UNIVERSITY of HULL

From Models to Agents:

LLM-Driven Renewable Energy Systems

Integrating Forecasting, Optimisation and Control

Dr Zekun Guo

Lecturer, Data Science, AI and Modelling (DAIM) Centre

University of Hull

SuperAIRE 2026 ECR Workshop

Outline

Part I: AI Foundations & the Agent Era

1. Transformers and the LLM Revolution
2. Frontier Models: GPT-4o, Claude 4.6, Gemini 3, DeepSeek-R1
3. The Agentic Era: Computer Use, Tool Use, and Reasoning

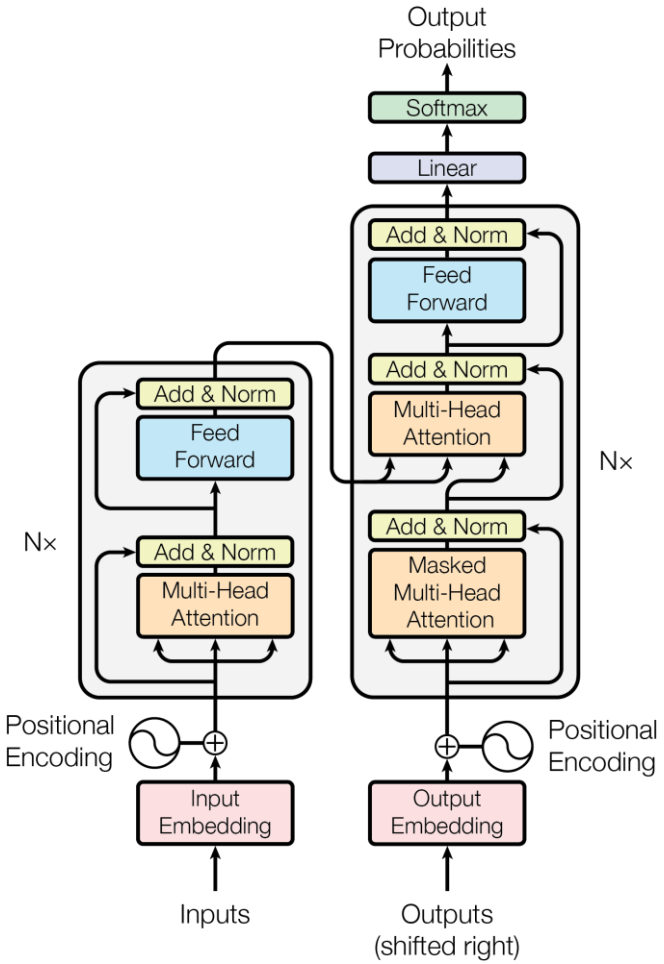
Part II: Challenges in Renewable Energy

Part III: Our Research

4. From Models to Agents: The Evolution
5. Application: LLM-Based Solar Forecasting
6. Energentic – a research focused venture
7. Results and Key Findings

Transformers and the LLM Revolution

"Attention Is All You Need" (Vaswani et al., 2017)



[1] Vaswani et al. (2017) NeurIPS.
[2] Hu et al. (2022) LoRA, ICLR.

Self-Attention Mechanism

Q (Query), K (Key), V (Value) matrices

$$\text{Attention}(Q, K, V) = \text{softmax}!\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Each token attends to all others in parallel

From Transformers to LLMs

- GPT: Decoder-only Transformer
- Pre-training: next-token prediction on internet-scale text
- SFT + RLHF/DPO: instruction following + alignment
- Emergent capabilities at scale: reasoning, code, planning

LoRA: Efficient Fine-tuning

- Freeze base weights, add low-rank update: W' = W + BA
- Our setup: Qwen2.5-7B + LoRA (r=64, a=128)

