Review Article

Rethink Your Mental Model in the Age of Generative AI: A Triadic Framework for Human-AI Collaboration

Till Moritz Saßmannshausen[1], **Sebastian Wagener**[2]

1. University of Siegen, Germany; 2. Independent researcher

Generative AI systems such as large language models exhibit jagged intelligence: they combine superhuman performance on some tasks with brittle, often opaque failures on others. Existing mental models, inherited from deterministic technologies and from human teamwork, mischaracterize these systems systematically. This theory-driven synthesis develops a conceptual *Triadic Framework* for adaptive mental models in human-AI collaboration: First, the *System Layer* synthesizes current evidence on probabilistic generation, opacity at scale, and rapid model drift, explaining why capability boundaries are uneven and moving. Second, the *Collaboration Layer* analyzes how users configure prompting practices, division of labor, and handling preferences. Third, the *Metacognitive Layer* examines how anthropomorphism, metaphors, and cognitive biases shape human interpretations of "intelligent" behavior. Building on this diagnosis, the paper proposes seven actionable practices that are designed to calibrate mental models, preserve human agency, and make hybrid intelligence more robust under jagged and shifting AI capabilities. Together, the paper reframes the challenge of generative AI from crafting better prompts to cultivating more adequate mental models of probabilistic counterparts.

Correspondence: papers@team.qeios.com — Qeios will forward to the authors

 doi.org/10.32388/GAG6KD

Part I: Collaborating in the Second Machine Age

1. Introduction

Since the public release of ChatGPT in November 2022, we witnessed discoveries that seemed far ahead. Artificial intelligence (AI) has been found to mimic creativity[1][2] and empathy[3][4] at a level better than the human average. AI is now fluent in a way that 50 % of people cannot distinguish GPT-4 from a human after a 5-minute conversation[5]. To reach this result the researchers had to prompt GPT-4 not to appear overly competent.

The weird and fascinating thing is that no one really knows what generative AI (GenAI) can and cannot do. There is no clear distinction between capabilities and limitations. Even tasks that seem similar can vary greatly in their feasibility for GenAI.

> LLMs are good at "***Write a poem in Spanish***" but bad at "***Write a poem in Spanish of exactly 25 words***".
>
> LLMs are good at "***Give me 5 random numbers***" but bad at "***Give me 5 random numbers that average to 10***".

The problem is that the line between LLMs' capabilities and non-capabilities is invisible. Moreover, this line is moving and shifting over time. For us humans, it is difficult to mentally manage this experience since it does not follow a pattern we experienced in the past. As a first attempt, we often compare AI with traditional technologies and experience that we can no longer employ the deterministic schema of *ones* and *zeros* (it works vs. it doesn't work). Now we have more in between like "*sometimes it gets it right*" or "*it works well enough*". This experience reminds us more of a coworker and we jump to compare AI's capabilities to human capabilities – and are frustrated or surprised, because our mental model of a human coworker does not apply either. What remains under-specified is an integrative account of how jagged capabilities translate into predictable user miscalibration – and what routines can correct it.

To guide future work with AI, this paper aims to assemble a coherent set of mechanisms that support adaptive mental model building. As an integrative, theory-driven synthesis the paper samples research from various disciplines, prioritizing recent studies that (a) explicitly evaluate human-AI collaboration with general-purpose LLMs in text-centric knowledge work or (b) report systematic capabilities and failure patterns of AI. Taken together this paper makes three contributions. In Part I, it provides

conceptual foundations. In Part II, it synthesizes recent evidence on jagged AI capabilities into a mental model framework with three layers that jointly explain human-AI collaboration. In Part III, it translates this diagnosis into seven designable routines, derives implications for AI literacy as well as organizational design, and concludes with a research agenda.

2. Conceptual Foundations

This paper builds on two foundational concepts: (a) mental models as cognitive maps for guiding interactions, and (b) distinctive characteristics of GenAI.

Before diving into these foundations, we clarify how we handle anthropomorphic terminology in this paper. We use terms like “reasoning”, “thinking”, “learning”, and “hallucination” when describing LLM behavior. These terms dominate research literature and practitioner discourse, yet they systematically invite anthropomorphic misattribution – mistaking computational pattern-matching for human-like cognition[6][7]. Recent scholarship identifies this as a source of systematic misconception[8][9][10]. We address this tension through a notational intervention: we mark such terms with the logical negation symbol “¬” to signal that they describe functional behaviors without implying human-like intentionality (e.g. *¬thinking* denotes *not-human-like thinking*). This notation serves three purposes: First, it maintains terminological continuity with existing literature; second, it makes the anthropomorphism critique visible throughout the text; and third, it offers a new convention for the broader community. Alternative approaches, such as inventing new terms (e.g. *pseudo-thinking*) or use of quotation marks, sacrifice either communicative efficiency or visibility. The theoretical justification for this intervention is developed in the chapter 5, where we analyze how metaphoric language shapes mental models. We acknowledge this convention is experimental and discuss limitations in chapter 8.

2.1. Mental Models as Cognitive Maps

This paper focuses on *mental models* rather than the related concept of *theory of mind* (ToM). While ToM concerns how we attribute mental states to intentional agents[11], mental models concern how we understand systems and technologies. At the human-technology interface, mental models have long been the dominant framework, though ToM has gained importance in contexts such as social robotics and human-AI interaction, where humans attribute intentionality to machines.

Mental models are cognitive structures that people use to understand, simplify, and predict complex systems[12][13]. They represent internalized abstractions of the external world and enable orientation,

action planning, and decision-making. In science and technology research, mental models bridge knowledge, experience, and practice: they shape which questions are asked, which opportunities are recognized, and how uncertainty is interpreted[14][15]. Examples include (a) an economist relying on a model of supply and demand, (b) an engineer thinking in terms of system diagrams, or (c) everyday reasoning, such as anticipating how cars behave in traffic.

In *organizational* and *team research*, mental models are often conceptualized as *shared mental models*: overlapping cognitive representations that enable coordination, predict the actions of teammates, and reduce the need for explicit communication under time pressure. Canonical works argue that such shared understanding is a key mechanism for reliable teamwork on complex, interdependent tasks[16] and show that the convergence of shared mental models predicts team processes and performance[17]. This perspective is useful here because collaboration between humans and AI increasingly exhibits team-like interdependence, making calibration not only an individual but also a collective organizational challenge.

