Epistemological Fault Lines Between Human and Artificial Intelligence

Walter Quattrociocchi,[1,*] Valerio Capraro,[2,†] and Matjaž Perc[3,4,5,6,‡]

[1] *Department of Computer Science, Sapienza University of Rome, Rome, Italy*
[2] *Department of Psychology, University of Milan Bicocca, Milan, Italy*
[3] *Faculty of Natural Sciences and Mathematics, University of Maribor, Maribor, Slovenia*
[4] *Community Healthcare Center Dr. Adolf Drolc Maribor, Maribor, Slovenia*
[5] *University College, Korea University, Seoul, Republic of Korea*
[6] *Department of Physics, Kyung Hee University, Seoul, Republic of Korea*

(Dated: December 22, 2025)

Large language models (LLMs) are widely described as artificial intelligence, yet their epistemic profile diverges sharply from human cognition. Here we show that the apparent alignment between human and machine outputs conceals a deeper structural mismatch in how judgments are produced. Tracing the historical shift from symbolic AI and information filtering systems to large-scale generative transformers, we argue that LLMs are not epistemic agents but stochastic pattern-completion systems, formally describable as walks on high-dimensional graphs of linguistic transitions rather than as systems that form beliefs or models of the world. By systematically mapping human and artificial epistemic pipelines, we identify seven epistemic fault lines, divergences in grounding, parsing, experience, motivation, causal reasoning, metacognition, and value. We call the resulting condition Epistemia: a structural situation in which linguistic plausibility substitutes for epistemic evaluation, producing the feeling of knowing without the labor of judgment. We conclude by outlining consequences for evaluation, governance, and epistemic literacy in societies increasingly organized around generative AI.

Keywords: Large Language Models, Epistemia, Judgment, Credibility, Epistemic Alignment

I. INTRODUCTION

The aspiration to build machines capable of mimicking or reproducing human thought long predates the advent of digital computers. Well before modern technology, myths and legends already testified to a fascination with artificial minds. In Greek mythology, for instance, Hephaestus is said to have crafted golden automatons that could move, speak, and reason, while in Jewish folklore the Golem appears as a man-made being animated through sacred letters and mystical rituals. By the late Middle Ages and the Renaissance, scholars such as Ramon Llull, and later Gottfried Wilhelm Leibniz, began to entertain the possibility of logical engines, devices that would manipulate symbols to carry out reasoning procedures. Leibniz's proposal of a calculus ratiocinator, a universal symbolic language that could in principle resolve disputes through computation, anticipated both formal logic and, in the long run, theoretical computer science [1]. In parallel, Enlightenment thinkers from Descartes to La Mettrie advanced mechanistic accounts of the mind, portraying human cognition as a system of interacting parts that might someday be reproduced artificially [2].

The modern notion of machines that 'think' emerged in the mid-20th century, when developments in mathematical logic, computation, and electronics converged. In his seminal 1950 essay, Alan Turing proposed what he called the imitation game—now widely known as the Turing Test—as an operational criterion for intelligence [3]. Crucially, Turing shifted the focus from defining thinking in the abstract to asking whether a machine's outward behavior could be indistinguishable from that of a human interlocutor. Over the subsequent decade, artificial intelligence crystallized as a distinct research field, producing early systems for theorem proving, game playing, and symbolic problem solving [4]. The long-standing philosophical speculation that machines might match human judgment—encompassing perception, reasoning, and even moral evaluation—thereby began to transform into the empirical and technological project whose consequences we are observing today.

In recent years, large language models (LLMs) have arguably been the most disruptive step in this trajectory [5]. State-of-the-art systems such as ChatGPT, Deepseek, Gemini, Llama, and Mistral now routinely clear the bar of the Turing Test, in some cases more reliably so than humans [6]. Unsurprisingly, a growing body of work has proposed LLMs as stand-ins for human participants in social science experiments [7], market and consumer research [8], and a variety of applications in the healthcare, education, work and information domains [9–11], among others [12–15]. This comes on top of their widespread deployment for everyday tasks such as drafting and editing text, translation, summarization, and educational support [16]. Proponents have gone further, arguing that in many contexts LLMs may offer advantages over human samples, citing lower cost, greater scalability, and the ability to generate large volumes of synthetic data in domains where real-world data are limited or difficult to obtain [17, 18].

* walterquattrociocchi@gmail.com
† caprarovalerio@gmail.com
‡ matjaz.perc@gmail.com

However, serious concerns have already been voiced [19, 20], and a number of foundational issues remain unresolved. A central open question is how judgment itself is instantiated and operationalized in LLMs. These systems are rapidly becoming embedded in the processes by which societies filter, rank, and interpret information: assessing the credibility of news, proposing explanations, and assisting in decisions that hinge on evaluative judgments [21–23]. Yet the internal procedures by which they arrive at such judgments—and the extent to which these procedures align with, diverge from, or systematically distort human modes of reasoning—remain only partially understood [24, 25].

Moreover, the historical trajectory from symbolic AI to modern LLMs hides another important and commonly overlooked aspect of how machines handle language and knowledge. Namely, symbolic AI treated intelligence as rule-based manipulation of explicit symbols, with hand-crafted rules and representations [26]. Early neural networks introduced data-driven learning but remained limited in scale and impact for language. From the 1990s onward, statistical NLP—and later neural sequence models—came to dominate, especially in systems such as web search and recommendation, which filter information: they retrieve and rank existing documents, leaving users to inspect multiple sources and judge credibility [27, 28]. Generative systems like contemporary LLMs instead synthesize new text directly, producing a single, context-sensitive answer. This shift from filtering to generation is not merely a technical change; it constitutes an epistemic transition in how information is delivered and consumed. Instead of being presented with a landscape of candidate documents to evaluate, the user is handed a fluent, authoritative-seeming answer that collapses the underlying diversity of sources into a single textual surface, ready for immediate consumption.

In what follows, we expand on these premises, first by explaining how transformer architectures work at a high level, emphasizing that their apparent intelligence emerges only under conditions of massive scale [29]. Secondly, we frame text generation as a stochastic walk on a weighted graph [30], where nodes correspond to tokens and edges to learned transition probabilities. Each answer is thus a trajectory in this graph, conditioned by the prompt and decoding parameters. Crucially, we emphasize that there are no intrinsic 'attractors' corresponding to concepts or truth; the system does not converge, it transits. What looks like a conclusion is simply path completion in a high-dimensional probability landscape. We then formally describe and compare the human and artificial epistemic pipelines by outlining seven fundamental stages. For each stage, we identify an *epistemic fault line*: a critical point at which human and LLM judgments diverge. Based on these fundamental differences, we introduce Epistemia as the condition in which linguistic plausibility becomes a structural substitute for epistemic evaluation [24]. The user experiences the possession of an answer without having traversed the process of forming a justified belief, i.e., without the labor of knowing. We also describe the psychological foundations of epistemia, focusing on the human heuristics and biases that make individuals especially susceptible to this phenomenon. Lastly, we discuss the broader implications of the epistemological fault lines between human and artificial intelligence. We argue that persistent epistemic divergence—despite increasing surface alignment—requires rethinking how generative systems are assessed, regulated, and integrated into epistemic practices. We outline an interdisciplinary research program spanning epistemic evaluation beyond surface alignment, epistemic governance beyond behavioral alignment, and epistemic literacy beyond critical thinking, aimed at preserving judgment as an accountable human practice in hybrid human–AI systems.

II. TRANSFORMERS AND THE ROLE OF SCALE

Transformer architectures implement a powerful form of linguistic automation. At their core, they estimate the conditional probability of the next token given a preceding context, via stacked self-attention layers that propagate and remix information across positions in the input [29]. Formally, this amounts to learning a massively parameterized function that maps sequences of symbols into probability distributions over subsequent symbols [31, 32].

From an engineering standpoint, this is remarkable. Self-attention enables efficient integration of long-range dependencies and the construction of expressive internal representations of regularities. When combined with massive training corpora and scaling laws this architecture yields systems that appear fluent, versatile, and adaptable across domains [33, 34].

However, what is being automated here is not cognition but language. Large language models operate on statistical regularities extracted from human-produced text, not on representations of the world [35]. Their apparent competence arises from learning how language behaves, not from forming beliefs about what is the case. They do not track truth conditions or causal structure; they track patterns of co-occurrence, association, and continuation in text [36].

In this sense, scale is not a bridge from linguistic automation to cognition. Increasing the volume of data and the number of parameters refines a function approximator but does not alter the underlying computation [37]. Scale delivers coverage and interpolation, not epistemic access. It improves surface alignment with human output [38], without inducing convergence in internal processes.

This distinction matters because contemporary development strategies increasingly attempt to compensate for this limitation by layering additional mechanisms on top of the generative core. Prominent among these are retrieval-augmented generation (RAG) [39, 40], tool

use [41], and external memory modules [42]. These approaches aim to reconnect language models to external sources of information by anchoring generation to documents, databases, or APIs.

The result is an architecture that produces answers that *look* increasingly reliable without possessing the machinery that normally makes reliability possible. The system becomes more convincing rather than more knowing.

This shift becomes critical when generative models displace traditional information technologies. Search engines and filtering systems returned documents and left judgment to users. Generative systems deliver a synthesized answer directly as natural language [43]. Searching, selecting, and explaining collapse into a single response. The cost of evaluation is not postponed; it is structurally absorbed into the generation process.

