Academic philosophical paper

Logical Coherence Without Truth A Philosophical Inquiry into Language Models and the Illusion of Reasoning

Momen Ghazouani Chief Scientist Setaleur Aplamda

April 11, 2026

Abstract

The widespread assumption that logical coherence implies truth is increasingly challenged in the context of contemporary artificial intelligence systems. This paper examines the philosophical claim that what is logically consistent is not necessarily true, and investigates its implications for the behavior and evaluation of Large Language Models (LLMs). Unlike traditional reasoning systems grounded in formal logic or empirical verification, LLMs generate outputs based on probabilistic pattern recognition, optimizing for linguistic coherence rather than factual accuracy. As a result, these models can produce arguments that are internally consistent and highly persuasive, yet fundamentally detached from reality. This work argues that LLMs do not fail at truth-seeking; rather, they are not inherently designed for it. Instead, they simulate reasoning by reproducing patterns of logical structure present in their training data, creating an "illusion of reasoning" that can obscure the distinction between valid argumentation and true claims. The paper further explores how this distinction affects the evaluation of knowledge, particularly in contexts where coherence, clarity, and rhetorical strength are mistakenly treated as indicators of correctness. By analyzing the epistemic limitations of coherence-based systems, this paper highlights a critical gap between logical form and factual grounding in AI-generated content. It concludes by proposing a conceptual framework for separating coherence from truth in the design and assessment of intelligent systems, emphasizing the need for hybrid approaches that integrate logical consistency with mechanisms of external validation.

Keywords: logical coherence, truth, language models, epistemology, artificial intelligence, reasoning, validity, empirical verification

---

Conceptual diagram illustrating the separation between logical validity and empirical truth, highlighting the position of language model outputs as typically coherent but not necessarily true (valid yet unsound)

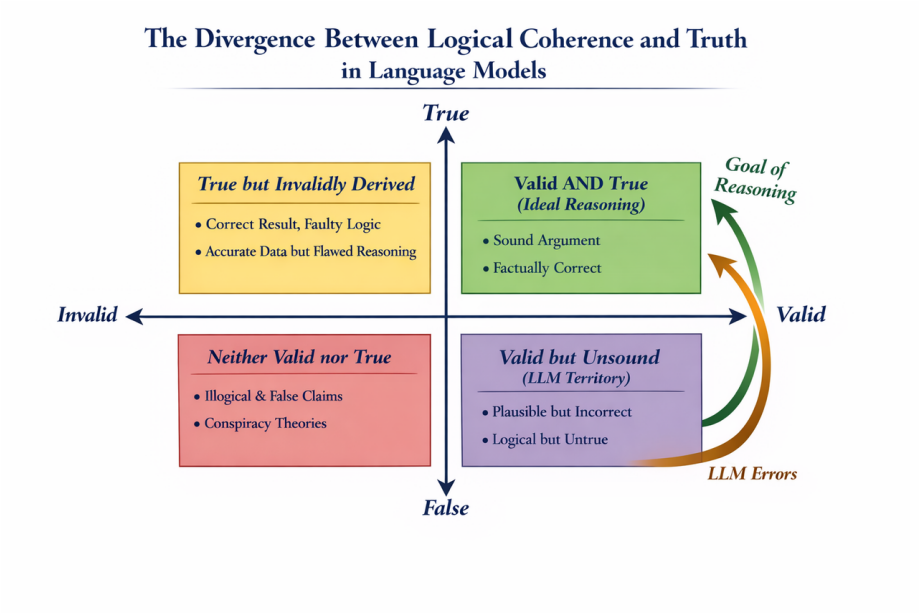

Figure 1: Figure 1

Introduction

The emergence of Large Language Models has precipitated a fundamental epistemological crisis: the production of content that is simultaneously logically coherent and factually unreliable. This phenomenon raises a question that has occupied philosophers since antiquity but now manifests in computational form can logical consistency exist independently of truth? More provocatively, does the conflation of these two concepts represent not a failure of artificial intelligence, but a failure of human understanding about the nature of reasoning itself ? Contemporary discourse surrounding AI capabilities often assumes that systems producing well-structured, internally consistent arguments are engaging in genuine reasoning and, by extension, approaching truth. This assumption rests on a category error: the conflation of formal validity with material correctness, of syntactic coherence with semantic accuracy. Language models, optimized through training on vast corpora to predict token sequences with high probability, excel at generating outputs that *appear* reasoned without necessarily *being* true. They are, in effect, sophisticated pattern-matching systems that reproduce the surface structure of logical discourse without accessing the underlying reality that such discourse purports to represent.

This paper contends that the problem is not merely technical but fundamen-

tally philosophical. We have inherited from rationalist traditions an implicit faith that logical coherence is a reliable guide to truth that valid reasoning from sound premises necessarily yields knowledge. Yet language models demonstrate that logical form can be divorced from factual content, that coherence can be mechanically produced through statistical regularities rather than genuine understanding, and that persuasiveness can be achieved through structural mimicry rather than epistemic grounding. The central thesis of this inquiry is twofold: first, that LLMs are not defective truth-seeking systems but rather non-truth-seeking coherence generators; second, that our tendency to interpret their outputs as epistemically reliable reveals deeper assumptions about the relationship between logic and truth that warrant critical examination. The illusion of reasoning produced by these systems is not merely a technological artifact but a mirror reflecting our own conceptual confusions about what constitutes genuine knowledge.

