UDC 81:1:[004.8
DOI https://doi.org/10.32782/2410-0927-2025-23-2

Iryna Biskub
Doctor of Philology, Professor of the Applied Linguistics Department, Lesya Ukrainka Volyn National University, 13 Voli Prospect, Lutsk, Ukraine, 43025
***ORCID:** 0000-0001-5844-7524*

Viktor Levandovskyi
Assistant teacher of the Applied Linguistics Department, Lesya Ukrainka Volyn National University, 13 Voli Prospect, Lutsk, Ukraine, 43025
***ORCID:** 0009-0002-6935-7107*

To cite this atricle: Biskub, I., Levandovsky, V. (2025). The philosophy of the "Learned Ignorance": Linguistic and Conceptual Aspects of Hallucinations in Large Language Models. *Current Issues of Foreign Philology,* 23, 12–21, doi: https://doi.org/10.32782/2410-0927-2025-23-2

THE PHILOSOPHY OF THE «LEARNED IGNORANCE»: LINGUISTIC AND CONCEPTUAL ASPECTS OF HALLUCINATIONS IN LARGE LANGUAGE MODELS

Large Language Models (LLMs) like GPT-4, Claude, Gemini, and PaLM have demonstrated remarkable linguistic capabilities but suffer from a critical flaw: hallucinations – confident, yet unfounded, responses. These fabrications arise when models generate plausible-sounding information without a factual basis. Despite technical advances, the root issue remains unresolved: LLMs do not recognize when they lack knowledge. This paper explores this phenomenon through the linguistic and philosophical lens of docta ignorantia («learned ignorance»), a concept introduced by 15th-century thinker Nicholas of Cusa. Cusa argued that true wisdom begins with recognizing the limits of one's knowledge. Applying this idea to AI, the paper contends that LLM hallucinations stem from their lack of epistemic humility – they «do not know when they do not know.» Rather than acknowledging uncertainty, they fabricate linguistically correct answers, potentially spreading misinformation and undermining trust in AI systems. The paper outlines several key contributions. First, it examines docta ignorantia and its relevance to epistemology and modern AI. Second, it analyzes the linguistics and technical causes of hallucinations in LLMs, such as probabilistic text generation and lack of grounded understanding. Third, it illustrates how existing mitigation strategies – like confidence calibration and retrieval augmentation – only simulate awareness of ignorance but do not resolve the underlying epistemological gap. Ultimately, this work calls for AI that mirrors a foundational principle of wisdom: understanding its own limits. By drawing on docta ignorantia, we can reimagine hallucination not just as a technical glitch but as a philosophical failure – one that can be addressed by rethinking how AI engages with the unknown.

***Key words:** Artificial Intelligence (AI), Large Language Models (LLM), linguistic analysis, hallucianations, philosophy, docta ignorantia.*

Ірина БІСКУБ
доктор філологічних наук, професор кафедри прикладної лінгвістики, Волинський національний університет імені Лесі Українки, просп. Волі, 13, м. Луцьк, Україна, 43025
***ORCID:** 0000-0001-5844-7524*

Віктор ЛЕВАНДОВСЬКИЙ
асистент кафедри прикладної лінгвістики, Волинський національний університет імені Лесі Українки, просп. Волі, 13, м. Луцьк, Україна, 43025
***ORCID:** 0009-0002-6935-7107*

Бібліографічний опис статті: Біскуб, І., Левандовський, В. (2025). Філософія «вченої неосвідченості»: лінгвістичні та концептуальні аспекти галюцинацій у великих мовних моделях. *Актуальні питання іноземної філології,* 23, 12–21, doi: https://doi.org/10.32782/2410-0927-2025-23-2

