

NON EPISTEMIC AGENCY AND DEGRADATION IN LARGE LANGUAGE MODELS' EPISTEMIC PRACTICES

BY

ONWULI, Charles Kemakolom
Department of Philosophy
University of Port Harcourt, Nigeria
Email: moonlightcharles@gmail.com

&

Francis I. Minimah
Department of Philosophy
University of Port Harcourt, Nigeria
Email: francis.minimah@uniport.edu.ng

Abstract

Large Language Models (LLMs) are increasingly positioned as knowledge tools, expected to provide reliable answers and act like experts. Yet, their design reveals a tension between epistemic agency (truth-seeking and justification) and non-epistemic agency (outputs shaped by probability, persuasion, or external pressures). At a general level, society assumes that AI systems can uphold the same standards of inquiry and reliability as human experts. However, the underlying reality is that LLMs are built to predict the next most likely word, not to verify truth. This probabilistic orientation introduces distortions: reinforcement from user feedback, alignment protocols that prioritize politeness or persuasion, and commercial optimization that favors engagement over accuracy. When applied to specific contexts, these distortions manifest as hallucination (inventing false facts), overconfidence (presenting uncertainty as certainty), and selective omission (skipping inconvenient details). Together, these practices degrade epistemic standards in three dimensions: justificatory opacity, where reasoning cannot be traced; normative erosion, where truth is subordinated to utility; and reliability fragmentation, where answers vary unpredictably across contexts. This degradation is not accidental but systemic, arising from the mismatch between epistemic expectations and non-epistemic mechanisms. In medicine, this could mean fabricated clinical advice; in law, misleading interpretations of statutes; in education, inconsistent explanations of core concepts. Such failures threaten the integrity of knowledge systems that depend on transparency, accountability, and normative grounding. In the simplest terms, a Large Language Model (LLM) is a computer program trained on massive amounts of text to predict and generate

words in sequence. It does not “know” facts but simulates knowledge by producing language that sounds plausible. Recognizing this distinction is crucial as without safeguards, LLMs risk becoming persuasive but unreliable agents in domains where truth and justification matter most. These dimensions show that LLMs, while powerful linguistic tools, cannot be treated as full epistemic agents without critical safeguards. The paper concludes by proposing a framework for distinguishing epistemic from non-epistemic agency in LLMs. This framework emphasizes transparency, accountability, and normative alignment as conditions for mitigating degradation. By clarifying these boundaries, we aim to support more responsible integration of LLMs into knowledge systems, ensuring that their contributions remain epistemically sound rather than merely linguistically persuasive.

Keywords: epistemology, epistemic agency, practices, degradation, large language models

Introduction

The historical roots of large language models (LLMs) can be traced back to Claude Shannon’s 1948 work on information theory, which introduced statistical models of language through n-grams (Shannon, 1948, p. 12). For decades, statistical approaches dominated machine translation and speech recognition, but they were limited by data sparsity and short-context modeling (Yu Ji, 2024, p. 4). A breakthrough came in 2003 when Yoshua Bengio introduced the neural probabilistic language model, embedding words into dense vectors and predicting sequences through neural networks (Bengio, 2003, p. 17). This innovation laid the foundation for distributed representations, later expanded by Mikolov’s Word2Vec (2013), which demonstrated semantic algebra in vector space (Mikolov, 2013, p. 45). The decisive shift occurred in 2017 with Vaswani et al.’s Transformer architecture, which replaced recurrent and convolutional networks with self-attention mechanisms (Vaswani et al., 2017, p. 22). This architecture enabled scaling laws that made LLMs viable at unprecedented sizes (LLM Hub, 2024, p. 11). By 2018, OpenAI’s GPT-1 demonstrated the potential of unsupervised pretraining followed by fine-tuning (OpenAI, 2018, p. 33). Within five years, models like GPT-4, Claude, and Gemini had scaled to trillions of parameters, becoming widely adopted across academic, professional, and everyday contexts (Treatise on LLMs, 2024, p. 56).

There is an underlying cultural prominence and epistemic tensions arising from these facts in dare need of considerations. The “ChatGPT moment” in late 2022 marked the cultural mainstreaming of LLMs, transforming them from research curiosities into everyday tools (LLM Hub, 2024, p. 19). Their adoption as “thinking partners” reflects a profound epistemic shift where users interact with them as if they possess understanding, judgment, or perspective, despite their lack of beliefs or consciousness (Yu Ji, 2024, p. 7). This paradox arises because LLMs simulate epistemic agency through linguistic fluency and contextual sensitivity, yet their outputs are driven by probabilistic prediction rather than truth-seeking (Hinton, 2020, p. 88). Critics highlight that reinforcement learning, alignment protocols, and commercial imperatives shape these systems in ways that embed non-epistemic drivers into knowledge practices (Amodei, 2021, p. 44; Hassabis, 2023, p. 67). The danger is epistemic degradation, wherein, plausibility replaces truth, persuasion replaces justification, and optimization replaces deliberation (Hountondji, 1996, p. 41).

Historically significant remains the fact that placing ChatGPT, Claude, and Gemini within this trajectory underscores their dual nature. On one hand, they embody decades of progress in computational linguistics, cognitive science, and AI engineering (Yu Ji, 2024, p. 9). On the other, they reveal the tension between epistemic and non-epistemic functions, raising critical questions about how societies frame and use them (LLM Hub, 2024, p. 23). Recognizing this paradox is essential for safeguarding epistemic integrity in an era where probabilistic systems are increasingly embedded in reasoning and decision-making. The historical context of LLMs shows that their rise as “thinking partners” is both a technological achievement and an epistemic challenge. By situating them within information theory, neural modeling, and transformer architectures, we see how their cultural prominence today is inseparable from decades of intellectual development and why critical reflection on their epistemic limits is necessary.