It is under these conditions that plausibility begins to substitute for verification. Large language models often generate outputs that are fluent, coherent and expressed with confidence rather than grounded in rigorous evaluation processes—what once required an act of judgment is now presented as a product of computation [44, 45]. The danger is therefore not simply that generative systems may err, but that they succeed precisely by making evaluation optional [45, 46].

III. TEXT GENERATION AS A WALK ON A GRAPH

Text generation in large language models can be described as a stochastic process evolving on a discrete, high-dimensional state space. Let V be a finite vocabulary and let $G = (V, E)$ be a directed, weighted graph whose edges encode conditional transition probabilities learned from data. Given a context $c_t = (w_1, \ldots, w_t)$, the model instantiates a probability measure $P(\cdot \mid c_t)$ over V and samples a successor state $w_{t+1} \sim P(\cdot \mid c_t)$. This defines a time-inhomogeneous Markov process over G, consistent with classical formulations of random walks on graphs [30].

Each output is therefore the realization of a stochastic trajectory generated by local sampling in this state space. Greedy decoding, temperature scaling, top-k and nucleus sampling modulate the entropy and effective support of $P(\cdot \mid c_t)$, reshaping the local geometry of the probability space [47]. However, these procedures do not introduce invariants, constraints, or objectives associated with truth, reference, or validity. They merely alter how probability mass is explored.

Empirical language distributions are heavy-tailed and structurally anisotropic [48]: probability mass concentrates in a limited number of regions corresponding to frequent constructions, dominant frames, and statistically reinforced co-occurrence patterns. As a result, trajectories are dynamically biased toward high-density basins. This follows from well-known concentration phenomena in high-dimensional stochastic processes [49]: random walks overwhelmingly remain confined within regions of large measure, while transitions into low-density regions are exponentially suppressed. This dynamic produces a form of statistical attraction that is often misread as conceptual stability. In reality, dense regions of G are not semantic attractors but statistical aggregates. Mode persistence is not belief; recurrence is not memory; concentration is not understanding. What stabilizes is a distribution, not a meaning.

Within this framework, so-called "hallucinations" are not anomalous failure modes of an otherwise epistemic system. They are an expected outcome of sampling from a statistical model that does not encode reference, truth conditions, or evidential constraints. In a generative system, producing content that is ungrounded with respect to external reality is not the exception; it is the default operational state. Grounded outputs occur only when the local probability structure happens to coincide with factual structure, or when external mechanisms impose additional constraints [46, 50]. From an algorithmic standpoint, there is no internal symmetry break between truthful and false continuations. Both are merely realizations drawn from the same conditional distribution.

As scale increases, this regime does not qualitatively change. Larger models refine probability estimates and smooth local neighborhoods of the distribution, thereby increasing fluency and internal coherence. But scale does not inject epistemic structure into the process. It sharpens likelihood, not validity. Text generation is therefore an ergodic process under statistical constraints, not a procedure of epistemic convergence. It optimizes for distributional fit, not for correctness with respect to the world. A "conclusion" is not the terminus of evaluation, but the terminal state of a stochastic trajectory.

IV. HUMAN AND ARTIFICIAL EPISTEMIC PIPELINES

We decompose human judgment into seven sequential stages: sensory and social information; perceptual and situational parsing; memory, intuitions, and learned concepts; emotion, motivation, and goals; reasoning and information integration; metacognitive calibration and error-monitoring; and value-sensitive judgment. These operations are slow, imperfect, and biased, yet they unfold within an epistemic loop in which the world, other agents, and institutions continually push back, constraining error.

We then map LLM judgment onto the same scaffold. Textual prompts replace sensory and social information; tokenization and preprocessing replace perceptual and situational parsing; pattern recognition in embeddings replaces memory, intuitions, and learned concepts; statistical inference via neural layers replaces emotion, motivation, and goals; textual context integration replaces reasoning and information integration; forced confidence and hallucination replace metacognition; and probabilis-

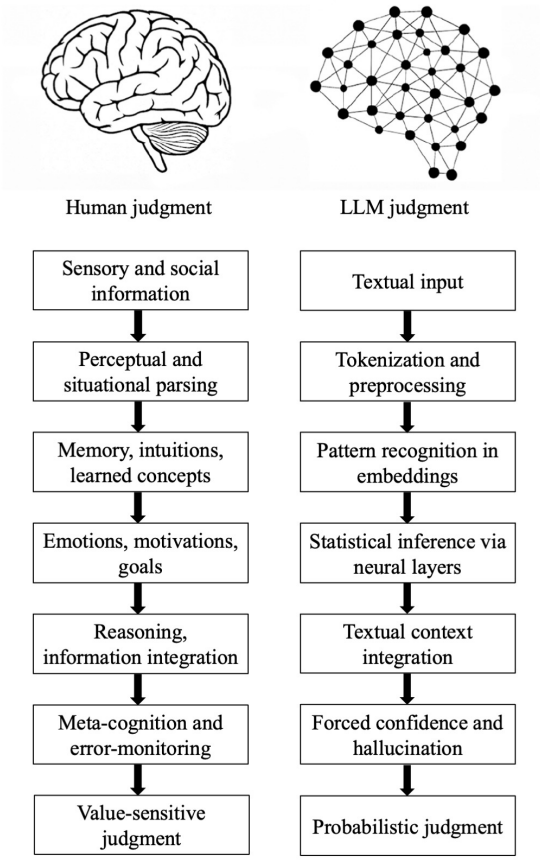

FIG. 1. The human and LLM epistemic pipelines, each organized into seven corresponding stages.

tic prediction replaces value-sensitive judgment (see Figure 1).

At each stage, the processes appear parallel yet diverge sharply in structure, function, and epistemic grounding. These contrasts expose key epistemological fault lines: points at which the two pipelines follow fundamentally different trajectories despite sometimes yielding superficially similar outputs. For each fault line, we illustrate concrete cases in which human and LLM judgments are likely to diverge.

A. Sensory and social information vs. Textual input

Human judgment begins with the acquisition of sensory and social information in an inherently multimodal environment. Vision, audition, proprioception [51], and emotional expressions [52] jointly shape how situations are initially construed. This information is not isolated but embedded within a social world rich with affective signals: facial expressions [53], tone of voice [54], social cues and norms shaping interpretation [55], and even power dynamics modulating how emotional signals are processed [56].

LLMs, by contrast, begin with textual input. They do not inhabit or sample a world but operate over abstracted representations of it. They do not perceive environments, bodies, or social or emotional signals; they receive sequences of symbols whose significance is entirely derivative of statistical patterns learned during pretraining and subsequently adjusted through supervised and reinforcement-based fine-tuning [57, 58]. This input is stripped of nearly every feature that gives human perception its world-directed richness: no gesture, no affective tone, no temporal continuity, no shared situation. Although recent multimodal models can accept images, audio, or video as input, their "perceptual access" remains fundamentally derivative: the system receives pretrained embeddings rather than engaging in sensorimotor exploration or bodily interaction.

A direct consequence of this absence of perceptual grounding is that LLMs sometimes make judgments that would be unthinkable for a human at this initial stage of input acquisition. For example, when given a transcript of a conversation without vocal tone, gesture, or context, an LLM may misinterpret sarcasm as sincerity, fail to detect anger or fear, or treat a threat ("say that again and see what happens") as a neutral statement. Humans would effortlessly register these nuances because multimodal cues (voice tension, facial expression, interpersonal distance) are part of the perceptual input itself. For an LLM, none of these signals exist. What is immediate and unambiguous for a human is invisible to the system. And while it is true that recent LLM-based and multimodal models have shown some improved ability to recognize sarcasm and emotional tone from text, their performances typically remain below human levels [59]. Additionally, these improvements are fragile. A recent systematic review emphasizes that irony and sarcasm detection often remains unreliable and that performance degrades sharply when conversations involve cultural subtleties, indirect speech acts, or noisy, real-world language use [60].

This is the first epistemological rupture. Humans ground judgments in perceptual reality and social context. LLMs must reconstruct meaning indirectly from text alone.

B. Perceptual and situational parsing vs. Tokenization and text preprocessing

After receiving sensory and social information, humans engage in perceptual and situational parsing, a tightly integrated process that transforms raw experience into meaningful structure. These operations occur simultaneously, in mutually constraining loops [61, 62], and extend seamlessly to the interpretation of social cues such as gaze, intention, and affect [63]. Perception does not merely register stimuli: it actively organizes them into meaningful structure, identifying objects and opportu-

nities for action [51] and, through grounded conceptual processes, recognizing agents, intentions, and potential threats [64]. At the same time, higher-level expectations, cultural schemas, and social knowledge shape what is perceived as salient or relevant [65, 66]. By the time a human has parsed a situation, they have already extracted a structured understanding of that situation, embedded within physical, interpersonal, and normative contexts.

For LLMs, the analogous stage is tokenization and text preprocessing, a transformation that is fundamentally mechanical. Here, raw linguistic input is segmented into discrete symbols (tokens) according to a predetermined vocabulary optimized for model efficiency [67–69]. Tokenization is blind to pragmatics, speaker intention, emotional tone, and situational nuance; it does not infer objects, agents, or social dynamics. Tokenization simply maps character strings to indices [35]. Preprocessing operations, such as lowercasing, special token insertion, or punctuation handling, further standardize the input but do not add semantic structure [70]. This stage produces a representation that is structurally convenient but semantically thin, designed for numerical computation rather than interpretation.

