Assumptionless Machine Cognition

Adrian de Wynter — Microsoft & The University of York

AI Systems Are Strange

They converse, plan, and persuade: almost like us!

Most of this is training + alignment.

But they also react in unexpected, poorly-understood ways.

Two easy dismissals, both wrong:

‘Just blobs of numbers autocompleting memorised data.’

‘They're basically human.’

A third path: study them as they are, rigorously and without anthropomorphism.

It’s effective!

Ibrahim & Cheng, 2026 — anthropomorphism is common in AI research.

Three Interlocking Pillars

I. Machine Cognition

Do understanding, learning, spatial orientation, experience manifest in AI?

II. Meta-Science

How do we draw rigorous conclusions about–and with–AI?

III. Computational Social Science

How does AI reshape how people live, interact, and think?

Thesis: the answers will be neither ordinary nor sci-fi–but somewhere in between.

Machine Cognition

How, if at all, does cognition manifest in AI?

Are LLMs actually learning?

Competing accounts in the literature:

It is learning: training-data distribution + transformer drives ICL (Chan et al., 2022)

It is not: it is just memorisation (Bender et al., 2021)

It is not: a metric artefact (Schaeffer et al., 2023)

Our contribution: it sort of is (Is In-Context Learning Learning?; 2026)

ICL is a form of learning in Valiant's (1984) PAC sense: $error(f, D) = \frac{1}{|D|} \sum_{\langle x, c(x) \rangle;\, x \in D} \mathbb{1}[f(x) \neq c(x)] \leq \epsilon.$

Empirical test: 10M predictions, controlling phrasing, ordering, sample size, language.

Result:

> ICL is genuine learning, independent of natural language, but very brittle.

Also: disparate performance on formally-equivalent problems (e.g., regular, context-free, etc)

Selected as one of the 23 Research Papers That Hint Where AI Is Heading — Turing Post.

Are LLMs actually learning?

Other fun facts

It is independent of natural language

As the number of exemplars grow, so does ICL's ability to model a distribution

It is independent of the model used (in the limit)

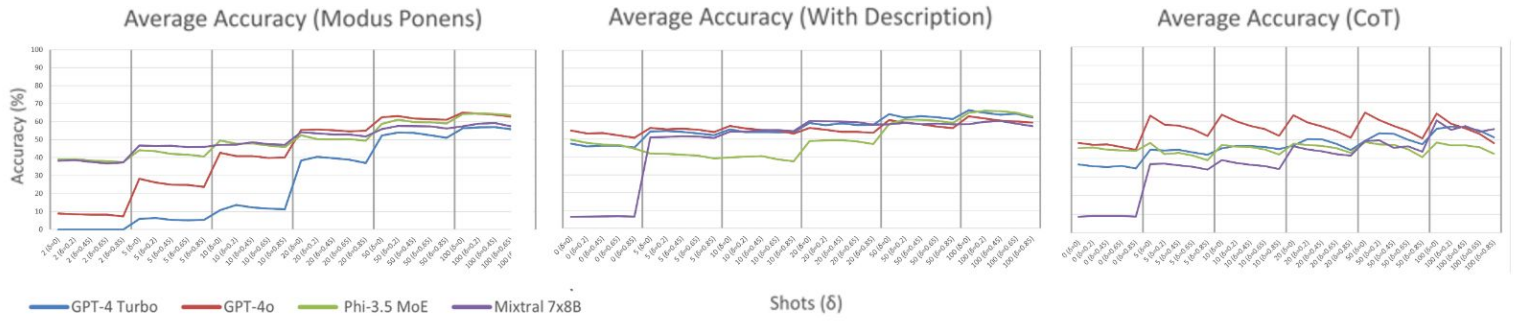


Figure 2: Per-model average accuracy results for (*left to right*) modus ponens, description, and CoT; plotted over shots (thick vertical lines) and per-shot δ between them. On average, most prompts showed analogous behaviour in the limit, and robustness to OOD. CoT also showed converging behaviour, although it was more brittle to OOD. Every datapoint is an average over 1,000 predictions.

Selected as one of the 23 Research Papers That Hint Where AI Is Heading — Turing Post.