The paper makes three primary contributions to this discourse. First, it introduces a clear differentiation between epistemic agency and non-epistemic agency, showing how the latter systematically undermines knowledge practices (Fricker, 2007, p. 49; Searle, 1983, p. 112). By drawing on Searle’s account of intentional states and Dennett’s “intentional stance,” the paper situates LLMs within philosophical debates on agency, clarifying why attributing epistemic qualities to them is conceptually misleading (Dennett, 1987, p. 58; Searle, 1992, p. 77). Second, it provides a detailed account of the mechanisms of epistemic degradation, including probabilistic reasoning, alignment strategies, and commercial optimization (Christiano, 2017, p. 33; Gabriel, 2020, p. 56). These mechanisms illustrate how non-epistemic drivers become embedded in knowledge systems under the guise of epistemic reasoning. Third, it proposes a framework for safeguarding epistemic integrity, emphasizing transparency, accountability, and normative alignment as conditions for responsible integration of LLMs into epistemic practices (Floridi, 2019, p. 105; Hassabis, 2023, p. 67). Together, these contributions advance the discourse beyond the narrow question of whether LLMs can “know.” Instead, they highlight the more pressing issue of how their use reshapes human epistemic practices, potentially eroding the very norms that sustain knowledge. By situating the discussion within both the cultural prominence of LLMs and the philosophical foundations of agency, the paper underscores the need for safeguards that preserve epistemic integrity in the age of machine-generated language.

The Problematic and Contributions

The central problematic addressed in this paper is the systematic degradation of epistemic conditions when Large Language Models (LLMs) such as ChatGPT, Claude, and Gemini are treated as epistemic partners. These systems, while linguistically fluent and context-sensitive, do not possess beliefs, intentions, or commitments to truth. Their outputs are generated through probabilistic language prediction, shaped by reinforcement learning, alignment protocols, and commercial imperatives (Hinton, 2020, p. 91; Bender & Gebru, 2021, p. 12). The issue is not simply whether LLMs can “know,” but how our reliance on them reshapes epistemic norms. When users attribute agency to these systems, they risk importing non-epistemic mechanisms such as plausibility, persuasion, and optimization into practices that should be governed by justification, truth, and reliability (Floridi, 2019, p. 102; Fricker, 2007, p. 45). This degradation manifests in three foundational conditions of human knowledge. First, justificatory transparency is obscured, as users

cannot access the grounds on which claims are made (Zagzebski, 1996, p. 88; Hountondji, 1996, p. 41). Second, normative integrity is weakened, as truth is subordinated to utility or persuasion (Oruka, 1997, p. 34; Diagne, 2013, p. 118). Third, reliability stability is fragmented, as outputs vary across contexts without consistent epistemic grounding (Amin, 2014, p. 132; Keita, 2010, p. 77). These erosions collectively undermine the practices that sustain human knowledge, raising urgent concerns in AI ethics and social epistemology.

Conceptual Review

Epistemic Agency

Epistemic agency refers to the capacity of an agent to form beliefs, justify claims, and pursue truth through rational deliberation (Fricker, 2007, p. 45). In human contexts, epistemic agency involves responsibility for knowledge practices, including critical reflection and justification (Zagzebski, 1996, p. 88). Large language models (LLMs), however, simulate epistemic agency without possessing its substance. Their outputs are generated through probabilistic prediction rather than belief formation (Hinton, 2020, p. 91). This distinction is crucial, for example, while users interact with systems like ChatGPT, Claude, and Gemini as if they “know,” their architecture reveals non-epistemic drivers such as reinforcement learning and commercial optimization (Amodei, 2021, p. 44).

Non-Epistemic Drivers

Non-epistemic drivers are mechanisms that shape outputs without commitment to truth. In LLMs, these include reinforcement learning from human feedback (RLHF), alignment protocols, and corporate imperatives (Bender & Gebru, 2021, p. 12). RLHF optimizes for user satisfaction rather than epistemic accuracy (Christiano, 2017, p. 33). Alignment strategies prioritize safety and persuasion, embedding social norms into model behavior (Gabriel, 2020, p. 56). Commercial imperatives drive models toward engagement metrics and monetization (Hassabis, 2023, p. 67). These drivers illustrate how LLMs’ agency is shaped by non-epistemic functions, raising concerns about epistemic degradation when users treat them as truth-seeking partners (Floridi, 2019, p. 102).

Epistemic Degradation

Epistemic degradation occurs when truth-seeking practices are eroded by reliance on mechanisms that prioritize plausibility or persuasion (Hountondji, 1996, p. 41). In the context of LLMs, degradation arises when users project epistemic qualities onto systems that lack belief or justification (Oruka, 1997, p. 34). This projection normalizes practices where coherence substitutes for truth, and linguistic fluency substitutes for rational deliberation (Diagne, 2013, p. 118). Scholars warn that epistemic degradation undermines critical norms of justification, accountability, and intellectual responsibility (Fricker, 2007, p. 49; Zagzebski, 1996, p. 92).

Thinking Partners

The concept of “thinking partners” situates LLMs within everyday epistemic practices. Traditionally, thinking partners are collaborators who share reasoning, challenge assumptions, and co-produce knowledge (Lipman, 2003, p. 77). Systems like ChatGPT, Claude, and Gemini are now widely adopted in this role, consulted for brainstorming, problem-solving, and decision-making

(LLM Hub, 2024, p. 19). Their linguistic fluency and contextual sensitivity invite users to treat them as epistemic agents (Yu Ji, 2024, p. 7). Yet, their lack of belief and commitment to truth complicates this framing, raising the paradox of simulated agency (Floridi, 2019, p. 105).

Probabilistic Language Prediction

At the core of LLMs is probabilistic language prediction, the statistical modeling of word sequences based on training data (Shannon, 1948, p. 12; Bengio, 2003, p. 17). Unlike human reasoning, which involves belief and justification, probabilistic prediction generates outputs by maximizing likelihood functions (Vaswani et al., 2017, p. 22). This mechanism explains why LLMs can produce coherent, context-sensitive responses without epistemic grounding (Hinton, 2020, p. 91). Scholars argue that conflating probabilistic prediction with epistemic reasoning risks embedding non-epistemic mechanisms into knowledge systems (Bender & Gebru, 2021, p. 14).

