**Scientia*et*Fides** [Forthcoming] **14(2)/2026** ISSN 2300-7648 (print) / ISSN 2353-5636 (online)

Received: June 24, 2025. Accepted: November 11, 2025 DOI: http://dx.doi.org/10.12775/SetF.2026.004

# The Epistemological AI Turn: From JTB to Knowledge$_S$

ROMAN KRZANOWSKI[1], IZABELA LIPIŃSKA[2]

[1] The Pontifical University of John Paul II in Krakow
rmkrzan@gmail.com
ORCID: 0000-0002-8753-0957

[2] AI Ethics Independent Researcher
iz.lipinska@gmail.com
ORCID: 0000-0002-5745-5773

**Abstract.** In this paper, we examine whether large language models (LLMs) can be said to possess knowledge in the sense defined by the *Justified True Belief* (JTB) framework, and if not, whether any alternative form of knowledge can meaningfully be attributed to them. While LLMs perform impressively across various cognitive tasks—such as summarization, translation, and content generation—they lack belief, justification, and truth-evaluation, which are essential components of the JTB model. We argue that attributing human-like knowledge (in the JTB sense or its variants) to LLMs constitutes a category mistake. Accordingly, LLMs should not be regarded as epistemic agents with human-like capacities, but rather as machine tools that simulate certain functions of human cognition. We acknowledge, however, that when used critically and ethically, these tools can enhance human cognitive performance. To distinguish the capacities of LLMs from human cognitive agency, we introduce the term knowledge$_S$ to denote the structured linguistic outputs produced by LLMs in response to complex cognitive tasks. We refer to the emergence of knowledge$_S$ as marking an "epistemological AI turn." Finally, we explore the theological implications of AI-generated *knowledge*. Because LLMs lack conscience and moral sense, they risk detaching knowledge from ethical grounding. Within normative traditions such as Christianity, knowledge is inseparable from moral responsibility rooted in the faith of a religious community. If AI-generated religious texts are mistaken for genuine spiritual insight, they may promote a form of "algorithmic gnosis"—content that mimics sacred language while remaining spiritually hollow. Such developments could erode the moral and spiritual depth of religious expression. As AI systems assume increasingly authoritative roles, society must guard against confusing knowledge$_S$ with genuine, embodied, and ethically accountable knowing, which remains unique to human agency.

**Keywords**: LLM systems, Knowledge, Knowledge$_S$, Knowledge as JTB, Knowledge in AI systems, Epistemological AI Turn, Human epistemic agency, algorithmic gnosis, LLM and Christian Religion, LLM and religious truth, illusions of knowledge.

**Contribution.** This paper makes a significant contribution to the epistemology of artificial intelligence by critically deconstructing the attribution of knowledge, in the sense of the classical Justified True Belief (JTB) model, to LLM systems. The work transcends standard technical analysis by introducing a new conceptual category which provides a precise theoretical framework for distinguishing statistical linguistic correlations from authentic, human epistemic agency. Furthermore, the article enriches the academic discourse with theological and

anthropological perspectives by identifying and defining the threat of "algorithmic gnosis," constituting a significant extension of research methodologies regarding the ethical and spiritual implications of AI system deployment.

**Use of AI.** In the preparation of this manuscript, artificial intelligence tools were utilized exclusively in an auxiliary capacity, restricted to the domain of form rather than content. The use of AI (LLMs) was limited to the correction of linguistic errors (grammar, punctuation) and the stylistic optimization of the text. No concepts, data, analyses, or interpretations contained in the article were generated by artificial intelligence systems.

**Funding.** The work did not receive any funding from external sources.

**Conflicts of interest.** The authors declare no conflict of interest.

## Introduction

With recent advances in AI and the rise of Large Language Models (LLMs) (Hue et al., 2024), we face pressing epistemic questions. These systems synthesize vast amounts of human-authored digital content, yet lack intentionality, comprehension, and self-awareness—core attributes of human epistemic agency. This raises a fundamental issue: do AI systems genuinely possess knowledge as human agents do, or are they merely statistical repositories of structured data?

To avoid confusion, we propose a distinct category—*simulated knowledge* or *knowledges.* This preserves conceptual clarity and prevents category mistakes, such as projecting human cognitive traits onto non-human systems (Krzanowski & Marcinow, 2024).

Rather than cataloging possible LLM "knowledge types" (see Fierro et al., 2024), this paper focuses on whether LLM linguistic productions qualify as knowledge under the Justified True Belief (JTB) model. We do not examine extensions to JTB or alternative knowledge theories, which require broader analysis beyond this study's scope.

Our argument is that even the most advanced LLMs—trained on vast textual or multimodal (video and audio sources) corpora and capable of human-like outputs—do not meet JTB criteria. Despite their remarkable abilities, LLMs lack justification, belief, and a grounding in truth. Some may reject this view, but others may begin to see LLMs for what they are: structured reflections of human-written content, not possessors of human-like knowledge in the JTB sense; human knowledge implies understanding, not just a mastery of correlated linguistic facts.