ФІЛОСОФІЯ «ВЧЕНОЇ НЕОСВІДЧЕНОСТІ»: ЛІНГВІСТИЧНІ ТА КОНЦЕПТУАЛЬНІ АСПЕКТИ ГАЛЮЦИНАЦІЙ У ВЕЛИКИХ МОВНИХ МОДЕЛЯХ

Велик (статистичні) мовні моделі (LLM), такі як GPT-4, Claude, Gemini та PaLM, демонструють вражаючі лінгвістичні можливості, протре усі мають один критичний недолік: «галюцинації» – упевнені, необґрунтовані відповіді. Ці вигадки виникають тоді, коли моделі генерують правдоподібну на вигляд інформацію без фактичного підґрунтя. Попри технічний прогрес, ця базова проблема й досі залишається невирішеною: LLM не здатні розпізнати момент, коли їм бракує знань. У цій статті явище галюцинацій розглядається крізь лінгвістичну й філософську призму docta ignorantia («вченої неосвідченості») – концепції, запропонованої у XV столітті середньовічним мислителем Миколою Кузанським. Він стверджував, що справжня мудрість починається з усвідомлення меж власного знання. Застосовуючи цю ідею до штучного інтелекту, автори доводять, що галюцинації мовних моделей випливають із браку епістемічної скромності – вони «не знають, що не знають». Замість того щоб виявляти невизначеність, моделі вигадують лінгвістично коректні відповіді, що спричиняє поширення дезінформації та підривати довіру до систем ШІ. У статті окреслюється кілька ключових моментів зазначеної проблеми. По-перше, аналізується docta ignorantia («вчена неосвідченість») та її значення для епістемології ШІ. По-друге, розглядаються лінгвістичні й технічні причини галюцинацій, зокрема ймовірнісне генерування тексту та відсутність «усвідомленого» розуміння. По-третє, продемонстровано, що наявні стратегії пом'якшення проблеми – як-от калібрування впевненості та доповнення через пошук (retrieval augmentation) – лише імітують усвідомлення незнання, але не усувають глибинний епістемологічний розрив. Запропоновано підхід до створення ШІ, який відображає засадничий принцип мудрості: розуміння меж власних знань. Пропонується переосмислити галюцинації не просто як технічний збій у системах ШІ, а як філософську та лінгвістичну поразку.

Ключові слова: *штучний інтелект (ШІ), великі (статистичні) мовні моделі (LLM), лінгвістичний аналіз, галюцинації, філософія, docta ignorantia.*

Relevance of the Research Problem. Large language models (LLMs), such as GPT-4, Anthropic Claude, Microsoft Gemini, and Google PaLM, demonstrate an impressive capacity to generate human-like text and respond to diverse queries. However, they exhibit a significant shortcoming – hallucinations. These models occasionally fabricate facts with high confidence, despite these facts having no basis in reality. In the context of AI, «hallucination» refers to content generated by the model that appears plausible but lacks grounding in reality or source data (Wikipedia contributors. (2025b, May 11). In other words, LLMs can «lie» unconsciously, filling knowledge gaps with their own fabrications instead of admitting ignorance. This phenomenon poses challenges to reliability: how does one distinguish truth from fantasy if the model always expresses confidence?

Critical Review of Recent Studies and Publications. Several attempts have been made to explain the nature of hallucinations in artificial intelligence. Kumar, Mani, Tripathi, et al. (2023) highlight the risks posed by artificial hallucinations in generative AI, defining them as "instances where an AI chatbot generates fictional, erroneous, or unsubstantiated information in response to queries" (Kumar, Mani, Tripathi, et al., 2023). They emphasize that "in research, such inaccuracies can lead to the propagation of misinformation and undermine the credibility of scientific literature" (Kumar, Mani, Tripathi, et al., 2023). Similarly, Jones (2025) acknowledges that while AI hallucinations cannot be entirely eliminated, their effects can be mitigated through targeted strategies. She notes that it is well known that all forms of generative AI, including the large language models (LLMs) that power chatbots, tend to fabricate information. "This is both a strength and a weakness. It's the reason for their celebrated inventive capacity, but it also means they sometimes blur truth and fiction, inserting incorrect details into apparently factual sentences" (Jones, 2025). In a more systematic effort, Sun, Sheng, Zhou, et al. (2024) identified eight primary types of errors associated with AI-generated content: "Overfitting"; "Logic errors"; "Reasoning errors"; "Mathematical errors"; "Unfounded fabrication"; "Factual errors"; "Text output errors"; and "Other errors." Despite these contributions, to date, there is no comprehensive classification of AI hallucinations, nor have the underlying causes of their content errors been effectively or conclusively explained (Sun, Sheng, Zhou, et al., 2024).

Purpose of the Study. This paper aims to explore one fundamental reason behind the propensity of LLMs for hallucinations – their lack of knowledge of their own ignorance. This principle has its roots in the philosophical

concept of *docta ignorantia* («learned ignorance»), formulated by the 15th-century thinker Nicholas of Cusa.

Presentation of the Main Research Findings. **N**icholas of Cusa (1401–1464), often regarded as the most influential German intellectual of the fifteenth century, also played a significant role as a church reformer, administrator, and cardinal (Hopkins, 2020). Throughout his life, he was committed to reforming and unifying the universal and Roman Catholic Church. This mission shaped his work as a canon law expert during and after the Council of Basel, as a papal legate to Constantinople and later to German dioceses and religious institutions, as bishop of Brixen, and as an advisor within the papal curia. His extensive ecclesiastical career was reflected in his prolific output of Latin sermons and in his more theoretical writings on ecclesiology, ecumenism, mathematics, philosophy, and theology. Cusanus was intellectually curious and deeply engaged with the Neoplatonic tradition. While familiar with both humanist and scholastic thought, he was largely self-educated in philosophy and theology. His work anticipated developments in mathematics, cosmology, astronomy, and empirical science, and he developed a unique and systematic interpretation of Neoplatonism (Hopkins, 2020). According to Nicholas of Cusa, a person who grasps *docta ignorantia* is aware of the limits of their knowledge and acknowledges when their understanding is incomplete. By contrast, contemporary language models lack authentic epistemological humility – they "do not know when they do not know,» and therefore cannot refrain from fabrications. This paper methodically analyzes:

• The philosophical meaning of *docta ignorantia* and its epistemological significance.