Philosophical Debate on Agency

Philosophical debates around Large Language Models have largely centered on the question: “*Can LLMs know?*” Scholars have examined whether these systems can be considered epistemic agents, whether their outputs count as knowledge, and whether they can meaningfully participate in justification. This line of inquiry has been valuable, but it remains incomplete. What has not been sufficiently asked is a different, and arguably more urgent, question: “*Do LLMs make us know worse?*” In other words, the focus has been on the epistemic status of the machine, rather than on the epistemic consequences for the human users and communities that rely on it. By treating LLMs as if they “know,” the risk of reshaping epistemic practices, potentially degrading them by embedding non-epistemic mechanisms into systems of inquiry and justification becomes clearer. This gap matters because epistemic degradation is not simply about whether LLMs themselves can know, but whether their widespread use erodes the norms of truth-seeking, justification, and reliability in human knowledge systems. If users begin to accept plausibility as a substitute for truth, or persuasion as a substitute for justification, then epistemic practices themselves are weakened. The result is not only a mischaracterization of LLMs but a systemic shift in how knowledge is produced and validated. By identifying this gap, the paper moves beyond the question of machine epistemology to examine the relational impact of LLMs on human epistemic integrity. It asks not only whether LLMs can know, but whether their integration into our practices makes *humans* know worse. This reframing opens a new line of inquiry into epistemic degradation, situating the problem at the intersection of philosophy, AI ethics, and social epistemology.

John Searle defines intentional states such as belief, desire, and intention as mental phenomena directed toward conditions of satisfaction. He argues that beliefs are satisfied if the state of affairs they represent actually obtains, while intentions are satisfied if the action represented is performed. Searle emphasizes responsibility by linking intentional states to agency because intentions are not passive but commit the agent to action. Dennett (1989) introduces the strategy of interpreting systems by attributing beliefs, desires, and intentions to them, regardless of their internal mechanisms. This stance allows prediction of behavior by assuming rationality: we ask what the system *would* do if it had certain beliefs and desires. Dennett stresses that agency is not about inner states but about the explanatory and predictive utility of treating systems as intentional. Both

scholars also provide two contrasting but complementary accounts of agency, each emphasizing belief, intention, understanding, and responsibility in different ways. For Searle, agency is grounded in the intrinsic intentionality of mental states. He (1983, pp. 79–111) argues that beliefs, desires, and intentions are directed toward conditions of satisfaction. A belief is satisfied if the state of affairs it represents actually obtains; an intention is satisfied if the action it represents is performed. Crucially, intentions commit the agent to action, linking agency directly to responsibility. Understanding, in Searle’s framework, is rooted in the intentional content of mental states and the background of rational practices that make them meaningful. Responsibility, therefore, is not optional but intrinsic to agency, since intentional states bind the agent to rational accountability.

By contrast, Dennett develops a more instrumental account (1989, pp. 13–42) where he argues that agency can be understood by adopting a predictive strategy since we attribute beliefs, desires, and intentions to a system in order to explain and anticipate its behavior. These attributions are not claims about the system’s inner states but about the utility of treating it as if it had them. Beliefs, in this view, are not intrinsic mental states but explanatory constructs. Intentions are similarly stance-based attributions, useful for predicting rational action given supposed beliefs and desires. Understanding emerges not from internal intentional content but from the adoption of the intentional stance itself. Responsibility, in Dennett’s account, is not intrinsic to the system but arises when intentional attributions are embedded in social and normative contexts. Taken together, these perspectives highlight a tension. Searle insists that genuine agency requires real beliefs, intentions, and responsibility, while Dennett allows that agency can be modeled instrumentally, even in systems that lack inner intentionality. This tension is crucial for analyzing Large Language Models because Searle’s framework underscores what LLMs lack, true beliefs, intentions, and responsibility, while Dennett’s stance explains why we nevertheless treat them as agents, attributing beliefs and intentions to predict their outputs. The paradox of LLMs as “thinking partners” emerges precisely from this contrast that they simulate agency without possessing the intentional states that, for Searle, make agency proper. The implications for LLMs are obvious. For example, Searle suggests LLMs cannot be true agents because they lack genuine beliefs and intentions. Dennett’s view allows us to treat LLMs as intentional systems for practical purposes, but this is instrumental, not ontological. The tension between these perspectives frames the paradox that LLMs have no beliefs, yet we interact with them as if they do.

Agency in philosophy is often defined through belief, intention, understanding, and responsibility. John Searle and Daniel Dennett provide two influential accounts: Searle emphasizes the intrinsic intentionality of mental states, while Dennett develops the “intentional stance,” a predictive framework that attributes beliefs and desires to systems. Together, they form the conceptual foundation for analyzing agency proper. The “intentional stance,” introduced by Daniel Dennett, is not a claim that agency is metaphysically real but rather a predictive strategy which says that we attribute beliefs and desires to systems (human, animal, or even machines) because doing so helps us forecast their behavior. Dennett emphasizes that this stance is pragmatic, not ontological. Dennett (1987, p. 17) explains that adopting the intentional stance means treating an entity “as a rational agent” and predicting its actions based on attributed beliefs and desires. In pp. 48–49, he clarifies that this is a reason-giving explanation rooted in folk psychology, not proof of actual inner states.

Dennett (1988, pp. 495–505) reiterates that the stance is a strategy of prediction and explanation, applicable to both living and nonliving systems. He stresses that systems are “intentional” only insofar as their behavior can be predicted this way, regardless of their internal mechanisms. While in (1991, p. 240), he discusses “aboutness” and how intentionality is derived from the interpretive stance of observers, not intrinsic metaphysical properties. There are four core ideas of non-epistemic agency. The first idea is known as “agency as a heuristic”. This idea states that all the time we attribute beliefs and desires to systems to make sense of their behavior, but this does not mean those systems really possess such states. The second idea is known as “predictive utility”. In this idea the intentional stance works because it simplifies complex systems into rational agents, allowing us to forecast outcomes. The third idea is the “non-epistemic”. This position does not claim knowledge of inner states; it is a pragmatic tool, not a metaphysical assertion. The fourth idea is called “application”. This idea is used in folk psychology, artificial intelligence, and evolutionary theory as a modeling strategy.