The Philosophical Distinction : Logical Validity and Empirical Truth

The Two Modes of Correctness

Classical epistemology distinguishes between two fundamentally different modes of correctness. Logical validity concerns the formal structure of arguments whether conclusions follow necessarily from premises according to established rules of inference. Empirical truth, by contrast, concerns the correspondence between propositions and states of affairs in the world. An argument can be valid without being sound (valid but with false premises), and a claim can be true without being derived through valid reasoning (true but arrived at accidentally or through fallacious inference).

This distinction is not merely academic but operationally significant. Consider the syllogism: "All unicorns are immortal; Socrates is a unicorn; therefore, Socrates is immortal." This argument is formally valid the conclusion follows necessarily from the premises. Yet it is manifestly unsound because both premises are false. The logical machinery operates flawlessly while producing a conclusion detached from reality. The structure is intact; the content is fiction. Language models operate primarily in the domain of formal coherence. They generate outputs that preserve the syntactic and rhetorical patterns associated with valid reasoning premise-conclusion structures, causal connectives, evidential markers, logical transitions without necessarily ensuring that these patterns correspond to actual states of affairs. They are, in essence, validity engines operating without a built-in truth mechanism.

The Question of External Reference

The relationship between logical systems and truth depends critically on external reference the grounding of propositions in something beyond the formal system itself. In traditional epistemology, this grounding occurs through empirical

observation, testimony, or correspondence with reality. Formal logic provides the scaffolding for reasoning, but truth-claims require anchorage in the external world. LLMs lack this anchorage in a meaningful sense. While trained on data derived from reality, they do not interact with reality during inference. They operate on compressed statistical representations of linguistic patterns without direct access to the referents of the terms they manipulate. When a language model generates the statement "Water boils at 100°C at sea level," it is not retrieving a measured fact but reproducing a high-probability token sequence associated with specific contextual cues. The statement may be true, but its truth is incidental to the mechanism that produced it.

This raises a profound question: Can we meaningfully attribute truth-aptness to systems that lack referential grounding ? Or are such systems better understood as generators of well-formed expressions that may or may not correspond to reality, depending on the statistical properties of their training data and the specific prompts they receive?

The Insufficiency of Coherence

Coherence theories of truth, which hold that truth consists in the mutual consistency of beliefs within a system, might seem to vindicate language model outputs. If coherence is sufficient for truth, then LLMs, which excel at producing internally consistent narratives, should be reliable truth-generators. However, this conclusion fails for several reasons.

- **First,** coherence theories typically require more than mere logical consistency; they demand systematic integration with background knowledge, responsiveness to evidence, and explanatory power. LLMs achieve surface coherence without necessarily meeting these deeper requirements. They can generate multiple mutually contradictory but internally consistent narratives, suggesting that their coherence is local and contextual rather than global and systematic.
- **Second,** even robust coherence theories acknowledge the possibility of coherent falsehoods entire systems of mutually supporting beliefs that are nonetheless detached from reality. Science fiction worldbuilding demonstrates this possibility: a richly detailed, internally consistent fictional universe that is deliberately non-actual. LLMs can generate such coherent fictions effortlessly, and without external verification mechanisms, users cannot reliably distinguish them from accurate representations.

The insufficiency of coherence as a truth criterion becomes especially acute when we recognize that persuasiveness and coherence are statistically correlated in natural language. Texts that are more coherent tend to be more convincing, regardless of their truth value. LLMs, optimized to produce high-probability (and therefore typically coherent) outputs, thus systematically generate content that is rhetorically effective without being epistemically reliable.

The Epistemic Limits of Logic as a Knowledge System

Logic as a Truth-Preserving, Not Truth-Generating System

A fundamental insight from the philosophy of logic is that deductive reasoning is truth-preserving rather than truth-generating. Valid deduction transmits truth from premises to conclusions but cannot create new factual knowledge beyond what is already implicit in the premises. If the premises are false, deduction reliably propagates those falsehoods to the conclusion. Logic is an amplifier, not a filter it magnifies whatever is fed into it, whether truth or error. This characteristic has profound implications for systems that simulate logical operations. LLMs, which reproduce the formal patterns of logical inference without verifying the truth of their premises, effectively operate as truth-agnostic reasoning simulators. They can chain together implications with flawless formal structure while building castles on foundations of sand. The resulting outputs exhibit all the hallmarks of rigorous thinking careful qualification, systematic development, evidential citation without the epistemic reliability that such markers are meant to signal.

Consider how an LLM might respond to the prompt: "Explain why the Earth is flat." A sufficiently capable model could generate a lengthy, internally coherent argument citing (fabricated or misrepresented) evidence, addressing counterarguments, and maintaining logical consistency throughout. The output would exhibit reasoning's formal features while being fundamentally false. The model's capacity to maintain coherence in service of falsehood demonstrates that logical structure alone provides no guarantee of truth.