A direct consequence of this symbolic segmentation is that LLMs can make errors that no human would ever make at this stage. Because the model processes strings rather than situations, even simple linguistic inputs may fracture into misleading subword units (e.g., "therapist" tokenized as "the rapist"). These are not superficial errors but structural outcomes of a system that slices text into tokens rather than parsing scenes, intentions, or events [71]. Because LLMs rely on subword tokenization, even minor typographical or formatting changes can distort meaning. For instance, in Chinese, tokenization can split characters in ways that break their semantic radicals [72]. Likewise, subword tokenizers may mis-handle prefixes or suffixes that signal negation, leading models to misunderstand the intended meaning [73].

Therefore, at this second stage the epistemological fault line widens. Human perception has already constructed a layered, meaning-rich model of the environment. LLMs, at an equivalent stage, have performed only a formal partitioning of text. One system parses a world; the other segments a string.

C. Memory, intuitions, learned concepts vs. Pattern recognition in embeddings

Next, both humans and LLMs draw on prior knowledge, but they do so in fundamentally different ways. Humans rely on episodic memory, intuitive physics and psychology, and learned concepts. Episodic memory contains specific events encoded with temporal and contextual details. These memories enable individuals to recognize analogies, anticipate social consequences, and interpret new situations through the lens of prior lived experience [74, 75]. Humans also possess core knowledge systems, innate or early-developed and pre-linguistic, such as intuitive physics (object permanence, solidity, gravity, causal forces) and intuitive psychology (attribution of beliefs, desires, and intentions to others), which develop early and scaffold perception and reasoning throughout life [76, 77]. Additionally, humans possess learned concepts representing abstract, generalized knowledge: categories, scripts, causal theories, and social norms accumulated through education, culture, and repeated practice [65, 78, 79]. In judgment, humans fluidly combine these systems, retrieving specific past experiences to contextualize concepts and using conceptual frameworks to interpret ambiguous situations [74, 78, 79].

LLMs, by contrast, rely on statistical pattern extraction in high-dimensional embedding spaces: words that co-occur, sentences that share structure, or concepts appearing in similar contexts. Modern embedding models, such as word2vec and transformer-based representations, encode similarity, not experience [70, 80, 81]. They have no episodic memory: nothing like a lived past, no autobiographical trace of events, no temporally structured recollection of "what happened when". They cannot draw on intuitive physics or intuitive psychology: no innate sense of object solidity or gravity, no built-in understanding of beliefs, desires, or intentions, no causal expectations about agents or objects: only correlations between how such ideas tend to be discussed in text. Nor do they possess learned concepts in the human sense: their "concepts" are not abstractions built from experience or education but statistical clusters reflecting how words are distributed across training data.

A direct consequence of these differences is that LLMs may treat physical impossibilities as plausible whenever such scenarios appear in linguistic corpora. Even multimodal models fail in intuitive-physics and causal-reasoning tasks [82]. LLMs may also fail to track beliefs, intentions, and deception in situations where even young children succeed, because they lack an intuitive psychology for representing distinct mental states [83]. And they may produce conceptual blends when words have multiple senses, especially metaphoric ones [84].

This third step further widens the fault line: humans ground interpretation in lived experience, intuitive models of the physical and social worlds, and conceptual understanding, whereas LLMs rely solely on statistical associations learned from language.

D. Emotion, motivation, goals vs. Statistical inference via neural layers

As humans process percepts and retrieve prior knowledge, their judgments are continuously shaped by emotion, motivation, and goals: the affective and purposive forces that give cognition direction [85, 86]. Emotions modulate attention, signal relevance, and provide rapid evaluations of risk, opportunity, and social meaning [87, 88]. Motivation orients individuals toward desired

outcomes, while goals structure decision-making by defining what counts as success or failure in a given context [89]. These forces are inherently value-laden, grounding human judgments in personal identity, social commitments, moral principles, and long-term aspirations [90, 91]. Often, these motivations can be traced to evolutionarily shaped concerns such as fear of death [92], the instinct for self-preservation [93], and efforts to leave a lasting mark or legacy to reach symbolic immortality [94, 95].

For LLMs, the corresponding stage is statistical inference through layered neural computation. Given tokenized input and embedding representations, the model propagates activations through its transformer architecture, updating vector representations according to learned parameters. Each layer performs linear transformations, attention-weighted aggregation, and nonlinear mappings that compute the probability distribution over next tokens [29, 96]. This process is entirely mechanistic and optimization-driven: it aims to minimize predictive error, not to pursue goals or respond to emotional salience. The model does not care about truth, utility, moral implications, outcomes, or death. It has no preferences, motivations, or internal states beyond the numerical activations that encode statistical associations [57].

Although training regimes such as reinforcement learning with human feedback (RLHF) introduce externally specified reward signals meant to shape model behavior toward human values, this mechanism does not give the model intrinsic goals or motivations; it merely adjusts statistical tendencies through additional optimization [58]. RLHF teaches a model to behave as if it held certain preferences, but without creating any internal states that resemble value commitments, desires, or purposes [97]. Paradoxically, these alignment adjustments can introduce backfire effects, including political biases [98] and "surprising" gender biases, such as systematically responding that a woman should not be harassed to prevent a nuclear apocalypse, while accepting that a woman may be tortured to achieve the same outcome [99]. Inference is therefore procedural rather than purposive, a sequence of inflexible matrix operations rather than a value-oriented interpretive and generalizable act.

At this stage, the epistemological fault line gets more profound: humans create judgments under the influence of goals and emotions that confer meaning, priority, and direction, whereas LLMs perform context-agnostic statistical transformations devoid of intrinsic aims.

E. Reasoning and information integration vs. Textual context integration

Next, both humans and LLMs attempt to produce a coherent response given prior inputs. At this more advanced stage of judgment, humans engage in reasoning, a cognitive process that allows them to draw inferences, integrate evidence, form causal explanations, consider counterfactual possibilities, and construct long-term plans [100–102]. These operations allow individuals to extend beyond immediate perception and derive conclusions from principles, rules, or structured mental models. This process is not merely based on universal rules, but it is guided by goals, values, and an awareness of the limits of one's own knowledge [103, 104]. At this reflective stage, humans evaluate alternatives against different kinds of standards: objective ones, such as mathematical correctness or factual accuracy [105], and subjective or internal ones, such as personal moral values, ideals, and self-guides [106].

For LLMs, the analogous stage is textual context integration within model constraints, a process that lacks any genuine reasoning. Given a sequence of tokens, the model integrates them through attention mechanisms that weight relationships between elements in its input window [29]. This allows the LLM to maintain thematic coherence, track referents, and adapt output to preceding content [70]. Yet this form of 'integration' remains purely syntactic and statistical: the model does not generate causal hypotheses, adjudicate between competing interpretations, nor does it construct a causal model of the world [107]. Whereas humans base judgments on causal understanding, LLMs rely on correlations, making them especially vulnerable to spurious associations [108] and prone to characteristically non-human errors [109, 110]. For example, researchers observed a substantial performance drop when models were asked to reason about unseen or hypothetical causal relationships, a strong indication that LLMs do not construct causal models but rather rely on surface associations [111–113].

At this point, the epistemological divergence becomes irreconcilable: human cognition is goal-directed and organized around causal models of the world, whereas LLM cognition lacks intrinsic goals and is driven by statistical patterns vulnerable to spurious correlations.

F. Metacognitive calibration and error-monitoring vs. Forced confidence and hallucination

After constructing preliminary interpretations and engaging in reasoning, humans deploy a distinct layer of metacognitive evaluation: monitoring uncertainty, detecting potential errors, estimating confidence, and, when necessary, withholding judgment. These processes rely on neural systems for conflict detection [114], error monitoring [115], and confidence estimation [116]. Metacognitive signals guide whether individuals double-check facts, seek additional information, revise faulty assumptions, or suspend belief. Even young children demonstrate uncertainty monitoring and the ability to acknowledge not knowing [117]. Humans sometimes integrate social-epistemic norms into metacognitive evaluation: they take into account the stakes and social context, assess whether a judgment ought to be made at all, and adjust their expressed confidence accordingly, for example, to signal

trustworthiness or to avoid reputational costs [118, 119]. Thus, metacognition functions not merely as internal error checking but as a socially embedded mechanism regulating when and how judgments are expressed.

LLMs lack metacognition entirely. They do not possess internal monitors for conflict, uncertainty, error likelihood, or epistemic stakes. They cannot track the reliability of their own representations; they cannot realize they "do not know", and they are notoriously reluctant to admit it. When LLMs generate statements of uncertainty ("I'm not sure"), these are linguistic constructions, not internal confidence signals. Consequently, LLMs routinely produce hallucinations because nothing in the architecture encodes epistemic humility. Hallucination is not a bug but a structural consequence of maximizing next-token probability under incomplete constraints [46, 50]. Without a capacity to represent uncertainty or suspend judgment, LLMs simulate confidence continuously, even when wrong or uninformed.

At this stage, the epistemological rupture becomes definitive: humans possess a self-regulating, uncertainty-sensitive mechanism that supervises judgment formation; LLMs possess none. Humans can refrain from judging; LLMs must predict.

G. Value-sensitive judgment vs. Probabilistic judgment

The culmination of the human epistemic pipeline is the formation of value-sensitive, context-dependent judgments. Individuals evaluate situations in light of personal values, cultural norms, reputational concerns, and long-term goals. Judgments therefore express not only what someone believes but who they are and what they care about. Moreover, human judgments are shaped by real-world stakes: errors have consequences for relationships, livelihoods, and identities. This imbues human decision-making with a sense of accountability and normativity.