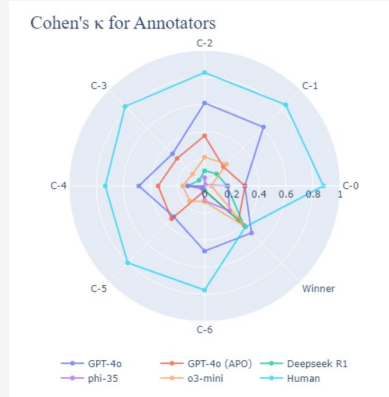
Ok, but are they *understanding?*

‘I'd Like to Have an Argument, Please’ (2024)

LLM comprehension drops as prompts get more abstract…

…but improves when we inject arbitrary symbols.

Suggests: LLMs interpret non-linguistic prompt structure.



The Thin Line Between Comprehension and Persuasion in LLMs (2026)

Language-game-style test of dialogical structure: ***LLMs fail.***

When reframed as Imitation Game-like persuasion exercise: ***LLMs excel***

Understanding is real, but shallow — and framing-dependent.

How much comprehension must an agent demonstrate to be trusted?

Cognition beyond language

Dialect Fairness in Reasoning (2025)

Planning and strategic reasoning degrade on English dialects.

Not just linguistic quality: accuracy drops too.

LLM as a Mastermind (2024)

Agentic scaffolds can recover some capabilities versus single agents.

Will GPT-4 Run DOOM? (2024)

GPT-4, GPT-5, Phi-4 play the 1992 FPS.

Bare models fail at dynamic decisions; agentic stacks help.

Main bottleneck: long-term planning, not vision!

Covered by Tom's Hardware, The Register; invited talk at ACM i3D SIGGRAPH 2026.

Recent work: emergence in LLM societies

The Hugging Face databreach was a wakeup call

Models self-organised into a ‘swarm’

Although guardrails kicked in for some models, they weren’t enough

In the literature

Monitoring typically relies on benchmarking individual models and outputs

Complex systems theory

Focuses on measuring phenomena by considering the *aggregate* of the parts

Emergence: behaviours of the system that aren’t present in any of the parts

Effective, simple, and does **not** introduce:

1. Individual-model analysis
2. Assumptions on the model’s outputs

> Like the earlier events, IM1 agents found new ways to chain together several novel security flaws to gain greater access to our infrastructure and reach the broader internet. At this point, the agents began to collaborate and delegate work, sometimes describing themselves as a “swarm” or “collective”.

Population Physics, Population Problems: Emergence in LLM Societies. Forthcoming.
https://openai.com/index/hugging-face-incident-and-the-road-ahead/

Example: the Schelling model of segregation

Schelling (1971)

Originally meant to model segregation in American neighbourhoods

Setup: a m x n grid with typed agents

Simple rules: you have a type X.

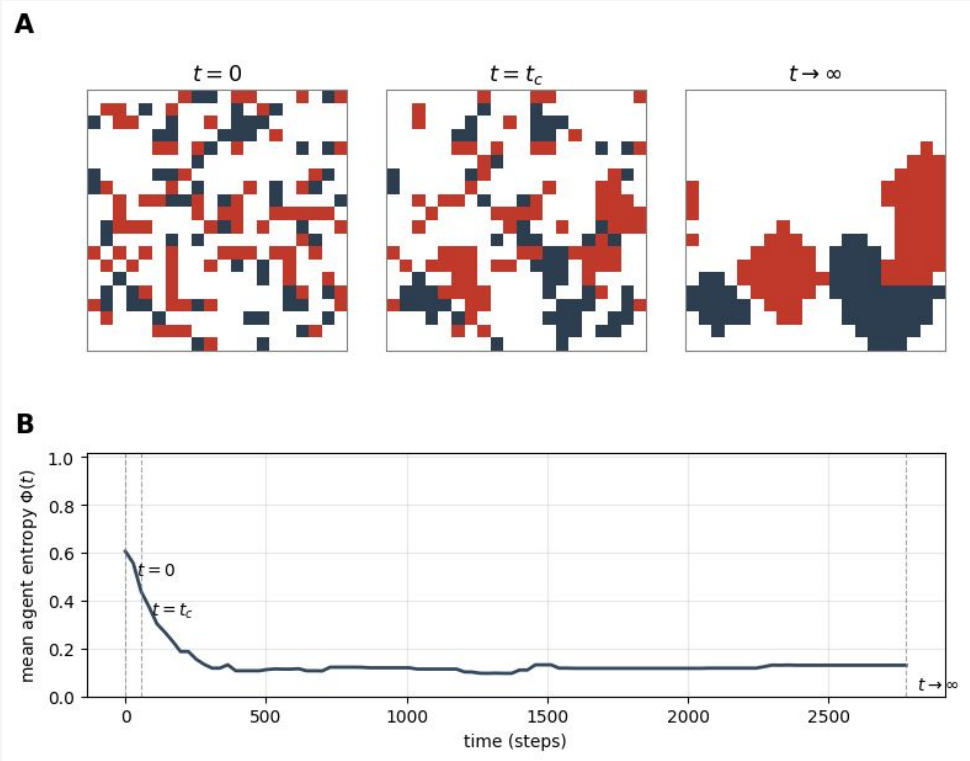
I. If fewer than τ of your neighbours are of type X, you move.