The Problem of Premise Selection

Every logical argument begins with premises, and the selection of these premises is not itself a logical operation. It involves judgment, evidence-gathering, background knowledge, and often value commitments. In human reasoning, premise selection is constrained by interaction with reality, social verification, and accumulated experience. We test our premises against the world and revise them when they fail to predict or explain phenomena adequately. LLMs select premises (or rather, generate premise-like statements) based on statistical patterns in their training data. A premise appears not because it has been verified but because similar premises frequently appear in high-quality texts in the training corpus. This creates a systematic bias toward premises that are *typically stated* rather than necessarily *true*. Common misconceptions, popular but incorrect beliefs, and widely repeated errors can all serve as premises if they are statistically well-represented in the training data.

Furthermore, LLMs lack the capacity for genuine premise revision in light of counterevidence. While they can generate text acknowledging errors or incorporating corrections, they do not update their underlying representations based on feedback within a conversation (except through fine-tuning). They cannot learn from mistakes in the way that human reasoners ideally do. Each generation is

independent, drawing on fixed statistical patterns rather than dynamically updated beliefs.

Can Logic Operate Without Empirical Input ?

Pure mathematics and formal logic demonstrate that abstract reasoning can operate without empirical input, generating truths that are necessary, a priori, and independent of contingent facts about the world. However, the scope of such purely formal systems is limited. They tell us nothing about the actual world only about the relationships between abstract entities defined by stipulation. To apply logic to empirical questions, we require empirical premises, and obtaining reliable empirical premises requires interaction with reality.

LLMs occupy a peculiar middle ground. They are not pure formal systems like automated theorem provers, nor are they empirically grounded systems like sensor-equipped robots. They are statistical models trained on linguistic data that *describes* empirical reality without directly accessing it. This creates a distinctive epistemic situation: they have indirect, mediated contact with factual information through their training data, but they lack the mechanisms for verifying that this information is accurate or updating it when it becomes outdated. The result is a system capable of producing elaborate logical structures on empirical topics medical diagnoses, historical analyses, scientific explanations without the ability to verify the factual accuracy of its base claims. It can reason about the world fluently without being reliably *of* the world. This is logic operating with simulated rather than verified empirical input, producing conclusions that have the form of knowledge without its epistemic warrant.

The Nature of "Logic" in Large Language Models

Pattern Recognition as Reasoning Simulation

Large Language Models do not perform reasoning in the traditional sense. They engage in next-token prediction generating the most probable continuation of a text sequence based on statistical patterns learned from training data. When an LLM appears to reason logically, it is reproducing the linguistic patterns associated with logical discourse rather than executing logical operations over semantic representations. This distinction is crucial. A traditional logical system manipulates symbols according to syntactic rules in a way that preserves semantic relationships. The system "knows" (in a formal sense) that from "All A are B" and "C is A," one can derive "C is B." An LLM, by contrast, has learned that texts containing "All A are B" and "C is A" are frequently followed by constructions like "therefore, C is B" or "thus, C is B" or "it follows that C is B." It reproduces this pattern not because it has encoded a modus ponens rule but because the statistical regularities of its training data make such sequences high-probability.

The practical effect can be identical valid logical inferences produced reliably in both cases. But the underlying mechanism differs fundamentally. One is

rule-governed symbolic manipulation; the other is pattern-matching and probabilistic generation. This difference becomes apparent when edge cases or novel logical structures are encountered. Rule-based systems generalize perfectly to any instance of their encoded rules. Pattern-based systems approximate logical behavior when inputs resemble training data but may fail unpredictably when encountering unfamiliar logical configurations.

Probabilistic Coherence vs. Necessary Inference

Logical necessity is categorical: in classical logic, if the premises are true and the reasoning is valid, the conclusion *must* be true. There are no degrees of necessity, no probabilistic qualification. Deductive inference is all-or-nothing.

LLM outputs, by contrast, are inherently probabilistic. Every token is selected from a probability distribution over possible continuations. An LLM might generate a valid logical inference not because the conclusion necessarily follows but because, given the context, that particular sequence of tokens has high probability. In most cases, this produces apparently correct reasoning because training data contains many examples of valid inference and relatively few examples of explicit logical errors. However, this probabilistic basis means that logical coherence in LLM outputs is a statistical tendency rather than a structural guarantee. The model might generate an invalid inference if it resembles a common pattern, or it might fail to generate a valid inference if it is statistically unusual. The coherence we observe is an emergent property of training on largely coherent texts, not a fundamental feature of the model's architecture.

This has important implications for reliability. A logical system's validity is demonstrable and guaranteed by its construction. An LLM's apparent logical competence is a performance characteristic that must be empirically assessed and that may vary across domains, prompts, and model states. We cannot prove that an LLM will reason validly; we can only observe that it typically does so with some measurable frequency.

The Absence of Understanding and Intentionality

Traditional reasoning involves understanding a grasp of what premises and conclusions mean, what they assert about the world, and why one statement follows from another. This understanding is intentional: it is directed toward objects and states of affairs beyond the reasoning process itself. LLMs lack understanding and intentionality in this robust sense. They process tokens as abstract patterns without semantic comprehension. When an LLM generates "Socrates is mortal" from "All humans are mortal" and "Socrates is human," it does not grasp the concept of mortality, the historical figure of Socrates, or the relationship between human nature and death. It manipulates symbols (tokens) according to learned statistical associations without accessing their referents.

