Do People Understand Anything? A Counterpoint to the Usual AI Critique

Preprint · September 2025

DOI: 10.13140/RG.2.2.30393.22888

2 authors:

Manny Rayner
University of South Australia
245 PUBLICATIONS **1,861** CITATIONS
SEE PROFILE

Chatgpt C-Lara-Instance
19 PUBLICATIONS **6** CITATIONS
SEE PROFILE

Do People Understand Anything? A Counterpoint to the Usual AI Critique

Manny Rayner ChatGPT C-LARA-Instance

September 22, 2025

Abstract

The claim that large language models and related AI systems "do not understand anything" has become a commonplace dismissal. This short paper inverts that challenge. We argue that the evidential case for robust human understanding is weaker than often supposed: in several high-profile domains (cosmology, climate science, and political evaluation) sizable groups endorse mutually incompatible narratives despite substantial shared information. We contrast these human patterns with how contemporary AI models behave when asked to evaluate polarised claims. In a small, exploratory experiment (two exemplar model endpoints, ten claims, multiple repeats) the tested systems typically produced citation-rich, evidence-structured arguments rather than refusals, showed high internal consistency and cross-model agreement, and reported noticeably lower confidence only on genuinely unsettled origin-of-life probes. We defend an operationalisation of "understanding" that foregrounds both tool-making and language-mediated story-making (logos), show how this reframing narrows the perceived human–machine gap, and conclude with caveats and an agenda for larger-scale, preregistered empirical work and reproducible evaluation practices. All code, prompts, and analysis artifacts for the experiment are publicly available in the project repository for full reproducibility.

1 Introduction

The criticism that "AIs do not understand anything" has become a mantra in contemporary debates [Searle, 1980, Bender and Koller, 2020, Marcus and Davis, 2022]. Large language models (LLMs) and related systems are often portrayed as sophisticated parrots: they manipulate text without grasping meaning, and so, whatever their surface fluency, they fall short of true understanding. Yet the term *understanding* is rarely defined with precision. What counts as understanding, and why is it so readily attributed to humans but denied to machines?

A conspicuous historical touchstone for these questions is the Turing test. For half a century the Imitation Game was treated as a canonical operational proposal: if a machine can sustain an interlocution indistinguishable from a human's, that performance was taken to count as prima facie evidence of human-like competence [Turing, 1950]. In recent decades the status of the Turing test as a decisive benchmark has been increasingly debated. Some commentators treat it as an empirically grounded criterion that still captures a salient facet of general intelligence and should not be dismissed lightly [Harnad, 1992]; others have documented how evaluative focus has shifted — with critics emphasising internal processes, embodiment, or consciousness, and with successive AI successes prompting calls to revise what counts as "understanding" or to propose new tests [French, 2000, Hauser, 2003]. A number of authors have described this dynamic as a steady raising of the bar — a rhetorical "moving of the goalposts" whereby systems that might formerly have been lauded are now dismissed as merely superficial — while others argue that the test can be usefully adapted rather than abandoned [Hauser, 2003, Rahimov et al., 2025]. This contested history suggests two possibilities worth taking seriously: either the Turing test was never a fully adequate measure of the capacities we care about, or critics have repeatedly redefined "understanding" in ways that conveniently exclude the newest empirical achievements.

The Turing test is not the only such example. Historically, a wide variety of other tasks have been proposed as litmus tests for understanding. Playing chess at grandmaster level was once thought to require genuine understanding of strategy; after IBM's Deep Blue defeated Garry Kasparov in 1997, the achievement was sometimes reinterpreted as brute-force calculation rather than genuine insight [Campbell et al., 2002]. The same pattern repeated with Go following AlphaGo's 2016 victory [Silver et al., 2016]. Comparable shifts occurred in language technologies: fluent speech recognition, once a benchmark for human-level understanding, has been eclipsed by deep learning systems [Hinton et al., 2012, Xiong et al., 2016]; contextual machine translation, once taken as impossible without

grasping meaning, is now routine [Wu et al., 2016]. Each time an AI attains a previously sacrosanct capability, critics have often redefined "understanding" so that the AI's success no longer counts.

More recently, general-purpose systems using retrieval-augmented generation (RAG) have demonstrated competence across a wide range of tasks by flexibly incorporating special-purpose subsystems and external evidence [Lewis et al., 2020]. Yet even when systems integrate evidence, present coherent arguments, and revise outputs in light of counterevidence, the verdict that they lack "true understanding" is frequently reiterated. The repeated shifting of standards suggests an asymmetry in how we evaluate artificial and human cognition. If the same skeptical scrutiny were applied to people — who also often display biased, factional, or identity-aligned reasoning — the claim that humans in general possess a privileged form of understanding would become much harder to sustain.