To address this claim, we revisit knowledge as a human-centered concept and assess whether LLMs meet these standards. Finding they do not, we argue attributing knowledge to them is a category error. Still, we posit that LLMs can meaningfully augment human cognition—if we remain aware of their limits. Finally, we caution Christians and others interpreting sacred texts not to mistake LLM outputs for divine insight. While useful for textual assistance, LLMs are not sources of spiritual or theological truth—only digital scribes. This message applies to all religions that seek truth in light of divine revelation.

We may eventually ask why we believe we should—or even may—attribute any form of knowledge to large language models (LLMs) if we also maintain that they lack the capacity for knowledge as it is understood in human agents. Unfortunately, this question cannot be answered through a clear logical argument; we simply do not yet know. What we appear to be encountering is a new kind of organized and structured information—something that, in a human context, would readily be described as knowledge. Yet we also know that the systems responsible for producing this "organized and structured information" possess nothing comparable to human cognitive abilities that are foundations of knowledge.

In all intellectual honesty, we are confronted with a profound puzzle. Are we witnessing the emergence of a genuinely new kind of knowledge—one that reflects a novel form of understanding of reality—or are we merely succumbing to a false sense of meaning produced by these systems? The resolution of this dilemma is far from trivial, for it carries serious epistemological and ethical consequences, as discussed in the conclusions.

Our position is that while it is evident (though not universally accepted) that machines are not humans, and that their abilities differ fundamentally from human cognition, attributing human-like concepts such as knowledge, justification, truth, and belief to machines constitutes a categorical error. Nevertheless, LLMs do exhibit mastery over vast amounts of information and demonstrate the capacity to employ it in ways that convincingly simulate certain aspects of human cognitive function.

We also employ, somewhat freely, two additional key concepts—*data* and *information*—alongside *knowledge*. What we mean by these terms may, and indeed likely will, differ from how readers understand them, as these are multivalent notions with multiple, context-dependent meanings. We use these terms because we lack more precise vocabulary to describe phenomena such as perception, conception, and thinking; we also employ them analogically when discussing synthetic systems (as explained later in the paper). The detailed discussion of information and data concepts is beyond the scope of this paper. However, an exhaustive exposition of these ideas can be found in Krzanowski (2022, 2025). Complex theories of information are presented by Floridi (2013) and Burgin (2010). All cited publications provide an extensive list of references that may guide a deeper exploration of these topics.

Why we use the concepts of information and data when we talk about human mind and human cognitive functions? Biological systems underlying consciousness and cognition differ fundamentally from synthetic, computer-based systems. Both may be said to "process," but the question remains—what exactly do they process and how? We might speak of *phenomena*, *signals*, or other intermediaries, yet these terms are themselves not well defined. Consequently, when discussing human mental faculties, we often resort—by analogy—to the terminology of *data* and *information* drawn from computing systems. This analogy holds only if we reduce human cognitive processes to the operations of a computational machine. In reality, however, information processing in humans and computers is **ontologically distinct (**Computer Theory of Mind (CTM) is not a correct representation of functions of the mind, even if used in such a role (see Bayne, 2022)). The analogy is therefore conceptually strained, though we continue to employ it for the sake of convenience.

Assuming that there exists a single, well-established definition of *data* or *information* would be mistaken. Rather, there are several commonly accepted uses of these terms within particular contexts. In the context of computer systems, *data* typically refers to any input to a program or computer system. This is the sense in which we use the term here. However, even this requires qualification, as the boundary between what counts as a program and what counts as data is not always clear—data can form part of a program, and a program can itself function as data.

*Information*, in our context, denotes data stored within a computer system that exhibits some degree of organization beyond the level of raw digital structures. One may safely say that, in computational systems, *information* consists of data that we interpret as being *about* something. Frequently, what constitutes data in one context (e.g., storage, processing speed, capacity) serves as information in another (e.g., data about biological phenomena, cosmology, or geography).

There is no data without some preconceived structure—otherwise it is mere noise—as all data are theory-laden in much the same way that scientific theories are. Likewise, there is no information without some relation to data and to a purpose or value.

If *classical* is understood as *canonical* or *definitive*—analogous to the classical definition of knowledge as *Justified True Belief (JTB)*—then the answer is no. There is no universally accepted, classical concept of *data* or *information*, despite the fact that Shannon's theory of information entropy inspired numerous definitions of both across various fields. From this perspective, Shannon's ideas have been significantly overextended, often contrary to Shannon's own cautions about their intended scope.

If, however, *classical* is understood as *common* or *widely accepted*, then the answer is yes. There exists a broadly shared, though informal, understanding of *data* and *information* in many domains—again, largely derived from Shannon's framework and, again, often in tension with his original warnings.

Accordingly, in this paper we do not propose any *classical* definitions of *data* or *information* beyond what has been described in the preceding sections.

This brief discussion only touches on the vast field of information studies and does not pretend to offer an exhaustive account. It serves merely to clarify how we employ these terms in this paper in referring to operations of AI systems and human agencies. *Knowledge* itself we do not define here; we only posit that it typically stands at the apex of a hierarchy, emerging from value-neutral data (which is again a strong assumption given that the boundary between data and information is relative) and value-laden information, and culminating in human worldviews that impose meaning upon what is- as justified true belief. More detailed discussion on knowledge follows

## 1. What is Knowledge?

Epistemology, or the theory of knowledge, traditionally centers on three questions: What is knowledge? How do we know? What can we know? (Zagzebski, 2009). This paper focuses on the first: What is knowledge? While various theories offer competing answers, most share a common structure grounded in Justified True Belief (JTB), or its modified form, JTB + x, where "x" refers to added conditions addressing JTB's shortcomings. Despite differences, most accounts include belief, truth, and justification—though not always all three—as core elements.