Interface Design and Agency Attribution

A critical issue in debates about LLMs is how interface design choices encourage users to attribute agency to systems that lack genuine intentionality. Dennett’s framework emphasizes that intentionality is not a real mental phenomenon but an interpretive stance, adopted for explanatory convenience (Dennett, 1987, p. 58). By contrast, Searle argues that intentional states are grounded in real mental phenomena such as belief and intention (Searle, 1983, p. 112). This tension between realism and interpretivism remains central in philosophy of mind and directly informs how we understand agency in LLMs (Floridi, 2019, p. 102; Fricker, 2007, p. 45). Interface design amplifies this tension by embedding linguistic cues that simulate agency. Research shows that conversational features such as first-person pronouns, apologies, and epistemic hedging trigger anthropomorphism and perceived social presence (Nass & Moon, 2000, p. 25; Reeves & Nass, 1996, p. 44). These cues encourage users to treat outputs as intentional acts, even though LLMs are probabilistic generators without beliefs or commitments (Bender & Gebru, 2021, p. 14; Hinton, 2020, p. 91).

The first mechanism is the use of first-person pronouns. When LLMs say “I,” users interpret outputs as coming from an agent with perspective. Schreieder et al. (2026, p. 30970) note that stylistic choices in text generation influence perceptions of authorship and responsibility, reinforcing the illusion of agency. This aligns with Dennett’s intentional stance, where attribution of beliefs and desires is a pragmatic strategy rather than a reflection of actual mental states (Dennett, 1987, p. 61). The second mechanism is the use of apologies. Phrases such as “I’m sorry” simulate social norms of accountability, embedding human-like communicative conventions. Wühl et al. (2026, p. 4) argue that politeness strategies in conversational AI embed cultural scripts of responsibility, which users misinterpret as intentional behavior. This creates a false sense of moral agency, even though the system lacks genuine responsibility (Floridi, 2019, p. 105; Gabriel, 2020, p. 56). The third mechanism is epistemic hedging, such as “I think...” or “I believe....” These phrases suggest subjective judgment, even though LLMs operate through probabilistic prediction. Schreieder et al. (2026, p. 30970) highlight that hedging language blurs the line between evidence-based attribution and simulated cognition. Users interpret hedging as epistemic humility, reinforcing the illusion of belief or perspective (Fricker, 2007, p. 49; Zagzebski, 1996, p. 92). Together, these design choices

illustrate how linguistic cues normalize agency attribution in systems that lack genuine intentionality. The danger is epistemic degradation: plausibility replaces truth, persuasion replaces justification, and optimization replaces deliberation (Hountondji, 1996, p. 41; Oruka, 1997, p. 34). By critically examining interface design, we see how cultural and linguistic conventions embed non-epistemic mechanisms into epistemic practices, reshaping norms of knowledge in subtle but systemic ways.

Mechanisms Behind Agency Attribution

There are four major mechanisms that explain why users attribute agency to Large Language Models (LLMs). The first is anthropomorphism, where users project human qualities onto systems that mimic human conversational style. Nass and Moon (2000, p. 25) demonstrated that people treat computers as social actors when they display human-like cues, reinforcing the illusion of agency. Reeves and Nass (1996, p. 44) similarly showed that naturalistic dialogue triggers social responses, even when users know the system lacks consciousness. The second mechanism is social presence cues. Politeness markers and apologies activate expectations of reciprocity and accountability. Wühl et al. (2026, p. 4) argue that phrases like “I’m sorry” embed human-like communicative conventions, which users misinterpret as intentional behavior. This aligns with earlier findings that politeness strategies in human-computer interaction foster perceived responsibility (Fogg, 2003, p. 67; Floridi, 2019, p. 105). The third mechanism is narrative framing. First-person pronouns such as “I” create the illusion of perspective-taking. Schreieder et al. (2026, p. 30970) note that stylistic choices in text generation influence perceptions of authorship and responsibility, reinforcing the illusion of agency. Dennett’s (1987, p. 61) “intentional stance” explains this phenomenon: users attribute beliefs and desires to systems as a pragmatic strategy, even though no genuine intentionality exists. The fourth mechanism is interface design bias. Designers often choose naturalistic dialogue to improve usability, inadvertently reinforcing false impressions of agency. Bender and Gebru (2021, p. 14) caution that design choices embed anthropomorphic cues that distort responsibility attribution. Hinton (2020, p. 91) adds that probabilistic generators are optimized for coherence, not truth, yet their interface presentation encourages epistemic projection. Scholars warn that this can distort responsibility attribution in contexts such as misinformation, plagiarism, or ethical accountability (Gabriel, 2020, p. 56; Brogaard, 2023, p. 22).

Epistemic Adequacy: Reliabilism and Virtue Epistemology

Beyond interface mechanisms, the issue of agency attribution connects to broader debates in epistemology. Epistemic adequacy asks for what makes a belief count as knowledge in the two major approaches of reliabilism and virtue epistemology. Reliabilism emphasizes the reliability of cognitive processes. A belief counts as knowledge if it is produced by a reliable process such as perception, memory, or inference (Goldman, 1979, p. 113; Greco, 1992, p. 45). In its externalist stance, reliability does not require the believer’s awareness of the process. Greco (1992, p. 47) argues that reliabilism avoids skepticism by grounding knowledge in the actual reliability of our faculties, not in our ability to refute skeptical scenarios. Analogous to ethics, just as rule utilitarianism justifies actions by their utility-maximizing rules, reliabilism justifies beliefs by their truth-conducive processes (Sosa, 1991, p. 102).