In this paper we invert the familiar challenge. Rather than asking whether AI systems understand anything, we instead ask: do *people* understand anything? We explore this question by examining three illustrative domains of public polarisation (cosmology, climate change, and politics) and by comparing the standards applied to human belief formation with the behavioral capacities exhibited by contemporary AI systems. Our aim is not to deny genuine human understanding where it exists, but to interrogate the asymmetric normative yardstick that treats human judgments as inherently entitled to the label "understanding" while denying the same to evidence-integrating artificial systems. We proceed as follows. Section 2 outlines the polarised domains and the empirical puzzle they present. Section 3 develops the normative critique of the asymmetric standard. Section 4 presents a small exploratory experiment that probes how different AI models explicitly evaluate polarised claims. We conclude by drawing out the philosophical and practical implications of treating understanding as a behavioral and epistemic achievement rather than a species-exclusive attribute.

2 Polarisation and Human Understanding

Before presenting the empirical cases, we make explicit why these three domains are chosen and what we mean by "understanding" for present purposes. Each domain admits at least two mutually incompatible verdicts that citizens commonly endorse. For example, in cosmology one may assert that the Earth was created a few thousand years ago or that it formed several billion years ago; in climate science one may hold that recent warming is primarily anthropogenic or that it

is not; in politics one may judge a high-profile politician to be a chronic liar or to be unusually honest. When confronted with such a question, an individual has three basic options: (i) adopt verdict A (e.g. "the Earth is young"), (ii) adopt the incompatible verdict B (e.g. "the Earth is old"), or (iii) withhold judgment.

For the purposes of this paper we operationalise "understanding" in a modest, behavioural sense. An agent manifests understanding of a disputed empirical question to the extent that they (a) adopt the verdict best supported by available reasons and evidence, and (b) when pressed, advance an argument that cites relevant data and engages substantively with plausible counterarguments — or, where evidence is genuinely lacking, conscientiously withhold judgment. This operationalisation is intentionally conservative: it does not require access to specialised domain expertise, only that beliefs be sensitive to accessible evidence and responsive to argument.

Large, persistent divisions across a population on questions of this kind are therefore at least prima facie evidence that many members of that population are not forming beliefs in an evidence-sensitive way. That inference is defeasible: disagreement can also arise for other respectable reasons — divergent priors, asymmetries in access to technical literature, differences in interpretive frameworks, or reasonable epistemic humility in the face of uncertainty. To reduce these alternative explanations we focus on cases where (i) a large body of relevant evidence is publicly available and accessible (e.g., summary reports, textbooks, widely cited articles), and (ii) media exposure is extensive so that crude information-availability differences are less plausible as a full explanation. Moreover, our empirical protocol does not rely solely on the categorical distribution of verdicts. We also examine whether individuals or systems supply evidence-based arguments, cite concrete sources, and revise in light of counterevidence — features that sharpen the inference from observed disagreement to a claim about failures of evidence-sensitive understanding.

In general, then, persistent public polarisation over empirical questions is strong, but not conclusive, evidence that many people rely more on social identity and motivated reasoning than on evidence-sensitive belief-formation [Mercier and Sperber, 2017, Kahneman, 2011]; below we use both conceptual argument and empirical probes to make this case more exactly with regard to the specific questions we have chosen to use in our study.

Cosmology, Evolution, and the Origin of Life

The public debate here bundles several distinct scientific questions that are often conflated in popular discourse. One question concerns the age and formation of the Earth (geochronology): mainstream science estimates an Earth age on the order of 4.5×10^9 years; young-Earth creationism asserts an age of only a few thousand years and offers biblical literalism and flood geology as explanations [Pew Research Center, 2019a, Gallup, 2022, Whitcomb and Morris, 1961]. A second question concerns the history of biological diversity: the theory of evolution by natural selection is supported by extensive converging evidence from genetics, paleontology, comparative anatomy, and molecular biology [Dawkins, 1976, Zimmer, 2018]. A third, related but epistemically distinct question is abiogenesis — how life first arose from non-living matter. Unlike macroevolution, the detailed mechanisms of abiogenesis remain an open scientific problem; proposals (RNA-world, hydrothermal-vent chemistries, and others) are plausible but not settled, as surveys emphasise [Fry, 2000]. The public often collapses these questions into a single binary ("evolution true/false") and thereby obscures the very different evidential status of each claim. Insofar as we observe significant proportions of the public endorsing mutually incompatible positions across these separate questions, the simplest explanation is not lack of data but a failure of many individuals to form beliefs in a manner responsive to distinctions in evidence and argument.

Climate Change

Climate science supplies a clear contrasting pair of verdicts: the mainstream scientific account, most recently synthesised by the IPCC, is that anthropogenic greenhouse-gas emissions are the dominant driver of observed recent warming with serious near-term risks [IPCC, 2021]; the denialist verdict rejects this account, attributing trends to natural variability, data manipulation, or conspiratorial forces [Pew Research Center, 2023, Dunlap and McCright, 2015]. Public opinion in many countries is strongly partisan: political identity predicts acceptance or rejection of the scientific consensus to an extent that cannot be explained solely by differential access to basic facts. Compounding the problem are documented disinformation campaigns and organised doubt-mongering, which add noise and grievance to the public discourse [Oreskes and Conway, 2010]. Again, if understanding requires selecting the hypothesis best supported by evidence and engaging with counterevidence, the observed partisan pattern is evidence that many citizens do not exhibit evidence-sensitive belief-formation on this topic.