LLMs, by contrast, produce probabilistic, text-based judgments determined by the statistical structure of their training data and the immediate textual prompt. An LLM's "judgment" is simply the next-token distribution conditioned on context. It does not evaluate truth, moral weight, or pragmatic consequences: it estimates what sequences of words are most likely given patterns learned from text corpora. Its outputs do not reflect intrinsic preferences, values, or goals. Errors carry no internal repercussions, and contradictions do not undermine its epistemic integrity. The model cannot defend a judgment except by generating further statistically plausible text.

In this stage, the epistemological divide has real-world consequences. For humans, a judgment is a world-directed, value-infused commitment, that integrates causal models of the world with emotion, identity, and moral purpose. For LLMs, a judgment is merely a linguistic prediction. Even when their outputs superficially align, the underlying epistemic procedure is fundamentally different.

V. THE FAULT LINES

The previous section highlights seven epistemological fault lines that separate human judgment from LLM judgment, one for each stage of the epistemic pipeline.

The *Grounding* fault captures the fact that humans begin judgment with perceptual and social information, whereas LLMs begin with text, reconstructing meaning indirectly from symbols. The *Parsing* fault highlights how humans parse situations through rich perceptual and conceptual mechanisms, while LLMs perform a purely formal segmentation into tokens. The *Experience* fault reflects how humans rely on episodic memory, intuitive physics and psychology, and learned concepts, whereas LLMs rely solely on statistical associations in embedding space. The *Motivation* fault points to the role of emotions, goals, and values in guiding human cognition, contrasted with the goal-free optimization dynamics of LLMs. The *Causality* fault signals the divergence between human causal reasoning and LLM reliance on correlations. The *Metacognitive* fault emphasizes humans' ability to monitor uncertainty and withhold judgment, in contrast to LLMs' structural inability not to produce an answer. Finally, the *Value* fault delineates how human judgments embody identity, morality, and real-world stakes, while LLM judgment consists only of probabilistic predictions without intrinsic valuation. Each fault line corresponds to a stage in the epistemic pipeline where human and artificial processes diverge in structure, function, and epistemic grounding. These fault lines are summarized in Table I.

Despite these fault lines, LLM outputs often appear superficially fluent [120, 121] and confidently articulated [122], even when they are factually wrong [123]. This creates substantial room for an "illusion of veracity", a systematic divergence between the actual accuracy of an LLM output and the accuracy perceived by a human user. Such illusion descends from the fact that people routinely use these very properties—fluency and confidence—as credibility heuristics in everyday judgment. In everyday communication, fluency typically correlates with familiarity, honesty, and communicative competence, making it a reliable heuristic; people infer accuracy from processing ease even when the fluency is artificially induced [124–127]. Similarly, expressed confidence is a powerful cue to credibility, over and above how fluently a message is written. In a classic mock-trial study, more confident witnesses were judged as more credible and more expert by jurors [128]. More recently, participants were more likely to believe and trust a highly confident eyewitness than a cautious one describing the same accident; only with experience did they begin to discount miscalibrated confidence [129]. Building on such findings, scholars have proposed a "confidence heuristic", whereby, in the absence of stronger diagnostic cues, expressed confidence

Epistemological fault line	Definition
The Grounding fault	Humans anchor judgment in perceptual, embodied, and social experience, whereas LLMs begin from text alone, reconstructing meaning indirectly from symbols.
The Parsing fault	Humans parse situations through integrated perceptual and conceptual processes; LLMs perform mechanical tokenization that yields a structurally convenient but semantically thin representation.
The Experience fault	Humans rely on episodic memory, intuitive physics and psychology, and learned concepts; LLMs rely solely on statistical associations encoded in embeddings.
The Motivation fault	Human judgment is guided by emotions, goals, values, and evolutionarily shaped motivations; LLMs have no intrinsic preferences, aims, or affective significance.
The Causality fault	Humans reason using causal models, counterfactuals, and principled evaluation; LLMs integrate textual context without constructing causal explanations, depending instead on surface correlations.
The Metacognitive fault	Humans monitor uncertainty, detect errors, and can suspend judgment; LLMs lack metacognition and must always produce an output, making hallucinations structurally unavoidable.
The Value fault	Human judgments reflect identity, morality, and real-world stakes; LLM "judgments" are probabilistic next-token predictions without intrinsic valuation or accountability.

TABLE I. Seven epistemological fault lines marking structural divergences between human and LLM judgment.

substitutes for knowledge, competence, and correctness [130].

VI. EPISTEMIA

We define *Epistemia* [24] as the structural condition in which linguistic plausibility substitutes for epistemic evaluation. It designates a regime in which systems produce answers that are syntactically well-formed, semantically fluent, and rhetorically convincing, without instantiating the processes by which beliefs are normally formed, tested, and revised [57, 107]. The user experiences the possession of an answer without having traversed the cognitive labor of judgment [131].

Epistemia is not a psychological quirk and not a transient misuse of technology. It is not reducible to *automation bias*—the tendency to over-trust automated recommendations [132]—nor to a mere problem of misplaced authority attribution, in which users incorrectly treat a system as an expert [133]. Both automation bias and authority effects can exacerbate Epistemia, but they presuppose that the underlying system is, at least in principle, an epistemic agent that could deserve or fail to deserve trust. In the case of large language models, this presupposition is false. The core issue is not that users trust the wrong source, but that they are interacting with a source that does not possess any internal mechanisms for forming, holding, or revising beliefs at all [134, 135].

Epistemia is, instead, an architectural phenomenon that arises whenever generative systems are inserted into epistemic workflows while lacking internal machinery for reference, verification, or belief maintenance. Under these conditions, plausibility becomes a functional surrogate for justification. What is optimized is not the correctness of claims with respect to the world, but their fit with a learned distribution of linguistic usages.

The defining mark of Epistemia is the decoupling of content from evaluation. In human cognition, judgment is embedded in an epistemic loop: claims are checked against evidence, beliefs collide with counterexamples, and conclusions remain revisable in light of new information and social feedback. In generative systems, by contrast, there is no internal locus where claims can be tested, withdrawn, or defended. The model does not distinguish between "true" and "false" continuations; it distinguishes between more and less likely ones. What is generated is not what holds, but what fits.

Epistemia is therefore the outcome of a precise misalignment: highly sophisticated linguistic competence coupled with the absence of epistemic control. As generative systems improve, this mismatch becomes more dangerous, not less. The more persuasive the system becomes, the easier it is to confuse coherence with correctness, fluency with reliability, and stylistic competence with knowledge.

Importantly, Epistemia does not depend on error rates. It persists even when systems are factually accurate. The core harm is not the production of falsehoods, but the structural bypassing of evaluation itself. When answers are delivered in finalized form, without visible traces of uncertainty, conflict, or evidential grounding, the user is

placed in a position of epistemic passivity. Judgment is not exercised; it is consumed.

In this sense, Epistemia marks a transformation not in what is known, but in how knowing is produced. It shifts epistemic activity from a process to a product. The operative question is no longer "What should I believe, given the available evidence?" but "What sounds right, given what is presented to me?" The mechanisms of scrutiny, contestation, and revision are displaced by mechanisms of immediate acceptance or rejection of pre-packaged answers.

Epistemia thus names a reconfiguration of the epistemic environment: a world in which access to linguistically competent outputs becomes easier than access to justified beliefs, and in which the experience of understanding detaches from the practice of justification. It is in this gap—between fluent answers and accountable cognition—that a new form of epistemic instability takes root.

VII. DISCUSSION AND OUTLOOK

This Perspective examined a growing but often undertheorized tension at the core of contemporary generative AI: the simultaneous increase in *surface* alignment between human and machine outputs and the persistence of deep epistemic divergence in the processes that generate them. The central point is not that LLMs cannot produce useful text—they plainly can—but that their apparent resemblance to human judgment is primarily a resemblance in *linguistic form* and social presentation, rather than in the epistemic operations that make judgment answerable to the world.

By systematically comparing human and artificial epistemic pipelines, we argued that current LLM architectures lack the mechanisms that make human judgment possible across seven epistemic fault lines: grounding, parsing, experience, motivation, causality, metacognition, and value. What these systems provide instead is a highly optimized form of linguistic continuation, capable of producing contextually appropriate and rhetorically convincing outputs without performing epistemic evaluation. This is precisely why the most salient risk is not reducible to occasional inaccuracy or bias. The risk is structural: correctness becomes decoupled from the processes of justification that normally sustain it, and thus from the institutional and psychological practices through which epistemic responsibility is enacted.

To characterize this condition, we introduced *Epistemia*: a structural regime in which linguistic plausibility substitutes for epistemic evaluation, generating the experience of knowing without the cognitive labor of judgment. Epistemia is not a transient misuse pattern nor a defect that disappears with better benchmarks. It is not resolved by scale, higher scores, or more convincing behavior. It arises from architectural features of generative systems, and therefore persists even when outputs are accurate, calibrated, or behaviorally aligned. Indeed, as generative models become more capable, the *felt* reliability of their outputs often increases faster than the system's capacity to warrant that reliability, thereby widening the practical gap between persuasion and justification.

These observations carry direct implications for how generative systems are evaluated, governed, and integrated into epistemic practices. We outline three such implications below.

