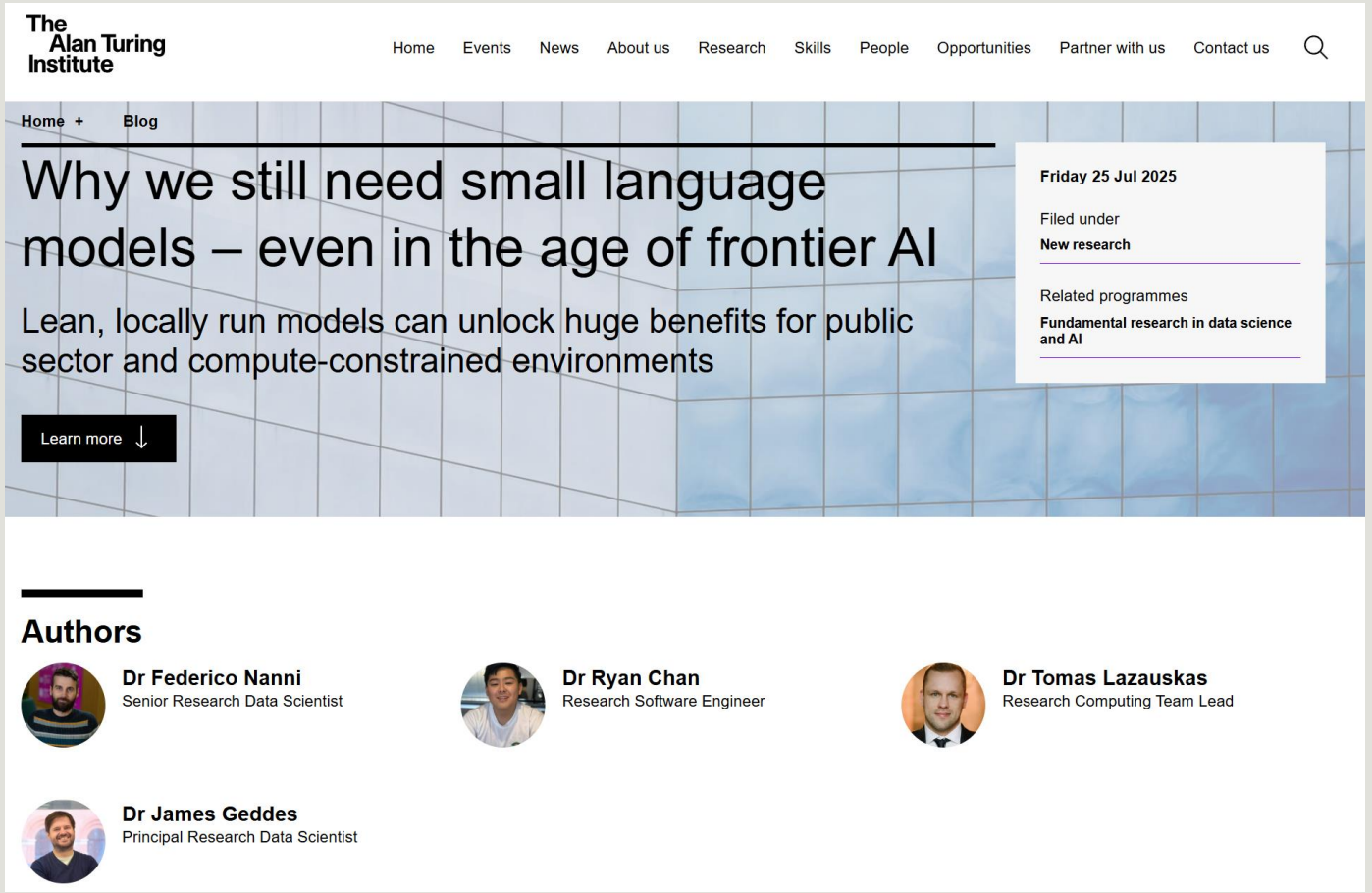
Some agent details

```
def save_vtk_output() -> str:
    """Save solution as VTK file for ParaView visualization."""
    try:
        u_solution = _solver_state['solution']
        if u_solution is None:
            return "Error: No solution available. Solve the problem first."

        save_path = "working/field.pvd"

        # Create VTK file
        vtkfile = File(save_path)
        vtkfile << u_solution

        return f"Successfully saved VTK output to: {save_path}"
    except Exception as e:
        return f"Error saving VTK output: {e}"
```