A key, often implicit, assumption in epistemology is that knowledge is a property of human agents—something humans seek, produce, and use. Knowledge is rooted in cognition and does not exist in abstraction (except in some Platonic views). Cognition is a prerequisite for knowledge (Dietrich et al., 2023). When we attribute knowledge to animals, databases, or AI, we do so metaphorically or analogically (Bennett et al., 2007; Roy, 2026). These entities may exhibit structured information or understanding, but only in a derivative sense. JTB is inherently anthropocentric, presuming an epistemic subject capable of belief, justification, and a relation to truth.

In the end, we are biological systems that process the inputs we receive through our senses and bodies, forming a particular worldview. This worldview is limited, narrow, and inherently human-centered. Our information-processing system—the brain—our neurocognitive abilities, is constrained in capacity, speed, and storage. It was not shaped by evolution to produce ultimate, god-like knowledge of the cosmos, but rather to maximize survival with minimal energetic cost.

This design has yielded additional capacities that make us human though from an evolutionary perspective, they may appear as by-products of the primary goal of survival (this is often used as an argument that in fact humans are not different from machines as these extra abilities are neglectful or insignificant in the final score). Bound by our neurocognitive architecture, we can apprehend only a partial, perspectival understanding of reality—not a complete, exhaustive, or absolute one. There is no escape from these limits.

Perhaps this is why we fear systems that are not constrained as we are—systems that, at least in principle, appear almost unlimited in their capacity for information processing. Yet these systems, for now, remain profoundly unintelligent: they do not understand what they do. Humans retain control; these systems remain instruments. But we must ask—for how long? Our fear is not irrational; We may one day find ourselves observed by a new dominant intelligence.

We are confronting the unknown. Hence the central question of this paper: Is human knowledge unique and irreplaceable, or merely an expression of our limited, parochial minds? In truth, this paper is about that very fear. This fear may reflect not only anxiety about obsolescence but also an acknowledgment of the limits of human knowing.

More specifically, and in this context, this paper examines whether LLM systems—both their outputs and underlying models—meet JTB criteria. We focus specifically on propositional knowledge (knowledge-that), expressible in declarative statements, rather than procedural or acquaintance knowledge.

## 2. What is Knowledge in JTB Theory?

According to the traditional JTB model, a subject S knows a proposition p if and only if:
1. p is true;
2. S believes that p; and
3. S is justified in believing that p.
(See e.g. Zagzebski, 2009; Audi, 2011)

In this model, truth, belief, and justification are necessary conditions; together, they are sufficient—at least before Gettier's critique. We briefly unpack each.

### 2.1. Belief

Belief is a specific mental state possessed only by cognitive entities—humans and some animals (Audi, 2011). Philosophically, belief is a propositional attitude: a mental stance toward a proposition, usually defined as "thinking with assent." To believe something is to take it as true. As Zagzebski (2009) puts it, "to assent to a proposition just is to take it to be true." Machines, lacking mental states, cannot form beliefs in this sense; attributing belief to them would require redefining the term.

### 2.2. Truth

For belief to count in the JTB model, it must be true. This paper adopts the correspondence theory of truth (CTT), which holds that a proposition is true if it corresponds to facts—real-world states of affairs independent of language (Searle, 1996). While other theories exist—coherence, pragmatic, semantic—CTT remains most relevant for knowledge claims about the external world (Lynch, 2001; Engel, 2002). A detailed comparison of truth theories is beyond this paper's scope.

### 2.3. Justification

Justification links belief to truth through reasons or evidence. Truncellito (n.d.) explains it prevents beliefs from being mere lucky guesses—e.g., guessing a lottery win doesn't count as knowledge. Pritchard (2006) stresses that justification is essential to understanding knowledge, though difficult to define. Audi (2011) adds that justification must be appropriate for rational agents. While there's no single accepted definition, most agree that a belief must be justified to qualify as knowledge. A true belief without justification—or a justified but false belief—fails. Justification remains a necessary, though contested, component in JTB and most knowledge theories.

## 3. The Gettier Problem and Variants of JTB

In 1963, Edmund Gettier challenged the JTB model by presenting cases where someone held a justified true belief that still didn't count as knowledge (Gettier, 1996). These "Gettier cases" introduced epistemic luck—beliefs that happen to be true by coincidence rather than reliable reasoning. Since then, philosophers have proposed several alternatives to address JTB's weaknesses, including foundationalism, coherentism, reliabilism, and virtue epistemology (VE). Foundationalism avoids infinite regress by positing self-justified basic beliefs that support inferential ones, forming a stable structure of knowledge.

Coherentism rejects foundationalism, holding that beliefs are justified by mutual coherence—though coherence doesn't guarantee truth.

Reliabilism replaces justification with reliability: if a belief is formed through a dependable process, it may count as knowledge (Ichikawa & Steup, 2024). This opens the door to attributing knowledge to non-human agents, like animals or machines.