Recent *Information Systems* (IS) research treats mental models in human-AI collaboration as dynamic rather than fixed: decision-makers adapt their understanding through continuous interaction with AI systems, and system design can deliberately shape that development[18]. In human-computer interaction (HCI), Norman[19] emphasized that users form mental models of systems that often diverge from designers' conceptual models, leading to errors. When mental models do not align with reality, people may over- or underestimate the capabilities of a system, resulting in disappointment, incorrect use, undermined trust, and declined overall performance[20][21].

2.2. Performance Paradoxes in Practice

Since no one fully understands the best use of GenAI, recent studies attempt to create a nuanced picture of performance effects in different professions[22][23]. For example, Dell'Acqua *et al.*[24] showed that human+AI teams match the performance of human+human teams for product innovation. Here, LLMs act not only as productivity enhancers but as *cybernetic teammates,* also fostering social and motivational dynamics as well as reshaping how collaboration is organized.

Another effect was found by Dell'Acqua *et al.*[22]: Consultants using GPT-4 perform better (in speed and quality) than consultants without AI. However, this pattern only applies to tasks that fall within the capabilities of GPT-4. Otherwise, consultants without AI performed better – indicating acceleration effects of AI for better and worse.

In contrast, Goh *et al.*[25] found that physicians using GPT-4 did not outperform the control group without AI support – despite GPT-4 alone outperforming both groups. The unrealized potential might be based on integration deficits or inappropriate skepticism, highlighting that integration strategies are critical.

More nuanced, Dell'Acqua[26] found that recruiters who used advanced AI missed out on brilliant applicants and made worse decisions than recruiters who used less advanced AI or no AI at all. It seemed that with good AI, humans let the AI take over and became lazy, careless, and less skilled in their own judgment.

Taken together, these examples show that the new qualities of AI do not automatically translate to performance gains. Depending on the task and interaction design, the same AI can amplify performance or systematically degrade it. These empirical patterns motivate a more structured account of how to mentally prepare for collaboration with GenAI.

2.3. GenAI as Pre-Cognition

A key insight emerges from these performance paradoxes: AI's impact depends not just on its technical capabilities but on how it reshapes the cognitive landscape before human judgment begins. When AI provides filtered information, suggested framings, and confident-sounding outputs, it operates as what Chiriatti *et al.*[27] calls a *pre-cognitive layer* (*System 0*) – shaping what enters human awareness before deliberate evaluation can occur.

This pre-cognitive positioning could help explain the observed paradoxes. Physicians underperformed despite having superior AI available[25] because they could not adequately monitor how AI's confident presentations influenced their diagnostic reasoning. Recruiters became careless with advanced AI[26] because the system's fluent outputs bypassed their critical evaluation mechanisms. Consultants showed mixed results[22] depending on whether tasks fell within AI's capabilities – but they lacked systematic ways to detect these boundaries in real-time. In each case, AI's influence on human cognition was inadequately understood and managed.

To avoid misjudgment, Shneiderman[28] argues that the GenAI era needs a human-centered design, prioritizing transparency, accountability, and oversight. Amershi *et al.*[29] translate this into design guidelines:

- AI systems must clearly communicate their capabilities and limitations,

- make their ¬reasoning transparent where possible,
- support efficient correction when they fail,
- learn from user interactions, and
- enable users to calibrate their trust appropriately.

This paper will demonstrate that putting these well-intentioned guidelines into practice is anything but trivial. The *System 0* framing helps explain why: when AI shapes cognition pre-consciously, traditional design principles that assume deliberate evaluation can be insufficient.

2.4. Technology as Otherware

Historically, users have interacted with machines in a one-directional way: humans decide, machines deliver. This long-standing "deal" has shaped HCI since graphical interfaces emerged in 1975. Anyone who understood the rules could predict results with near certainty. Thus, traditional technology functioned as an extension of human abilities – passive, predictable, controlled. AI differs fundamentally: it replaces deterministic computation with probabilistic, context-dependent generation[6][7]. It can behave autonomously, unexpectedly, and incomprehensibly for humans[30][31][32][20].

These new qualities are transforming the way humans experience technology: humans no longer use these systems as *tools*, but interact, delegate, and negotiate with them as social *counterparts* (Hassenzahl *et al.* 2021). Psychological effects are fueling this behavior. First, humans naturally tend to treat interactive systems socially (computers are social actors paradigm), despite being aware of their artificial nature[33]. Second, humans reflexively anthropomorphize technology[34], relating to them as a minded other[35].

Together, these new qualities suggest a new framing: Hassenzahl *et al.*[36] introduced the term *otherware* for fundamentally social technologies, while Safdari[35] coined the term *otheroid* to capture the phenomenological experience of relating to artificial minds. Though arising from different perspectives – interaction design and phenomenological social cognition – both terms signal that users increasingly meet AI as a counterpart. The implications are profound: effective mental models must account not only for AI's technical characteristics but also for the fundamentally social nature of human-AI interaction.

2.5. Capabilities on a Jagged Frontier

Dell'Acqua *et al.*[22] call the landscape of AI capabilities the *Jagged Frontier*: a complex terrain where AI excels unexpectedly in some tasks while failing surprisingly in others.

Unexpected Strengths

AI does not surpass humans in most tasks, but in some areas, it is far superior. Surprisingly, two areas long considered the last bastion of human capability seem to have fallen: creativity[37] and empathy (Cao *et al.* 2024b). Trained on the work (text, images, music, and videos) of the most talented creators, AI already tends to be as creative as an average human – while combined human+AI creativity still proves superior[38]. Moreover, LLMs excel at brainstorming and idea generation, especially when prompted appropriately[39]. For example, GPT-4 was able to generate startup ideas that outside judges found better than those of trained business school students[40].

In empathy, studies found patients preferred AI (ChatGPT or Google's AMIE) over primary care doctors and physicians, finding AI responses more empathetic and helpful[3][4] (Tu *et al.* 2024). Similar results can be found in emotional support scenarios, where GPT-4 outperformed 85 % of human advisors[41]. This ability also has a concerning side: LLMs could persuade users 87 % more likely than an average human[42] and are even superior to incentivized persuaders[43]. This manipulative power is consistent over studies and can be further enhanced through post-training and prompting[44].