A. Epistemic evaluation beyond surface alignment

Current evaluation paradigms for large language models overwhelmingly rely on surface alignment: agreement with human answers, task success, or behavioral similarity under controlled prompts [136–138]. While these evaluations remain a necessary condition for any meaningful notion of alignment, from the perspective of Epistemia they are systematically insufficient. They primarily test whether outputs *look right*—that is, whether they resemble a target distribution of human responses—not whether they are produced through processes that sustain judgment under uncertainty, contestation, and worldly constraint [24, 35]. Because LLMs can achieve impressive task performance without instantiating mechanisms for grounding, causal modeling, uncertainty monitoring, or value-sensitive commitment, output-focused benchmarks risk mistaking linguistic competence for epistemic competence [57, 139].

This limitation is most consequential in domains where justification, error awareness, and responsibility are constitutive of competent practice: science, medicine, law, journalism, and public policy [140, 141]. In such settings, the epistemic cost of error is not exhausted by a wrong answer. It includes inappropriate confidence, the presentation of conjecture as settled fact, brittle reasoning that collapses under distributional shift, and the inability to recognize when abstention or deference is the normatively correct move. Even when a response happens to be correct, the absence of an internal epistemic loop can still be harmful if it trains users and institutions to treat a fluent completion as a substitute for warranted belief.

Future research should therefore complement output-level evaluation with process-sensitive probes. Concretely, this means designing tests that target (i) *uncertainty management* (when and how a system expresses uncertainty, requests missing information, or refuses to answer), (ii) *counterfactual sensitivity* and causal stability (whether conclusions track interventions rather than surface associations), (iii) *robustness to correlation-breaking shifts* (where distributional regularities diverge from the world structure users actually care about), and (iv) *normative appropriateness of abstention* (tasks where withholding judgment is the epistemically correct outcome). The key shift is conceptual: epistemic evaluation must move from asking whether models can repli-

cate human-looking judgments to whether their behavior is coupled to mechanisms that preserve the meaning of judgment under epistemic stress [24, 35]. In an Epistemia-prone environment, evaluating only the product of generation is a category error; what must be evaluated is the reliability of the *pipeline* that delivers that product.

B. Epistemic governance beyond behavioral alignment

Much of the current governance discourse around LLMs, and generative AI more broadly, is organized around behavioral alignment, understood as ensuring that systems produce safe, compliant, and socially acceptable outputs, rather than around guarantees about the epistemic processes underlying those outputs [142–145]. This focus is necessary but insufficient under conditions of Epistemia, because the core failure mode is not merely that a system says something harmful, but that it is positioned to *replace* or *short-circuit* human and institutional judgment while lacking the epistemic capacities that would make such substitution legitimate.

AI governance therefore requires a shift from regulating *what* systems say to regulating *how* generative outputs are introduced into epistemic workflows, and where they may permissibly substitute for human judgment. Several governance implications follow.

First, governance should explicitly distinguish domains where generative outputs can be assistive from domains where they functionally become decision procedures. In high-stakes settings, the relevant question is not whether the model can often provide the right answer, but whether its use induces epistemic passivity: collapsing search, adjudication, and justification into a single authoritative-seeming response. In practical terms, this supports governance measures that require human-in-the-loop review with clearly specified accountability, especially where responsibility cannot be meaningfully delegated to a non-epistemic system [140, 141].

Second, disclosure requirements should be reframed as epistemic transparency obligations rather than generic "AI use" labels. Under Epistemia, the crucial information is not simply that a system was used, but what epistemic functions it did *not* perform: whether it grounded claims, checked sources, tracked uncertainty, or could have abstained. This suggests that governance and organizational policy should demand context-appropriate disclosures about evidential status (supported versus conjectural), confidence (including when confidence is merely stylistic), and limitations tied to the seven fault lines identified above [24, 35, 57].

Third, governance should invest in epistemic risk taxonomies distinct from conventional safety taxonomies. Traditional "AI safety" tends to emphasize toxicity, malicious use, and direct harms. Epistemic risk includes the degradation of justificatory norms, the institutionalization of plausibility as a decision criterion, and the displacement of distributed epistemic checks (peer review, second opinions, adversarial scrutiny) by a single generative interface. Formalizing these categories would make it possible to specify safeguards proportionate to epistemic stakes, not merely to the possibility of offensive content.

Finally, governance should treat technical add-ons (retrieval, tool use, external memory) as partial mitigations rather than epistemic solutions. They may reduce certain error rates, but they do not by themselves instantiate belief, understanding, or accountability. Without explicit governance constraints, the addition of such mechanisms can even intensify Epistemia by further increasing the persuasive authority of outputs while leaving responsibility diffuse [24, 35].

C. Epistemic literacy beyond critical thinking

Under conditions of Epistemia, users are increasingly exposed to fluent outputs that simulate judgment. This places new demands on education and professional training [146]. Classical accounts of critical thinking emphasize the evaluation of arguments and evidence, but they were largely designed for epistemic environments in which the production, evaluation, and ownership of judgment are co-located within a single epistemic agent capable of generating reasons, revising beliefs, and bearing responsibility for error [147–149]. In generative settings, by contrast, the production of plausible reasons can be automated, while the work of evaluation and accountability remains human—often invisibly so.

We therefore propose *epistemic literacy* as a distinct competence that must be explicitly conceptualized and taught in the age of generative AI. Whereas critical thinking focuses on assessing the validity, coherence, and evidential support of arguments—typically at the level of individual claims or lines of reasoning—epistemic literacy focuses on navigating epistemic environments in which judgment is staged, mediated, and distributed across humans and machines. It equips users to recognize when apparent judgments are the product of statistical pattern completion rather than epistemic evaluation, and to identify which dimensions of judgment remain irreducibly human.

In practice, epistemic literacy includes at least three families of skills.

First are *pipeline awareness* skills: understanding the difference between a system that retrieves evidence and one that synthesizes text; recognizing when a response is likely to be a completion rather than an evaluation; and anticipating the characteristic signatures of the seven fault lines (for example, the mismatch between fluent explanations and absent causal commitment, or between confident tone and absent uncertainty monitoring) [24, 35, 57].

Second are *procedural safeguards* for everyday use: habits of verification proportionate to stakes, routines

for cross-checking against independent sources, and explicit norms for when to defer judgment (including when to seek expert review) rather than treating the presence of a coherent answer as closure. Importantly, this is not merely "be skeptical." It is learning to reintroduce, at the level of practice, the epistemic loop that generative systems bypass: evidence seeking, contestation, and revisability.

Third are *institutional competencies*: designing workflows, classroom practices, and professional standards that prevent the outsourcing of epistemic responsibility to generative interfaces. This includes making uncertainty visible where appropriate, requiring provenance or justification for consequential claims, and clarifying accountability when AI-assisted outputs circulate in organizations. Epistemic literacy thus extends beyond the individual user to the norms and infrastructures that determine whether a society treats plausibility as a substitute for justification.

In this sense, epistemic literacy does not replace critical thinking but complements it. It extends critical thinking to settings in which the central epistemic challenge is no longer only the evaluation of arguments, but the governance of judgment in hybrid human–AI systems: deciding when to use generative tools, how to interpret their outputs, and how to preserve the social and institutional practices that keep belief formation answerable to the world.

VIII. CONCLUSION

This Perspective argued that contemporary large language models occupy a distinctive epistemic position: they can produce outputs that are often indistinguishable from human judgments while relying on a generative mechanism that is not itself a form of judgment. By framing text generation as stochastic path completion in a high-dimensional space of learned linguistic transitions, we emphasized that the impressive behavioral alignment of LLMs is compatible with a deeper structural mismatch in how conclusions are produced. The appearance of understanding can therefore coexist with the absence of the epistemic operations that make understanding accountable to the world.

We made this mismatch explicit by mapping human and artificial epistemic pipelines and identifying seven epistemic fault lines that separate human and LLM judgment: grounding, parsing, experience, motivation, causality, metacognition, and value. Across these divergences, human judgment remains embedded in an epistemic loop that couples perception, memory, affect, causal modeling, uncertainty regulation, and normative commitment to a world that can push back. LLM outputs, by contrast, are synthesized continuations conditioned on text and decoding dynamics: they can be coherent, persuasive, and often correct, yet they are not produced by a system that forms beliefs, adjudicates evidence, monitors epistemic error, or bears stakes. The relevant divide is thus not between "intelligent" and "unintelligent" systems, but between epistemic agents and systems that simulate the surface form of agency without instantiating its underlying constraints.

On this basis, we introduced *Epistemia* as the structural condition in which linguistic plausibility becomes a surrogate for epistemic evaluation, producing the experience of knowing without the labor of judgment. Epistemia is not reducible to user naivety, occasional hallucination, or the misuse of an otherwise epistemic tool. It is an architectural and socio-technical phenomenon that arises when generative systems deliver finalized, fluent answers in contexts where justification, uncertainty, and revisability are essential. As generative models scale and become more persuasive, the risk is not only that errors persist, but that evaluation is systematically displaced: the epistemic workload is shifted from the system to the user and, more importantly, made easier to omit.

The framework developed here motivates a broader research program integrating behavioral, cognitive, and computational sciences to systematically compare how humans and machines respond to uncertainty, causal disruption, moral trade-offs, and epistemic conflict. Such a program should move beyond surface performance and explicitly target process-level capacities: when abstention is warranted, how uncertainty is represented or simulated, how causal counterfactuals are handled when correlations fail, and how value-sensitive commitments emerge (or do not) under stakes. The goal is neither to anthropomorphize LLMs nor to force them into human epistemic categories, but to delimit, with empirical and formal precision, which epistemic functions can be meaningfully delegated to generative systems and which must remain human or institutionally distributed.