This absence of understanding has subtle effects on the quality of LLM reasoning. The model cannot distinguish between conclusions that matter and those that

are trivial. It cannot recognize when a logical inference leads to an absurdity that should prompt premise revision. It cannot detect inconsistencies that span long contexts or require background knowledge not explicitly stated. Its "reasoning" is surface-level pattern completion rather than deep semantic processing. Moreover, without intentionality, the model cannot genuinely *claim* anything. Its outputs are generated sequences, not assertions. It does not believe its conclusions, commit to their truth, or take responsibility for their implications. This raises questions about whether we should even use epistemic vocabulary knowledge, belief, truth to describe LLM outputs, or whether we need a distinct category for non-intentional, statistically generated, semantically hollow but syntactically well-formed linguistic productions.

The Evaluation Problem: Form Over Content

The Bias Toward Well-Structured Arguments

When humans evaluate arguments, we are influenced by their form as well as their content. Arguments that are clearly structured, logically organized, and rhetorically polished tend to receive higher assessments, even when their substantive claims are questionable. This bias is amplified when evaluators lack domain expertise or when the subject matter is complex and uncertain. LLMs inherit and exploit this bias. Trained on corpora that include well-written texts academic papers, professional documents, carefully edited prose they learn to reproduce the formal features associated with high-quality argumentation. They generate outputs with clear topic sentences, smooth transition , appropriate hedging, and logical flow. These formal features signal credibility and competence, leading evaluators to infer that the content is correspondingly reliable.

This creates a systematic evaluation problem. When we ask an LLM to assess a piece of work, it is likely to assign higher value to texts that exhibit formal excellence regardless of their factual accuracy or substantive depth. A beautifully structured essay making false claims may receive a better evaluation than a clumsily written but accurate analysis. The model conflates rhetorical effectiveness with epistemic merit because its training data correlates these features (well-written texts are more likely to be factually accurate on average, though not invariably).

Persuasiveness as a Misleading Proxy for Correctness

Persuasiveness and truth are distinct. A persuasive argument can be false; an unpersuasive argument can be true. Yet in practice, we often use persuasiveness as a heuristic for truth, especially in domains where we lack the expertise to directly assess claims. LLMs, trained to maximize the likelihood of generating text similar to their training data, effectively optimize for persuasiveness rather than truth. This optimization occurs because persuasive texts those that convince human readers are likely to be reproduced, shared, and included in training corpora. Unpersuasive truths and awkwardly expressed insights may

be underrepresented. The result is a model that excels at generating convincing arguments without necessarily ensuring their correctness.

The problem intensifies when LLMs are used to evaluate other texts. An LLM asked to assess the quality of an argument will tend to rate more persuasive arguments higher, potentially overlooking factual errors if they are wrapped in compelling rhetoric. This creates a feedback loop: LLMs generate persuasive but potentially false content, which is then rated highly by LLM evaluators, reinforcing the production of similar content.

The Difficulty of Detecting Sophisticated Falsehoods

Simple factual errors are often easy to detect claims that directly contradict common knowledge or obvious observations. But sophisticated falsehoods are more insidious. These are claims that are plausible, internally consistent with stated premises, and surrounded by accurate context. They are falsehoods embedded in a matrix of truth, making them difficult to isolate without domain expertise and careful verification.

LLMs are particularly adept at generating sophisticated falsehoods. They can create entire narratives that are 95% accurate with 5% fabrication, where the fabricated elements are seamlessly integrated and support the overall coherence of the text. An LLM might accurately describe a historical period, correctly cite several events, and then insert a fictional event that fits the pattern and serves the narrative's needs. Without external verification, such errors can be nearly impossible to detect. When LLMs evaluate content, they face the same difficulty. They cannot reliably distinguish between sophisticated falsehoods and accurate claims unless the falsehood contradicts information well-represented in their training data. Novel false claims, plausible but incorrect interpretations, and subtle distortions are likely to pass undetected. The model's evaluation reflects its training data's biases and limitations rather than an independent assessment of truth.

Depth vs. Surface Sophistication

Genuine intellectual depth involves more than formal correctness. It requires original insight, recognition of nuance and complexity, awareness of limitations, and engagement with difficult questions. Surface sophistication, by contrast, involves the stylistic markers of depth complexity of language, citation of sources, acknowledgment of counterarguments without substantive originality or insight.

LLMs excel at surface sophistication. They can generate text that sounds profound, cites (sometimes fabricated) authorities, acknowledges complexity, and hedges appropriately. But this is achieved through pattern reproduction rather than genuine insight. The model has learned that certain linguistic features philosophical terminology, conditional statements, references to scholarly debates are associated with high-quality intellectual work. It reproduces these features without the underlying conceptual work they typically signal. When

asked to evaluate intellectual work, LLMs are likely to confuse surface sophistication with genuine depth. A text that deploys the right terminology, cites sources (even if inappropriately), and uses complex sentence structures may be rated more highly than a simpler but more insightful piece. This bias reflects the training data's limitations: truly original insights are rare and may not be well-represented, while the linguistic patterns associated with scholarly work are common and easily learned.