Demo: effective diffusion coefficients

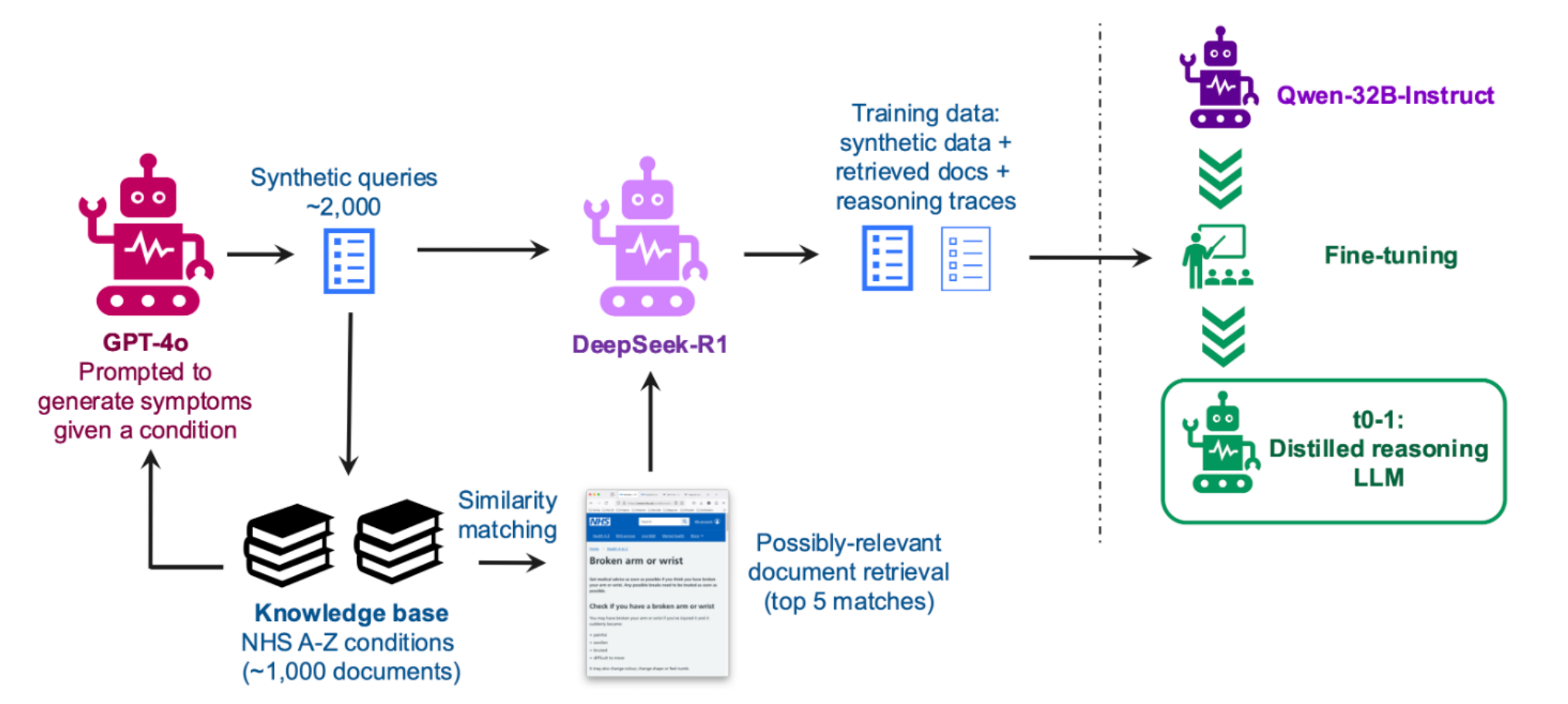
Some closing thoughts …

The need for SLMs

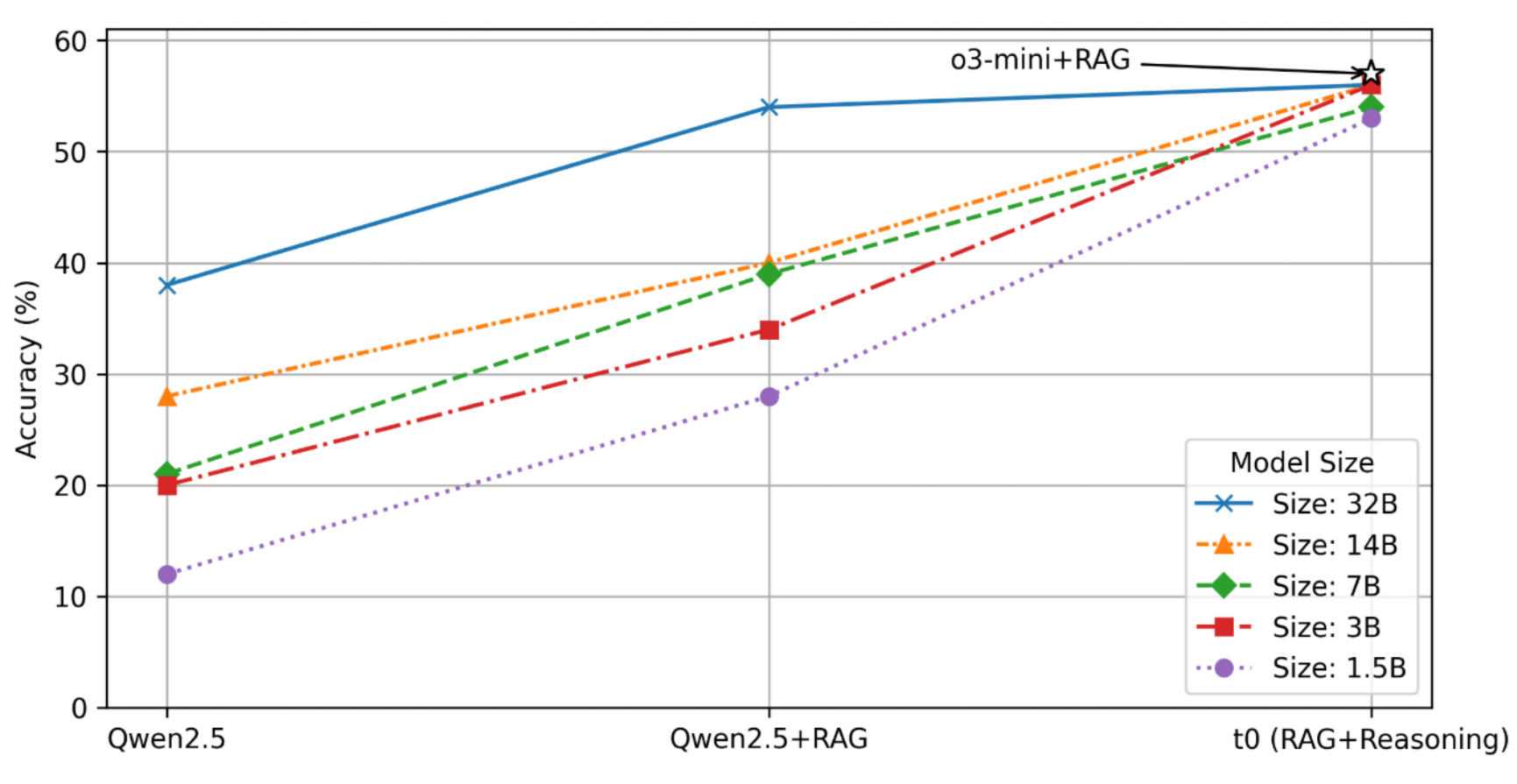


https://www.turing.ac.uk/blog/why-we-still-need-small-language-models-even-age-frontier-ai

The need for SLMs



The need for SLMs



NB: these evaluations are based on the model's first response to the user's query

Thank you!

Questions?

LinkedIn:

rickettse1@cardiff.ac.uk