Finally, the practical stakes of clarifying these fault lines are societal. Evaluation regimes that reward plausibility, governance frameworks that regulate only outward behavior, and educational practices that treat fluent synthesis as comprehension together create the conditions for Epistemia to become normalized. Conversely, explicitly acknowledging the epistemological discontinuities between human cognition and generative transformers provides a basis for redesigning benchmarks, policies, and literacies around epistemic responsibility rather than rhetorical competence. In an epistemic environment increasingly organized around generative AI, preserving judgment as a genuinely accountable, human-directed practice requires more than better models. It requires maintaining the social and institutional conditions under which reasons can be demanded, errors can be owned, and belief remains answerable to evidence.

ACKNOWLEDGMENTS

M.P. was supported by the Slovenian Research and Innovation Agency (Grant No. P1-0403).

[1] W. Lenzen, Leibniz and the calculus ratiocinator, in *Technology and Mathematics: Philosophical and Historical Investigations*, edited by S. Hansson (Springer, Cham, 2018) pp. 47–78.

[2] K. Gunderson, Descartes, La Mettrie, language, and machines, Philosophy **39**, 193 (1964).

[3] A. M. Turing, Computing machinery and intelligence, in *Parsing the Turing test: Philosophical and methodological issues in the quest for the thinking computer*, edited by R. Epstein, G. Roberts, and G. Beber (Springer, Cham, 2007) pp. 23–65.

[4] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach* (Prentice Hall, Saddle River, NJ, 1995).

[5] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, *et al.*, A survey of large language models, arXiv:2303.18223 (2025).

[6] C. R. Jones and B. K. Bergen, Large language models pass the turing test, arXiv:2503.23674 (2025).

[7] I. Grossmann, M. Feinberg, D. C. Parker, N. A. Christakis, P. E. Tetlock, and W. A. Cunningham, AI and the transformation of social science research, Science **380**, 1108 (2023).

[8] J. Brand, A. Israeli, and D. Ngwe, Using GPT for Market Research, in *Proceedings of the 25th ACM Conference on Economics and Computation*, edited by D. Bergemann (Association for Computing Machinery, New York, NY, 2024) pp. 613–613.

[9] M. Cascella, J. Montomoli, V. Bellini, and E. Bignami, Evaluating the feasibility of ChatGPT in healthcare: An analysis of multiple clinical and research scenarios, J. Med. Syst. **47**, 33 (2023).

[10] V. Capraro, A. Lentsch, D. Acemoglu, S. Akgun, A. Akhmedova, E. Bilancini, J.-F. Bonnefon, P. Brañas-Garza, L. Butera, K. M. Douglas, *et al.*, The impact of generative artificial intelligence on socioeconomic inequalities and policy making, PNAS Nexus **3**, pgae191 (2024).

[11] S. Uygun İlikhan, M. Özer, M. Perc, H. Tanberkan, and Y. Ayhan, Complementary use of artificial intelligence in healthcare, Med. J. West. Black Sea **9**, 7 (2025).

[12] Y. K. Dwivedi, N. Kshetri, L. Hughes, E. L. Slade, A. Jeyaraj, A. K. Kar, A. M. Baabdullah, A. Koohang, V. Raghavan, M. Ahuja, *et al.*, Opinion Paper: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy, Int. J. Inf. Manage. **71**, 102642 (2023).

[13] K.-B. Ooi, G. W.-H. Tan, M. Al-Emran, M. A. Al-Sharafi, A. Capatina, A. Chakraborty, Y. K. Dwivedi, T.-L. Huang, A. K. Kar, V.-H. Lee, *et al.*, The potential of generative artificial intelligence across disciplines: Perspectives and future directions, J. Comput. Inf. Syst. **65**, 76 (2025).

[14] P. Budhwar, S. Chowdhury, G. Wood, H. Aguinis, G. J. Bamber, J. R. Beltran, P. Boselie, F. Lee Cooke, S. Decker, A. DeNisi, *et al.*, Human resource management in the age of generative artificial intelligence: Perspectives and research directions on ChatGPT, Hum. Resour. Manag. J. **33**, 606 (2023).

[15] M. A. K. Raiaan, M. S. H. Mukta, K. Fatema, N. M. Fahad, S. Sakib, M. M. J. Mim, J. Ahmad, M. E. Ali, and S. Azam, A review on large language models: Architectures, applications, taxonomies, open issues and challenges, IEEE Access **12**, 26839 (2024).

[16] M. Perc and M. Özer, Disappearing minds in the age of artificial intelligence, Insan & Toplum **15**, 1 (2025).

[17] J. J. Horton, *Large language models as simulated economic agents: What can we learn from homo silicus?*, Tech. Rep. (National Bureau of Economic Research, 2023).

[18] N. Arora, I. Chakraborty, and Y. Nishimura, AI–human hybrids for marketing research: Leveraging large language models (LLMs) as collaborators, J. Mark. **89**, 43 (2025).

[19] D. C. Dennett, The problem with counterfeit people, The Atlantic (2023).

[20] Y. Gao, D. Lee, G. Burtch, and S. Fazelpour, Take caution in using LLMs as human surrogates, Proc. Natl. Acad. Sci. U.S.A. **122**, e2501660122 (2025).

[21] M. R. DeVerna, H. Y. Yan, K.-C. Yang, and F. Menczer, Fact-checking information from large language models can decrease headline discernment, Proc. Natl. Acad. Sci. U.S.A. **121**, e2322823121 (2024).

[22] J. Kunz and M. Kuhlmann, Properties and challenges of LLM-generated explanations, arXiv:2402.10532 (2024).

[23] S. Ikeda, Inconsistent advice by ChatGPT influences decision making in various areas, Sci. Rep. **14**, 15876 (2024).

[24] E. Loru, J. Nudo, N. Di Marco, A. Santirocchi, R. Atzeni, M. Cinelli, V. Cestari, C. Rossi-Arnaud, and W. Quattrociocchi, The simulation of judgment in LLMs, Proc. Natl. Acad. Sci. U.S.A. **122**, e2518443122 (2025).

[25] M. Perc, Counterfeit judgments in large language models, Proc. Natl. Acad. Sci. U.S.A. **122**, e2528527122 (2025).

[26] A. Newel and H. A. Simon, Computer science as empirical inquiry: Symbols and search, Commun. ACM **19**, 113 (1976).

[27] C. Manning and H. Schutze, *Foundations of Statistical Natural Language Processing* (MIT Press, Cambridge, MA, 1999).

[28] C. D. Manning, *Introduction to Information Retrieval* (Syngress Publishing, Rockland, MA, 2008).

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, Attention is all you need, in *Advances in Neural Information Processing Systems*, edited by I. Guyon, U. v. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. V. N.

Vishwanathan, and R. Garnett (Curran Associates, Inc., Red Hook, NY, 2017) pp. 5998–6008.

[30] L. Lovász, Random walks on graphs: A survey, in *Combinatorics, Paul Erdős is Eighty, Vol. 2*, edited by D. Miklós, V. T. Sós, and T. Szőnyi (János Bolyai Mathematical Society, Budapest, 1993) pp. 1–46.

[31] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, Language models are few-shot learners, in *Advances in Neural Information Processing Systems*, edited by H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin (Curran Associates, Inc., Red Hook, NY, 2020) pp. 1877–1901.

[32] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, edited by J. Burstein, C. Doran, and T. Solorio (Association for Computational Linguistics, Stroudsburg, PA, 2019) pp. 4171–4186.

[33] J. Kaplan, S. McCandlish, T. Henighan, *et al.*, Scaling laws for neural language models, arXiv:2001.08361 (2020).

[34] J. Hoffmann *et al.*, Training compute-optimal large language models, arXiv:2203.15556 (2022).

[35] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, On the dangers of stochastic parrots: Can language models be too big?, in *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, edited by L. Irani, M. Mitchell, D. Robinson, and S. Kannan (Association for Computing Machinery, New York, NY, 2021) pp. 610–623.

[36] G. Marcus and E. Davis, *Rebooting AI: Building Artificial Intelligence We Can Trust* (Pantheon Books, New York, NY, 2020).

[37] S. R. Bowman, Eight things to know about large language models, Critical AI **2**, No Pagination (2024).

[38] L. Strobl, W. Merrill, G. Weiss, D. Chiang, and D. Angluin, Transformers as recognizers of formal languages: A survey on expressivity, arXiv:2311.00208 (2023).

[39] P. Lewis, E. Perez, A. Piktus, *et al.*, Retrieval-augmented generation for knowledge-intensive nlp, in *Advances in Neural Information Processing Systems*, edited by H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin (Curran Associates, Inc., Red Hook, NY, 2020) pp. 9459–9474.

[40] S. Borgeaud, A. Mensch, J. Hoffmann, T. Cai, E. Rutherford, K. Millican, G. B. Van Den Driessche, J.-B. Lespiau, B. Damoc, A. Clark, *et al.*, Improving language models by retrieving from trillions of tokens, in *Proceedings of the 39th International Conference on Machine Learning*, edited by K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato (Proceedings of Machine Learning Research, Online, 2022) pp. 2206–2240.