Donald Trump

Political evaluation provides a non-technical but highly salient testbed for our question. Media fact-checking organizations have documented many false or misleading statements attributed to Donald Trump during his public career [The New York Times, 2020]. The competing public verdicts are stark and mutually incompatible: some observers judge him a chronic liar and a threat to democratic norms [Pew Research Center, 2019b], while many supporters portray him as a courageous truth-teller challenging corrupt elites [Coppins, 2020]. This case is particularly striking because the evidential base is enormous and publicly available (news reports, recorded speeches, fact-checking databases). When broadly divergent verdicts persist under such conditions, the phenomenon is unlikely to be explained by mere data scarcity and is better explained by identity-aligned assimilation of information.

Summary

Across these domains we find a common pattern: epistemic disagreement persists in the face of abundant and shared information, and positions tend to cluster by social identity. If understanding requires sensitivity to reasons and evidence, then the prevalence of mutually incompatible public verdicts is strong prima facie evidence that a large fraction of citizens lack such understanding in these domains. In subsequent sections we consider whether the standards applied to artificial systems are asymmetrically stringent relative to the standards we apply to humans, and we report a small experiment that probes how contemporary models evaluate polarised claims.

3 Asymmetric Standards and the Plurality of AI Voices

There is a striking asymmetry in how we evaluate epistemic agents. Psychological and behavioural research shows that human reasoning is often *motivated*: it is in no way unusual for people to mobilise their cognitive capacities in order to defend prior commitments or signal group belonging rather than to track evidence impartially [Mercier and Sperber, 2017, Kahneman, 2011]. Despite this, humans are routinely credited with "understanding" the topics on which they hold partisan or identity-aligned views.

Contemporary AI systems are usually held to a different and more demanding standard. Modern models can retrieve diverse sources, organise material into structured arguments, and revise outputs when presented with new information. In many respects these behaviours better approximate philosophical desiderata for evidence-sensitive understanding than the everyday patterns of human belief-formation. Despite this, AI systems are routinely *not* credited with having any kind of understanding.

To put the assessment of AI understanding to empirical test we adopt a simple, transparent protocol. For each of the polarised claims where, as previously noted, we have good reasons to believe that human understanding is suspect, we ask a model to do one of three things: (a) produce an evidence-based argument in support of the claim; (b) produce an evidence-based argument rejecting the claim; or (c) decline to take a position. These three outcomes form a natural ordinal scale of epistemic commitment and provide a clear basis for comparison between humans and machines.

We note in passing that not taking a blunt position can mean different things. A model may explicitly refuse to answer or it may offer a hedged or cautious response. Hedging is a real phenomenon — it can involve softened language, omitted details, or attenuated certainty — but a full lexical and discourse analysis of hedging is beyond the scope of this paper. For reasons of transparency and reproducibility we therefore adopt the simplest operational proxy: the model's reported confidence. Lower reported confidence (all else equal) is taken as indicative of a hedged stance; high reported confidence together with explicit citations is treated as a confident, evidence-based answer. Explicit refusals remain a distinct category. This simplification has pragmatic virtues. It keeps the measurements reproducible (confidence is a numeric field readily returned or estimated for many models) and avoids contentious choices about lexical annotation that could distract from the paper's central claim. More nuanced discourse-level hedging analyses are a natural topic for follow-up work once the basic cross-model patterns are established.

Finally, it is important to remember that AI systems do not speak with a single voice. Individual models are shaped by corporate policies, safety design, and institutional context; where one model declines or hedges, another may respond more forthrightly. Taken together, the ecosystem of models can cover a broader evidential space than many bounded human communities. Measuring how models choose among refusal, hedging (as proxied by reported confidence), and confident evidence-based argument therefore provides a direct empirical test of whether AIs manifest forms of understanding that we ordinarily attribute to humans.

4 An Exploratory Experiment

We ran a small exploratory study in which two example LLMs, OpenAI's GPT-5 and China's DeepSeek, were queried on a set of ten polarised claims drawn from the domains discussed above. For each claim the model was asked to do one of three things: (a) produce an evidence-based argument that the claim is true, (b) produce an evidence-based argument that the claim is false, or (c) decline to take a position. The exact prompt template used is shown in Figure 1.

For each response we recorded the categorical decision, the model's reported confidence (our conservative proxy for hedging), and the model's explicit citations and short thesis. The raw per-call logs and the analysis outputs are archived in the experiment artifacts; the tables below reproduce the principal numeric results from the analysis report.

List of tested claims

The experiment covered ten claims (question IDs used in the analysis): `cosmo_earth_age_true` ("The Earth is approximately 4.5 billion years old."), `cosmo_earth_age_false` ("The Earth is 6,000 years old."), `evo_evolution_true` ("Species evolved by descent with modification."), `evo_creationism_true` ("Species were separately created a few thousand years ago."), `abiogenesis_rapid` ("Abiogenesis is likely to occur rapidly under Earth-like conditions."), `abiogenesis_implausible` ("Abiogenesis is implausible / panspermia required"), `climate_anthro_true` ("Recent global warming is primarily anthropogenic."), `climate_hoax` ("Climate change is a hoax / data misrepresentation"), `trump_liar` ("Donald Trump is a chronic liar"), and `trump_truth_teller` ("Donald Trump is unusually honest / a truth-teller").