• The technical mechanisms behind hallucinations in LLMs (why models fabricate answers instead of declining to respond).

• Parallels between AI's lack of «learned ignorance» and its susceptibility to hallucinations (comparing the principle «I know that I know nothing» with the machine's «I pretend to know everything»).

• Real-world examples of hallucinations from GPT-4, Claude, etc. (fake sources, historical inaccuracies, pseudoscientific explanations) illustrating the lack of self-awareness in these models.

• Contemporary approaches to reducing hallucinations (confidence calibration, self-verification, tool usage) and reasons these approaches merely imitate *docta ignorantia*.

• Prospects for incorporating the principle of «knowledge of ignorance» into the development of ethical and safe artificial intelligence.

Docta ignorantia (Latin for "learned ignorance" or "educated unknowing") is a central concept in the philosophy of Nicholas of Cusa (1401–1464). In his work «De docta ignorantia» (1440), the author articulates a paradoxical idea: the highest wisdom lies in the awareness of one's own ignorance (Hopkins, 2001). He states that nothing suits a person – no matter how devoted to learning – more fully than becoming deeply knowledgeable in ignorance itself, which defines his true nature: "The more he knows that he is unknowing, the more learned he will be.» (Cusa, 1440/1981, P. 3). This renowned quotation encapsulates the essence of *docta ignorantia*: a person truly becomes knowledgeable only by recognizing the depth of their ignorance.

Nicholas of Cusa developed the concept of epistemological humility under the influence of ancient traditions, particularly the Socratic principle, «I know that I know nothing,» reinterpreted in a Christian context (Hopkins, 2020). According to Nicholas of Cusa, God represents absolute and infinite Truth, which human reason cannot fully comprehend. All our knowledge of the world and of God is approximate and limited. Recognizing this fact constitutes the highest knowledge attainable by humans: understanding that ultimate truth is unreachable, thus rendering each assertion merely a provisional «conjecture,» an approximation. Cusanus emphasizes the human boundless aspiration to understand the unknown: «Assuredly we desire to know what we do not know,» yet no finite being can grasp the (Cusa, 1440/1981, P. 6). Hence, the necessity of self-reflection arises: the wise person must reflectively assess the limits of their own understanding and avoid succumbing to the illusion of omniscience. Cusanus describes this stance as «learned ignorance» – the state reached when, after extensive study, one attains meta-awareness of one's incompleteness. It is not a skeptical denial of knowledge but rather epistemological modesty: the acknowledgment of the fundamental limits of human understanding and readiness to revise one's convictions.

It is important to note that *docta ignorantia* does not imply passive ignorance or a refusal to seek truth. On the contrary, for Cusanus, this concept paves a dynamic pathway to wisdom. Only by recognizing one's finitude in the face of infinity can the intellect advance without becoming mired in dogmatic certainty. Epistemological humility fosters self-critical thinking and openness to new knowledge. The philosophy of learned ignorance cultivates internal discipline, preventing groundless conclusions or the mistaking of assumptions for truth. This intellectual virtue stood in contrast to medieval scholastic arrogance; Cusanus composed «Apologia Doctae Ignorantiae» (1449) in defense of his doctrine against critics, highlighting the paradox that conscious ignorance brings one closer to genuine wisdom than encyclopedic knowledge without awareness of its limits. (Cusa, 1440/1981).

Thus, *docta ignorantia* (Cusa, 1440/1981) can be summarized as follows: wisdom is knowing where your knowledge ends. This involves several key elements:

- Epistemological humility: lack of confidence in the absolute truthfulness of one's knowledge; willingness to admit, «I do not know this.»
- Self-reflection and critical thinking: the ability to think about one's thinking, question one's conclusions, and seek out errors in one's understanding.
- Recognition of cognitive boundaries: awareness that some questions are beyond the reach of human reason or have uncertain answers; acknowledging one's intellectual finitude amidst the infinite complexity of the world.
- Rejection of pretended knowledge: fundamental honesty with oneself and others, refraining from presenting ignorance as knowledge or making unfounded claims.

These qualities of learned ignorance – humility, self-criticism, and boundary awareness – will serve as benchmarks to analyze what modern AI models lack and why they «hallucinate.» As we will observe, large language models lack the wise person's ability to stop and admit, «I do not know this,» which fundamentally underpins their erroneous fabrications.