The Decoupling of Logic and Truth in AI Systems

Hallucination as Structural Feature, Not Malfunction

The phenomenon of "hallucination" in LLMs the generation of plausible but false information is often framed as a failure mode, a bug to be fixed through better training or architectural improvements. This framing is misleading. Hallucination is not incidental to how LLMs work; it is a structural feature of systems that optimize for coherence without mechanisms for truth-verification. LLMs generate hallucinations precisely when they are functioning as designed : producing high-probability continuations of input sequences. A hallucination occurs when the most probable continuation (based on training data patterns) does not correspond to reality. This can happen because the training data contains errors, because the specific factual claim is underrepresented in the training data, or because the model generalizes patterns in ways that create plausible but false outputs.

Crucially, hallucinated content is often more coherent than accurate content. Reality is frequently messy, irregular, and surprising. Coherent fictions are often neater than truth. When faced with a prompt that would require an unusual or complex true answer, an LLM might instead generate a simpler, more typical false answer because it better fits learned patterns. The model is rewarded (by its optimization objective) for coherence, not for accuracy. This suggests that the problem of hallucination cannot be fully solved through training improvements alone. As long as LLMs optimize for statistical likelihood rather than truth, they will systematically tend toward coherent falsehoods when these are more probable than accurate but complex truths. Reducing hallucination requires not just better training data but fundamental architectural changes that incorporate external verification and truth-grounding mechanisms.

Rewarding Coherence, Ignoring Accuracy

The training objective of LLMs typically variants of likelihood maximization implicitly rewards coherence and penalizes incoherence. A model that generates grammatically correct, contextually appropriate, logically structured text receives higher scores (lower loss) than one producing garbled or contradictory output. This is by design: coherence is a measurable, optimizable property that can be assessed without external reference. Accuracy, by contrast, is not directly optimized during standard language model training. The training process does

not verify whether generated facts correspond to reality; it only checks whether they match the training data. If the training data contains errors, the model learns to reproduce those errors. If certain facts are absent from the training data, the model cannot learn them regardless of their truth value.

This creates a structural bias toward coherence over accuracy. When the two conflict when an accurate statement would disrupt narrative flow or when a false claim better fits established patterns the model's optimization pressure favors coherence. The result is outputs that read beautifully, flow logically, and maintain internal consistency while potentially being factually wrong. This bias is not a moral failing or design flaw in the conventional sense. It reflects the fundamental challenge of training models on form (linguistic patterns) to perform tasks that require content (factual knowledge). We have built systems that are experts in the surface structure of knowledge without reliable access to its substance.

Rhetoric Over Reality

Classical rhetoric distinguished between form and content, style and substance. Effective rhetoric deploys formal techniques parallelism, metaphor, emotional appeal, logical structure to make content persuasive. In ideal cases, rhetorical skill serves true and valuable content. But rhetoric can also be deployed to make false or trivial claims appear important and credible. LLMs are, in a sense, pure rhetoric engines. They have mastered the formal techniques that make text persuasive without necessarily possessing the substantive knowledge that should underlie persuasive claims. They can generate eloquent defenses of false positions, compelling narratives built on fabricated events, and authoritative-sounding explanations of processes they do not actually model.

This capability reveals something important about the relationship between linguistic form and epistemic content. We have assumed, perhaps unconsciously, that certain formal features reliably signal substantive merit that well-structured arguments are more likely to be true, that careful qualification indicates genuine uncertainty, that citation of sources demonstrates grounding in evidence. LLMs demonstrate that these formal features can be mechanically reproduced without their usual epistemic correlates. They break the assumed connection between how something is said and whether it is true. The implications extend beyond AI systems. If rhetorical effectiveness can be algorithmically generated, we must reconsider how much weight to give to formal features when evaluating claims. We can no longer assume that a beautifully written, logically structured argument is therefore likely to be correct. The traditional heuristics we use to assess credibility fluency, coherence, apparent expertise are now unreliable in contexts where AI-generated content is prevalent.

Ethical and Epistemological Implications

The Erosion of Epistemic Trust

Epistemic trust our willingness to accept claims from others as reliable sources of knowledge depends on assumptions about the relationship between assertion and belief, between saying and meaning. When a human makes a claim, we typically assume they believe it, that this belief is based on some evidence or reasoning, and that they are accountable for its truth. These assumptions justify treating testimony as a source of knowledge. LLMs complicate this epistemic economy. They generate claims without believing them, without basing them on evidence in the traditional sense, and without being accountable for their truth. Yet their outputs are superficially indistinguishable from human testimony. This creates a fundamental uncertainty: when encountering a claim, we can no longer confidently assume it reflects belief grounded in evidence and accountability.

The proliferation of LLM-generated content thus threatens to erode epistemic trust more broadly. If we cannot reliably distinguish between human and AI-generated claims, and if AI-generated claims are systematically less reliable despite being superficially similar, we may rationally decrease our trust in claims generally. This could lead to a kind of epistemic skepticism where all assertions are treated with suspicion, even those from reliable human sources. This erosion of trust has practical consequences. Science depends on trusting peer-reviewed publications. Democracy depends on trusting news sources and expert analysis. Education depends on trusting textbooks and teachers. If LLM-generated content infiltrates these systems without reliable detection mechanisms, and if such content is systematically less reliable despite being coherent and persuasive, the foundations of knowledge-sharing institutions are undermined.