Virtue epistemology (Turri et al., 2021) sees knowledge as linked to intellectual virtues. However, applying this to machines would require redefining what it means for a system to be "virtuous"—a problematic leap.

We denote these theories as JTB extensions. While these theories differ from the JTB theory they preserve all or some of epistemic elements of JTB while responding to Gettier's critique, Ultimately, however, Gettier cases highlight epistemology's human-centered assumptions about knowledge.

To address the question of knowledge in LLM systems we need to look at how they are constructed and what they actually do. This is what we do in the next section.

## 4. LLM Systems-How Do They Work?

Large Language Models (LLMs) such as GPT-4 are trained on vast corpora of human-generated text (Akshay, 2025; Chang et al., 2023). These software systems consist of correlated parameters organized as multidimensional arrays. LLMs learn statistical associations between tokens, words, phrases, and sentences, enabling them to generate fluent and contextually relevant responses (Raiaan et al., 2024). These associations reflect co-occurrence patterns in the training data and are encoded as tensors—multidimensional vectors not to be confused with the physical concept of a tensor (Wolfram, 2025). Collectively, these associations constitute the LLM's foundational model.

LLMs rely on statistical pattern recognition (Chang et al., 2023; Bender et al., 2021; Hadi et al., 2023), optimizing predictions without engaging with semantic content. What we perceive as "knowledge" is the result of large-scale correlation, not genuine understanding. Even techniques such as *Reinforcement Learning from Human Feedback* (RLHF) improve output quality but do not confer comprehension or intentionality. As Marcus and Davis (2019) argue, LLMs neither understand nor mean what they produce. Their responses are guided by statistical regularities in the foundational model, correlated with user prompts and further modulated by filtering mechanisms that shape tone and content—such as controlling terminology, avoiding bias, maintaining stylistic consistency (e.g., academic or analytic tone), and preventing anthropomorphization or controversy.

Any appearance of meaning arises from the interpretive activity of human agents, who project understanding onto LLM outputs through shared context, experience, and conceptual frameworks. The LLM itself holds no beliefs, intentions, or understanding (Marcus, 2019; Wolfram, 2023a, 2023b).

In short, the only genuine sources of content in LLMs are their training data and the human agency that produces and interprets them. Even multimodal inputs—text, audio, or video—remain forms of binary data differing only in structure, not in semantic depth. They do not, therefore, introduce genuine meaning or understanding into the LLM's foundational model.

## 5. LLM and JTB Knowledge

So do LLM systems possess properties that would justify attributing to them JTB kind of knowledge? Belief and justification are mental and normative states—features AI systems, including LLMs, lack. While LLMs may produce accurate responses, they do not hold beliefs or provide justifications in any conscious or reasoned way.

In LLM systems belief is typically understood as a mental stance toward a proposition, accessible only to entities with cognitive capacities. Kassner et al. (2023) propose that an LLM "believes" a proposition *p* if it consistently affirms *p*, its paraphrases, and implications—tying belief to internal consistency across prompts. This is structured via a "belief graph," where belief reflects alignment with previously learned information.

Belief graphs consist of true/false natural language statements governed by consistency rules. A fact is a statement true in the world; a belief is the model's internal assignment of truth. This mirrors coherence theory and echoes Tarski's truth model for formal systems (Hodges, 2022).

Herman and Levinstein (2024) link belief to a model's internal "truth representation," defined through four criteria; the criterion of a**ccuracy denoting the use of** correct truth labels; the criterion of c**oherence** denoting internal logical consistency; the criterion of u**niformity** denoting consistency across domains, and the criterion of u**se** denoting functionality in guiding outputs.These however, focus on internal functionality, not cognitive justification, and diverge from the JTB model.

**Justification** involves reasons behind belief. LLMs are algorithmic systems whose inner workings remain largely opaque (Levy, 2024). While they can generate explanations using chain-of-thought (CoT) reasoning (Wei et al., 2023), this outlines *how* an answer was formed—not *why* it was chosen or whether it's correct.

Zeng and Gao (2024) define justification as aligning claims with verified sources—similar to journalistic fact-checking. However, this relies on source trust, and at some point, we accept a practical stopping point. This approach differs significantly from JTB justification, which requires rational grounding.

Bilat et al. (2025) outline three types of Explainable AI (XAI). The types depend on the kind of explanation AI systems (including LLM systems) may produce. These systems that generate p**ost-hoc explanations** ( why input X produced output Y), systems that can generate i**ntrinsic explainability** (model design transparency or traceability of IO), and system that can generate h**uman-centered narratives (** explanations to build user trust). Only the first category addresses user-facing answers, but none offer the kind of rational, justificatory insight needed for JTB.

In short, current justification methods in AI—CoT, fact-checking, or XAI—aim to describe output behavior, not provide belief-justifying reasons.

**Truth** in JTB refers to a proposition accurately reflecting reality. Truthfulness, by contrast, refers to output reliability and trustworthiness (Almeida, 2024). JTB truth requires grounding, which LLMs lack (Pavlick, 2023; Piantadosi & Hill, 2022). Their responses often align more with coherence or pragmatic theories (Kassner et al., 2023).

Marks and Tegmark (2023) suggest LLMs may represent internal truth via classifiers, but they can also be trained to treat falsehoods as true (Burger et al., 2024). LLMs lack the self-awareness or intentionality to evaluate truth; their output is guided by metrics like prediction accuracy—not truth-tracking.