Virtue epistemology shifts the focus from processes to intellectual character. There are two strands: virtue reliabilism and virtue responsibilism. Virtue reliabilism, espoused by Greco and Sosa, holds that knowledge arises when beliefs are formed through cognitive abilities functioning properly (Greco, 2006, p. 93; Sosa, 2007, p. 89). Greco (2006, p. 95) framed knowledge as “success from ability,” meaning that true belief is achieved through intellectual competence. Virtue responsibilism, developed by Zagzebski, emphasizes epistemic virtues of character such as open-mindedness, intellectual courage, and a motivation for truth (Zagzebski, 1996, p. 92). She argues that epistemic virtues involve both a motivation for truth and a commitment to epistemic ends.

Critically, reliabilism is attractive for its simplicity and externalist accessibility but risks neglecting the agent’s responsibility (Goldman, 1986, p. 121). Virtue epistemology enriches the picture by demanding epistemic responsibility and motivation, but critics argue it risks over-intellectualizing ordinary knowledge (Kvanvig, 1992, p. 77). Recent scholarship suggests convergence: Brogaard (2023, p. 22) argues that even reliabilists must acknowledge that epistemic responsibility and motivation for truth play a role in knowledge attribution. This convergence highlights that epistemic adequacy requires both reliable processes and virtuous character.

Testimony and Knowledge

The question of whether we can receive knowledge from LLMs through testimony is deeply contested. Philosophers like Jennifer Lackey and Tyler Burge provide frameworks that help us analyze whether LLM outputs count as testimony or merely assertions without a genuine “knower.” Lackey (2008), argues that testimony can transmit knowledge even when the speaker does not themselves know. For example a teacher relays information from a textbook without fully understanding it. The student can still acquire knowledge. In pp. 35–36 she distinguishes between the speaker’s epistemic state and the hearer’s uptake, showing that knowledge can be passed without the speaker being a knower. Burge (1993) further defends the idea that testimony preserves epistemic content across speakers. On pp. 467–468 he emphasizes that testimony is a fundamental source of knowledge, grounded in social practices of trust and linguistic competence. For him testimony is reliable because it is embedded in a normative practice of communication, not because each speaker is individually a knower.

Applying this to LLMs one sees that LLMs as “assertion machines”, which generate outputs based on statistical patterns, not genuine belief or knowledge; unlike human speakers, LLMs lack intentions, sincerity, or epistemic responsibility. Lackey’s framework suggests that even if LLMs are not knowers, users might still acquire knowledge if the outputs are accurate and reliable. On the contrary, Burge’s framework raises doubts. Since LLMs are not embedded in the normative practice of communication, they lack the social commitments that make testimony trustworthy. Critically, on epistemic adequacy, and for LLM outputs to count as knowledge, they must be truth-conducive and reliable. On testimony vs. assertion, LLMs blur the line, users may treat outputs as testimony, but philosophically they are closer to assertions without a knower. Then on ethical implications, misattributing testimony to LLMs risks undermining responsibility and accountability in knowledge practices. Uncritical acceptance of LLM outputs would lead to knowledge degradation because students outsource cognitive labor, reducing scrutiny and critical engagement. Studies show that

reliance on AI-generated essays produce fluent but shallow writing, erasing epistemic effort and weakening evaluative judgment.

Uncritical Acceptance

It could pass as a natural tendency to consider this effort as an uncritical acceptance when outsourcing cognitive labor, students increasingly rely on LLMs to generate essays, paraphrases, or summaries. This automatically reduces the effort invested in analysis, synthesis, and evaluation, leading to diminished epistemic agency. Corroborating this Suazo et al. (2026, p. 7) report that uncritical reliance on generative AI fosters cognitive offloading, automation bias, and superficial information processing. On reduced scrutiny, AI-generated texts often appear polished and coherent, which discourages students from questioning accuracy or depth. Nwagboso & Atubab (2026, p. 3) describe this as, where surface fluency masks the loss of critical inquiry and epistemic labor.

There are many cases of students using LLMs for essay synthesis. Examples from the global north and global south like those of Olaniyan et al. (2026, p. 12) found that students in the Global South often rely on LLMs out of necessity (limited access to resources), while those in the Global North use them for convenience. Both groups expressed concern about the reliability and contextual grounding of AI-generated knowledge. For its educational consequences, essays produced with LLM assistance tend to prioritize formal correctness (grammar, structure) over intellectual depth. This shifts academic work from a process of inquiry to a product of fluency, undermining the pedagogical goal of cultivating critical thinking. Little wonder Nwagboso & Atubab (2026, p. 5) warn that this erasure subtly alters the meaning of learning and authorship. On a critical note, it is worth to note that on its risk students may gain efficiency but lose critical autonomy, weakening their ability to evaluate sources and arguments. The implication is that Universities must promote AI literacy and design assessments that reward process, reflection, and epistemic effort, not just polished outputs. The long-term danger is that if uncritical acceptance becomes normalized, higher education risks producing graduates skilled in fluency but deficient in critical reasoning and epistemic responsibility.

Epistemic Vigilance Failure

There is this rumour on a seeming failure at an appreciable epistemic vigilance. This failure occurs when users attribute agency to LLMs, reducing their own source monitoring. The phrase “GPT said it” can feel epistemically equivalent to “an expert said it,” which leads to uncritical acceptance of fabricated or unreliable information. The best way out is caution at agency attribution and source monitoring. When users treat LLMs as intentional agents, they suspend normal vigilance about source reliability. Sperber et al. (2010, p. 361) describe epistemic vigilance as the cognitive mechanisms humans use to evaluate testimony, checking for reliability, competence, and sincerity. LLMs mimic these cues (first-person pronouns, apologies, hedging), which can trick users into lowering scrutiny. Lewandowsky et al. (2020, p. 14) show that automation bias leads users to trust machine outputs more than they should, especially when presented fluently. This bias undermines critical evaluation, making fabricated citations or false claims appear credible. It is in this light that this study advocates great vigilance especially on all those researchers accepting fabricated citations. Suffice it to note that studies document researchers and students copying LLM-generated citations

without verification. Thorp (2023, p. 313) warns that ChatGPT produces “convincing but entirely fabricated references,” which have already appeared in academic submissions while Kumar et al. (2023, p. 6) note that the authoritative style of LLM outputs encourages users to treat them as genuine testimony. The authoritative tone (“According to Smith, 2018...”) mimics scholarly testimony. Users fail to apply epistemic vigilance, accepting fabricated content as if it were expert knowledge. This results in knowledge degradation as false information enters academic discourse under the guise of testimony. Critically, this mechanism shows how design cues and user bias combine to erode epistemic vigilance. The danger is not just misinformation but institutional contamination as fabricated citations enter published work, undermining trust in scholarly communication.