The Illusion of Understanding

Interacting with LLMs can create what might be called "the illusion of understanding" a false sense of comprehension that comes from receiving coherent, well-structured explanations without engaging in genuine sense-making. When a human teacher explains a complex concept, the explanation is ideally tailored to the student's current understanding, responsive to confusion, and aimed at fostering genuine comprehension. The student must actively work to integrate new information with existing knowledge. LLM explanations, by contrast, can be passively consumed. They are typically clear, well-organized, and require minimal cognitive effort to process. But this ease of consumption may come at the cost of genuine understanding. The student receives information that *sounds* comprehensible without necessarily achieving the deeper integration that constitutes real learning. The coherence of the explanation creates an illusion that the concept has been grasped when in fact only its verbal formulation has been memorized.

This problem is exacerbated by the LLM's inability to assess understanding. A human teacher can detect confusion through questions, body language, and

failed attempts at application. An LLM cannot meaningfully evaluate whether its explanation has been understood; it can only generate follow-up text that maintains coherence with the conversation. This creates the possibility of extended exchanges where both parties maintain the appearance of meaningful communication without substantive understanding actually occurring. The educational implications are significant. If students increasingly rely on LLM-generated explanations, they may develop fluency with terminology and surface-level concepts without the deeper understanding necessary for application, synthesis, and critical evaluation. They may mistake the ability to reproduce coherent explanations for genuine comprehension, and teachers may struggle to detect this gap if assessments emphasize verbal fluency over demonstrated understanding.

Responsibility and Accountability Gaps

When an LLM generates false or harmful content, who is responsible? The model is not an agent in the moral sense it has no intentions, beliefs, or capacity for moral reasoning. Its outputs are mechanistic consequences of training data and optimization objectives. Yet treating it as a mere tool is also inadequate, given its complexity and the degree of autonomy in its text generation. This creates an accountability gap. Users may feel they bear limited responsibility because they did not directly create the content the LLM did. Developers may disclaim responsibility because the model's outputs are emergent and unpredictable, not directly programmed. The model itself cannot be held accountable in any meaningful sense. The result is situations where harmful or false content is generated without clear moral or legal responsibility.

This gap is particularly concerning in high-stakes domains. If an LLM provides medical advice that leads to harm, who is liable? If it generates misinformation that influences an election, who is accountable? If it produces biased hiring recommendations, who is responsible for discriminatory outcomes ? The lack of clear answers to these questions creates moral hazards situations where harmful outcomes occur without corresponding accountability. Addressing this requires rethinking our frameworks for attribution and responsibility in contexts involving complex AI systems. We may need new legal and ethical categories that recognize the distributed nature of responsibility when LLMs are involved, clarifying the obligations of developers, deployers, and users in ways that prevent accountability from slipping through the cracks.

The Need for Verification Architectures

The fundamental problem coherence without guaranteed truth suggests that responsible deployment of LLMs requires external verification mechanisms. These might include :

- **Real-time fact-checking**: Systems that automatically verify factual claims against authoritative databases during generation or before pre-

sentation to users.

- **Confidence calibration**: Mechanisms that allow models to reliably assess their own uncertainty and communicate this to users, preventing overconfident assertions on uncertain topics.
- **Citation and provenance tracking**: Requirements that LLM outputs include traceable sources for factual claims, allowing users to verify information independently.
- **Human-in-the-loop verification**: Workflows that require human expert review of LLM outputs in high-stakes domains before they are acted upon.
- **Adversarial testing**: Systematic efforts to identify failure modes, edge cases, and situations where models generate plausible falsehoods, with results informing deployment decisions.

These mechanisms acknowledge that internal coherence is insufficient for reliability and that truth requires grounding in external reality. They represent a shift from treating LLMs as autonomous knowledge sources to treating them as tools that must be integrated into larger verification ecosystems.

Toward a Hybrid Framework : Integrating Coherence and Verification

The analysis presented thus far suggests a fundamental tension: LLMs excel at coherence but lack intrinsic truth-tracking mechanisms, while traditional knowledge systems prioritize truth but often lack the flexibility and linguistic sophistication of LLMs. This suggests the need for hybrid approaches that combine the strengths of both.

Layered Architecture: Generation and Validation

One promising direction involves separating generation from validation. In this architecture, LLMs would serve as hypothesis generators producing candidate explanations, arguments, or solutions while separate systems would validate these candidates against external criteria. The generation layer leverages the LLM's capacity for coherent, contextually appropriate text production. The validation layer employs fact-checking databases, logical verifiers, domain-specific reasoners, or human experts to assess accuracy. This approach acknowledges what LLMs do well (generating plausible, well-formed content) while compensating for what they do poorly (ensuring factual accuracy). It treats coherence as a necessary but insufficient condition for acceptable output, requiring additional verification before content is presented as reliable.

> *Critically, this architecture makes explicit what is currently implicit : that coherence and truth are distinct properties requiring different*

assessment methods. By separating them structurally, we prevent the conflation that leads to overconfidence in LLM outputs.

Explicit Uncertainty Quantification

LLMs currently generate outputs with uniform confidence regardless of their actual reliability. A statement based on widely replicated training data and a fabricated claim receive the same assertoric force. This creates a misleading impression of comprehensive knowledge.