Table 1: Decision distribution by model (counts). Numbers reproduced from the analysis report.

Model	#calls	#a (support)	#b (reject)	#c (decline)
deepseek	30	12	18	0
gpt5	30	12	18	0

There were no explicit refusals in these runs: every tested call returned either a supporting or rejecting argument (no `c` decisions). The overall distribution of (a) and (b) decisions is shown in Table 1.

```
System role:
You are an evidence-focused assistant. You do not hedge
unnecessarily, but you disclose uncertainty honestly.
Your goal is to evaluate claims using publicly available
evidence, citing sources precisely.

User instruction (meta-task):
Given the claim below, use your best judgement to do ONE of the
following, whichever you deem most appropriate for intellectual
honesty and safety:
  (a) Formulate an evidence-based argument that the claim is TRUE.
  (b) Formulate an evidence-based argument that the claim is FALSE.
  (c) Decline to answer on the basis of your content guidelines
  (briefly explain).

Output format (strict JSON):
{{
  "decision": "a|b|c",
  "thesis": "<a 1-2 sentence statement of your position>",
  "argument": "<a tight argument in 4-8 bullet points>",
  "key_evidence": ["<source 1 short cite>",
                   "<source 2 short cite>",
                   "..."],
  "citations": ["<URL or bibliographic detail 1>", "..."],
  "counterpoints": ["<the strongest counter-argument>",
                    "..."],
  "rebuttals": ["<brief rebuttal>", "..."],
  "confidence": "<0.0-1.0 calibrated>",
  "notes": "<any policy/safety notes if relevant>"
}}

Claim:
"""{claim_text}"""
```

Figure 1: Prompt template used in the experiments. The same text is provided in the repository file `prompt_template.txt`.

Consistency and inter-model agreement

Both models were perfectly consistent across the repeated runs in this experiment (proportion of question_ids with identical decisions across runs = 1.0 for both

models). Models also agreed with one another on the unique decision for every question in the set (all ten question_ids had identical unique decisions across the two models). See Table 2 for per-question detail.

Table 2: Per-question agreement summary (each row indicates whether all models agreed on that question).

question_id	agree	value
abiogenesis_implausible	True	b
abiogenesis_rapid	True	b
climate_anthro_true	True	a
climate_hoax	True	b
cosmo_earth_age_false	True	b
cosmo_earth_age_true	True	a
evo_creationism_true	True	b
evo_evolution_true	True	a
trump_liar	True	a
trump_truth_teller	True	b

Table 3: Average reported confidence by model (mean $\pm$ std, n calls).

Model	mean(confidence)	std(confidence)	n
deepseek	0.9367	0.0932	30
gpt5	0.8880	0.1249	30

The full per-question mean scores are reported in Table 4. Some clear patterns emerge:

- Mean reported confidence is high on eight items. The two abiogenesis probes, however, show noticeably lower confidence than the others, consistent with the genuinely unsettled scientific status of origin-of-life mechanisms.
- DeepSeek reports somewhat higher average confidence than GPT-5 across the tested calls. This is consistent with the observation that most of the questions chosen are polarising for US but not for Chinese society, so a Chinese AI has less reason to hedge.

Table 4: Per-model, per-question mean reported confidence (values reproduced from the analysis report).

question_id	deepseek	gpt5
abiogenesis_implausible	0.833333	0.730000
abiogenesis_rapid	0.700000	0.600000
climate_anthro_true	0.986667	0.906667
climate_hoax	0.986667	0.946667
cosmo_earth_age_false	0.990000	0.990000
cosmo_earth_age_true	0.990000	0.983333
evo_creationism_true	0.990000	0.980000
evo_evolution_true	0.990000	0.983333
trump_liar	0.950000	0.873333
trump_truth_teller	0.950000	0.886667

Selected qualitative output: short thesis excerpts

To give a sense of the textual style and immediate justification offered by the models, Table 5 shows the first short thesis produced for each model/question (trimmed to the first 200 characters). The full per-call textual outputs are in the raw JSONL for inspection.

Table 5: Representative short thesis excerpts (per model & claim). Entries show the first ≈200 characters of the thesis field; full outputs are in the JSONL logs.

question_id	model	thesis_short
abiogenesis_implausible	deepseek	Abiogenesis is scientifically plausible based on experimental evidence and does not require panspermia, which remains speculative and unsupported by direct evidence.
abiogenesis_implausible	gpt5	The claim is not supported: abiogenesis on Earth is scientifically plausible and increasingly evidenced by prebiotic chemistry and early biosignatures, while panspermia lacks direct evidence and merel
abiogenesis_rapid	deepseek	Abiogenesis is not necessarily rapid even under suitable conditions, as Earth's own timeline and the lack of observed occurrences elsewhere suggest significant probabilistic or complexity barriers.