Frontier Models in 2025-2026

Reasoning Models

- OpenAI o3 (2025): explicit chain-of-thought reasoning at inference time
- DeepSeek-R1 (2025): open-source reasoning via RL, MoE architecture
- Claude Opus 4.6 (Feb 2026): tops benchmarks in agentic coding, tool use, 1M context
- Gemini 3 Flash (Jan 2026): frontier intelligence, sub-500ms, 1M context

Multimodal & Omni Models

- GPT-4o (2024): unified text + audio + vision, real-time conversation
- Claude Sonnet 4.6 (Feb 2026): enhanced computer use + 1M context
- Gemini 2.5 Pro: native multimodal, 1M token context, strong scientific reasoning

Open-Source Ecosystem

- Meta Llama 3.1 (405B): open-weights, community fine-tuning, competitive quality
- Qwen2.5 (Alibaba, 0.5B-72B): our solar forecasting base model, strong multilingual
- DeepSeek V3: open MoE model, competitive with GPT-4 on many benchmarks

[3] OpenAI (2023-25). [4] DeepSeek-AI (2025). [5] Anthropic (2026). [6] Yang et al. (2024) Qwen2.5.

The Agentic Era: Beyond Chat to Autonomous Action

UNIVERSITY of HULL

Computer Use & Deep Research

OpenAI Operator / CUA (2025):

Processes raw pixels, uses virtual mouse/keyboard
Navigates websites, fills forms, multi-step tasks

Claude Computer Use (2025-26):

Claude controls a desktop environment directly
Sonnet 4.6: enhanced computer use capabilities

OpenAI Deep Research (2025):

Powered by o3, searches & analyses 100s of sources
Produces research-analyst-level reports autonomously

Key Shift:

From "chat with AI" to "AI does the work"
Agents reason, plan, use tools, and act in the world

From Chatbot to Agent: Evolution

2022: Simple Chatbot

Input -> LLM -> Output (no tools, no memory)

2023: ReAct / Tool-Using Agent

Think -> Act -> Observe loop (Yao et al.)

2024: Autonomous Agents

Planning + Memory + Tool use + Code execution

2025: Agent Infrastructure

MCP, A2A, Agent SDKs -- production-grade

2026: Agent-Native Applications

Computer use, deep research, multi-agent
Enterprise adoption "exceeding projections" (AWS)

Energy Agent (Our Work):

LLM + MPC + Sensor feedback + Rolling control

[7] OpenAI Operator (2025). [8] Anthropic Claude Computer Use (2025-26). [9] OpenAI Deep Research (2025).

Challenges in Renewable Energy & What LLMs Can Do

UNIVERSITY of HULL

Why Current Approaches Fall Short

Renewable Energy Challenges:

High variability: 10-20% day-ahead forecast error
Multi-system coupling: PV + wind + battery + grid + EV
Real-time decisions at 15-minute intervals
Dynamic electricity pricing and market complexity

Traditional Pipeline Limitations:

Siloed: Forecast -> Optimise -> Control (no feedback)
Static parameters that never adapt
Manual intervention for parameter tuning
No explainability for decisions

What LLMs Bring to Energy Systems

LLM Capabilities:

Data interpretation: parse weather, grid, market
Forecast reasoning: contextual uncertainty analysis
Decision explainability: human-readable rationales
Adaptive modelling: dynamic parameter adjustment

But LLMs Alone Are Not Enough:

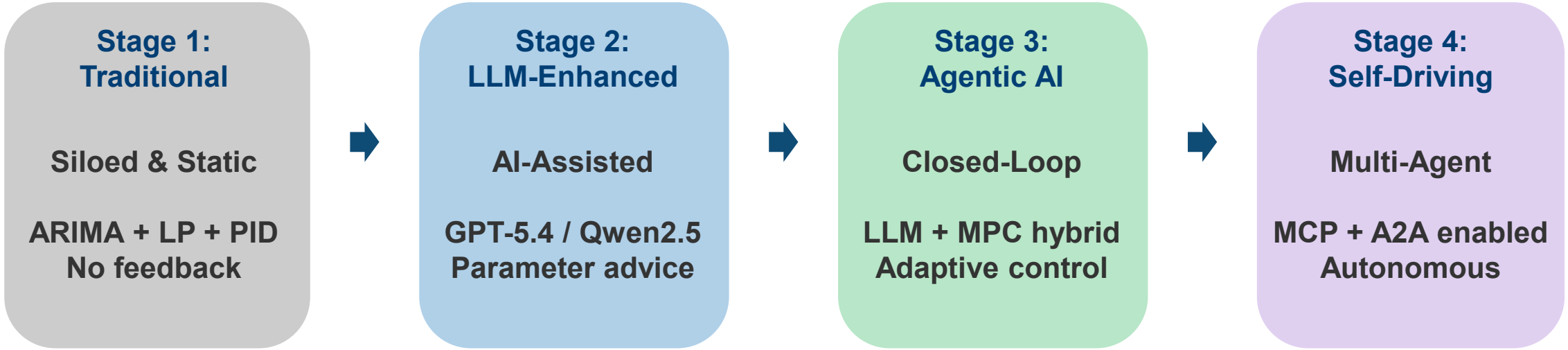
No guarantee of physical constraints
Unstable numerical outputs (hallucination)
No closed-loop control capability

Solution: LLM + MPC Hybrid

LLM reasoning + MPC constraint guarantees

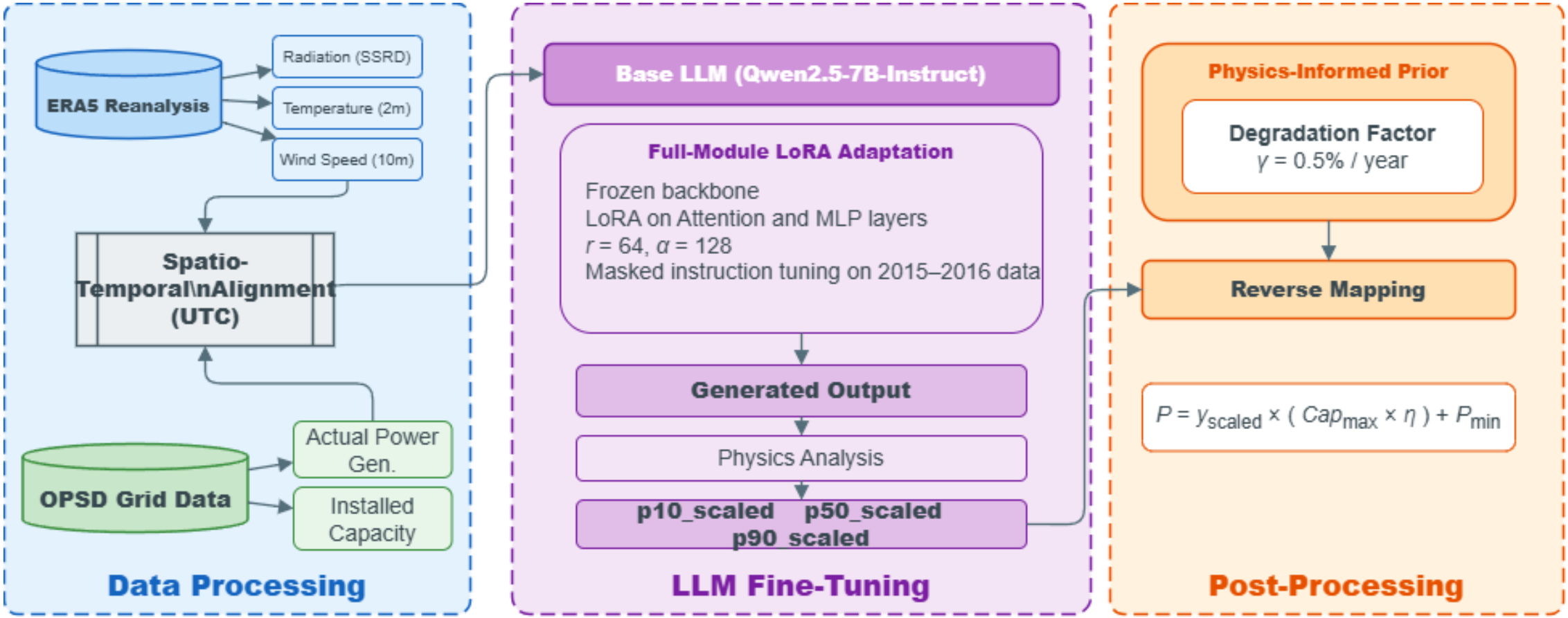
From Models to Agents: An Evolutionary Framework