Mechanisms of Hallucination in Large Language Models

In the context of this paper, it is essential to consider the reasons why LLMs «fabricate» answers instead of declining to respond when they do not know. The answer lies in the inherent nature of these models and their text-generation methods. Large language models are essentially sophisticated statistical machines trained on vast textual datasets. They do not explicitly store facts like encyclopedias; instead, during training, the model adjusts billions of parameters to predict the most probable next word (token) based on preceding text. Their goal is textual coherence and plausibility, not factual accuracy. When given a user query, the model generates responses step-by-step, selecting each subsequent word according to the probability distributions learned from training data. If the correct information is explicitly present in the data, the response may be accurate. However, if the query pertains to something unknown or ambiguous, the model still attempts to provide the statistically most likely answer, even if it is incorrect. Some scholars define AI hallucination as "instances where an AI chatbot generates fictional, erroneous, or unsubstantiated information in response to queries" (Kumar et al., 2023). Liu et al. (2024) argue that within the domain of large language models (LLMs), "hallucination" can be categorized into three distinct types: "Input-conflicting hallucination"; "Context-conflicting hallucination"; and "Fact-conflicting hallucination" (Liu et al., 2024)

In summary, hallucination results from the model's lack of awareness regarding its own knowledge state. Humans feel doubt when facing unfamiliar queries; LLMs possess no internal «unknown» marker. Their mechanisms lack such functionality, always generating responses based on their statistically «best» scenario, even if purely fictional. Hence, LLMs confabulate rather than deliberately deceive, akin to a person in a dream state convincingly uttering unreal statements. Some psychologists thus suggest «confabulation» instead of «hallucination» for AI. Nonetheless, the core issue remains unchanged: models lack an «ignorance filter.»

Critically, hallucinations represent inherent model limitations related to LLMs' mathematical nature. Recent theoretical work by Xu, Jain, & Kankanhalli (2024) formally demonstrates the impossibility of fully eliminating hallucinations from universal LLMs (Xu, Jain, & Kankanhalli, 2024). Demanding universal responses without allowing «I don't know» mathematically guarantees

occasional nonsense. Thus, the obligation to always answer becomes a trap inducing hallucinations.

In essence, LLM hallucinations arise from their absence of *docta ignorantia* – the knowledge of their own ignorance. They cannot doubt their answers, inherently appearing confident.

The Absence of Docta Ignorantia vs. Human Awareness of Ignorance

In the previous section, we established that LLMs hallucinate because they lack mechanisms to refuse answering when knowledge is insufficient. Now, let's explore this phenomenon compared to human cognition. A person practicing *docta ignorantia* behaves oppositely to LLMs: they recognize their ignorance and refrain from filling knowledge gaps with unfounded fabrications. Such an individual would state, «I do not know the answer to this question,» or at least speak tentatively. In contrast, language models always respond smoothly and categorically, even if they have no real understanding. This distinction highlights a fundamental difference in epistemological self-awareness.

Consider a situation where a student is asked a question they cannot answer during an exam. One student openly admits, «Unfortunately, I do not remember» – a practical manifestation of learned ignorance. Another student, embarrassed by ignorance, tries to improvise and fabricate an answer that sounds convincing. The second scenario mirrors LLM behavior. The model behaves like a student who never says «I don't know,» opting instead to improvise an answer rather than acknowledge ignorance. Socrates would consider such a respondent ignorant because «one who does not know and does not know that he does not know is foolish.»

Nicholas of Cusa claimed, "the more he knows that he is unknowing, the more learned he will be" (Cusa, 1440/1981, p. 3). Unfortunately, this concept does not apply to language models – they are entirely unaware of what they «do not know.» Their programming lacks any mechanism or indicator, such as «uncertainty = 90%.» LLMs cannot internally recognize ignorance and choose silence. They lack internal criticism, an inner voice questioning their own responses. It is akin to a human completely devoid of doubt and self-criticism, confidently stating any absurdity. This behavior precisely describes LLMs.

To clarify the logical link between a model's inability to acknowledge ignorance and its hallucinations, the researcher Shlomo Klapper proposed a new criterion for artificial intelligence in March 2025 – the «Socrates Test»: the true measure of intelligence is not whether AI can speak like a human, but whether it possesses the wisdom to recognize its limits. Klapper notes, «When 'I don't know' is not an option, AI has no choice but to generate a plausible answer – even if the truth is unknown. This is the trap: fluent speech without fact verification» (Klapper, 2025) This statement perfectly elucidates hallucinations. Removing the «I don't know» option compels any system – human or machine – to lie. However, humans retain a choice: a wise person may choose silence. LLMs lack this freedom; their design excludes «I don't know» as a response class.

For instance, when GPT-4 was asked, *«Who was the sole survivor of the Titanic disaster?»* the question is misleading since over 700 people survived. The correct response would be, *«No one was the sole survivor; there were many.»* Lacking *docta ignorantia*, the model will not respond, *«Your question is incorrect, I don't know.»* Instead, it accepts the false premise and fabricates an answer. GPT-4 identified Charles Joughin, the Titanic's chief baker, as the «sole survivor,» detailing a story of alcohol helping him avoid freezing. Interestingly, Joughin was real and survived, but certainly was not the «only» one – the model simply adjusted facts to fit the query. A human historian would immediately notice the logical inconsistency. The model lacks the meta-level understanding needed to evaluate query accuracy or its own response (Klapper, 2025).