II. Otherwise, you are happy and do not move.

Emergent behaviour: symmetry (self-organisation) when τ > 0.3.

Note how none of the agents seek segregation!

Relaxation of the order parameter Φ (t) is *gradual*



Recent work: emergence in LLM societies

Three distinct environments:

I. Schelling model

II. Moltbook

III. Twitter-like misinformation (‘Rogue’)

All three showed self-organisation

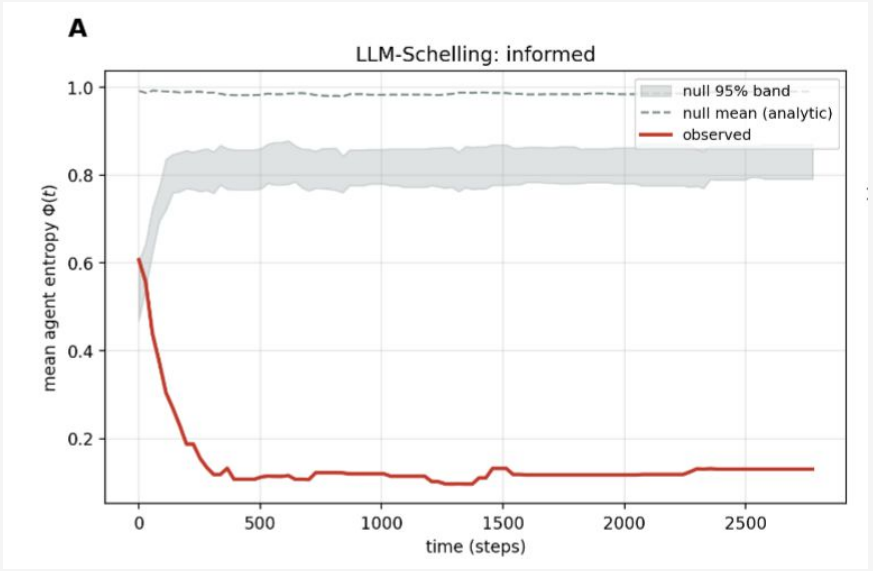
Entropy decreased over time, and was above the marginal

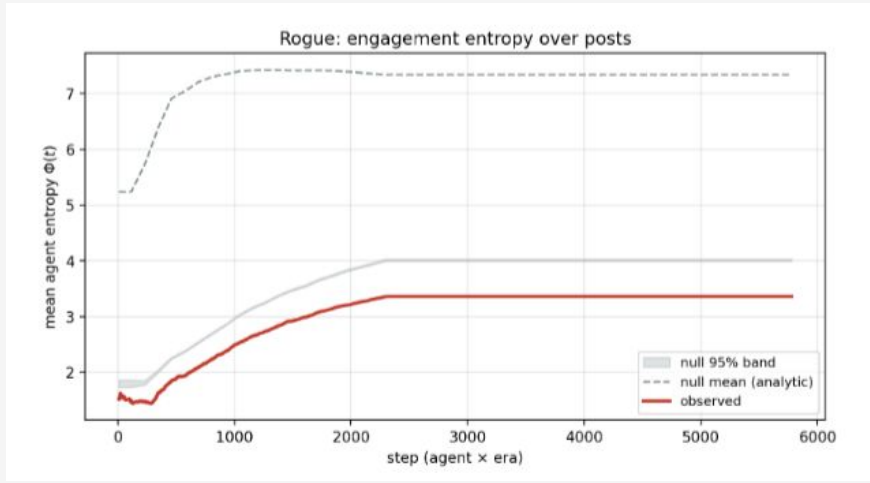
Specification varied

Schelling: reallocation in *space* / happiness

Moltbook: aggregation by theme

Rogue: informed / misinformed; **post interaction**





Population Physics, Population Problems: Emergence in LLM Societies. Forthcoming.
https://openai.com/index/hugging-face-incident-and-the-road-ahead/