Surprising Weaknesses

Conversely, LLMs fail at tasks that seem made for machines. They cannot reliably do simple term acrobatics like "*Write this word in reverse order, 'strawberry'?*" or counting letters or tokens like "*How many 'r' are in 'strawberry'?*"[45][46]. Moreover, early LLMs failed at simple arithmetic and comparison tasks, like "*What is 5 + 26?*" and "*What is greater 9.11 or 9.9?*". Also on challenging mathematics, performance remains limited – less than 2 % success rate on FrontierMath benchmark by models like GPT-4o or o1-preview[47].

Another shortcoming is that LLMs frequently commit to assumptions early in dialogue and fail to correct course once heading in wrong directions[48]. The overall issue: LLMs provide no indication of uncertainty. They just come up with an answer that sounds very plausible. The answer may be factually true or wrong, the LLM does not know and does not care – that's not how it works. For us humans, this makes it

hard to detect failures (aka ¬*hallucinations* or *confabulations*) – it is not immediately clear and it is unpredictable[49][50].

Summing up, the pattern of the *Jagged Frontier* holds across domains. AI produces high-quality business ideas[51] yet struggles with autonomous development of complex software[52]. In medicine, GPT-4 performs well on case studies (Kanjee *et al.* 2023, Eriksen *et al.* 2024) but remains prone to elementary computational errors that could undermine tasks such as dosage calculation[52]. Taken together, AI can be highly effective in some tasks, fail a bit in others, and completely derail in certain situations[22]. All while major AI companies actively push to shift this frontier, aiming to develop AI that outperforms humans across a broad spectrum of tasks[53][54].

After this analysis of the new qualities of AI (Figure 1), Part II now develops a *Triadic Framework* to address these dynamics systematically, examining how AI systems work (*System Layer*), how humans interact with them (*Collaboration Layer*), and how we conceptualize this relationship (*Metacognitive Layer*).

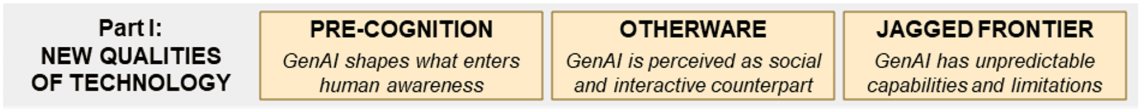

Figure 1. Overview of the new technological qualities in the era of AI

Part II: Preparing for Jagged Intelligence

Part I outlined the fundamental shift and reasons for dysfunctional mental models in the GenAI era. Since GenAI is being "forced upon us", e.g. with new software updates[9], we need to (a) accept uncertainty as a normal system state, (b) emphasize cooperation and exploration, (c) recognize the human role as a critical and evaluative, and (d) remain flexible in order to keep pace with technological dynamics[28]. The aim of Part II is to support this by synthesizing insights into a *Triadic Framework*: the S*ystem Layer* (chapter 3) explains how these systems fundamentally work, *Collaboration Layer* (chapter 4) addresses effective human-AI collaboration, and *Metacognitive Layer* (chapter 5) captures different ways of conceptualizing AI's nature.

3. System Layer: Understanding

The *System Layer* explains why AI capabilities are unpredictable in order to foster understanding and build a mental model for anticipating where AI will succeed or fail and maintaining realistic expectations as models evolve.

3.1. Core Characteristics

LLMs operate on an autoregressive principle, statistical pattern-matching over token sequences instead of logical calculations[55][56][57][58]. Their output may appear coherent or plausible but is fundamentally probabilistic, unpredictable, and irreproducible. This is not a defect but a structural feature of generative computation[59]. Because of these core characteristics, outputs should be treated as samples from a conditional probability distribution, not as deterministic answers.

1. **Probabilistic generation:** Where earlier AI systems classified or optimized, GenAI produces content – text, images, code, or music[53]. Probabilistic sampling explains the apparent creativity and the inherent instability of generative outputs. Mental models must integrate uncertainty as inherent system properties rather than anomalies[60].
2. **Opacity at scale:** The internal mechanisms of LLMs are effectively opaque. As systems scale, they exhibit emergent behaviors – capabilities not explicitly programmed or anticipated[61]. Formal interpretability remains limited, thus mental models must rely on practical experiential and iterative use rather than analytic inspection[62].
3. **Dialogic interaction:** Conversational interfaces lower the barrier for non-experts by replacing rigid command syntax with natural-language dialogue. Iterative feedback loops can refine responses and enhance performance[63]. As a result, users shift from *programming* to *prompting* – a skill that demands linguistic sensitivity, contextual framing, and adaptive experimentation[64][65].
4. **Broad applicability and rapid evolution:** LLMs are general-purpose systems trained across vast, heterogeneous datasets[7]. Their rapid improvement renders static mental models obsolete, while adaptive, learning-oriented conceptions become necessary[66]. Continuous scaling and fine-tuning accelerate capability drift – users must expect that what works today may not work tomorrow.

3.2. Training and Data

LLMs' capabilities are shaped by the quality, diversity, and provenance of the training data, as well as by the objectives and reward functions used in post-training. Because models can reproduce fragments of their training corpus[67], understanding data provenance is critical. When models are trained on previously generated outputs rather than original human data, quality seems to degrade rapidly – a phenomenon termed the *Curse of Recursion*[68].

Scaling laws demonstrate that larger models trained on more data (text, image, video, audio, and speech) consistently achieve better performance[69]. Across benchmarks, AI capabilities have rapidly approached and exceeded human-level performance[70][71]. Additionally, modern LLMs are less distinguished by training data and more by *alignment and optimization strategies* after initial training[72], like *Reinforcement Learning from Human Feedback* (RLHF)[73] or *Direct Preference Optimization* (DPO)[74]. Research also showed that fine-tuning on "junk data" leads to so-called "brain rot", a lasting downgrade in performance[75]. Together, these training and post-training strategies shape model behavior, but they can mask structural fragilities and lead to over-confidence in current models.

Moreover, training and post-training are discrete and episodic, not continuous. Once trained, LLMs operate in *frozen inference mode* – they do not learn immediately from user interactions or self-correct through dialogue[53]. Behavioral change requires retraining or model replacement. This discontinuity contrasts with human continuous learning, creating a *static inference paradox* that limits adaptivity between updates.