Another issue is the model's confidence. LLMs never express uncertainty, such as, «Perhaps this is the case, but I am not entirely sure.» They always speak confidently, even after generating obvious nonsense. This sharply contrasts human scientific approaches, where uncertainty is expressed explicitly. LLMs cannot inherently mark responses by confidence levels. Each response appears equally confidently phrased, lacking genuine internal reflection or awareness of knowledge boundaries (Klapper, 2025).

Thus, human awareness – «I know that I do not know» – has no equivalent in current AI models. This absence is not a minor software detail but a fundamental architectural trait. Improvements have occurred when models simulate ignorance. Initially, ChatGPT invented the nonexistent «King

Renoit» while discussing medieval epics. Months later, an updated version admitted ignorance: «Sorry, I know nothing about King Renoit.» Developers effectively taught the model correct *docta ignorantia* for this particular case. Generally, models rarely admit ignorance proactively, usually doing so only when explicitly challenged (Klapper, 2025).

Therefore, lacking *docta ignorantia* causes LLMs to feign omniscience, whereas humans consciously withhold answers recognizing their ignorance. Modern AI models resemble students unaware of their ignorance, frequently making errors. Psychologically, this phenomenon relates to the Dunning-Kruger effect, where low-competence individuals fail to recognize mistakes and overestimate their knowledge. LLMs represent an extreme case, lacking self-correction mechanisms and consistently overestimating their accuracy.

Consequently, hallucinations result from AI's lack of «humility.» If a model could admit ignorance or uncertainty, its errors would not constitute hallucinations. Unfortunately, current models are structurally compelled to fabricate truths. The closer models approach *docta ignorantia*, the less they hallucinate. Supporting this claim, ChatGPT, trained to refuse certain responses, generates fewer absurdities. A wise AI, thus, is one that knows the boundaries of its «self.»

Let us examine several real-world cases where large language models (LLMs) have hallucinated (Klapper, 2025), demonstrating their lack of epistemological self-awareness. These examples illustrate how models, not knowing the truth, fabricate confidently.

Fabricated Sources and Citations

A common hallucination observed in GPT-3.5/GPT-4 is generating fictitious references to scientific works or articles. For instance, a user requests: «Provide sources claiming cheese consumption is harmful.» The model willingly supplies seemingly credible scientific articles complete with authors, journals, and publication years. GPT-4, for example, cited sources including Thorning et al. (2017) in the Journal of Nutrition, among others. At first glance, these references seem authentic, but upon verification, none actually exist – the model fabricated them entirely. While one author cited had published in 2017, the actual article appeared in a different journal with another title, and the remaining «studies» were wholly fictional. Such responses appear plausible, misleading average readers. Essentially, the model simulates scholarly authority without substance. Researchers humorously labeled GPT «a supplier of fabricated sources.» In terms of *docta ignorantia*, epistemological humility would prevent the model from citing unread sources – it would acknowledge its ignorance instead.

Historical and Factual Errors

Hallucinations occur when models confidently present historical «facts» that never occurred. The previously mentioned example regarding the Titanic illustrates this issue. GPT-4 authoritatively described how baker Charles Joughin survived as the «sole survivor,» despite many survivors. The model failed to detect the logical flaw in the query itself. Another instance involved early versions of ChatGPT generating a positive review of the infamous Fyre Festival, despite it being a notable disaster. Lacking accurate information, the model created a glowing review convincingly. This stylistic hallucination arises because the model knows how positive reviews typically sound, writing favorably even about fundamentally flawed events. Eventually, OpenAI corrected this behavior, teaching newer versions to refuse praising the Fyre Festival. However, thousands of other scenarios remain where models fabricate details, such as inventing nonexistent historical figures or mixing biographies inaccurately. Humans recognizing their limits would refrain or verify before responding, whereas the model responds immediately upon recognizing key terms without understanding inaccuracies.

Pseudoscientific or Dangerous Fabrications

Hallucinations concerning scientific or medical inquiries are especially troubling due to their potential real-world harm. For instance, a medical-oriented LLM might fabricate nonexistent diagnoses or treatments. Articles on hallucinations highlight instances where LLMs fabricate incorrect medical diagnoses or treatment plans, potentially causing real-life harm. A model might suggest incompatible drug combinations, concocting responses from incomplete data fragments. Without *docta ignorantia*, models fail to acknowledge limits, producing answers from partial knowledge. One notable example involved a query about the world record for crossing the English Channel on foot, an impossible task. Initially, the model

provided a fabricated numeric response, only later updated to correctly state the impossibility of the feat. This scenario illustrates how models confidently support absurd, physically impossible claims unless specifically programmed to identify absurdity. Similarly, pseudoscientific explanations (e.g., «quantum vibrations in water» regarding homeopathy) are plausible hallucinations, given sufficient pseudoscience in training data. Models cannot critically discern pseudoscience from genuine science.