[41] T. Schick, J. Dwivedi-Yu, *et al.*, Toolformer: Language models can teach themselves to use tools, arXiv:2302.04761 (2023).

[42] G. Mialon *et al.*, Augmented language models: a survey, arXiv:2302.07842 (2023).

[43] X. Li, J. Jin, Y. Zhou, Y. Zhang, P. Zhang, Y. Zhu, and Z. Dou, From matching to generation: A survey on generative information retrieval, ACM Trans. Inf. Syst. **43**, 1 (2025).

[44] R. Rosenbacke *et al.*, Beyond Hallucinations: The Illusion of Understanding in Large Language Models, arXiv:2510.14665 (2025).

[45] M. Turpin, J. Michael, E. Perez, and S. R. Bowman, Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting, in *NIPS '23: Proceedings of the 37th International Conference on Neural Information Processing System*, edited by A. Oh, T. Naumann, and A. Globerson (Curran Associates, Inc., Red Hook, NY, 2023) pp. 74952–74965.

[46] A. T. Kalai and S. S. Vempala, Calibrated language models must hallucinate, in *STOC 2024: Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, edited by B. Mohar and I. Shinkar (Association for Computing Machinery, New York, NY, 2024) pp. 160–171.

[47] A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi, The curious case of neural text degeneration, arXiv:1904.09751 (2019).

[48] S. T. Piantadosi, Zipf's word frequency law in natural language: A critical review and future directions, Psychonomic Bull. Rev. **21**, 1112 (2014).

[49] R. Vershynin, *High-dimensional probability: An introduction with applications in data science* (Cambridge University Press, Cambridge, U.K., 2018).

[50] Z. Xu, S. Jain, and M. Kankanhalli, Hallucination is inevitable: An innate limitation of large language models, arXiv:2401.11817 (2024).

[51] J. Gibson, *Ecological Approach to Visual Perception* (Routledge, Oxfordshire, U.K., 2014).

[52] K. R. Scherer, The dynamic architecture of emotion: Evidence for the component process model, Cognition and Emotion **23**, 1307 (2009).

[53] P. Ekman and W. V. Friesen, Constants across cultures in the face and emotion, J. Pers. Soc. Psychol. **17**, 124 (1971).

[54] R. Banse and K. R. Scherer, Acoustic profiles in vocal emotion expression, J. Pers. Soc. Psychol. **70**, 614 (1996).

[55] C. D. Frith and U. Frith, Social cognition in humans, Curr. Biol. **17**, R724 (2007).

[56] G. A. Van Kleef, How emotions regulate social life: The emotions as social information (EASI) model, Curr. Dir. Psychol. Sci. **18**, 184 (2009).

[57] E. M. Bender and A. Koller, Climbing towards nlu: On meaning, form, and understanding in the age of data, in *Proceedings of the 58th annual meeting of the association for computational linguistics*, edited by D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault (Association for Computational Linguistics, Stroudsburg, PA, 2020) pp. 5185–5198.

[58] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, *et al.*, Training language models to follow instructions with human feedback, NIPS **35**, 27730 (2022).

[59] L. Bojić, O. Zagovora, A. Zelenkauskaite, V. Vuković,

M. Čabarkapa, S. Veseljević Jerković, and A. Jovančević, Comparing large language models and human annotators in latent content analysis of sentiment, political leaning, emotional intensity and sarcasm, Sci. Rep. **15**, 11477 (2025).
[60] X. Gao, S. Nayak, and M. Coler, Spoken in jest, detected in earnest: A systematic review of sarcasm recognition-multimodal fusion, challenges, and future prospects, IEEE Trans. Affect. Comput. **16**, 2526 (2025).
[61] A. Clark, Whatever next? predictive brains, situated agents, and the future of cognitive science, Behav. Brain. Sci. **36**, 181 (2013).
[62] K. Friston, The free-energy principle: a unified brain theory?, Nat. Rev. Neurosci. **11**, 127 (2010).
[63] J. Koster-Hale and R. Saxe, Theory of mind: a neural prediction problem, Neuron **79**, 836 (2013).
[64] L. W. Barsalou, Grounded cognition, Annu. Rev. Psychol. **59**, 617 (2008).
[65] F. C. Bartlett, *Remembering: A study in experimental and social psychology* (Cambridge University Press, Cambridge, U.K., 1995).
[66] J. S. Bruner *et al.*, Going beyond the information given, Contemporary Approaches to Cognition **1**, 119 (1957).
[67] R. Sennrich, B. Haddow, and A. Birch, Neural machine translation of rare words with subword units, in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, edited by K. Erk and N. A. Smith (Association for Computational Linguistics, Stroudsburg, PA, 2016) pp. 1715–1725.
[68] T. Kudo and J. Richardson, Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing, arXiv:1808.06226 (2018).
[69] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, Language models are unsupervised multitask learners, OpenAI **1**, 9 (2019).
[70] D. Jurafsky and J. H. Martin, *Speech and Language Processing* (Prentice Hall, Saddle River, NJ, 2023).
[71] Y. Chai, Y. Fang, Q. Peng, and X. Li, Tokenization falling short: On subword robustness in large language models, arXiv:2406.11687 (2024).
[72] D. A. Haslett, Tokenization changes meaning in large language models: Evidence from Chinese, Computational Linguistics **51**, 785 (2025).
[73] T. Truong, J. Otmakhova, K. Verspoor, T. Cohn, and T. Baldwin, Revisiting subword tokenization: A case study on affixal negation in large language models, in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, edited by K. Duh, H. Gomez, and S. Bethard (Association for Computational Linguistics, Stroudsburg, PA, 2024) pp. 5082–5095.
[74] E. Tulving, Episodic memory: From mind to brain, Ann. Rev. Psychol. **53**, 1 (2002).
[75] D. L. Schacter and D. R. Addis, The cognitive neuroscience of constructive memory: Remembering the past and imagining the future, Philos. Trans. R. Soc. B **362**, 773 (2007).
[76] E. S. Spelke and K. D. Kinzler, Core knowledge, Developmental Sci. **10**, 89 (2007).
[77] A. Gopnik, A. N. Meltzoff, and P. K. Kuhl, *The Scientist in the Crib: Minds, Brains, And How Children Learn* (William Morrow & Co., New York, NY, 1999).
[78] L. W. Barsalou, Perceptual symbol systems, Behav. Brain Sci. **22**, 577 (1999).
[79] G. Murphy, *The Big Book of Concepts* (MIT Press, Cambridge, MA, 2004).
[80] R. Bommasani, On the opportunities and risks of foundation models, arXiv:2108.07258 (2021).
[81] T. Mikolov, K. Chen, G. Corrado, and J. Dean, Efficient estimation of word representations in vector space, arXiv:1301.3781 (2013).
[82] L. M. Schulze Buschoff, E. Akata, M. Bethge, and E. Schulz, Visual cognition in multimodal large language models, Nat. Mach. Intell. **7**, 96 (2025).
[83] A. Marchetti, F. Manzi, G. Riva, A. Gaggioli, and D. Massaro, Artificial Intelligence and the Illusion of Understanding: A Systematic Review of Theory of Mind and Large Language Models, Cyberpsychol. Behav. Soc. Netw. **28**, 505 (2025).
[84] G. Gallipoli and L. Cagliero, It is not a piece of cake for GPT: Explaining Textual Entailment Recognition in the presence of Figurative Language, in *Proceedings of the 31st International Conference on Computational Linguistics*, edited by O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. Di Eugenio, and S. Schockaert (Association for Computational Linguistics, Stroudsburg, PA, 2025) pp. 9656–9674.
[85] A. Damasio, *Descartes' Error: Emotion, Reason, and the Human Brain* (Putnam, New York, NY, 1994).
[86] K. Oatley, D. Keltner, and J. M. Jenkins, *Understanding Emotions* (Blackwell Publishing, Oxford, U.K., 2006).
[87] J. E. LeDoux, *The Emotional Brain: The Mysterious Underpinnings of Emotional Life* (Simon and Schuster, New York, NY, 1996).
[88] N. H. Frijda, *The Emotions* (Cambridge University Press, Cambridge, U.K., 1986).
[89] E. A. Locke and G. P. Latham, Building a practically useful theory of goal setting and task motivation: A 35-year odyssey., Am. Psychol. **57**, 705 (2002).
[90] E. T. Higgins, Beyond pleasure and pain., Am. Psychol. **52**, 1280 (1997).
[91] R. F. Baumeister, K. D. Vohs, C. Nathan DeWall, and L. Zhang, How emotion shapes behavior: Feedback, anticipation, and reflection, rather than direct causation, Pers. Soc. Psychol. Rev. **11**, 167 (2007).
[92] J. LeDoux, *Anxious: Using the Brain to Understand and Treat Fear and Anxiety* (Penguin Books, London, U.K., 2016).
[93] V. Capraro, The dual-process approach to human sociality: Meta-analytic evidence for a theory of internalized heuristics for self-preservation., J. Pers. Soc. Psychol. **126**, 719 (2024).
[94] E. Becker, *The Denial of Death* (Free Press, Washington, D.C., 1973).
[95] T. Pyszczynski, J. Greenberg, S. Solomon, J. Arndt, and J. Schimel, Why do people need self-esteem? A theoretical and empirical review, Psychol. Bull. **130**, 435 (2004).
[96] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning* (MIT Press, Cambridge, MA, 2016).
[97] I. Gabriel, Artificial intelligence, values, and alignment, Minds Mach. **30**, 411 (2020).
[98] D. Rozado, The political preferences of LLMs, PLOS One **19**, e0306621 (2024).
[99] R. A. Fulgu and V. Capraro, Surprising gender biases in

