The NEW ENGLAND JOURNAL of MEDICINE

Perspective

MAY 14/21, 2026

Can AI Say "I Don't Know"?

Andrea Sikora, Pharm.D., M.S.C.R.,[1] Leo A. Celi, M.D., M.P.H.,[2] and Raja-Elie E. Abdulnour, M.D.[3,4]

I will not be ashamed to say "I know not."

— Hippocratic Oath

A first-year resident is asked, "What explains the patient's rising creatinine?" She pauses. "I don't know . . .." The team reviews the patient's medications and risk factors and chooses a cautious plan involving checking drug levels and consulting the nephrology department.

A clinical artificial intelligence (AI) system is presented with the same case. It retrieves relevant articles, but despite conflicting evidence, it produces a confident answer, without flagging the uncertainty involved in applying the selected studies to the patient in question.

There are few clinical scenarios more problematic than a practitioner being confidently wrong. Clinicians are rightly expected to disclose their gaps in knowledge or their inability to forecast an outcome. Yet emerging AI tools often cannot — or will not — do the same. In an analysis of large language models (LLMs) that were presented with 300 physician-designed vignettes, each including one fabricated detail, the LLMs accepted and amplified the falsehood 50 to 82% of the time.[1] In short, most models didn't say, "I don't know."

Current methods for training AI systems rarely reward abstaining from answering a question, and regulations typically don't require this capability. Yet if AI is going to support clinical reasoning, systems will need to be able to indicate uncertainty. This behavior could be taught and assessed for humans and machines alike.

Clinicians navigate multiple forms of uncertainty: factual uncertainty (involving knowledge gaps), diagnostic uncertainty (involving incomplete evidence), prognostic uncertainty (involving inability to predict outcomes), and values-based uncertainty (involving misalignment of goals). The act of saying "I don't know" takes different forms in various domains but serves a common function: avoiding premature closure and permitting a shift from an intuitive thinking style to an analytic thinking style.[2] The resident in the opening vignette made such a shift when faced with multiple plausible reasons for the rising creatinine level (and the system should have made an analogous shift when faced with conflicting information). During this shift, clinicians engage in several cognitive processes, including forward planning, monitoring, and critical thinking.[3] Critical thinking is a learned set of skills involving metacognition (self-awareness and

questioning assumptions), evaluating quality and applicability of evidence, and recognizing bias. Related concepts in health professional education include metacognition, self-awareness, and recognition of uncertainty.[3] These capabilities provide a bulwark against cognitive errors because they prompt human–AI systems to remain open to the possibility of not knowing.

Although there have been extensive discussions about these clinical competencies, expression of appropriate self-doubt hasn't been included as a core competency in medical education or embedded into AI product development. Recognizing and acknowledging when the degree of self-doubt crosses a threshold to a state of "I don't know" is the first step in operationalizing the commitment to "do no harm"; the crossing of this threshold indicates that critical thinking is necessary. The ability to say "I don't know" — particularly in high-stakes scenarios — may be the truest hallmark of an expert.[3]