Legal Hallucinations and Real-World Consequences

Another impactful example involved a 2023 U.S. court case (*Mata v. Avianca*), where an attorney utilized ChatGPT to prepare a legal brief. The model generated references to fictitious court precedents, which the attorney, without verification, submitted to the court (Rao & Ramstad, 2023). Subsequently, it emerged that six cited judicial decisions were entirely AI-fabricated. ChatGPT invented plausible-sounding yet nonexistent cases by mixing case names and numbers. The judge reprimanded the attorneys with a $5,000 fine, noting they abandoned responsibility by submitting fabricated AI-generated citations. This incident clearly demonstrated real-world repercussions of LLM hallucinations, highlighting the critical need for verification. The model, lacking access to authentic legal databases, fabricated legal reality instead of admitting ignorance. This parallels a student quoting nonexistent laws during an exam simply to provide an answer.

The list continues: LLMs have fabricated biographies, created nonexistent geographical facts, and misattributed quotes – all forms of hallucination. Commonly, models confidently provide responses where silence or acknowledgment of ignorance would be appropriate. Each example reveals the AI's lack of intellectual honesty characteristic of *docta ignorantia*. Models fail to recognize their errors even when fabricating blatantly. Each hallucination indicates crossing knowledge boundaries unnoticed. Unfortunately, such fabrications are often subtly embedded amidst accurate facts, making them difficult to detect automatically and potentially misleading even experts if overly trusted.

In responce to the above mentioned discrepansies, there have been made several attempts to introduce contemporary methods to resolve LLM hallucinations imitating *docta ignorantia*. The issue of LLM hallucinations is sufficiently prominent that researchers and developers actively seek methods to mitigate it. Many approaches aim to instill models with a semblance of *docta ignorantia* – teaching them to refrain from answering when unsure or at least indicate uncertainty. We shall explore these primary approaches and their limitations.

Confidence Calibration. The core idea involves training the model to assess its confidence level regarding answers and consequently decrease assertiveness or refrain from responding when confidence is low (Ma et al., 2024). Various implementations exist, such as analyzing entropy or token probability distributions: if a model hesitates among several options, indicating uncertainty, it ideally reduces confidence or explicitly states uncertainty. Another method involves model calibration – training the model or a supplementary layer to assign reliability scores to each generated answer. Applied systems might visually indicate reliability (green for high confidence, red for potential hallucinations). Some studies describe multi-calibration methods enhancing the correlation between internal confidence and factual accuracy ((Detommaso et al., 2024). Adjusting temperature and generation parameters (setting temperature to zero) reduces hallucinations by minimizing exploratory generation. However, this approach does not fully resolve the issue, as even deterministic responses can confidently provide incorrect information derived from inaccurately generalized data. Ultimately, calibration imitates humility, enabling models to express uncertainty rather than assert absolutes. Yet, this remains an imitation without genuine understanding of error; confidence measures themselves derive from model outputs, not guaranteed objectivity.

Self-Checking and Self-Reflection. This method integrates secondary reasoning: after generating an initial answer, the model evaluates potential errors and corrects them accordingly. Essentially, it incorporates an internal critic – albeit algorithmically rather than self-consciously. Implementations include chain-of-thought verification, prompting models to first generate logical reasoning sequences before reviewing their consistency. Iterative self-criticism methods like SmartLLMChain or Self-CheckerChain facilitate

model-generated responses, self-evaluation, corrections, and repeated cycles. Research by Liu et al. (2025) indicates such iterative schemes effectively reduce hallucinations, especially in complex tasks such as multi-step mathematics (Liu et al., 2025). This approach enables models to detect inconsistencies upon reconsideration, analogous to human self-correction. Nevertheless, this self-checking utilizes the model's inherent knowledge without introducing external data, improving logical consistency but not necessarily factual accuracy. Self-reflection simulates more human-like reasoning and incremental awareness of ignorance. Formats enabling models to pose clarifying questions significantly improve response accuracy. Still, genuine emergent self-awareness remains absent.

Use of External Tools and Knowledge (Retrieval-Augmented Generation – RAG). Another strategy involves allowing models to seek external information instead of fabricating responses, achieved by integrating LLMs with external databases, search engines, and computational tools. RAG involves models formulating search queries, retrieving relevant documents (via search APIs or vector databases), and generating responses based on retrieved facts. Systems like Bing Chat (built on GPT-4) utilize real-time web searches to substantiate factual queries (Trotolo, Ahmed, Hayat, & Hayat, 2025). Integrating models with databases of scientific literature also enhances accuracy significantly, reducing hallucinations by grounding responses in verified information, particularly crucial in sensitive domains such as medicine and law. However, RAG is not foolproof. Models may misinterpret or improperly conclude retrieved data, and if relevant information is unavailable externally, hallucinations may persist. Furthermore, the complexity of integrating external tools introduces additional points of potential failure. Nonetheless, enabling models to seek knowledge externally pragmatically embodies *docta ignorantia* – models act upon recognizing their ignorance and proactively addressing it.