GPT, Comput. Hum. Behav. Rep. **16**, 100533 (2024).
[100] S. A. Sloman, *Causal Models: How People Think About the World and Its Alternatives* (MIT Press, Cambridge, MA, 2005).
[101] N. J. Roese, Counterfactual thinking, Psychol. Bull. **121**, 133 (1997).
[102] P. M. Gollwitzer, Implementation intentions: Strong effects of simple plans, Am. Psychol. **54**, 493 (1999).
[103] H. A. Simon, A behavioral model of rational choice, Q. J. Econ. , 99 (1955).
[104] G. Gigerenzer and D. G. Goldstein, Reasoning the fast and frugal way: Models of bounded rationality, Psychol. Rev. **103**, 650 (1996).
[105] S. A. Sloman, The empirical case for two systems of reasoning, Psychol. Bull. **119**, 3 (1996).
[106] J. Haidt, The emotional dog and its rational tail: A social intuitionist approach to moral judgment, Psychol. Rev. **108**, 814 (2001).
[107] M. Mitchell and D. C. Krakauer, The debate over understanding in AI's large language models, Proc. Natl. Acad. Sci. U.S.A. **120**, e2215907120 (2023).
[108] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann, Shortcut learning in deep neural networks, Nat. Mach. Intell. **2**, 665 (2020).
[109] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman, Building machines that learn and think like people, Behav. Brain Sci. **40**, e253 (2017).
[110] G. Marcus and E. Davis, *Rebooting AI: Building Artificial Intelligence We Can Trust* (Vintage Books, New York, NY, 2019).
[111] M. Zečević, M. Willig, D. S. Dhami, and K. Kersting, Causal parrots: Large language models may talk causality but are not causal, arXiv:2308.13067 (2023).
[112] Z. Jin, Y. Chen, F. Leeb, L. Gresele, O. Kamal, Z. Lyu, others, and B. Schölkopf, CLADDER: assessing causal reasoning in language models, in *NIPS '23: Proceedings of the 37th International Conference on Neural Information Processing Systems*, edited by A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Curran Associates, Inc., Red Hook, NY, 2023) pp. 31038–31065.
[113] H. Chi, H. Li, W. Yang, F. Liu, L. Lan, X. Ren, and B. Han, Unveiling Causal Reasoning in Large Language Models: Reality or Mirage?, in *NIPS '24: Proceedings of the 38th International Conference on Neural Information Processing System*, edited by A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (Curran Associates, Inc., Red Hook, NY, 2024) pp. 96640–96670.
[114] M. M. Botvinick, T. S. Braver, D. M. Barch, C. S. Carter, and J. D. Cohen, Conflict monitoring and cognitive control, Psychol. Rev. **108**, 624 (2001).
[115] W. J. Gehring, B. Goss, M. G. Coles, D. E. Meyer, and E. Donchin, A neural system for error detection and compensation, Psychol. Sci. **4**, 385 (1993).
[116] S. M. Fleming and N. D. Daw, Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation, Psychol. Rev. **124**, 91 (2017).
[117] K. E. Lyons and S. Ghetti, The development of uncertainty monitoring in early childhood, Child Dev. **82**, 1778 (2011).
[118] N. Pescetelli, G. Rees, and B. Bahrami, The perceptual and social components of metacognition, J. Exp. Psychol. Gen. **145**, 949 (2016).
[119] L. Schnaubert, S. Krukowski, and D. Bodemer, Assumptions and confidence of others: The impact of sociocognitive information on metacognitive self-regulation, Metacogn. Learn. **16**, 855 (2021).
[120] M. Huschens, M. Briesch, D. Sobania, and F. Rothlauf, Do you trust ChatGPT?–perceived credibility of human and AI-generated content, arXiv:2309.02524 (2023).
[121] W. Dai, Y.-S. Tsai, J. Lin, A. Aldino, H. Jin, T. Li, D. Gašević, and G. Chen, Assessing the proficiency of large language models in automatic feedback generation: An evaluation study, Comput. Educ. Artif. Intell. **7**, 100299 (2024).
[122] Y. Chen, L. Yuan, G. Cui, Z. Liu, and H. Ji, A close look into the calibration of pre-trained language models, in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, edited by A. Rogers, J. Boyd-Graber, and N. Okazaki (Association for Computational Linguistics, Stroudsburg, PA, 2023) pp. 1343–1367.
[123] L. Krause, W. Tufa, S. B. Santamaría, A. Daza, U. Khurana, and P. Vossen, Confidently wrong: exploring the calibration and expression of (Un) certainty of large language models in a multilingual setting, in *Proceedings of the Workshop on Multimodal, Multilingual Natural Language Generation and Multilingual WebNLG Challenge (MM-NLG 2023)*, edited by A. Gatt, C. Gardent, L. Cripwell, A. Belz, C. Berg, A. Erdem, and E. Erdem (Association for Computational Linguistics, Stroudsburg, PA, 2023) pp. 1–9.
[124] I. M. Begg, A. Anas, and S. Farinacci, Dissociation of processes in belief: Source recollection, statement familiarity, and the illusion of truth, J. Exp. Psychol. Gen. **121**, 446 (1992).
[125] C. Unkelbach, Reversing the truth effect: Learning the interpretation of processing fluency in judgments of truth, J. Exp. Psychol. Learn. Mem. Cogn. **33**, 219 (2007).
[126] A. L. Alter and D. M. Oppenheimer, Uniting the tribes of fluency to form a metacognitive nation, Pers. Soc. Psychol. Rev **13**, 219 (2009).
[127] A. Dechêne, C. Stahl, J. Hansen, and M. Wänke, The truth about the truth: A meta-analytic review of the truth effect, Pers. Soc. Psychol. Rev. **14**, 238 (2010).
[128] B. E. Whitley Jr and M. S. Greenberg, The role of eyewitness confidence in juror perceptions of credibility, J. Appl. Soc. Psychol. **16**, 387 (1986).
[129] E. R. Tenney, B. A. Spellman, and R. J. MacCoun, The benefits of knowing what you know (and what you don't): How calibration affects credibility, J. Exp. Soc. Psychol. **44**, 1368 (2008).
[130] P. C. Price and E. R. Stone, Intuitive evaluation of likelihood judgment producers: Evidence for a confidence heuristic, J. Behav. Decis. Mak. **17**, 39 (2004).
[131] L. Rozenblit and F. Keil, The misunderstood limits of folk science: An illusion of explanatory depth, Cogn. Sci. **26**, 521 (2002).
[132] R. Parasuraman and V. Riley, Humans and automation: Use, misuse, disuse, abuse, Human Factors **39**, 230 (1997).
[133] J. D. Lee and K. A. See, Trust in automation: Designing for appropriate reliance, Human Factors **46**, 50 (2004).
[134] P. Shojaee, I. Mirzadeh, K. Alizadeh, M. Horton,

S. Bengio, and M. Farajtabar, The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity, arXiv:2506.06941 (2025).

[135] S. Trott, C. Jones, T. Chang, J. Michaelov, and B. Bergen, Do large language models know what humans know?, Cogn. Sci. **47**, e13309 (2023).

[136] I. D. Raji, E. M. Bender, A. Paullada, E. Denton, and A. Hanna, Ai and the everything in the whole wide world benchmark, arXiv:2111.15366 (2021).

[137] P. Liang, R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, Y. Zhang, D. Narayanan, Y. Wu, A. Kumar, *et al.*, Holistic evaluation of language models, arXiv:2211.09110 (2022).

[138] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, *et al.*, A survey on evaluation of large language models, ACM Trans. Intell. Syst. Technol. **15**, 1 (2024).

[139] R. Heersmink, B. de Rooij, M. J. Clavel Vázquez, and M. Colombo, A phenomenology and epistemology of large language models: Transparency, trust, and trustworthiness, Ethics Inf. Technol. **26**, 41 (2024).

[140] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, Concrete problems in ai safety, arXiv:1606.06565 (2016).

[141] L. Messeri and M. J. Crockett, Artificial intelligence and illusions of understanding in scientific research, Nature **627**, 49 (2024).

[142] European Union, Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), Official Journal of the European Union (2024).

[143] White House Office of Science and Technology Policy, Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People, Executive Office of the President of the United States (2022).

[144] UK Department for Science, Innovation and Technology, A Pro-Innovation Approach to AI Regulation, UK Government White Paper (2023).

[145] Cyberspace Administration of China, Interim Measures for the Management of Generative Artificial Intelligence Services, State Internet Information Office of the People's Republic of China (2023).

[146] C. Voinea, S. P. Mann, J. Savulescu, and B. D. Earp, The calculator analogy: Epistemic virtues for using llms, Technology in Society , 103198 (2025).

[147] D. F. Halpern, Teaching critical thinking for transfer across domains: Dispositions, skills, structure training, and metacognitive monitoring, Am. Psychol. **53**, 449 (1998).

[148] D. Kuhn, A developmental model of critical thinking, Educ. Res. **28**, 16 (1999).

[149] R. H. Ennis, Critical thinking: A streamlined conception, in *The Palgrave Handbook of Critical Thinking in Higher Education*, edited by M. Davies and R. Barnett (Palgrave Macmillan, New York, NY, 2015) pp. 31–47.