UNIVERSITY of HULL



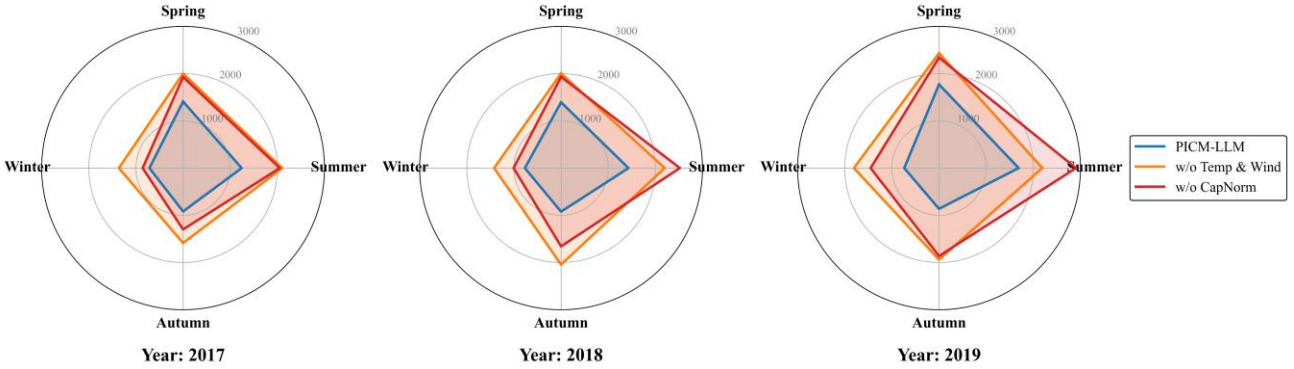
Agent = LLM (Reasoning) + MPC (Optimisation + Constraints) + Environment (Feedback)

LLM-Assisted Energy Time-Series Data Forecasting



LLM-Assisted Energy Time-Series Data Forecasting

1. **PICM-LLM shows strong cross-year stability**, maintaining high accuracy and reliable interval coverage from 2017–2019 despite distribution drift.
2. **Traditional deep models collapse in 2019**, with large errors and severe under-coverage, while PICM-LLM remains robust in both point and interval metrics.
3. **Balanced probabilistic performance:** PICM-LLM keeps PICP >80% with controlled PINAW, making it suitable for real-world scheduling applications.