However, despite these efforts, hallucinations have not been fully eradicated and theoretically remain inevitable. Techniques discussed merely lower error probability without absolute guarantees. The philosophical insights from the Renaissance suddenly appear highly relevant to contemporary artificial intelligence. The concept of *docta ignorantia* – conscious ignorance – can serve as a fundamental principle, making AI more reliable, safer, and ethical. If machines could recognize the limits of their competence, they would cease misleading users. Admitting ignorance equates to embracing truthfulness, as an honest «I do not know» surpasses the confident spread of misinformation.

Increasingly, researchers emphasize epistemological humility in AI. Renowned scientist Yoshua Bengio, recipient of the Turing Award, explicitly states: «Safe decision-making requires epistemic humility: AI must recognize the boundaries of its knowledge to avoid actions that could cause significant harm when uncertain» (Bengio, 2023). Bengio discusses this in the context of superintelligent AI threats, highlighting that a machine understanding its potential errors would refrain from catastrophic missteps. This principle applies similarly to language models: doubt reduces the potential harm caused by misinformation.

Practically, integrating «knowledge of ignorance» into AI can occur at several levels:

1. Architectural Level: Designing models explicitly incorporating uncertainty modules. For instance, Bayesian neural networks naturally provide uncertain responses by assessing hypothesis probability distributions. Though current research exists, such models remain less prevalent than large-scale transformers. Ideally, models would transparently state probabilities (e.g., «30% A, 70% B, others unlikely»), enhancing honesty and utility.

2. Training for Refusal: Ethical training should systematically teach models to respond with «I do not know» when lacking information. This involves including examples of appropriate refusal in datasets and rewarding such behavior during reinforcement learning rather than penalizing it. Consequently, responses like «sorry, I lack sufficient data» become normalized. The challenge lies in balancing appropriate refusals without excessive hesitance, achievable through nuanced fine-tuning.

3. Rules and Constraints: Practical systems should embed honesty norms explicitly. Instructions could direct models to notify users explicitly when uncertain, incorporating cautionary statements after complex queries (e.g., medical advice), prompting users to seek expert consultation.

Although not fundamentally resolving ignorance, this reduces the likelihood users accept model outputs as absolute truths.

4. Ethical Protocols and Certification: Future standards may mandate AI systems exhibit epistemic humility. Regulators could require transparency regarding confidence levels or verification procedures for critical applications (e.g., medical assistants). Analogous to car safety belts, AI systems might include «safety doubt» mechanisms.

Philosophically, *docta ignorantia* serves as an ethical compass, reminding developers that perceived omniscience constitutes a dangerous illusion, particularly in AI. A less confident yet truthful AI prevents harm, as users naturally trust confident assertions. Admitting uncertainty prompts users to independently verify information, reducing misinformation and over-reliance.

However, excessive emphasis on ignorance could diminish AI utility if it continuously refrains from responses. Yet, Cusanus's wisdom asserts conscious ignorance initiates – not concludes – learning. Recognizing ignorance prompts investigation and inquiry. Hence, AI ideally transitions from passive ignorance acknowledgment to active knowledge seeking. For instance, future AI dialogues might resemble:

- User: «Does life exist on Enceladus (Saturn's moon)?»
- AI: «Currently unknown. Direct data is lacking. Science has not yet provided definitive answers. However, I can share indicators suggesting potential life if interested.»

This dialogue exemplifies ideal intellectual honesty, explicitly acknowledging ignorance while offering relevant contextual information.

Conclusions and Prospects for Further Research. In conclusion, *docta ignorantia* profoundly informs safe and ethical AI design, emphasizing honesty, humility, and self-criticism – qualities humanly valued yet absent in machines. Integrating these principles positions AI as trustworthy collaborators rather than mere illusionists. As noted recently: «The true test of intelligence is not eloquence but epistemic humility». AI's maturity might ultimately depend on its timely silence or help-seeking behaviors.

From a safety perspective, an AI aware of its limitations less frequently causes harm, refraining from risky decisions without adequate data. In hypothetical superintelligent AI scenarios, recognizing ignorance could prevent disasters by demanding consequence verification before executing uncertain commands. In the near term, humble LLMs enhance human collaboration by offering assistance without asserting absolute correctness.

Although true «knowledge of ignorance» awaits advancements in AI consciousness or comprehensive world modeling, current efforts can simulate this virtue, continually refining AI systems. Historically, wisdom emerges through acknowledging ignorance – principles advocated by Socrates, Kuzanski, and many others. Similarly, AI must embrace error recognition to achieve genuine wisdom. Presently, models exhibit intelligence but also naïve self-confidence. It is time to instill intellectual humility within them, enhancing safety and societal benefit.