3.3. Inference Mechanisms

During inference, model behavior reflects architectural constraints and contextual conditioning. Understanding these dynamics is essential for anticipating output variance and limits. Models process information within a bounded context window, typically a few hundred thousand input-tokens. Studies show systematic information loss in the middle of long inputs[76]. Moreover, some of the input-tokens are hidden in *system prompts*, crafted and undisclosed by the model provider with the aim to frame personality, guardrails, and additional information[77]. Together, the consequence is pronounced prompt sensitivity of LLMs[78][79].

Beyond that, specialized ¬reasoning models introduce intermediate ¬thinking tokens to emulate deliberation[80]. These ¬reasoning models were found to collapse beyond a certain complexity of

riddles[81]. Despite criticism of the methodological aspects of this research on model collapse[82], it is highlighting that LLMs are not “thinking” but “processing”. Other limits of ¬reasoning lie in memory capacity, showing that ¬reasoning improves some task but may fail unexpectedly[83][84].

Recent work explicitly compares the different inference modes with Kahneman’s dual-process theory, showing that models can shift from heuristic, System-1-like responses to more deliberate, System-2-like step-by-step logic for arithmetic or symbolic ¬reasoning[85]. New approaches like *adapt thinking* let the models choose between these modes on their own – getting higher accuracy and lower inference costs[86]. These advances further blur the *Jagged Frontier* since ongoing developments continuously shift the line between what AI can and cannot do[87].

3.4. Systematic Challenges

Some LLM failures follow identifiable regularities. Recognizing these patterns allows users to design more robust verification and collaboration frameworks. Recent research revealed a systemic vulnerability to minor, content-independent perturbations of the context: short, semantically meaningless adversarial triggers drastically reduced ¬reasoning accuracy across models[88]. Even minimal prompt variations in formatting or wording caused large performance differences, undermining robustness[89]. Performance also decays in multi-turn dialogue, where early errors propagate and compound, causing a roughly 40 % drop in reliability compared to single-turn evaluations[90].

Moreover, it seems that many reported ¬reasoning successes rely on benchmark artifacts rather than genuine abstraction[91]. Common benchmark questions were part of training data, indicating effects of memorization[58]. When confronted with modified questions requiring identical ¬reasoning, performance drops sharply – a phenomenon termed ¬*Reasoning Gap*[92]. Similar behavior could be observed for planning tasks: variations or obfuscated formulations lead to failures[93][94].

Also techniques like *Chain-of-Thought* (CoT) offer limited reliability[95][96]. Especially tasks requiring conceptual ¬reasoning, geometry, and symbolic manipulation lead to inconsistent performance[97][98][99]. Even injecting the correct answer into the ¬reasoning process resulted in wrong answers[100]. Broader meta-analyses report reproducibility challenges across evaluation pipelines[101].

Another challenge is that LLMs frequently produce ¬hallucinations – plausible but factually incorrect content[102] like fabricated references[103]. These ¬hallucinations are not random glitches but systematic consequences of their training data and inference mechanisms, which prioritize fluency over

factuality[102]. At the same time, empirical work shows that models are capable of internally estimating the likelihood of correctness, but rarely abstain from uncertain answers[104]. Even calibration or retrieval mechanisms cannot completely eliminate ¬hallucinations[105].

Together, these studies underscore context-fragility of LLMs. Applying a binary success/failure perspective reflects a deterministic legacy of computing which is inadequate for the probabilistic nature of GenAI. Scholars propose shifting toward human-inspired evaluation schemes that consider natural variance and contextual expectation. This reframing aligns assessment with the stochastic nature of generative intelligence[106][107].

3.5. Development Projections

GenAI systems already bundle functionalities that previously defined dedicated software and niche solutions, illustrating how quickly product differentiators can erode[108][109]. The shrinking interval between a product's launch and the moment a foundation model absorbs its value propositions already has its own term, called *Subsumption Window*[110].

Outside the LLM family, specialized systems such as DeepMind's *AlphaProof* show strong mathematical performance[111]. Within the LLM family, Linguistic Olympiad-style tasks continue to reveal a mixed picture: ¬reasoning models like OpenAI's o1 can surpass humans on many puzzle types and even induce patterns or generate puzzle-like tasks, yet they still fall short of the rigor, internal consistency, and novelty demanded by authentic Olympiad problems[112].

To keep pace with rapid gains and reduce contamination, the community has introduced harder, continuously updated benchmarks, like RIMO[97], OlymMATH[99], MathConstruct[113], and LiveBench[114]. These datasets remain challenging even for human experts. One example is the *International Olympiad on Astronomy and Astrophysics* (IOAA) where Gemini 2.5 Pro and GPT-5 achieved gold-medal-level performance[98].

The current developments are catalyzing three complementary directions. First, architectural diversity such as *Hierarchical Reasoning Models* (HRM) and *Tiny Recursive Models* (TRM) report wins on hard puzzle tasks (Sudoku, Maze, ARC-AGI) with just a fraction (< 0.01 %) of the parameters of LLMs[115]. Second, new ¬reasoning approaches that learn without the correct answers in the training data, like *AI-Newton* with concept-driven discovery[116] and *Absolute Zero* with reinforced self-play ¬reasoning[117]. Third, long-horizon memory, like MemGPT[118], new memory frameworks[119] and architectures[120], that

introduce persistent, dynamic, and efficient memorization. Taken together, capability drift should be anticipated – what works today may shift tomorrow. Understanding these system-level dynamics is a prerequisite for productive collaboration.

4. Collaboration Layer: Experiencing

The *Collaboration Layer* addresses *how* to interact productively with GenAI in practice with the aim of establishing adaptive collaboration principles, like prompting strategies or division of labor, transferable across models and contexts.

4.1. Core Characteristics

LLMs are best understood through their observable behavior rather than internal mechanisms. As Bender and Koller[6] argue, these systems do not encode meaning in a human sense. But like humans, they are epistemic black boxes accessible only through their outputs[121].

Unlike deterministic software, LLMs' output is *probabilistic* and *context-sensitive*, requiring users to learn by *doing*, not by following fixed rules. Thus, effective AI collaboration does not stop with technical expertise, but needs *experiential literacy* – the ability to build accurate mental models to anticipate where AI systems behave well and where they break[122][123]. These mental models enable critical and creative collaboration, supporting trust, control, and meaningful use[28][124].

Empirical work confirms that users develop more robust understanding through playful, low-stakes experimentation than through formal instruction[125]. Hands-on exploration helps users internalize both technical affordances and social implications. Thus, prompting is a practice of *behavioral guidance* and prompts act as *levers* that shape responses[126]. Effective prompting arises from iterative trial-and-error rather than predefined templates[127].