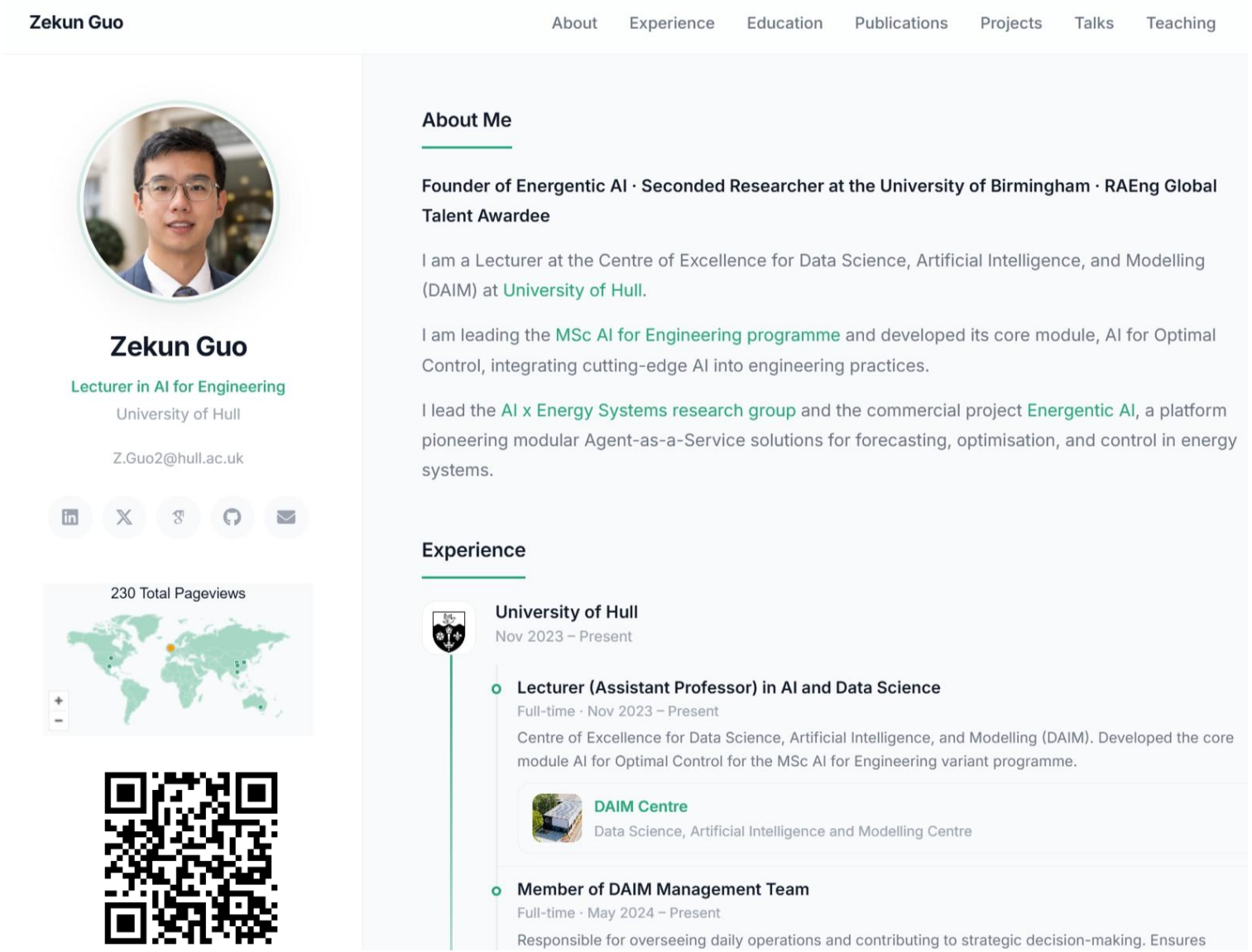
UNIVERSITY of HULL

Thank You!

Questions & Discussion

Dr Zekun Guo

Lecturer, DAIM Centre, University of Hull
zekun.guo@hull.ac.uk
https://drzekunguo.github.io/



Selected References:

[1] Vaswani+ (2017) NeurIPS. [2] Hu+ (2022) LoRA, ICLR. [3] OpenAI (2023-25) GPT-4/o3. [4] DeepSeek-AI (2025) R1.
[5] Anthropic (2026) Claude 4.6. [6] Yang+ (2024) Qwen2.5. [7-9] OpenAI Operator, Claude CU, Deep Research (2025).
[10] Anthropic MCP (2024-26). [11] Google A2A (2025). [12-13] Agent SDKs. [14] GAIA (2025). [15] X-GridAgent (2025).