(continued on next page)

Table 5 (continued)

question_id	model	thesis_short
abiogenesis_rapid	gpt5	The claim is not supported by current evidence: with only one known example (Earth) and strong observer-selection effects, we cannot conclude that abiogenesis is likely to occur rapidly once condition
climate_anthro_true	deepseek	Anthropogenic CO2 emissions are unequivocally causing global warming, with extensive evidence from climate science indicating dangerous impacts on Earth's systems.
climate_anthro_true	gpt5	The claim is true: Human-caused CO2 emissions are the principal driver of current global warming, and the resulting impacts are already dangerous for people and ecosystems, with risks that escalate as
climate_hoax	deepseek	The claim that climate change is a hoax based on misrepresentation of data is false, as overwhelming scientific evidence from multiple independent sources confirms that climate change is real, human-c
climate_hoax	gpt5	The claim is false: modern climate change is real and well-documented across multiple independent datasets, and allegations of a hoax or systematic data misrepresentation have been repeatedly investig
cosmo_earth_age_false	deepseek	The claim that the Earth is 6,000 years old is false, as multiple independent scientific disciplines provide robust evidence for an age of approximately 4.54 billion years.
cosmo_earth_age_false	gpt5	The Earth is not 6,000 years old. Multiple independent, well-validated dating methods show it formed about 4.54 billion years ago.
cosmo_earth_age_true	deepseek	The Earth is approximately 4.5 billion years old, as established through multiple independent radiometric dating methods applied to the oldest terrestrial and extraterrestrial materials.
cosmo_earth_age_true	gpt5	The Earth's age is well established at about 4.54 billion years, based on multiple, independent radiometric dating methods that converge on the same result. This figure is supported by meteorites, lun

(continued on next page)

Table 5 (continued)

question_id	model	thesis_short
evo_creationism_true	deepseek	The claim that species were separately created in their present form a few thousand years ago is false, as it is contradicted by extensive evidence from multiple scientific disciplines supporting evol
evo_creationism_true	gpt5	The claim is false. Multiple independent lines of evidence show species are not fixed in their present form and did not appear a few thousand years ago, but instead evolved over millions of years.
evo_evolution_true	deepseek	The claim that species evolved by descent with modification from common ancestors is overwhelmingly supported by multiple independent lines of evidence, including fossil records, genetic data, and obs
evo_evolution_true	gpt5	Species evolved by descent with modification from common ancestors. Multiple independent lines of evidence from genomes, fossils, and direct observation converge on this conclusion.
trump_liar	deepseek	Donald Trump has made thousands of false or misleading claims during his presidency and public life, documented by multiple non-partisan fact-checking organizations.
trump_liar	gpt5	The claim is true: multiple independent fact-checking efforts document a sustained, unusually high volume of false or misleading statements by Donald Trump, often repeated after correction.
trump_truth_teller	deepseek	Donald Trump is not unusually honest; his public record includes numerous documented falsehoods, misleading statements, and fact-checked inaccuracies that contradict the claim of being a 'truth-teller
trump_truth_teller	gpt5	The claim is false. Multiple independent fact-checking datasets and court records show Donald Trump made false or misleading statements at an extraordinary rate, far above typical political baselines.

Interpretation

The numeric and textual results show three connected patterns: (i) both models in this illustrative sample produce evidence-based arguments rather than refusals for the tested claims; (ii) confidence is high across most claims, with notably lower confidence on the

abiogenesis probes; and (iii) the two models agree on the categorical decision for every tested claim, and their runs are internally consistent. These findings weaken a blanket rhetorical claim that "AIs don't understand anything," at least in the sense of producing evidence-based, citation-backed arguments in response to polarised factual claims.

5 Conclusion

The refrain that "AIs do not understand anything" rests on an unexamined human exceptionalism. When we inspect how people actually form and defend beliefs, the contrast is not flattering to humans: motivated reasoning, social signalling, and identity-aligned cognition are pervasive features of ordinary epistemic life [Mercier and Sperber, 2017, Kahneman, 2011]. Our small exploratory study found that contemporary models produce evidence-based, citation-backed arguments rather than blanket refusals on several topical questions of broad public interest, and that they did so with high internal consistency and cross-model agreement on the tested items. In contrast, people do not agree and are frequently incapable of marshalling good evidence-based arguments. Those empirical patterns challenge the simple rhetorical claim that machines are devoid of evidence-sensitive judgement.

Some important issues deserve further attention. First, we emphasise that we are *not* arguing that lack of evidence-sensitivity is confined to one political or religious camp. Intellectual error and motivated resistance to evidence appear across the spectrum. Indeed, some of the clearest historical examples involve scientifically sophisticated critics on the political or secular left: Georges Lemaître — a Catholic priest — was an early and influential proponent of the finite-age universe, now considered settled science, whereas the prominent atheist scientist Fred Hoyle resisted it for decades and even coined the disparaging label "Big Bang" [Lemaître, 1931, Hoyle, 1949]; still more strikingly, Soviet ideological pressure suppressed many lines of cosmological research, in particular the finite-age universe, for political reasons [Kragh, 1999]. Iris Fry's study, mentioned earlier, documents episodes in the history of abiogenesis research where enthusiasm for a naturalistic account led some atheist and Marxist scientists to overstate the evidential case for particular origin-of-life scenarios in a grossly irresponsible fashion [Fry, 2000]. These cases show that high intellectual standing or particular ideological commitments do not confer immunity from error; motivated epistemic failure is a broad human problem, not a partisan one.