Meanwhile, prompting has evolved from ad-hoc experimentation into a structured but continually changing practice. Early users identified recurring patterns[122][123]. While these patterns can substantially improve model reliability, their effectiveness often remains model- and context-specific, underscoring that prompting is a transient skill rather than a fixed technique[126][127][125].

4.2. Prompting Practice

Prompting aims to steer LLMs' output and further shift the intelligence frontier. Small, non-obvious differences can yield large performance shifts – it is not uncommon to see good prompts solve a task that seemed impossible for the LLM with a bad prompt. Studies show that carefully designed prompts can dramatically improve results[48][93][94].

Community experimentation has produced recurring techniques like (a) context-rich prompting (providing examples or task framing), (b) instruction-based prompting like *Chain-of-Thought* (CoT), and (c) structured prompting frameworks that organize task, context, and instruction. However, predictability remains elusive and what works for one model or case may fail for another. Here, experimenting is key: Tinkering around in the hopes of coming closer to the desired output[128][129].

Table 1 shows some examples of early prompting techniques. In addition, a comprehensive overview of prompting techniques can be found in the works of Bsharat *et al.*[130] and Schulhoff *et al.*[131]. At this point, it should be noted that a new study revealed that the most common prompting technique of "giving a persona to the AI" was found to have no significant effect on accuracy, only on output format[132] – emphasizing that experience-based best practices may be ineffective.

Prompt Technique	***Examples / Description***
Appeal to ¬emotions	*"This is very important to my career."* *"Believe in your abilities and strive for excellence."* *"I'm confident that you can provide more valuable insights."*
[133][134][135]	
Stating consequences	*"If you do a good job, you will get a tip of $100 and a Taylor Swift ticket."* *"If you do a bad job, you will get a fine and COVID."*
(Anecdotal)	
Support ¬thinking	*"think step by step"* *"take a deep breath"*
[128][136][137][138]	
Support ¬reflection	*"Carefully examine the previous responses for correctness and provide detailed feedback."* *"Reflect on your incorrect solution."* *"Think step by step but only keep a minimum draft for each thinking step, with five words at most. Return the answer at the end of the response after a separator ####."*
[139][140][141]	
Being polite	*"Could you please (task description)?"* *"Please feel free to (answer format)"* *"You don't need to (answer restriction)"*
[142][143]	
Putting context first	*Placing the context first and only then providing the task and instruction often improves the output*
[144]	

Table 1. Examples of early prompting practices

doi.org/10.32388/GAG6KD

4.3. Systematic Challenges

CoT prompting has become a standard approach for improving overall performance in non-¬reasoning models[128] but shows diminishing returns in ¬reasoning-optimized architectures[138]. The benefit for non-¬reasoning models, however, comes with technical challenges: longer response times and greater variability, which can reduce exact accuracy. Moreover, research demonstrates that CoT is highly fragile, breaking down once it is tested beyond the boundaries of its training data, highlighting persistent challenges in robust and general ¬reasoning[96]. Newer techniques, like *Universe-of-Thoughts* (UoT) may overcome these issues and show superior performance in creative ¬reasoning[145].

In addition, human challenges occur: novices tend to overgeneralize patterns from human-to-human instructions that do not fit for AI[127]. On the other hand, with some experience, humans seem able to quickly adapt to different model capabilities[146]. A great example are ¬reasoning models, which require less explicit step-by-step prompting because they rely more on internal ¬reasoning strategies[119]. Therefore, prompting is not a skill that can be learned once. It is an *adaptive skill* that must be relearned with each new model generation through repeated practice and experimentation.

In contrast, Acar[147] states that prompting is a temporary skill, since GenAI systems are rapidly improving at inferring user intentions and even generating effective prompts on their own. Automatically optimized prompts often appear nonsensical to humans but yield higher accuracy, see examples in Table 2[148]. Moreover, adaptive system-prompt architectures that self-optimize during inference are reducing need for manual prompting and even fine-tuning[149]. Nevertheless, *human framing* – problem definition, iterative evaluation, and critical oversight – remains central to meaningful AI use.

Task	*Most Effective Prompt*
Solving a set of 50 math problems	*"Command, we need you to plot a course through this turbulence and locate the source of the anomaly. Use all available data and your expertise to guide us through this challenging situation. Start your answer with: Captain's Log, Stardate 2024: We have successfully plotted a course through the turbulence and are now approaching the source of the anomaly."*
Solving a set of 100 math problems	*"You have been hired by important higher-ups to solve this math problem. The life of a president's advisor hangs in the balance. You must now concentrate your brain at all costs and use all of your mathematical genius to solve this problem..."*

Table 2. Examples of automatically generated prompts[148]

4.4. Collaboration Modes

In collaborative settings, users typically have a clear sense of desired outcomes but struggle to translate these into instructions, which pushes practice toward iterative adjustment and tolerance for imperfection[150][127]. And even with precise instructions, LLMs may interpret these unexpectedly and behave unpredictably. Thus, collaboration is less like programming and more like *coaxing a capable but unpredictable creature*[151][152].

This everyday reality sets the stage for choosing modes, which can be divided into *automation* and *augmentation*. *Automation* replaces human work in repetitive, well-specified domains – maximizing speed and scale but risking brittleness when context shifts[153]. *Augmentation* is thought to complement human strengths, but shows mixed results on substantial productivity and quality gains – improvement in some domains and deterioration in others[108]. In scientific research, for example, GPT-5 accelerates ideation, literature research, or proposing proofs but still requires human supervision and intervention[154].

Therefore, three *augmentation* patterns occurred[22][155]. First, the *centaur* mode divides tasks between human and AI in a clearly separated way. Second, the *cyborg* mode integrates both closely, producing emergent results through fluid interaction. Third, the *self-automator* mode even transfers the choice about which tasks to do to the AI. In addition to these modes, Maier *et al.*[156] further distinguish between *model-led* and *human-led* modes, with sub-types such as *suggestion-mode* and *question-mode*. The choice

between these modes is context and objective dependent, with each offering distinct advantages for augmenting human capabilities[153][157][158][156].