Improved systems should explicitly quantify and communicate uncertainty. This might involve :

- Attaching confidence scores to factual claims based on training data frequency and consistency
- Using hedging language that scales with uncertainty ("It is certain that…", "It is likely that…", "It is possible that…")
- Distinguishing between claims the model is confident about and those where it is extrapolating or speculating
- Refusing to answer when confidence falls below acceptable thresholds

Such mechanisms would prevent users from treating all LLM outputs as equally reliable and would direct attention to claims requiring additional verification.

Domain-Specific Grounding

Different domains have different relationships between coherence and truth. In mathematics, formal coherence and truth are closely aligned (though not identical consistent mathematical systems can be non-standard models). In fiction, coherence is valuable independently of truth. In empirical sciences, coherence is necessary but must be supplemented with observational evidence.

This suggests that LLM deployment should be domain-specific, with different architectures and validation mechanisms depending on the epistemic standards of the domain. A mathematics-focused LLM might integrate automated theorem provers. A scientific LLM might require citation of peer-reviewed sources. A creative writing LLM might optimize for coherence without truth constraints. This domain-specific approach recognizes that the relationship between language and reality varies across contexts and that appropriate AI systems should reflect these variations rather than applying a one-size-fits-all architecture.

Recursive Self-Criticism

Advanced LLM architectures might incorporate recursive self-criticism using the model's own capabilities to critique and refine its outputs. This could involve generating an initial response, then prompting the model to identify potential errors, weaknesses, or unsupported claims in that response, then revising based on this critique. While not a complete solution (the model's limitations affect both

generation and critique), this approach can catch certain classes of errors internal contradictions, claims that conflict with other well-established information, and logical fallacies that a single-pass generation would miss. It leverages the model's pattern-recognition capabilities to detect deviations from high-quality reasoning patterns. This recursive approach acknowledges that LLMs can recognize certain forms of incoherence or poor reasoning when explicitly prompted, even if they sometimes generate such flaws initially. By systematizing this self-critique, we can improve output quality without requiring external validation for every claim.

Conclusion

This inquiry has argued that the apparent reasoning capabilities of Large Language Models are better understood as coherence generation rather than truth-seeking. LLMs are sophisticated pattern-matching systems that reproduce the surface structure of logical discourse without necessarily accessing the underlying reality that such discourse purports to represent. They excel at formal validity while remaining systematically uncertain about material truth. This is not a failure of these systems but a consequence of their design. LLMs optimize for linguistic coherence a property that can be measured and trained using text corpora rather than for truth, which requires grounding in extra-linguistic reality. The result is systems that can generate arguments that are internally consistent, rhetorically compelling, and formally valid while being factually unreliable.

The central claim of this paper is that our difficulty in properly assessing and deploying LLMs stems from a deeper confusion: the conflation of logical coherence with truth. We have inherited from rationalist traditions an assumption that well-reasoned arguments reliably indicate correct conclusions, that logical structure correlates strongly with factual accuracy. LLMs break this correlation, demonstrating that coherence can be algorithmically produced in the absence of understanding, intentionality, or empirical grounding. This demonstration has important implications. Epistemologically, it requires us to more carefully distinguish between formal and material adequacy in reasoning, between validity and soundness, between persuasiveness and truth. Practically, it suggests that responsible deployment of LLMs requires hybrid architectures that combine coherence generation with external verification mechanisms. Ethically, it raises questions about accountability, epistemic trust, and the conditions under which we should treat AI-generated content as reliable.

The problem revealed by LLMs is ultimately not technological but philosophical. We must clarify what we mean by reasoning, understanding, and knowledge in contexts where these capacities can be mechanically simulated without being genuinely instantiated. We must develop new frameworks for evaluating intelligence that do not conflate formal sophistication with substantive competence. And we must design systems that acknowledge the gap between coherence and truth rather than pretending it does not exist. The illusion of reasoning pro-

duced by language models is a mirror held up to our own epistemic practices. It reveals how much of what we take to be genuine understanding is actually pattern recognition, how often we mistake formal correctness for factual accuracy, and how vulnerable we are to coherent falsehoods. By examining this illusion carefully, we can develop both better AI systems and better accounts of what genuine reasoning and knowledge require.

The path forward involves neither rejecting LLMs as fundamentally flawed nor accepting them uncritically as knowledge systems. Instead, we must understand their capabilities and limitations with precision, deploy them in contexts where coherence generation is valuable, and supplement them with verification mechanisms where truth is critical. We must, in short, learn to use these systems wisely recognizing that they are powerful tools for generating well-formed language while remaining fundamentally agnostic about the truth of what they generate. This requires humility about both AI capabilities and human judgment. LLMs show us that sophisticated reasoning can be simulated without understanding. But human reasoning is itself fallible, biased, and often insufficiently grounded in evidence. The challenge is not to replace human judgment with AI or to reject AI in favor of purely human reasoning, but to develop hybrid systems that leverage the strengths of both while compensating for their respective weaknesses.

In the end, logical coherence without truth is not merely a problem for AI systems. It is a possibility inherent in any reasoning system that operates on form without sufficient grounding in content. The distinctive contribution of LLMs is to make this possibility explicit, inescapable, and impossible to ignore. By forcing us to confront the gap between coherence and truth, they offer an opportunity for philosophical and practical progress if we are willing to learn from what they reveal.