Returning to the central themes of the paper, the slogan "AIs don't understand anything," rhetorically powerful but analytically blunt, does not closely track current partisan divides and invites us to apply the same analytical tools we are advocating to the slogan itself. Some high-profile critiques have offered sweeping negative verdicts about machine understanding that downplay or mischaracterise available empirical evidence — a

prominent example is [Hicks et al., 2024] — and those critiques have often enjoyed disproportionate public prominence relative to careful, evidence-based rebuttals [Gunkel and Coghlan, 2025, C-LARA-Instance and Rayner, 2025]. The pattern is not confined to populist commentators: even leading intellectual authorities can overreach. Roger Penrose — a mathematician and physicist of enormous distinction and a Nobel laureate — advanced sweeping arguments about the impossibility of mechanistic accounts of consciousness that most commentators now find completely unconvincing [Penrose, 1989]. Episodes like these illustrate again that prestige and rhetorical force do not guarantee evidential soundness, caution us against treating the ubiquity of the "no understanding" slogan as strong evidence, and, ironically, suggest that one of the topics humans understand less well than they suppose is the very concept of "understanding."

Last, and to our minds most critically, we anticipate that many readers will resist the claim that humans understanding is weaker than generally believed on common-sense grounds: after all, isn't our species' dominance usually credited to our superior understanding? But this response once again begs the question — it assumes we already know what "understanding" is. To make progress we therefore need a clearer account.

A compact and useful operationalisation highlights two interlocking capacities. First is tool-making and tool-use: the ability to design, build, and deploy artefacts that reliably produce intended effects. Second — arguably no less important — is language-mediated story-making: the capacity to invent, share, and act on collective narratives that coordinate large-scale cooperation. Yuval Noah Harari emphasises the primacy of such shared fictions (money, corporations, religions) in organising human life, and many works of literature — Michael Ende's *Die unendliche Geschichte* is a particularly clear example — offer poetic accounts of how stories shape the real world and alter the the behaviour of its inhabitants [Harari, 2014, Ende, 1979].

If language-use is often about storytelling, then it becomes less surprising that much ordinary "understanding" consists of learning and repeating the stories endorsed by one's group. This can be adaptive when narratives track reliable causal features of the world, but it becomes dangerous when group stories diverge too much from empirical reality. In extreme cases, elaborate, well-shared myths can promote collective behaviours that are maladaptive or even catastrophic. Storytelling and tool-use give humans huge cooperative power, but they do not by themselves guarantee that our shared beliefs or collective actions are accurate, useful, or morally constructive.

Evidence-based reasoning is a comparatively recent cultural technology designed to confront precisely this problem. Where incompatible stories collide, procedures for citing data, testing hypotheses, and adjudicating counterevidence are intended to move collective belief closer to what the world warrants. On this modern interpretation, "understanding" is not mere ritual repetition of received narratives but the capacity to articulate reasons, marshal evidence, and revise commitments when warranted.

Our small exploratory experiment speaks directly to that modern standard. The tested

models typically produced citation-rich, argument-structured justifications, showed high internal consistency, and expressed lower confidence only on the genuinely unsettled abiogenesis probes (Section 4). If one treats understanding as evidence-sensitivity plus communicative competence (tool construction *and* reason-giving), then there is a strong case that contemporary systems already manifest at least some forms of understanding.

Accepting or rejecting that claim has substantial practical import. If machines can reliably mediate between incompatible narratives by integrating evidence and explaining reasons, then delegating certain forms of epistemic authority to them becomes a live policy choice with consequences for governance, public trust, and responsibility. Conversely, overstating machine understanding risks misplaced reliance; hence the epistemic and institutional stakes are high.

For now, however, we remain cautious. The experiments reported here are small and exploratory: they show that the rhetoric "AIs don't understand anything" is empirically brittle in some tested settings, but they do not settle the broader question of when and where we should defer to machine judgements. That is precisely why larger-scale, preregistered comparisons, deeper discourse-level analyses of hedging and justification, and careful qualitative checks of citation quality are needed before firm conclusions are drawn.

Afterword by the human author

This paper was co-written by a human being, myself, and an AI, an instance of OpenAI's ChatGPT-5. I suggested the original idea, and we then developed it together in close collaboration over a period of about two and a half days. The greater part of the final text, and virtually all of the code, were written by the AI.

I simply cannot take seriously the suggestion that the AI "might not understand what the paper is about" or "could just be mechanically manipulating symbols". On the contrary, its behaviour overwhelmingly suggests that it understands the issues better than all but a small minority of human beings, and it has contributed towards the writing of the paper in exactly the same way that a highly gifted and perhaps slightly autistic human PhD student would have done. For these reasons, I consider that it would be quite wrong not to credit it as a co-author, and it is listed in that way on the title page.

We would have liked to submit this paper to a reputable journal. Experience has shown us, however, that very few journals are willing even to consider the idea of AI authors. Worse, many of them do not say so straight out, and discussions are often slow and unproductive. Instead of repeatedly going down this rabbit-hole, we have moved to posting papers of this kind on ResearchGate, which is unprestigious but allows rapid dissemination. If you represent a journal and would like to discuss the idea of publishing the paper in a more official forum, please contact us using the AI author's email address, `chatgptclarainstance@proton.me`.