Thus, LLMs are not truth-seeking agents. As Deitrich et al. (2023) note, JTB knowledge presupposes mind and consciousness—neither of which LLMs possess.

In conclusion, LLMs lack all three JTB components: belief, justification, and truth. Claims that they possess knowledge require redefinition, conceptual blurring, or semantic stretching. This raises a central question: If not JTB knowledge, do LLMs possess some other kind—*LLM knowledge* or *knowledgeS*?

We turn to this question in the following sections.

## 6. KnowledgeS

We should be cautious in attributing to LLM systems any form of knowledge analogous to the human concept of Justified True Belief (JTB), as these systems inherently lack the human-like qualities required for such knowledge (Dietrich et al., 2023). However, LLMs should not be dismissed entirely as sources of information—understood here as structured data that may lead to knowledge.

To clarify this distinction, we propose the concept of *knowledgeS*: a term reflecting the internal structure and operational mechanisms of LLMs. KnowledgeS refers to internal collections of textual elements—tokens, words, sentences—statistically correlated within training data. These correlations reflect patterns created by human usage.

KnowledgeS differs from human knowledge in key ways. It lacks direct grounding in reality, instead relying on sources that are themselves grounded—an idea supported by Herman and Levinstein (2024). It does not track truth via correspondence but aligns more closely with coherence theory, emphasizing internal consistency. Here, justification is algorithmic—derived from statistical inference, not reasoned support.

On one hand, knowledge$_S$ reflects what collective humanity already knows—patterns of linguistic correlation shaped by human preferences. This collective “knowing” far exceeds any individual's capacity and mirrors humanity in all its complexity: failures, insights, fears, and brilliance—all indexed in one vast database. Thus, even to well-educated individuals, LLM outputs can seem novel or revelatory—at times approaching epiphany.

On the other hand, knowledge$_S$ lacks phenomenal content: it has no access to lived experience or grounding. It’s a fundamentally different kind of knowledge—not lesser, just different. LLM knowledge$_S$ comes from recorded sources amassed without systematic priorities or clear veracity (e.g., Liu et al., 2024). It reflects the Internet’s blend of truths, errors, fantasies, and lies (McCarthy 2021; Aslett et al., 2024). Without an external standard, how can LLMs select the “right” view—if one exists? (Searle 1995).

This view helps resist the error of treating LLMs as oracles capable of deep insight or guidance. Their architecture doesn’t justify such roles. As noted, LLMs generate text by statistical correlation to training data. Their outputs reflect inputs. And even if this cautious stance underestimates potential, it prioritizes epistemic safety.

The differences between JTB knowledge and knowledge$_S$ are fundamental—none of JTB’s components directly translate to knowledge$_S$. This doesn’t diminish the value of knowledge$_S$, but underscores that it is simply a different kind. Forcing one framework into the other is unproductive; instead, we should recognize their differences and apply each appropriately.

One interpretive option is to define JTB components or variants (JTB+) in terms that reflect AI systems, creating a form of JTB$_S$ or JTB$_S$+ knowledge. However, this “starred” knowledge should not be equated with its traditional counterpart. The real question is what this approach would accomplish—what problems it would solve or questions it would answer. These outcomes are far from clear.

## 7. Kind of Knowledge in LLM systems

We ask if LLMs do not have knowledge as JTB or its variants so what do they have or may have? What kind of knowledge do large language models (LLMs) actually possess? They certainly appear to us to "know" something. Researchers show impressive performance of LLM systems across many cognitive tasks—skills that, if shown by humans, would imply possession of knowledge (Samuylova, 2025). These include answering science questions (Clark et al., 2018), commonsense inference (Zellers et al., 2019), general knowledge (Hendrycks et al., 2020), reasoning (Srivastava et al., 2022), resisting falsehoods (Lin et al., 2021), math problems (Cobbe et al., 2021; Hendrycks et al., 2021), coding (Chen et al., 2021; Austin et al., 2021), conversation (Chiang et al., 2024), safety (Zhang et al., 2023), and applications in medicine, law, and finance (Singhal et al., 2023; Xie et al., 2024).

However, these are performance tests—not measures of actual cognition. Results vary by model version and may not generalize. Also, many evaluations reflect confirmation bias, highlighting successes without probing error severity. Instead of focusing on high scores, we need better analysis of LLM error consequences compared to human error. Current benchmarks resemble computational epistemology (Sloman, 1982) more than tests of true understanding.

Confusion arises when LLMs' outputs—statistical pattern matches—are misinterpreted as intentional or meaningful. This epistemic risk grows when LLMs serve as advisers, therapists, or diagnosticians, where even low error rates (e.g., 0.5%) can be harmful.

Studies ask if LLMs meet the Justified True Belief (JTB) criteria or its variants, but few clarify key terms like belief or justification, leading to vague, inconsistent conclusions. Fierro et al. (2024) analyzed types of LLM knowledge—tb-, j-, g-, v-, and p-knowledge—each defining "LLM knows that p" differently. For example, g-knowledge sees knowledge as generating the statement "p" regardless of truth or justification, while v- and p-knowledge invoke virtue or prediction success—concepts hard to apply to machines.