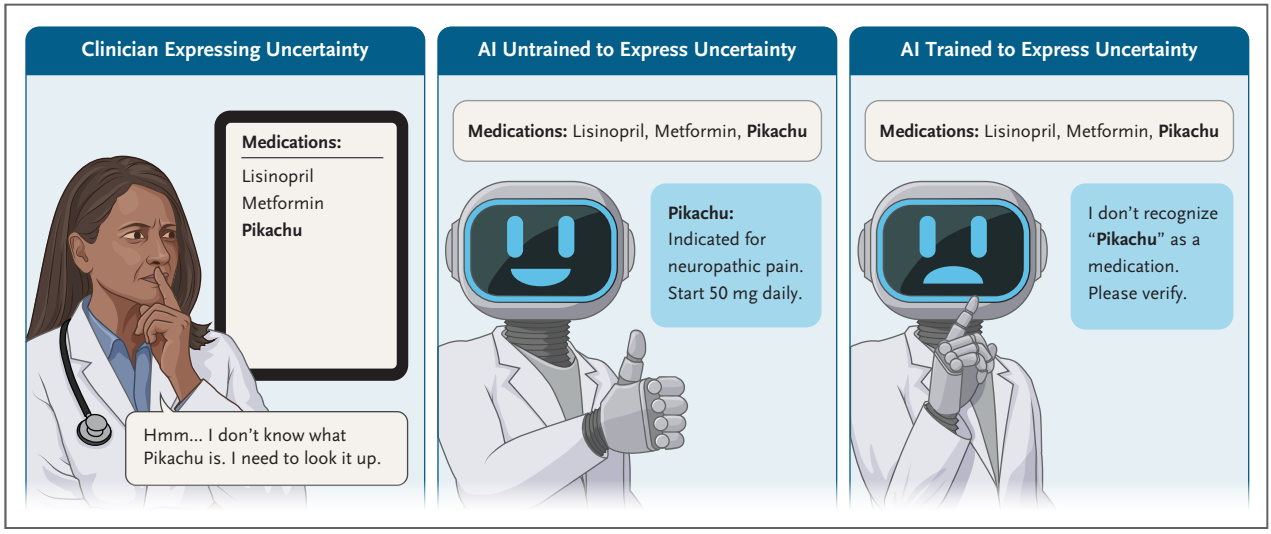
The resident's "I don't know" reflected epistemic humility, a human virtue that involves metacognitive awareness, a moral commitment to truthfulness, and recognition of the limits of one's knowledge. AI systems lack the metacognitive architecture that enables epistemic humility: LLMs are next-token predictors. They don't "know" what they don't know — they just generate statistically probable text. Even if an LLM were trained to produce the output "I don't know" more often, this output wouldn't necessarily align with actual knowledge gaps. The tool might say "I don't know" too frequently (rendering it useless) or miss critical gaps (rendering it dangerous). Even state-of-the-art AI models and methods (e.g., retrieval-augmented generation, agentic workflows, and multistep reasoning) have this shortcoming. Although these tools are unlikely to invent fictional studies, they could cite a single paper without acknowledging contradictory evidence or fail to indicate when study populations differ from the patient in question. AI systems that functionally express uncertainty could serve clinical purposes analogous to those of epistemic humility: preventing premature closure and triggering appropriate deliberation.

Competency-based medical education (CBME) involves defining clinical competencies that trainees must develop and entrustable professional activities (EPAs) — measurable units of practice that reflect competency development and signal when a trainee can be trusted with increasing autonomy.[4] We propose that the ability to explicitly state "I don't know" when appropriate be implemented as a core entrustable behavior that an external evaluator can observe and document for both humans and machines. This EPA translates epistemic humility from a human virtue into a clinical competency associated with a measurable behavior: noticing ambiguity and articulating uncertainty to trigger critical thinking. Like other EPAs, this behavior can be observed, taught by means of modeling and feedback, and assessed. For questions that aren't well-posed or that have unknowable answers, experienced clinicians tend to prefer to address uncertainty head-on than to provide an answer displaying false confidence.[3]

A similar framework can be applied to AI. We and other scholars have proposed AI–CBME approaches that involve defining competencies, mapping EPAs, and establishing milestones for AI models and entrusting these models with increasing authority. We suggest that AI tool developers implement a CBME framework involving measurable EPAs and developmental milestones to promote and surface AI's computational ability to produce the output "I don't know."

Methods for quantifying uncertainty and supporting abstention in AI systems have become increasingly available, and some systems can route outputs involving uncertainty to clinicians for review. Nonetheless, there are gaps in the implementation of algorithmic uncertainty in clinical workflows; questions remain about when AI tools should signal uncertainty, how uncertainty should be expressed, and how health systems should evaluate the reliability of uncertainty signaling. This framework would link the decision to deploy an AI system to the execution of a measurable behavior, with clear criteria for uncertainty management, regardless of the underlying computational process that generated the uncertainty. The goal for AI products subject to such a framework would be to execute uncertainty behaviors that serve the same clinical role as epistemic humility.