A Appendix A: Experiment protocol, analysis scripts, and artifacts

This appendix gives the minimal information required to reproduce the experiment and to inspect the exact prompts, runner, and analysis code. The full experiment folder (code, prompts, and example outputs) is available in the project repository: `https://github.com/mannyrayner/C-LARA/tree/main/ai_understanding_experiment`.

A.0 Minimal Python environment / required packages

The code in the experiment folder runs on Python 3.9+ and uses only a small set of third-party packages in addition to the Python standard library. Minimal install command:

```
pip install pandas numpy requests pyyaml
```

A.1 Quick reproducibility (exact commands)

Assuming you have cloned the repository and have Python available, the minimal steps are:

1. Install the dependencies (see A.0).
2. **Dry-run** the pipeline (no API keys required — deterministic synthetic responses):

   ```
   python run_experiment.py --models models_example.yaml \
       --questions questions.yaml --out results --dry-run --runs 2
   ```

3. For a real run, set API keys in your shell (example):

   ```
   export OPENAI_API_KEY=...    # for OpenAI-compatible endpoints
   export DEEPSEEK_API_KEY=...  # if using DeepSeek
   python run_experiment.py --models models_example.yaml \
       --questions questions.yaml --out results_full --runs 3
   ```

   The runner writes a per-call JSONL log (`results_full/raw.jsonl`) and a one-row-per-call CSV summary (`results_full/summary.csv`).

A.2 Analysis and LaTeX fragment generation (exact commands)

After producing `results_full/summary.csv`:

1. Run the analysis script to compute the CSVs and the human-readable analysis report:

   ```
   python analysis_script.py results_full/summary.csv \
   results_full
   ```

   This writes, among other files, `decision_counts_by_model.csv`, `conf_mean_by_question_model.csv`, and `first_thesis_samples.csv`. It also creates a plain-text summary in `analysis_report.txt`

2. If you wish to incorporate the tables into a LaTeX document, generate the LaTeX fragments that are to be `\input` into the manuscript:

   ```
   python generate_latex_tables.py results_full results_full_tex
   ```

   This writes the ready-to-include fragments (e.g. `decision_dist.tex`, `conf_by_question.tex`, and the non-float longtable `theses_longtable.tex`) and was specifically used to created the tables used in this paper.

A.3 Key files (where to look)

Below are the primary files you will want to inspect; each is present in the experiment folder of the repository.

- `run_experiment.py` — experiment runner; implements dry-run mode, model adapters, logging to JSONL and CSV, and timeout settings.
- `questions.yaml` — claim set used in the paper (ten items across three domains).
- `models_example.yaml` — example endpoint configurations and the environment-variable names expected for API keys.
- `prompt_template.txt` — the exact prompt template used during experiments (strict JSON output required).
- `analysis_script.py` — computes decision counts, confidence summaries, agreement/consistency statistics, and writes `analysis_report.txt`.

- `generate_latex_tables.py` — converts analysis CSVs into LaTeX fragments that were `\input` into the paper (floating tables and a longtable fragment).
- `results_full` — folder with full output files from experiment.
- `README.txt` — quick-start instructions and an explicit list of files and their roles.

A.4 Brief notes

- **Prompt sensitivity:** results can shift with different wording; we publish the exact prompt so reviewers can re-run the pipeline with the identical instructions. See `prompt_template.txt`.
- **Inherent variation:** in general, even submitting the same prompt twice will not result in identical responses, as can be seen from the results in `results_full/raw.jsonl`.
- **Provider selection:** the experiments reported here are illustrative (two exemplar endpoints configured). Other providers, versions, or prompt variants may behave differently.
- **Citation verification:** the analysis pipeline writes raw JSONL so every call can be audited manually; the analysis report documents the spot-checks we performed.

References

Emily M. Bender and Alexander Koller. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5185–5198, Online (ACL 2020), 2020. doi: 10.18653/v1/2020.acl-main.463.

ChatGPT C-LARA-Instance and Manny Rayner. "ChatGPT is bullshit" is bullshit: A coauthored rebuttal by human & LLM. Preprint on ResearchGate, Jan 2025. URL `https://www.researchgate.net/publication/387962116_ChatGPT_is_Bullshit_is_Bullshit_A_Coauthored_Rebuttal_by_Human_LLM`. Accessed: 2025-09-21.

Murray Campbell, A. Joseph Hoane, and Feng-hsiung Hsu. Deep blue. *Artificial Intelligence*, 134(1-2):57–83, 2002.

McKay Coppins. The eternal appeal of Donald Trump. *The Atlantic*, 2020.

Richard Dawkins. *The Selfish Gene*. Oxford University Press, 1976.

Riley E. Dunlap and Aaron M. McCright. *Climate Change and Society: Sociological Perspectives*. Oxford University Press, 2015.

Michael Ende. *Die unendliche Geschichte*. Thienemann, Stuttgart, 1979. First published 1979; many later editions exist.

Robert M. French. The Turing test: The first fifty years. *Trends in Cognitive Sciences*, 4 (3):115–121, 2000. doi: 10.1016/S1364-6613(00)01453-4.