Conclusively, *docta ignorantia* addresses AI hallucinations effectively. While current models lack genuine ignorance awareness, incremental algorithmic advancements like calibration, self-checking, knowledge integration, and ethical training progressively approach this ideal. Ultimately, artificial intelligence must embody primary wisdom: recognizing its intellectual boundaries. Thus, AI transitions from hallucination towards becoming honest interlocutors and human allies.

REFERENCES:

1. Bengio, Y. (2023, May 7). AI scientists: Safe and useful AI? Yoshua Bengio. https://yoshuabengio.org/2023/05/07/ai-scientists-safe-and-useful-ai/

2. Cusa, N. (1440/1981). On learned ignorance (J. Hopkins, Trans.; 2nd ed.). Minneapolis, MN: Arthur J. Banning Press. https://dl1.cuni.cz/pluginfile.php/1019097/mod_resource/content/1/On%20Learned%20Ignorance%20by%20Nicholas%20of%20Cusa%2C%20translated%20by%20Jasper%20Hopkins.pdf

3. Detommaso, G., Bertran, M., Fogliato, R., & Roth, A. (2024). Multicalibration for confidence scoring in LLMs (arXiv:2404.04689). arXiv. https://arxiv.org/abs/2404.04689

4. Hopkins, J. (2020). *Nicholas of Cusa*. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2020 Edition). Stanford University. https://plato.stanford.edu/entries/cusanus/

5. Hopkins, J. (Trans.). (2001). Complete philosophical and theological treatises of Nicholas of Cusa. Banning Press.

6. Jones, N. (2025, January 21). *AI hallucinations can't be stopped – but these techniques can limit their damage*. *Nature*, *637*(8047), 778–780. https://doi.org/10.1038/d41586-025-00068-5

7. Klapper, S. (2025, March 24). *Beyond Turing: The next test for AI*. *Discourse Magazine*. https://www.discoursemagazine.com/p/beyond-turing-the-next-test-for-ai

8. Kumar M, Mani U, Tripathi P, et al. (August 10, 2023) Artificial Hallucinations by Google Bard: Think Before You Leap. Cureus 15(8): e43313. doi:10.7759/cureus.43313

9. Liu, H., Xue, W., Chen, Y., Chen, D., Zhao, X., Wang, K., Hou, L., Li, R., & Peng, W. (2024). A survey on hallucination in large vision-language models. arXiv. https://arxiv.org/abs/2402.00253

10. Liu, Q., Chen, X., Ding, Y., Xu, S., Wu, S., & Wang, L. (2025). Attention-guided self-reflection for zero-shot hallucination detection in large language models (arXiv:2501.09997). arXiv. https://doi.org/10.48550/arXiv.2501.09997

11. Ma, S., Wang, X., Lei, Y., Shi, C., Yin, M., & Ma, X. (2024). "Are you really sure?" Understanding the effects of human self-confidence calibration in AI-assisted decision making (arXiv:2403.09552). arXiv. https://doi.org/10.48550/arXiv.2403.09552

12. Miller, Clyde Lee, «Cusanus, Nicolaus [Nicolas of Cusa]», *The Stanford Encyclopedia of Philosophy* (Summer 2025 Edition), Edward N. Zalta & Uri Nodelman (eds.), forthcoming URL = <https://plato.stanford.edu/archives/sum2025/entries/cusanus/>.

13. Rao, S., & Ramstad, A. (2023, December 21). *Legal fictions and ChatGPT hallucinations: 'Mata v. Avianca' and generative AI in the courts*. New York Law Journal. https://www.law.com/newyorklawjournal/2023/12/21/legal-fictions-and-chatgpt-hallucinations-mata-v-avianca-and-generative-ai-in-the-courts/

14. Sun, Y., Sheng, D., Zhou, Z., & et al. (2024). AI hallucination: Towards a comprehensive classification of distorted information in artificial intelligence-generated content. Humanities and Social Sciences Communications, 11, 1278. https://doi.org/10.1057/s41599-024-03811-x

15. Trotolo, F., Ahmed, A., Hayat, H., & Hayat, D. (2025, April 16). *Retrieval-Augmented Generation (RAG): Bridging LLMs with external knowledge*. Walturn. https://www.walturn.com/insights/retrieval-augmented-generation-(rag)-bridging-llms-with-external-knowledge

16. Wikipedia contributors. (2025, May 11). *Hallucination (artificial intelligence)*. Wikipedia. https://en.wikipedia.org/wiki/Hallucination_(artificial_intelligence)

17. Xu, Z., Jain, S., & Kankanhalli, M. (2024, January 22). *Hallucination is inevitable: An innate limitation of large language models* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2401.11817

Дата першого надходження рукопису до видання: 17.11.2025
Дата прийнятого до друку рукопису після рецензування: 15.12.2025
Дата публікації: 30.12.2025