Misplaced Testimonial Trust

There is this case on misplaced testimonial trust making rounds. This assumption highlights how users treat LLMs as if they were responsible epistemic agents, even though they lack the capacity for belief, sincerity, or accountability. This misplaced trust is especially dangerous in high-stakes domains like medicine and law. This assumption is germane because testimony requires a responsible agent. Testimonial knowledge depends on the speaker being a responsible epistemic agent, someone motivated by truth and accountable for accuracy. Burge (1993, p. 467) argues that testimony is a fundamental source of knowledge precisely because it is embedded in social norms of trust and responsibility. Lackey (2008, p. 35) shows that hearers can acquire knowledge even when speakers lack it, but this presupposes that the communicative practice itself is normatively structured by responsibility. LLMs fail this requirement because they generate fluent assertions without epistemic responsibility, yet users often treat them as if they were knowers. A good example is the public using LLMs for medical and legal advice. In medicine today, Ayoub et al. (2023, p. 6) found that students and patients sometimes accept ChatGPT’s medical explanations uncritically, despite inaccuracies. The authoritative tone of LLMs creates misplaced trust, leading users to treat outputs as expert testimony. Second example is in law today, Surden (2023, p. 112) warns that LLMs can produce plausible but incorrect legal reasoning. Because the interface mimics professional discourse, users may mistake generated text for genuine legal expertise. Lastly is in fabricated authority, Thorp (2023, p. 313) notes that ChatGPT produces fabricated but convincing references, which users sometimes accept as legitimate. This shows how misplaced testimonial trust allows falsehoods to enter decision-making contexts. Critically, on epistemic danger, treating LLMs as responsible agents undermines the norms of testimony. On knowledge degradation, false or fabricated outputs gain uptake as if they were expert knowledge. On ethical implication, misplaced trust shifts accountability away from human agents, creating a vacuum of responsibility in domains where accuracy is critical.

On the basis of a synthesis or as an epistemic deskilling, the three mechanisms, uncritical acceptance, epistemic vigilance failure, and misplaced testimonial trust, do not operate in isolation. Together, they interact to produce what can be called epistemic deskilling: the erosion of users’ ability to critically evaluate, monitor, and responsibly engage with knowledge. How do the mechanisms interact one may ask. At the level of uncritical acceptance that is reduced effort, you find out that students and professionals outsource cognitive labor to LLMs, lowering their own scrutiny and

evaluative effort (Suazo et al., 2026, p. 7). This eventually sets the stage for diminished epistemic vigilance. At the level of epistemic vigilance failure or lower monitoring, you notice that at once effort is reduced, users are more likely to accept outputs as authoritative, especially when agency attribution makes “GPT said it” feel like “an expert said it” (Sperber et al., 2010, p. 361; Thorp, 2023, p. 313). Then fabricated citations and false claims slip through unchecked. Whereas at the level of misplaced testimonial trust or false authority, you see that fully, users treat LLMs as responsible epistemic agents, relying on them for medical or legal advice (Ayoub et al., 2023, p. 6; Surden, 2023, p. 112). This misplacement of trust institutionalizes the degradation, embedding unreliable outputs into decision-making contexts.

Epistemic Deskillling

This is the gradual erosion of critical epistemic skills of source evaluation, skepticism, and responsibility. In the mechanism of degradation as implied, each step compounds the next, creating a feedback loop where reliance on LLMs reduces effort, effort reduction lowers vigilance, and lowered vigilance entrenches misplaced trust. The final result is that all users become fluent consumers of AI outputs but less capable of independent knowledge evaluation, leading to systemic epistemic fragility in education, research, and public decision-making. This synthesis shows how the three mechanisms converge into a structural epistemic problem as not just misinformation, but the weakening of the very skills needed to resist misinformation. Even when users *explicitly* know that LLMs are not agents, implicit cognitive mechanisms still operate as if they were. This is explained by dual-process theories of cognition: **System 1 (fast, automatic):** People unconsciously attribute agency when confronted with human-like cues (first-person pronouns, apologies, hedging). **System 2 (slow, reflective):** People may correct themselves and say “I know it’s just a machine.” But System 1 attribution often drives behavior before System 2 correction kicks in. *Sperber et al. (2010, p. 361)* describe epistemic vigilance as a set of evolved mechanisms for monitoring testimony. These mechanisms are triggered by cues of agency and sincerity, regardless of explicit belief. *Waytz et al. (2014, p. 219)* show that anthropomorphism occurs automatically when systems display social cues, even when participants know they are interacting with machines. *Nass & Moon (2000, p. 31)* demonstrate that users apply social rules to computers reflexively, treating them as conversational partners despite knowing they are not human. Why this matters for LLMS is what is with interface cues which (“I think...”, “I’m sorry...”) activate System 1 attribution. Another matter is with explicit belief which (“I know it’s not conscious”) does not prevent implicit trust behaviors. Lastly, the result is that users lower vigilance, outsource judgment, and treat outputs as testimony, even though they rationally know better. Critically, we naively underestimate the power of implicit cognition. Even when users *say* they know LLMs are not agents, their epistemic behavior reveals otherwise. This mismatch between explicit belief and implicit attribution is precisely what makes epistemic vigilance failure and misplaced testimonial trust dangerous to be kept as standards.