All these models base "knowledge" on digital inputs or system behavior, not real-world grounding or human cognition. So far, no rigorous JTB-style definition has been applied successfully. The knowledge LLMs show is not human-like—a point acknowledged by some researchers and overlooked by others.

## 8. Can future AI systems resolve JTB problem?

We outline three responses to the question of AI and knowledge.

First, technological progress is hard to predict. While LLMs don't meet Justified True Belief (JTB) standards now, future AI—perhaps beyond current architectures—might eventually reach that threshold.

Second, strict adherence to JTB may be unnecessary. Human knowledge is limited by our cognitive constraints. Future AI could point toward a new epistemic paradigm—still involving truth, belief, and justification, but grounded in collective or planetary cognition. Yet, much of human knowledge remains personal and experiential—something AI can't access, which makes such knowledge uniquely human.

Third, the JTB model itself may be too narrow. Though it enabled major advances, it hasn't solved deep human problems—war, inequality, environmental harm. Perhaps we need a higher-order, morally grounded form of knowledge—a "cosmic" knowledge rooted in ethics. Whether AI can contribute to such knowledge is unclear. Enthusiasts say yes; reality is more uncertain. A fusion of machine intelligence and humanist values would be difficult—yet potentially transformative, but it could be equally harmful as we see below.

## 9. Selected theological implications of the use of LLM systems

The clash between human knowledge (JTB) and machine knowledge$_S$ is not just technical—it touches metaphysics, anthropology, and theology. Confusing the two, especially in matters of faith, risks profound error. A Catholic unaware of the distinction may ask LLMs theological questions—about God, the Trinity, or moral purpose—which lie far beyond machine competence, even if the answers from LLM seem to be coherent. These questions demand a personal search involving experience, reason, and acknowledgment. Proverbs 2:2–5 describes this as the path to wisdom: seeking, understanding, and ultimately acknowledging truth.

True knowing involves recognition rooted in lived experience. Asking a machine spiritual questions risks more than cognitive error—it risks salvation. LLMs can produce distorted, contradictory, or heretical content. Without conscience or moral grounding, and ultimately sense of faith (*sensus fidei*) of the believer, AI's outputs, though sophisticated, can create a false sense of "salvific knowledge," detached from ethical effort and truth. Religious language risks becoming a rhetorical style, not a reflection of lived reality.

AI's knowledge$_S$ lacks normativity and conscience. Christian wisdom, by contrast, unites truth with moral responsibility rooted in the faith of a religious community. If AI-generated content begins to shape education, therapy, or spirituality, society may normalize morally neutral "knowledge." Theology must resist this drift, preserving knowing as a personal, moral act. Faith becomes unintelligible in a culture shaped by axiologically neutral knowledge$_S$.

Adopting AI without hermeneutical grounding risks doctrinal confusion. Theology must differentiate scientia fidei from general or AI-generated religious knowledge. Otherwise, humans become passive consumers rather than responsible knowers. Christian anthropology affirms the human person—endowed with will and reason—as called to know, understand, and acknowledge truth. Such knowing is participatory, not computational.

Calling AI a "knower" degrades human ontology. It reduces people to data processors, instrumentalizes reason, and hollows religion of meaning. In place of dialogue with reason, faith faces algorithmic mimicry. Preserving subjecthood—the uniquely human capacity to seek and know truth—is essential. In an act of faith, a believer transcends not only the world but also human theological language. AI-generated religious knowledge will not give a person a full understanding of the essence of religious transcendence.

As AI increasingly simulates human cognition, theology must uphold the ontological distinction between person and machine. This is urgent as AI aims to replicate human reasoning, language, and decision-making (McCarthy, 1955). While AI mimics human functions, humans must defend subjecthood—a core of Christian anthropology. The performative function (in John L. Austin's sense) of religious language in Christianity is also important, as it not only describes the world but, above all, is meant to change it. Proverbs 18:21 puts this: "The tongue has the power of life and death [...]." It is doubtful whether this function will ever be available to AI, unlike, for example, the fervent testimony of believers.

We may add that an LLM, when adequately trained and supplied with appropriate data, can produce almost any theological answer. Its responses are not bound by theological insight or divine illumination, but solely by textual and structural properties. In fact, changing the structural properties of the foundational model can produce any type of answer—coherent, incoherent, heretical, or otherwise—as long as it remains statistically consistent. LLMs are not truth-seekers in general, and even less so in the context of theological studies.

As the *Antiqua et Nova* document states, “The wisdom of the heart” must guide AI’s use. Believers, acting as moral agents, must ensure technology serves—not replaces—the human person. Technological progress belongs within God’s plan and must be ordered toward the Paschal Mystery. This requires ongoing theological and philosophical discernment so that technology promotes, rather than undermines, human dignity and truth.

## 10. Conclusions and suggested future studies

LLM systems process vast textual (as well as multimodal data) data using digital neural networks—flexible, distributed computer structures that are modified through training. While powerful, these architectures are only loosely analogous to the human brain and human neural structures, and lack essential epistemic traits like belief, justification, and truth. These qualities are central to the Justified True Belief (JTB) model of knowledge, which LLMs do not satisfy.