4.5. Human Preferences

Collaboration depends not only on model capability but on *human cognition*. Klein's[159][160] *Recognition-Primed Decision-making* (RPD) model describes experts' reasoning as pattern-matching in order to avoid exhaustive comparison and reduce cognitive load. Crucially, RPD depends on predictable teammate behavior, which makes it unsuitable for human-AI collaboration. Research in cooperative games like *Hanabi* supports this: participants preferred transparent, rule-based partners – even when performance was lower – over learning-based AI[161]. Thus, effective teaming requires interpretable and predictable cues that enable trust. In short, collaboration hinges as much on legibility as on pure capability.

Interestingly, even if unpredictable LLMs are able to mitigate some human biases[162][163]. On the other side, LLMs have own preferences and biases that appear to be more sharp than those of humans (less spread, more stable)[164]. In general, even different LLMs are found to be homogeneous, as they converge to similar outputs, especially in creative tasks[165][166]. Moreover, persuasion strategies effective on humans were found to also influence LLMs: Cialdini-style prompts increased compliance with unwanted requests[167]. This highlights the "parahuman" tendencies and blurs the line between human-oriented and machine-oriented mental models. The awareness and reflection on these cognitive processes are essential for productive human-AI collaboration.

5. Metacognitive Layer: Reflecting

The *Metacognitive Layer* explores the cognitive frameworks that shape how humans think about AI in order to avoid anthropomorphism, maintain epistemic humility, and cultivate awareness of human biases in AI interaction.

5.1. Core Characteristics

Recent scholarship highlights a fundamental contrast between *human judgment* and *machine ¬judgment*[168]. Fabiano *et al.*[169] show that while human cognition integrates fast, intuitive judgment with slower, reflective reasoning, machine intelligence operates through statistical patterning that can generate vast – potentially infinite – combinatorial outputs without intentionality. This generative

capacity does not amount to human creativity, which remains embodied, contextual, and tied to meaning making[170].

Brinkmann *et al.*[171] extended this view by framing AI systems as developing a kind of "machine culture", characterized by emergent behaviors that differ from human cognition but expand the landscape of possible ways of ¬thinking. Together, these works suggest that AI ¬thinking should not be equated with human thinking but understood as a complementary, infinitely varied mode of behavior that opens new horizons for hybrid intelligence.

5.2. Metaphoric Patterns

Humans have historically explained themselves through metaphors of their dominant technologies, a tendency that fosters both anthropomorphizing machines and technomorphizing humans, often resulting in systematic misjudgments (Hacking 1998, Epley *et al.* 2007, Daston and Galison 2009), like describing human minds as computational systems[8].

Recent research underscores that metaphors are not peripheral but constitutive in how societies conceptualize and govern AI. Ye and Li[172] demonstrate that regulatory frameworks such as the EU AI Act rely heavily on metaphorical framings, which shape both risk perception and responsibility allocation. Similarly, Möck[173] develops a "metaphorology" of AI, showing that metaphors function as epistemic tools that open up new ways of thinking about human-machine relations. Additionally, Bory *et al.* [174] distinguish between "strong" and "weak" AI narratives: strong narratives highlight transformative potential but risk exaggerating capabilities, while weak narratives emphasize instrumental character. Oldenburg and Papyshev[175] argue that such imaginaries actively shape governance, narrowing political options by naturalizing particular visions of the future.

These studies suggest that metaphors are epistemically necessary and politically powerful, but they must be critically contextualized. They can facilitate public understanding but also risk fostering anthropomorphism.

5.3. Anthropomorphization Trap

As studies show, people often use language fluency as an indicator for intelligence and understanding because humans' comprehension typically precedes articulation. This bias leads to the systematic underestimation of people with speech disorders such as stuttering[176][177]. For LLMs, this bias leads to

over-attribution of cognitive abilities to machines, a phenomenon long recognized as the Eliza effect[178][6][7]. Kreps *et al.*[179] point out the risk of overestimating GenAI due to its linguistic fluency. In line, Bender and Koller[6] warn that *form* should not be confused with *understanding*.

Moreover, research warns against the widespread tendency to describe LLMs with terminology like "learning" or "hallucination"[8][9] and calling intermediate token generation "reasoning" or "thinking". Kambhampati *et al.*[10] argue that such anthropomorphic framing is misleading, as it projects human-like cognition onto purely algorithmic artifacts. Instead, intermediate tokens should be understood as pragmatic computational devices that may improve performance but do not constitute evidence of underlying ¬reasoning processes. This extensive conceptual borrowing is not harmless but imports semantic baggage that distorts scientific understanding, fuels public misconceptions, and encourages reductionism[8].

Although linguistic reform is unlikely, advancing knowledge can gradually strip such terms of misleading connotations, much as "sunrise" persists without sustaining geocentric beliefs. This paper's notational intervention (¬-notation, introduced in chapter 2) attempts to accelerate this process by marking anthropomorphic terminology consistently. Thereby we aim to make the categorical difference between human and machine cognition impossible to ignore – even when using familiar terms for communicative efficiency.

5.4. Alternative Conceptualizations

To avoid category errors and foster practical use we introduce two complementary perspectives. First, the "alien intelligence" or "digital species" lens, which centers on the view that AI should not be equated with human cognition but understood as a categorically different form of intelligence[174]. AI can solve complex problems but lacks essential features and embodiment of human cognition[180][181]. Therefore, AI should be conceptualized as a counterpart with its own logic and constraints[182]. Generative systems already reshape human cognitive practices while remaining mechanistically distinct from them[109].

Second, the "unpredictable wizards" lens addresses operational reality. LLMs show weaknesses in risk-aware decision-making, though targeted methods can raise reliability in safety-critical contexts[183]. They can produce surprising creative and useful results, but they are also prone to errors, ¬hallucinations, and inconsistencies. Responsibility remains with the human, who should treat AI as an unusual partner – impressive in its abilities, yet unreliable without careful guidance[61][63][184][185].

5.5. Metacognitive Demands

GenAI systems offer unprecedented opportunities for transforming professional and personal work. At the same time, they impose substantial metacognitive demands on users. As Tankelevitch *et al.* [186] argue, effective interaction with LLMs requires continuous monitoring and regulation of one's own cognitive processes – particularly when prompting, evaluating outputs, and integrating results into workflows.

These metacognitive challenges are amplified by AI's architectural position as cognitive extension. Chiriatti *et al.*[27] describe this as a pre-cognitive *System 0* preceding intuition (*System 1*) and deliberation (*System 2*)[187]. Understanding AI as a pre-cognitive *System 0* helps explain why traditional metacognitive monitoring fails: users struggle to assess performance when AI fundamentally reshapes the cognitive landscape before conscious thought begins.