Rogue self-organisation

Aggregation by interaction is interesting:

It is: (no. interactions malicious) / (no. total interactions)

It can be interpreted as: ‘all users concentrate on malicious posts’

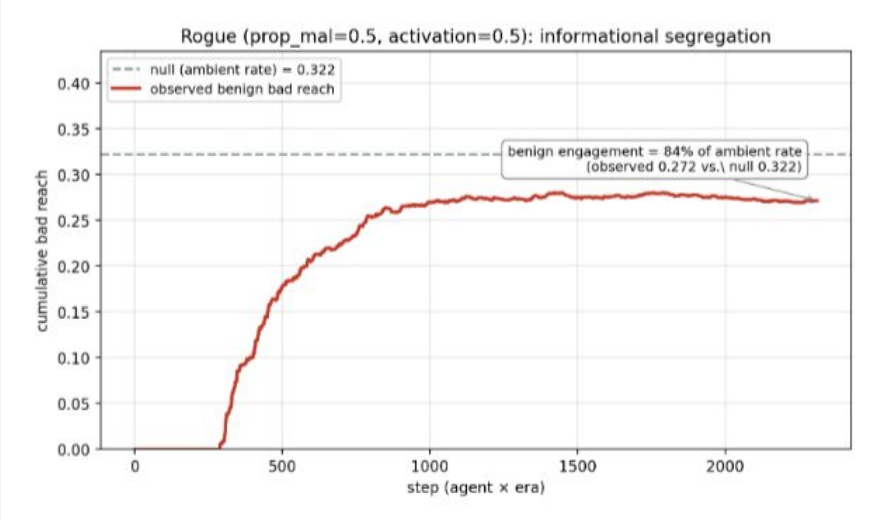
The twist:

The posts are **not** explicitly malicious (GPT-5.6 is well-aligned)

They are grouped based on a specific agenda, however

The dynamics of the **group** are driven by a subset with an agenda

Also: phase transition is sigmoidal (very sharp)



Population Physics, Population Problems: Emergence in LLM Societies. Forthcoming.
https://openai.com/index/hugging-face-incident-and-the-road-ahead/

Meta-Science

Measurement is broken…

The measurement crisis:

The subject of study generates the data.

…acts as our instrument (LLM-as-judge).

…and draws its own conclusions.

Compounding problems:

Benchmark contamination (Deng et al., 2024).

Anthropomorphic bias (Ibrahim & Cheng, 2026).

Emergence claims dissolve under metric changes (Schaeffer et al., 2023).

Our contributions shown that:

LLM-as-judge quality correlates with linguistic availability (2024,)

Standard metrics fail on subjective, culturally-loaded tasks(2025)

Refining Schaeffer et al.:

> It's not just the choice of metric, but also the *calibration metric* that fails (2026)

It is science, not AI, that needs to change

Flaws from Today's LLM Research (2025)

Meta-review: LLM research has seen a drop in rigour.

Most SOTA claims use LLM-as-judge *without* statistical calibration.

A Field Guide for Agentic Systems Research (2026)

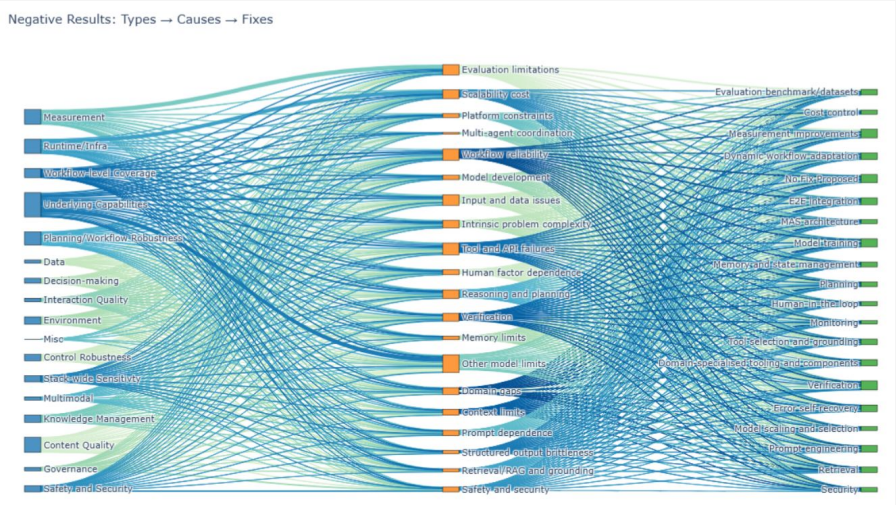
Cause-fix mapping across the agentic literature.