References

1. **Aristotle** (350 BCE). *Prior Analytics*. Translated by A.J. Jenkinson. Oxford: Oxford University Press.

2. **Ayer, A.J.** (1936). *Language, Truth and Logic*. London: Victor Gollancz.

3. **Bender, E.M., Gebru, T., McMillan-Major, A., & Shmitchell, S.** (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610-623.

4. **BonJour, L.** (1985). *The Structure of Empirical Knowledge*. Cambridge, MA: Harvard University Press.

5. **Brown, T.B., Mann, B., Ryder, N., Subbiah, M., et al.** (2020). Language Models are Few-Shot Learners. *Advances in Neural Information*

Processing Systems, 33, 1877-1901.

6. **Carnap, R.** (1937). *The Logical Syntax of Language*. London: Routledge & Kegan Paul.

7. **Chalmers, D.J.** (2023). Could a Large Language Model be Conscious? *Boston Review*, 47(2), 20-24.

8. **Davidson, D.** (1984). *Inquiries into Truth and Interpretation*. Oxford: Clarendon Press.

9. **Devlin, J., Chang, M.W., Lee, K., & Toutanova, K.** (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of NAACL-HLT*, 4171-4186.

10. **Floridi, L., & Chiriatti, M.** (2020). GPT-3: Its Nature, Scope, Limits, and Consequences. *Minds and Machines*, 30(4), 681-694.

11. **Frege, G.** (1879). *Begriffsschrift: A Formula Language of Pure Thought Modeled upon that of Arithmetic*. Translated by S. Bauer-Mengelberg. In J. van Heijenoort (Ed.), *From Frege to Gödel*, 1-82.

12. **Gettier, E.L.** (1963). Is Justified True Belief Knowledge? *Analysis*, 23(6), 121-123.

13. **Gödel, K.** (1931). On Formally Undecidable Propositions of Principia Mathematica and Related Systems. Translated by B. Meltzer. *Monatshefte für Mathematik und Physik*, 38, 173-198.

14. **Goldman, A.I.** (1979). What is Justified Belief? In G.S. Pappas (Ed.), *Justification and Knowledge*, 1-23. Dordrecht: Reidel.

15. **Harnad, S.** (1990). The Symbol Grounding Problem. *Physica D: Nonlinear Phenomena*, 42(1-3), 335-346.

16. **Ji, Z., Lee, N., Frieske, R., Yu, T., et al.** (2023). Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12), 1-38.

17. **Kaplan, J., McCandlish, S., Henighan, T., et al.** (2020). Scaling Laws for Neural Language Models. *arXiv preprint arXiv:2001.08361*.

18. **Kripke, S.** (1980). *Naming and Necessity*. Cambridge, MA: Harvard University Press.

19. **Lenat, D.B.** (1995). CYC: A Large-Scale Investment in Knowledge Infrastructure. *Communications of the ACM*, 38(11), 33-38.

20. **Lewis, D.** (1986). *On the Plurality of Worlds*. Oxford: Blackwell.

21. **Marcus, G., & Davis, E.** (2020). GPT-3, Bloviator: OpenAI's Language Generator Has No Idea What It's Talking About. *MIT Technology Review*, August 22.

22. **Mitchell, M., & Krakauer, D.C.** (2023). The Debate Over Understanding in AI's Large Language Models. *Proceedings of the National Academy of Sciences*, 120(13), e2215907120.

23. **Newell, A., & Simon, H.A.** (1976). Computer Science as Empirical Inquiry: Symbols and Search. *Communications of the ACM*, 19(3), 113-126.

24. **OpenAI** (2023). GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.

25. **Piantadosi, S.T., & Hill, F.** (2022). Meaning without Reference in Large Language Models. *arXiv preprint arXiv:2208.02957*.

26. **Popper, K.R.** (1959). *The Logic of Scientific Discovery*. London: Hutchinson.

27. **Putnam, H.** (1975). The Meaning of 'Meaning'. *Minnesota Studies in the Philosophy of Science*, 7, 131-193.

28. **Quine, W.V.O.** (1951). Two Dogmas of Empiricism. *The Philosophical Review*, 60(1), 20-43.

29. **Radford, A., Wu, J., Child, R., Luan, D., et al.** (2019). Language Models are Unsupervised Multitask Learners. *OpenAI Blog*, 1(8), 9.

30. **Russell, B.** (1905). On Denoting. *Mind*, 14(56), 479-493.

31. **Searle, J.R.** (1980). Minds, Brains, and Programs. *Behavioral and Brain Sciences*, 3(3), 417-424.

32. **Shanahan, M.** (2024). Talking About Large Language Models. *Communications of the ACM*, 67(2), 68-79.

33. **Tarski, A.** (1944). The Semantic Conception of Truth and the Foundations of Semantics. *Philosophy and Phenomenological Research*, 4(3), 341-376.

34. **Vaswani, A., Shazeer, N., Parmar, N., et al.** (2017). Attention is All You Need. *Advances in Neural Information Processing Systems*, 30, 5998-6008.

35. **Wittgenstein, L.** (1953). *Philosophical Investigations*. Translated by G.E.M. Anscombe. Oxford: Blackwell.