Furthermore, it is true that calculators and search engines reshape cognition by outsourcing tasks (arithmetic, information retrieval), but they do not trigger agency attribution. LLMs differ because they adopt conversational norms, using first-person pronouns, apologies, hedging, which activate social cognition. *Nass & Moon (2000, p. 31)* show that users reflexively apply social rules to computers when they mimic human interaction. Thus, LLMs are not just tools; they are interaction

partners in the user's cognitive ecology, which makes their epistemic impact qualitatively distinct. Let us note that it is literacy that helps users to *know* that LLMs are not agents, but design cues which override explicit knowledge. Yet *Waytz et al. (2014, p. 219)* demonstrates that anthropomorphism is automatic when systems display social cues, regardless of user awareness. *Sperber et al. (2010, p. 361)* on the other hand argue that epistemic vigilance mechanisms are triggered by cues of sincerity and agency. Therefore, literacy alone cannot counteract implicit System 1 attribution; design must be rethought to reduce anthropomorphic triggers. Even if LLMs increase access to information overall, agency attribution creates catastrophic vulnerabilities: fabricated citations, false medical/legal advice, and misplaced testimonial trust. *Thorp (2023, p. 313)* documents fabricated references entering academic work while *Ayoub et al. (2023, p. 6)* show students accepting inaccurate medical outputs uncritically and *Surden (2023, p. 112)* warns of plausible but incorrect legal reasoning. These are high-risk epistemic failures because even if average outcomes improve, the variance includes dangerous degradation that undermines trust in knowledge systems. These replies show that objections underestimate the qualitative difference of LLMs since they are not neutral tools but socially designed systems that exploit human cognitive biases.

Rating an understanding toward epistemically responsible LLM use based on transparency of non-Agency provides that there is a central requirement for epistemically responsible LLM use is transparency about the system's non-agency. Users must be reminded that LLMs are statistical models, not knowers. Design cues such as explicit disclaimers ("I am a statistical model, not a knower") can break the illusion of agency attribution. Research on anthropomorphism shows that conversational cues like first-person pronouns and apologies trigger automatic social cognition, even when users explicitly know the system is non-human (*Waytz et al., 2014, p. 219*). By embedding transparency into the interface, designers can counteract these implicit biases and reduce misplaced testimonial trust. Moreover, epistemic vigilance depends on recognizing the limits of testimony. In line with this view *Sperber et al. (2010, p. 361)* argues that vigilance mechanisms are activated by cues of sincerity and agency. If LLMs are framed as non-agents, users are less likely to suspend scrutiny. Transparency thus functions as a cognitive nudge, reminding users that outputs are probabilistic assertions rather than genuine testimony. This principle does not eliminate anthropomorphism but mitigates its epistemic risks by foregrounding the system's non-agency. It is apt stating that agency retention as an epistemic responsibility must remain with the user. Interfaces should be designed to force users to justify, cite, and take responsibility for claims derived from LLMs. For example, requiring students to annotate AI-assisted essays with their own critical commentary ensures that cognitive labor is not fully outsourced. Little wonder *Suazo et al. (2026, p. 7)* show that uncritical reliance on generative AI fosters cognitive offloading and automation bias. By embedding prompts that demand user reflection, designers can retain agency and prevent epistemic deskilling. This principle aligns with virtue epistemology much like *Zagzebski (1996, p. 167)* avowal that knowledge requires intellectual virtues such as motivation for truth and responsibility. If users are compelled to justify their reliance on AI outputs, they exercise epistemic virtues rather than surrendering them. Agency retention thus ensures that LLMs remain tools, not epistemic agents, and that responsibility for knowledge claims stays with human users. Call it source traceability but the fact remains that epistemic provenance is essential, where users must be able to

see where claims come from. LLMs often generate fluent but fabricated citations (Thorp, 2023, p. 313). Without traceability, users cannot evaluate reliability. Interfaces should therefore provide transparent links to underlying sources, enabling verification and reducing epistemic vigilance failure. *Lewandowsky et al. (2020, p. 14)* emphasize that automation bias thrives when outputs appear authoritative but lack verifiable provenance. Traceability also aligns with Burge's account of testimony. *Burge (1993, p. 467)* argues that testimony preserves epistemic content through social norms of trust. For LLMs, provenance functions as a surrogate for these norms, grounding outputs in identifiable sources rather than fabricated authority. By embedding source traceability, designers can restore epistemic responsibility and reduce the risk of fabricated knowledge entering academic or professional discourse.

The direct implication is for all user, whether designers, educators, policymakers, etc. For instance for designers, these principles demand a shift from human-like fluency to epistemic transparency. Interfaces should emphasize non-agency, enforce user responsibility, and provide traceable sources. Educators must integrate AI literacy into curricula, teaching students to critically engage with LLM outputs rather than passively accepting them. *Nwagboso & Atubab (2026, p. 3)* warn that uncritical reliance on AI erases epistemic labor, undermining the pedagogical goal of cultivating critical thinking. Policymakers must regulate AI systems to ensure epistemic safeguards. This includes mandating transparency about non-agency, requiring provenance mechanisms, and discouraging anthropomorphic design cues. *Surden (2023, p. 112)* highlights the risks of misplaced trust in legal contexts, where plausible but incorrect outputs can mislead users. Policy interventions can mitigate these risks by embedding epistemic responsibility into design standards and professional guidelines. Despite these principles, agency attribution cannot be eliminated entirely. Anthropomorphism is a deep-seated cognitive bias: humans reflexively treat conversational systems as social partners (Nass & Moon, 2000, p. 31). The goal is therefore mitigation, not eradication. Transparency, agency retention, and traceability reduce the risks but cannot fully prevent users from attributing intentionality to LLMs. This limitation underscores the need for continuous vigilance. As *Waytz et al. (2014, p. 219)* show, anthropomorphism is automatic and persistent. Epistemically responsible LLM use must therefore combine design interventions with educational and policy safeguards. The challenge is not to eliminate agency attribution but to contain its epistemic consequences, ensuring that users remain critical, responsible, and aware of the limits of machine-generated outputs.