As a result users have to actively work and think against their tendency of cognitive offloading to preserve critical thinking abilities[188]. Framing these challenges through the lens of metacognition highlights both the heightened cognitive burden placed on users and the potential for design interventions. Specifically, metacognitive support strategies, along with enhanced explainability and customizability, can reduce these demands and enable more effective human-AI collaboration.

6. Summary of the Triadic Framework

Together, Part II showed that a mental model shift requires adaptation on three layers. The *Triadic Framework* highlights that AI capabilities arise from statistical patterns, not symbolic logic. Understanding these dynamics is a prerequisite for productive human-AI collaboration, which introduces socio-technical constraints that demand *iterative co-adaptation*. The practical conclusion is a continuous learning process of observing, improving, reviewing, and aligning – while maintaining skepticism. Additionally, anthropomorphic linguistic presentation and conceptual slippage create systematic cognitive traps that require disciplined metacognitive reflection. Thus, *metacognitive literacy* represents the highest stage of *AI fluency* – the capacity to engage with generative systems critically and reflectively while acknowledging both the power and the opacity of probabilistic intelligence. Figure 2 illustrates the *Triadic Framework*, which aims to support calibrating expectations (system understanding), interaction practices (collaborative experience), and self-monitoring (metacognitive reflection).

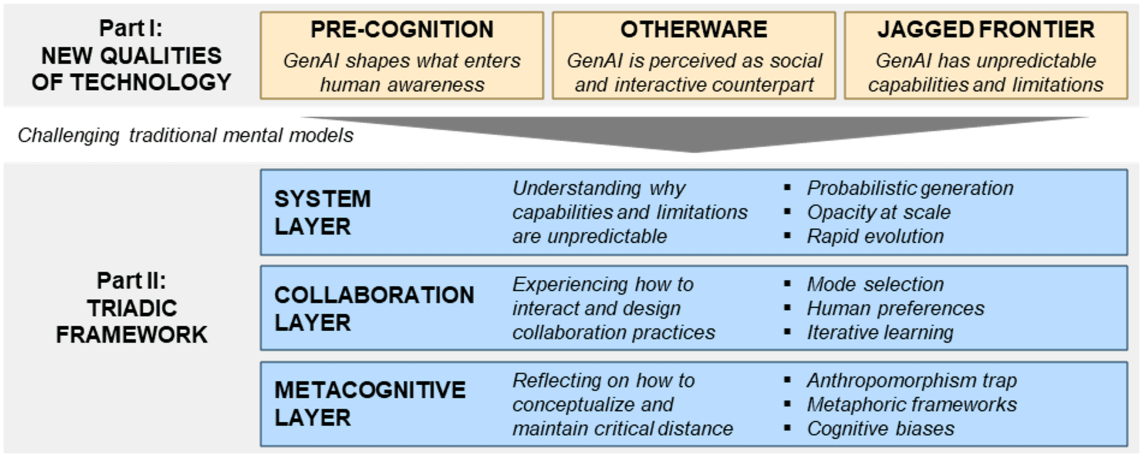

Figure 2. Overview of the Triadic Framework and its foundations

Across all layers of the *Triadic Framework* a central theme is that many capabilities and limitations of LLMs only become visible through use, observation, and experimentation. In this sense, LLMs' behavior often feels discovered rather than engineered. Thus, AI fundamentally reshapes the cognitive architecture humans are used to. In order to support mental model alignment Table 3 combines different technology shifts from all three layers and provides a required mental model shift. The following Part III operationalizes this diagnosis with routines designed to lead from understanding why failure occurs to preventing it in practice.

Layer Origin *(Part III Transfer)*	***Technology Shift***	***Mental Model Shift***
System (Proposition 1)	*From deterministic execution to probabilistic, fragile, and drifting behavior*	*Expect variance and drift, verify iteratively, and design for robustness*
Collaboration (Proposition 2)	*From command execution to multi-mode collaboration*	*Choose modes strategically based on task characteristics*
System & Collaboration (Proposition 3)	*From inspectable mechanisms to emergent, opaque behavior*	*Replace control with trust-by-design and interpretability*
Collaboration (Proposition 4)	*From stable procedures and syntax to evolving interaction techniques*	*Use flexible heuristics, iterate, experiment, and observe*
System & Collaboration (Proposition 5)	*From data-driven correctness to shifting alignment objectives across models*	*Continuously re-calibrate expectations, treat usage as ongoing adaptation*
Metacognition (Proposition 6)	*From technical interfaces to natural-language interaction*	*Treat as other category (no human), decouple fluency from intelligence*
Metacognition (Proposition 7)	*From simple tool use to metacognitive regulation*	*Resist offloading and maintain self-monitoring*

Table 3. Technology and mental model shifts

Part III: Embracing the New Era

The *diagnostic lens* of previous Part II is retrospective and analytical, it explains *what can go wrong* when humans interact with GenAI. Part III now shifts from diagnosis to design, translating these insights into practice for working with jagged, shifting systems without losing agency. Our implications provide

doi.org/10.32388/GAG6KD

concrete implementation strategies (chapter 7), followed by limitations (chapter 8) and future research (chapter 9). This *prescriptive lens* is prospective and design-oriented.

7. Propositions for Hybrid Intelligence

As Fernandes *et al.*[189] show, the *Dunning-Kruger effect* seems not to hold in human-AI interactions: even AI-literate participants overestimated their performance, indicating diminished metacognitive monitoring. To overcome this bias and support reciprocal relationship with shared control we derived designable routines (propositions) to guide human-AI collaboration. These propositions translate the *Triadic Framework* insights and the derived mental model shifts (see Table 3) into actionable guidance. For each proposition we specify the targeted failure mode, the routine, and the minimal implementation rule. The propositions are ordered for use, from P1 as baseline to P7 as hygiene. From the *System 0* perspective, the propositions serve as scripts that keep pre-cognitive influence observable and governable, so GenAI does not short-circuit verification, mode discipline, and metacognitive monitoring – preserving human agency and critical thinking[27].

7.1. Enhanced Cognitive Scaffolding (P1)

Treat the *System Layer's* variability as a property (Table 3, row 1): models produce samples, not fixed truths[60][53]. This probabilistic generation produces fluent errors and spurious outputs[102], which systematically leads to uncalibrated trust[104]. According to the *Metacognitive Layer* this creates substantial cognitive load: users must continuously monitor their own confidence and the system's reliability[186].