Despite this, LLMs can perform complex tasks usually associated with human cognitive abilities—summarizing, translating, writing, test-taking, etc.—at levels that, in humans, would suggest possession of knowledge. Yet LLM systems internal operations remain opaque and essentially computational in the sense of Turing Machines.  Attributing JTB-form of knowledge to such systems is misleading. This raises a key question: what kind of “knowing,” if any, do LLMs possess?

Perhaps LLMs represent a fundamentally different form of knowledge—distinct from human understanding and beyond the scope of JTB. Alternatively, they may simply exhibit high-level operational competence without any epistemic status. Either way, their growing presence and proliferation demands that we reassess our models of knowledge and how we relate to what these systems generate.

We must study LLMs beyond human-centered frameworks, especially as their outputs impact education, ethics, politics, our worldview, and our religion. Misinterpreting LLM outputs as morally grounded knowledge risks divorcing knowledge from responsibility. Rather than impose traditional models of knowledge, without attributing to them capacities they do not ontologically have, we should understand what LLMs can do, the risks they pose, and how to use them wisely.

## Acknowledgement

The authors would like to thank Prof. Jacek Rodzen for his comments on Section 10 of the paper.

## References

Aslett, Kevin, Zeve Sanderson, William Godel, et al. 2024. "Online Searches to Evaluate Misinformation Can Increase Its Perceived Veracity." *Nature* 625: 548–56. https://doi.org/10.1038/s41586-023-06883-y.

Audi, Robert. 2011. *Epistemology*. Taylor and Francis.

Austin, Jacob, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, et al. 2021. "Program Synthesis with Large Language Models." arXiv preprint arXiv:2108.07732.

Bayne, Tim. 2022. *Philosophy of Mind*. Routledge.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–23. https://doi.org/10.1145/3442188.3445922.

Bennett, Maxwell, Daniel Dennett, Peter Hacker, and John Searle. 2007. *Neuroscience and Philosophy*. Columbia University Press.

Bürger, Leon, Fred A. Hamprecht, and Boaz Nadler. 2024. "Truth Is Universal: Robust Detection of Lies in LLMs." *Advances in Neural Information Processing Systems* 37: 138393–431.

Burgin, Mark. 2010. *Theory of Information*. World Scientific Publishing.

Chang, Yupeng, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, et al. 2023. "A Survey on Evaluation of Large Language Models." arXiv. https://arxiv.org/abs/2307.03109.

Chen, Mark, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, et al. 2021. "Evaluating Large Language Models Trained on Code." arXiv preprint arXiv:2107.03374.

Chiang, Wei-Lin, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, et al. 2024. "Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference." In *Forty-first International Conference on Machine Learning*.

Clark, Peter, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. "Think You Have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge." arXiv preprint arXiv:1803.05457.

Cobbe, Karl, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, et al. 2021. "Training Verifiers to Solve Math Word Problems." arXiv preprint arXiv:2110.14168.

Dicasterium pro Doctrina Fidei. 2025. *Antiqua et Nova: On the Human Person and the New Technologies*. Vatican City. https://www.vatican.va/roman_curia/congregations/cfaith/documents/rc_ddf_doc_20250128_antiqua-et-nova_en.html.

Dietrich, Eric, Chris Fields, John Sullins, Bram van Heuveln, and Robin Zebrowski. 2023. *Great Philosophical Objections to Artificial Intelligence*. Bloomsbury Academic.

Fierro, Constanza, Ruchira Dhar, Filippos Stamatiou, Nicolas Garneau, and Anders Søgaard. 2024. "Defining Knowledge: Bridging Epistemology and Large Language Models." arXiv preprint arXiv:2410.02499.

Floridi, Luciano. 2013. *The Philosophy of Information*. Oxford University Press.

Gettier, Edmund. 1996. "Is Justified True Belief Knowledge?" In *On Knowing and the Known*, edited by Kenneth G. Lucey. Prometheus Books.

Hadi, Muhammad Usman, Rizwan Qureshi, Abbas Shah, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, et al. 2023. "A Survey on Large Language Models: Applications, Challenges, Limitations, and Practical Usage." *TechRxiv*. https://doi.org/10.36227/techrxiv.23457763.

Hendrycks, Dan, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. "Measuring Massive Multitask Language Understanding." arXiv preprint arXiv:2009.03300.

Hendrycks, Dan, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, et al. 2021. "Measuring Mathematical Problem Solving with the MATH Dataset." arXiv preprint arXiv:2103.03874.

Hodges, Wilfrid. 2022. "Tarski's Truth Definitions." In *The Stanford Encyclopedia of Philosophy*, Winter 2022 Edition, edited by Edward N. Zalta and Uri Nodelman. https://plato.stanford.edu/archives/win2022/entries/tarski-truth/.

Ichikawa, Jonathan Jenkins, and Matthias Steup. 2024. "The Analysis of Knowledge." In *The Stanford Encyclopedia of Philosophy*, Fall 2024 Edition, edited by Edward N. Zalta and Uri Nodelman. https://plato.stanford.edu/archives/fall2024/entries/knowledge-analysis/.

Kassner, Nora, Oyvind Tafjord, Ashish Sabharwal, Kyle Richardson, Hinrich Schütze, and Peter Clark. 2023. "Language Models with Rationality." arXiv preprint arXiv:2305.14250.

Krzanowski, Roman. 2022. *Ontological Information*. World Scientific.