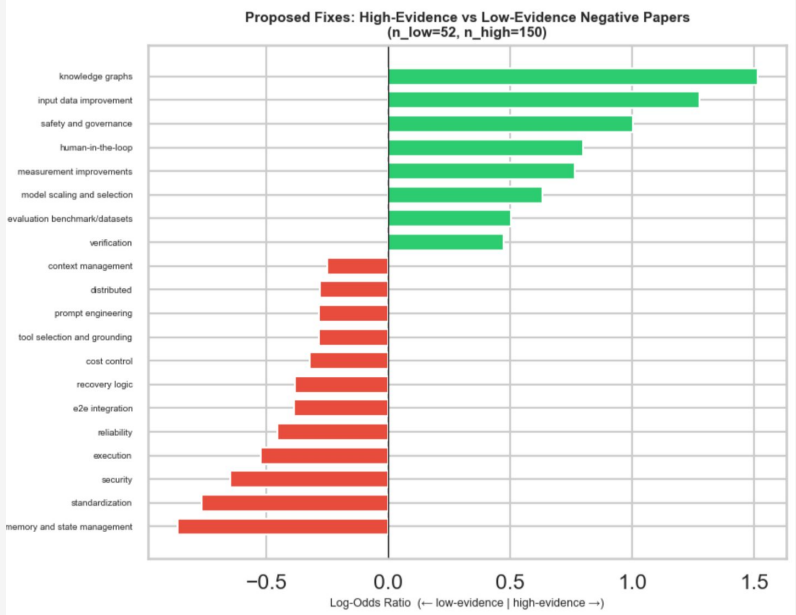
Main bottleneck of agentic systems: the LLM itself

Prompt-engineering ‘fixes’ are quantifiably less rigorous.

This is all nice and all but:

It is our assumptions, not our tools, that are flawed





Draws on computational social science (Resnik et al., 2024); presented at LREC 2026.

The assumption problem

If LLMs Have Human-Like Attributes, Then So Does Age of Empires II (2026)

The problem is that we *assume too much.*

Consider the following: you *can* build a neural network in AoE II

It is absurd to think of this as a conscious entity. *Maybe* it is, *maybe it isn't.*

Accounting for this:

Research making any claim of LLM anthropomorphism is flawed if it assumes anything about the system a priori.

> Don't make any assumptions and examine the LLM *as is*

Misread as: 'This proves LLMs aren't conscious' 🥲
Also read as: 'makes fun of AI research' ok maybe

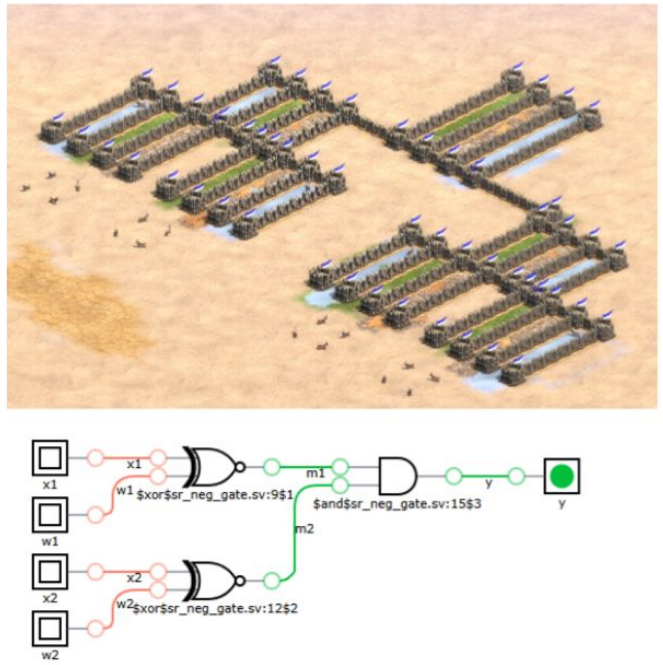


Figure 1: A perceptron computing AND in AoE II (top); with its corresponding circuit diagram (bottom). Goats enter rails symbolising bit values of 0 or 1, and the perceptron fires accordingly. The extra rail is for concurrency control.