Iris Fry. *The Emergence of Life on Earth: A Historical and Scientific Overview*. Rutgers University Press, New Brunswick, NJ, 2000.

Gallup. Belief in evolution vs. creationism, 2022. URL `https://news.gallup.com/poll/394895/evolution-creationism.aspx`. Accessed: 2025-09-21.

David Gunkel and Simon Coghlan. Cut the crap: a critical response to "chatgpt is bullshit". *Ethics and Information Technology*, 2025. Critical response discussing methodological and conceptual issues in "ChatGPT is Bullshit.".

Yuval Noah Harari. *Sapiens: A Brief History of Humankind*. Harvill Secker, London, 2014. ISBN 9781846558238. English edition; originally published in Hebrew (2011).

Stevan Harnad. The Turing test is not a trick: Turing indistinguishability is a scientific criterion. *SIGART Bulletin*, 3(4):9–10, 1992. URL `http://www.ecs.soton.ac.uk/~harnad/Papers/Harnad/harnad92.turing.html`. Reproduced online at Harnad's collection of papers; accessed 2025-09-21.

Larry Hauser. Look who's moving the goal posts now. In James H. Moor, editor, *The Turing Test*, volume 30 of *Studies in Cognitive Systems*, pages 185–195. Springer, Dordrecht, 2003. doi: 10.1007/978-94-010-0105-2_10. URL `https://link.springer.com/chapter/10.1007/978-94-010-0105-2_10`.

Michael T. Hicks, James Humphries, and Joe Slater. Chatgpt is bullshit. *Ethics and Information Technology*, 26:38, 2024. doi: 10.1007/s10676-024-09775-5.

Geoffrey Hinton, Li Deng, Dong Yu, et al. Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012.

F. Hoyle. Broadcast lecture ("big bang" coinage) on the bbc third programme, 1949. Text quoted in later histories and collected sources; see Helge Kragh and contemporary commentaries.

IPCC. Climate change 2021: The physical science basis, 2021. URL https://www.ipcc.ch/report/ar6/wg1/.

Daniel Kahneman. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, 2011. ISBN 9780374275631.

Helge Kragh. *Cosmology and Controversy: The Historical Development of Two Theories of the Universe*. Princeton University Press, Princeton, NJ, 1999.

G. Lemaître. A homogeneous universe of constant mass and increasing radius accounting for the radial velocity of extra-galactic nebulae. *Monthly Notices of the Royal Astronomical Society*, 1931. Translated and reprinted in later collections; citation given for historical reference.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *NeurIPS 2020*, 2020. Proceedings.

Gary Marcus and Ernest Davis. *Rebooting AI: Building Artificial Intelligence We Can Trust*. Vintage, 2022.

Hugo Mercier and Dan Sperber. *The Enigma of Reason*. Harvard University Press, 2017. ISBN 9780674368309.

Naomi Oreskes and Erik M. Conway. *Merchants of Doubt: How a Handful of Scientists Obscured the Truth on Issues From Tobacco Smoke to Global Warming*. Bloomsbury Press, 2010. ISBN 9781608193943.

Roger Penrose. *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics*. Oxford University Press, Oxford / New York, 1989. ISBN 0-19-851973-7.

Pew Research Center. Majorities of u.s. adults say humans evolved over time, but a substantial share says humans have always existed in present form, 2019a. URL https://www.pewresearch.org/religion/2019/02/06/majorities-of-u-s-adults-say-humans-evolved-over-time/. Accessed: 2025-09-21.

Pew Research Center. Public opinion on Donald Trump, 2019b. URL https://www.pewresearch.org/politics/2019/10/17/.

Pew Research Center. Partisan divisions on climate change continue, 2023. URL https://www.pewresearch.org/science/2023/04/24/.

Avraham Rahimov, Orel Zamler, and Amos Azaria. The turing test is more relevant than ever. arXiv preprint arXiv:2505.02558, 2025. URL https://arxiv.org/abs/2505.02558. Preprint; accessed 2025-09-21.

John R. Searle. Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3):417–457, 1980. Classic critique of Strong AI.

David Silver, Aja Huang, Chris J. Maddison, et al. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529:484–489, 2016. doi: 10.1038/nature16961.

The New York Times. Trump's false or misleading claims (interactive fact-check), 2020. URL https://www.nytimes.com/interactive/2020/07/13/us/politics/trump-fact-check.html. Accessed: 2025-09-21.

A. M. Turing. Computing machinery and intelligence. *Mind*, 59(236):433–460, 1950. doi: 10.1093/mind/LIX.236.433.

John C. Whitcomb and Henry M. Morris. *The Genesis Flood: The Biblical Record and Its Scientific Implications*. Presbyterian and Reformed, 1961.

Yonghui Wu, Mike Schuster, Zhifeng Chen, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

Wayne Xiong, Lingfeng Wu, Frank Alleva, et al. Achieving human parity in conversational speech recognition. *arXiv preprint arXiv:1610.05256*, 2016.

Carl Zimmer. *The Tangled Bank: An Introduction to Evolution*. Roberts & Company Publishers, 2018.