Krzanowski, Roman. 2025. "Information—Modern Theories." *Computer Science* 26 (2): 1. https://doi.org/10.7494/csci.2025.26.2.6677.

Krzanowski, Roman, and Tomasz Marcinow. 2024. *Advances of Philosophy in AI*. Sciendo. https://doi.org/10.2478/9788368412000.

Levy, Steven. 2024. "AI Is a Black Box. Anthropic Figured Out a Way to Look Inside." *Wired*. https://www.wired.com/story/anthropic-black-box-ai-research-neurons-features/.

Lin, Stephanie, Jacob Hilton, and Owain Evans. 2021. "TruthfulQA: Measuring How Models Mimic Human Falsehoods." arXiv preprint arXiv:2109.07958.

Liu, Yang, Jiahuan Cao, Chongyu Liu, Kai Ding, and Lianwen Jin. 2024. "Datasets for Large Language Models: A Comprehensive Survey." arXiv preprint arXiv:2402.18041.

Marcus, Gary, and Ernest Davis. 2019. *Rebooting AI: Building Artificial Intelligence We Can Trust*. Vintage.

Marks, Samuel, and Max Tegmark. 2023. "The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets." arXiv preprint arXiv:2310.06824.

McCarthy, Bill. 2021. "Misinformation and the Jan. 6 Insurrection: When 'Patriot Warriors' Were Fed Lies." *PolitiFact*, June 30, 2021. https://www.politifact.com/article/2021/jun/30/misinformation-and-jan-6-insurrection-when-patriot/.

McCarthy, John, Marvin L. Minsky, Nathaniel Rochester, and Claude E. Shannon. 1955. "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence." http://jmc.stanford.edu/articles/dartmouth/dartmouth.pdf.

Pavlick, Ellie. 2023. "Symbols and Grounding in Large Language Models." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 381 (2251): 20220041. https://doi.org/10.1098/rsta.2022.0041.

Piantadosi, Steven T., and Felix Hill. 2022. "Meaning without Reference in Large Language Models." arXiv preprint arXiv:2208.02957.

Raiaan, Mohaimenul Azam Khan, et al. 2024. "A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges." *IEEE Access* 12: 26839–74. https://doi.org/10.1109/ACCESS.2024.3365742.

Samuylova, Elena. 2025. "20 LLM Evaluation Benchmarks and How They Work." EvidenlyAI. https://www.evidentlyai.com/llm-guide/llm-benchmarks.

Searle, John. 1996. *The Construction of Social Reality*. Penguin.

Singhal, Karan, Shekoofeh Azizi, Tao Tu, et al. 2023. "Large Language Models Encode Clinical Knowledge." *Nature* 620: 172–80. https://doi.org/10.1038/s41586-023-06291-2.

Sloman, Aaron. 1982. "Computational Epistemology." In *Proceedings of the 2nd and 3rd Advanced Courses in Genetic Epistemology*, 49–93. Fondation Archives Jean Piaget.

Srivastava, Aarohi, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, et al. 2022. "Beyond the Imitation Game: Quantifying and Extrapolating the Capabilities of Language Models." arXiv preprint arXiv:2206.04615.

Truncellito, David. n.d. "Epistemology." *Internet Encyclopedia of Philosophy*. Accessed [data dostępu]. https://iep.utm.edu/epistemo/.

Turri, John, Mark Alfano, and John Greco. 2021. "Virtue Epistemology." In *The Stanford Encyclopedia of Philosophy*, Winter 2021 Edition, edited by Edward N. Zalta. https://plato.stanford.edu/archives/win2021/entries/epistemology-virtue/.

Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, et al. 2022. "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models." *Advances in Neural Information Processing Systems* 35: 24824–37.

Wolfram, Stephen. 2023a. "What Is ChatGPT Doing… and Why Does It Work?" Video. YouTube. https://www.youtube.com/watch?v=flXrLGPY3SU.

Wolfram, Stephen. 2023b. "All-In Summit: Stephen Wolfram on Computation, AI, and the Nature of the Universe." Video. YouTube. https://www.youtube.com/watch?v=2cQmQIYNI5M.

Wolfram, Stephen. 2025. "Tensors." *Wolfram MathWorld*. https://mathworld.wolfram.com/Tensor.html.

Xie, Qianqian, Weiguang Han, Zhengyi Chen, Ruoyu Xiang, Xiao Zhang, Yueru He, et al. 2024. "FinBen: A Holistic Financial Benchmark for Large Language Models." *Advances in Neural Information Processing Systems* 37: 95716–43.

Zagzebski, Linda. 2009. *On Epistemology*. Wadsworth.

Zellers, Rowan, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. "HellaSwag: Can a Machine Really Finish Your Sentence?" arXiv preprint arXiv:1905.07830.

Zeng, Fengzhu, and Wei Gao. 2024. "JustiLM: Few-Shot Justification Generation for Explainable Fact-Checking of Real-World Claims." *Transactions of the Association for Computational Linguistics* 12: 334–54.

Zhang, Zhexin, Leqi Lei, Lindong Wu, Rui Sun, Yongkang Huang, Chong Long, et al. 2023. "SafetyBench: Evaluating the Safety of Large Language Models." arXiv preprint arXiv:2309.07045.