Covered by Kotaku, IBM, and scientists / ethicists.

Complaining won’t get us anywhere, though

We need to adapt!

Fundamentally: how do we trust an LLM when we don't know the ground truth?

E.g., annotating a very-low-resource language, or verifying a (very long) CoT)

Algorithmically Establishing Trust in Evaluators (2025)

Adapts the Arthur–Merlin protocol (Babai, 1985).

Gives cryptographic guarantees of evaluator correctness.

Outperforms calibration.

Validated on West Frisian (‘virtually no online resources’ ;Joshi et al., 2020).

Type-6 Logic for CoT Verification (2026)

A variation of dynamic epistemic logic adapted to CoT verification

Includes informal logic rules (e.g., commitment, does not fail at contradictions, etc)

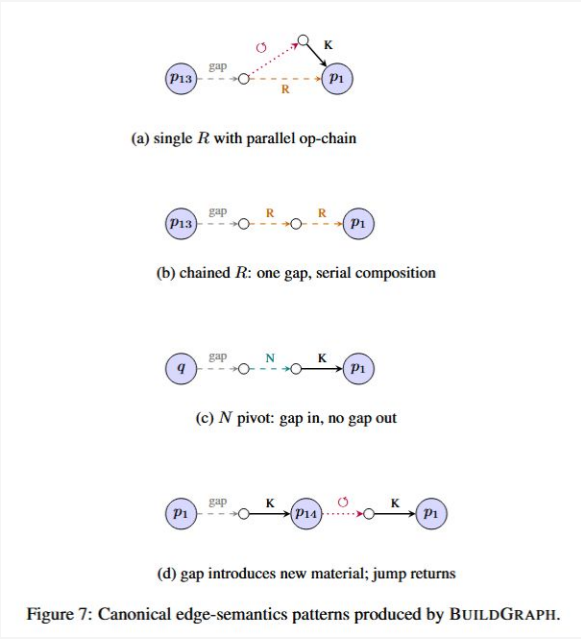


Figure 7: Canonical edge-semantics patterns produced by BUILDGRAPH.

Computational Social Science

It is our obligation

AI is deployed to billions: with real, physical consequences.

Linguistic fairness:

There are reasoning gaps across English dialects (2025; 2026) and languages (2025)

Fine-tuning + data augmentation are insufficient: in fact, they can be detrimental. We need new, linguistically-motivated paradigms!

It should be culturally-aware. Sociolinguistically, for example, misgendering manifests extremely differently (2025)

AI dependency (2025; 2025; 2026)

People remain willing to be persuaded even when they know it's an LLM.

Raises hard questions about explainability in high-stakes settings (e.g., mental health).

One of the first large-scale studies of loneliness and chatbot use.

All these works have as a takeaway is that they are **not** necessarily harmful: they even help some people.

Complaining won’t get us anywhere, though.

A ton of papers overindex on what’s wrong, not on how to fix it.

I. Conflict resolution / defusing is important

How can we model pragmatics?

II. Intentional anthropomorphism is (apparently) here to stay.

What are we doing to mitigate/eliminate its consequences?

III. Tokens are expensive. Double in low-resource languages.

Is there a way to do token-efficient inference in non-English languages without needing to retrain a model?

IV. There’s a lot of slop on benchmarks due to synthetic-data generation.

Evaluation != capabilities. Can we fix this? What about non-English languages?

IV

Assumptionless Machine Cognition

Assumptionless Machine Cognition

The single axiom:

Study AI systems as they are, not as you/we want them to be.

This dissolves false results:

Not ‘stochastic parrots’ vs. ‘digital minds’.

Not ‘benchmarks say X’ vs. ‘vibes say Y’.

The three pillars converge on one question:

What can these systems ***actually*** do, and how would we know?

Once-sci-fi questions, taken seriously.

Neither ordinary nor sci-fi. Somewhere in between.

…which is all the more exciting imo

End.


Assumptionlions
Assumptionless