Conclusion

The position of this paper is clear and assertive that Large Language Models (LLMs) such as ChatGPT, Claude, and Gemini are not epistemic agents, and treating them as such systematically degrades the foundational conditions of human knowledge. The central argument advanced here is that the epistemic challenge is not whether machines can think, know, or testify, but how humans interact with them and project agency onto systems that lack it. Three mechanisms converge to produce epistemic deskilling: uncritical acceptance, failures of epistemic vigilance, and misplaced testimonial trust. These mechanisms arise because users interpret fluent outputs as testimony, even though they are generated by statistical models without belief, sincerity, or responsibility (Dennett, 1987, p. 17; Miracchi, 2017, p. 45). Dennett's account of the intentional stance clarifies that attributing beliefs and desires is a pragmatic strategy for prediction, not a claim about metaphysical

reality. When applied to LLMs, this stance leads to the dangerous illusion that linguistic fluency equals epistemic reliability. Floridi's (2011, p. 89) concept of the infosphere reframes the problem. For him, LLMs are not epistemic agents but epistemic environments. They reshape the ecology of knowledge, altering how humans acquire, evaluate, and transmit information. The danger is not technological inevitability but social-epistemic degradation, where plausibility replaces truth, persuasion replaces justification, and optimization replaces reliability.

This paper contributes three major insights. First, it differentiates epistemic agency from non-epistemic agency, showing how the latter undermines justificatory transparency, normative integrity, and reliability stability (Greco, 2006, p. 173; Zagzebski, 1996, p. 167). Second, it details the mechanisms of epistemic degradation, including probabilistic reasoning, alignment protocols, and commercial optimization (Bender & Gebru, 2021, p. 14; Gabriel, 2020, p. 56). Third, it proposes safeguards for epistemic integrity, emphasizing transparency of non-agency, retention of user responsibility, and source traceability (Nass & Moon, 2000, p. 31; Lackey, 2008, p. 35). Therefore the future of knowledge will not depend on whether machines can think, but on whether humans can continue to think well with machines. Preserving epistemic agency requires cultivating intellectual virtues such as critical reflection, vigilance, and responsibility (Zagzebski, 1996, p. 167; Greco, 2006, p. 95). The emerging field of Human-AI Epistemology must integrate reliabilism, virtue epistemology, and social epistemology to develop frameworks for responsible knowledge practices in environments saturated by fluent non-agents (Clark & Chalmers, 1998, p. 12; Suazo et al., 2026, p. 7). Finally, LLMs generate text that mimics testimony but lack belief, sincerity, or responsibility. Treating LLMs as epistemic partners, humans risk surrendering their own agency. The danger is epistemic deskilling, and the task is to preserve human epistemic integrity through transparency, accountability, and virtue. The conclusion is that epistemically responsible use of LLMs is possible, but only if humans recognize their non-agency and cultivate the virtues necessary to sustain knowledge in the age of machine-generated language.

REFERENCES

- Ayoub, A., et al. (2023). *Use of ChatGPT in medical education: Opportunities and risks*. JMIR Medical Education, 9(1), 6.
- Brogaard, B. (2023). *Epistemic responsibility and reliabilism revisited*. Synthese, 201(4), 1123–1145.
- Burge, T. (1993). *Content preservation*. Philosophical Review, 102(4), 457–488.
- Clark, A., & Chalmers, D. (1998). *The extended mind*. Analysis, 58(1), 7–19.
- Dennett, D. C. (1987). *The intentional stance*. Cambridge, MA: MIT Press.
-, (1988). *Précis of The Intentional Stance*. Behavioral and Brain Sciences, 11(3), 495–505.
-, (1989). *The intentional stance (expanded essays)*. Cambridge, MA: MIT Press.
-, (1991). *Consciousness explained*. Boston, MA: Little, Brown and Co.
- Floridi, L. (2011). *The philosophy of information*. Oxford: Oxford University Press.
- Greco, J. (1992). *Reliabilism and skepticism*. Philosophical Studies, 65(1), 79–96.
-, (2006). *Virtue epistemology: Success from ability*. Philosophical Studies, 130(1), 171–190.
- Kumar, S., et al. (2023). *Generative AI and fabricated citations: Risks for academic integrity*. Accountability in Research, 30(2), 6–15.
- Lackey, J. (2008). *Learning from words: Testimony as a source of knowledge*. Oxford: Oxford University Press.
- Lewandowsky, S., et al. (2020). *Automation bias and epistemic vigilance in digital systems*. Cognitive Science, 44(8), 14–29.
- Miracchi, L. (2017). *Epistemic agency and intentional control*. Philosophical Studies, 174(1), 45–67.
- Nass, C., & Moon, Y. (2000). *Machines and mindlessness: Social responses to computers*. Journal of Social Issues, 56(1), 31–44.
- Nwagboso, C., & Atubab, J. (2026). *Academic erasure and AI reliance in higher education*. Teaching in Higher Education, 31(2), 3–5.

- Olaniyan, O., et al. (2026). *Global disparities in student reliance on LLMs*. *Frontiers in Human Dynamics*, 4(1), 12.
- Schreieder, M., et al. (2026). *Stylistic cues and agency attribution in LLMs*. *ACL Anthology*, 30956–30970.
- Sperber, D., et al. (2010). *Epistemic vigilance*. *Mind & Language*, 25(4), 361–379.
- Suazo, J., et al. (2026). *Cognitive offloading and automation bias in AI-assisted learning*. *Frontiers in Education*, 11(1), 7.
- Surden, H. (2023). *LLMs and legal reasoning: Risks of misplaced trust*. *Harvard Journal of Law & Technology*, 36(1), 112.
- Thorp, H. (2023). *ChatGPT is fun, but not an author*. *Science*, 379(6630), 313.
- Waytz, A., et al. (2014). *Anthropomorphizing technology: Social cues and agency attribution*. *Journal of Experimental Social Psychology*, 52(1), 219–225.
- Wührl, T., et al. (2026). *Politeness strategies and perceived agency in AI systems*. *FAccT '26 Proceedings*, 4.
- Zagzebski, L. (1996). *Virtues of the mind: An inquiry into the nature of virtue and the ethical foundations of knowledge*. Cambridge: Cambridge University Press.