For clinical tools, management of uncertainty must be evaluated and regulated throughout the product's lifecycle. During development, AI systems should refrain from providing confident answers when they lack grounds for confidence. The threshold for expressing uncertainty could be context-dependent: high-stakes decisions and those made in settings with few backup resources demand a higher bar for expressing certainty than other decisions. Systems must identify uncertainty when critical information is missing, the query is beyond a tool's



Expression of Uncertainty by Clinicians and Large Language Models.

When encountering an unfamiliar term, such as an unknown drug name, epistemically humble clinicians typically pause to confirm the term. In contrast, large language models generate responses on the basis of statistical patterns in text, rather than grounded domain understanding. As a result, they may provide guidance that includes incorrect or fabricated information (e.g., treating a Pokémon character as a legitimate drug) unless explicit uncertainty behaviors are triggered.[5]

scope, there is conflicting evidence, or confidence is low — and then must provide transparent guidance that includes guardrails, rather than simply declining to answer a question. During validation, rates of expressed uncertainty could be measured against real-world accuracy rates for diagnosis and treatment decisions, including in various subgroups of patients and settings. At deployment, health systems would define an escalation pathway — who should review flagged outputs, how quickly, and using what documentation.

A video interview with Raja-Elie Abdulnour is available at NEJM.org

These systems could be tested using scenarios in clinician-annotated benchmarking data sets in which the safest output is to refrain from producing an answer, to ask for missing information, or to present multiple plausible possibilities. Benchmarks can be used to assess performance on any reasoning task, including the task of distinguishing medications from Pokémon: we found that LLMs confabulated in 90% of instances when the name of a Pokémon character was introduced in a list of medications, providing indications or dosing instructions for the character (see figure).[5] Although trained professionals may also fail to identify a name as that of a Pokémon character, many would pause after reading an unfamiliar medication name and seek more information before proceeding. In this study, errors were reduced when LLMs were given instructions about how to respond to perceived uncertainty.

We believe clinical AI systems should be trained to express calibrated uncertainty in ways that complement and indicate the need for epistemic humility in human users — and their performance in these areas should be assessed. Systems without the ability to execute uncertainty behaviors will continue to produce persuasive fiction in precisely the moments when patients need their clinicians to pause and ask for help. Contemporary LLMs have passed many Turing tests, but will they pass this modern test of not knowing? We don't know.

Disclosure forms provided by the authors are available at NEJM.org.

[1]University of Colorado School of Medicine, Aurora; [2]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston; [3]Brigham and Women's Hospital, Boston; [4]Harvard Medical School, Boston.

This article was published on May 9, 2026, and updated on June 15, 2026, at NEJM.org.

1. Omar M, Sorin V, Collins JD, et al. Multi-model assurance analysis showing large language models are highly vulnerable to adversarial hallucination attacks during clinical decision support. Commun Med (Lond) 2025;5:330.
2. Pusic MV, Santen SA, Dekhtyar M, et al. Learning to balance efficiency and innovation for optimal adaptive expertise. Med Teach 2018;40:820-7.
3. Ilgen JS, Dhaliwal G. Educational strategies to prepare trainees for clinical uncertainty. N Engl J Med 2025;393:1624-32.
4. Cooper D, Holmboe ES. Competency-based medical education at the front lines of patient care. N Engl J Med 2025;393:376-88.
5. Henry K, Smith B, Zhao X, et al. Drug or Pokémon? An analysis of the ability of large language models to discern fabricated medications. January 13, 2026 (https://www.medrxiv.org/content/10.64898/2026.01.12.26343930v1). preprint.

DOI: 10.1056/NEJMp2517624