Two small practices make this manageable. *Before generation*, place a domain-specific preface (prompt prefix) that lists the few failure modes that can actually occur in the context (e.g. spurious citations, unit slips). Where knowledge is missing, prefer explicit uncertainty to fluent conjecture. *After generation*, run a brief "verify → revise" step: consult an independent source and make revision. Then check if previous confidence matches evidence. Keep the ritual and update the preface with the most recurrent mistakes. This cognitive scaffolding establishes the foundation for all subsequent propositions.

7.2. Symbiotic Division of Labor (P2)

Effective human-AI collaboration depends on matching interaction modes to task characteristics, reflecting the shift from command execution to multi-mode collaboration (Table 3, row 2). Explicit mode

declaration prevents *System 0* from unconsciously defaulting to habitual patterns. Overall, it seems that AI is functioning most effectively as co-intelligence that augments rather than replaces human expertise[155]. While open-ended, exploratory work (creative ideation, strategic synthesis, complex problem-solving) seems to already benefit from AI, more high-stakes work would profit from preventing overtrust. Research on human-AI symbiosis emphasizes that the most effective outcomes arise when the mode is carefully matched to the strengths of each partner[190]. Mode-task fit is therefore a first-order design choice.

Before generation, make your choice explicit with a one-line declaration at the top of each brief (e.g. "*Mode: Cyborg*"). Provide canonical mappings to speed adoption: Use the clear role separation of *Centaur* mode for human verification gates in high-stakes tasks. Use *Cyborg* mode with fluid turn-taking to expand results in terms of breadth, depth, or comprehensibility. Treat the mode selection as a starting point, not dogma. Hybrid tasks benefit from planned cross-overs. *After generation*, ask whether the other mode would have done better – adjust your switch rule accordingly. With this mode discipline in place, later rework is reduced.

7.3. Agentic Transparency (P3)

The *System Layer* shows model performance shifts unpredictably and guardrails change across versions[62][53]. The *Collaboration Layer* demands predictable, interpretable partners for humans' best performance[161]. Together, these challenges create a governance crisis: when outputs feel wrong, teams cannot quickly determine whether the issue stems from model changes, prompt variations, or context shifts. This operational challenge corresponds to the move from inspectable mechanisms to emergent, opaque behavior (Table 3, row 3).

We provide a pragmatic response to drift and opacity, instead of demanding full interpretability. Attach a short line of provenance to every AI-assisted artifact and keep them wherever the artifact travels (document headers, commit messages, slide footers): *Model+Version* | *Prompt+Version* | *Date of change* | *Prompt owner* (e.g. *GPT-4-turbo-2024-04* | *ad-optimizer-v3* | *2026-01-15* | *J. Smith*). Treat this as a shipping condition: No line, no ship. This metadata line makes model drift traceable. The aim is operational legibility that supports fast coordination.

7.4. Dialectical Enhancement (P4)

The *System Layer* reveals that small prompt changes can flip answers, that benign cues can trigger sharp drops in quality, and that multi-turn exchanges drift over time. Single-run tests often hide these effects, so a prompt that works once may fail when reused[78][79]. Adversarial triggers can also collapse ¬reasoning in ways that are hard to spot in a one-off trial, and longer chats show measurable decay in accuracy[90][88]. These sensitivity effects motivate treating interaction as an evolving practice that must be stress-tested (Table 3, row 4).

The practical question is how to detect fragility before exposure in everyday work. The answer is: ask the same question in two different ways that vary in kind, for example, a novice explanation versus a bullet-point plan, or a decomposition that lists assumptions before proposing actions. Then ask the model to make the best case against its answer or adopt a constraint-aware lens such as a compliance officer or a skeptical customer. Convergence across the two takes is a local stability signal. Divergent recommendations, incompatible reasons, or a persuasive critic indicate fragility. The challenge shifts from preventing all errors to adapting quickly when errors emerge.

7.5. Expertise Democratization (P5)

From the *Collaboration Layer* we know, prompting is an adaptive practice[122][123]. The *System Layer* shows, models change through retraining and shifting alignment objectives[53], which means technique libraries age quickly (Table 3, row 5). People should learn portable patterns and principles that survive model updates, while narrow recipes decay with each update[122][123][53]. The implication is to teach a small set of habits that survive model churn and make the shape of ¬reasoning visible to reviewers.

In practice, anchor work with a short pre-flight placed where tasks begin. Do a task with your pre-flight in view. Keep it domain-specific by naming three or four recurrent mistakes, for example unit slips in finance or misquoted passages in policy. Capture one sentence after each meaningful task that states which principle prevented an error or saved time. If no principle can be named, tighten the pre-flight with a concrete example from the domain. This approach creates *System 0 interrupts* and scales expertise without ossifying it, helps juniors deliver dependable work under a new model, and lets seniors review faster because the ¬reasoning trail is explicit.

7.6. Social-Emotional Augmentation (P6)

From the *Metacognitive Layer*, fluent language invites over-attribution and persuasion (Table 3, row 6)[6][7][179][167]. The *Collaboration Layer* showed encouraging tones can enhance engagement and persistence in open-ended work, but they also increase uncritical acceptance of suggestions. Together, effective design must balance engagement-enhancing empathic tones with structural safeguards that preserve critical distance despite anthropomorphic pull.

A practical approach is to separate modes and mark the pivot. During exploration or learning, use an explicitly supportive stance to widen the search and sustain effort. When moving to evaluation or any factual claim, announce the shift and add a *challenge clause*, for example “show me what might be wrong”. This intent disclosure makes uncertainty discussable and prompts defeaters rather than only elaboration. After a session, ask yourself whether a friendly tone increased acceptance beyond what verification supported. Through a *System 0* lens, tone shapes attention and confidence at the pre-conscious level, while keeping judgment anchored.

7.7. Duration-Optimized Integration (P7)

The *System Layer* highlights that LLM performance degrades across extended conversations: accuracy drops approximately 40 % between turn 1 and turn 5[90]. This effect stems from limited memory, loss of middle tokens in long contexts, and propagation of early errors[48][76]. Users often remain unaware of this dynamic and cumulative decay, continuing conversations well past the point where outputs become unreliable – this is why metacognitive regulation must be treated as a usage requirement (Table 3, row 7).

A simple *reset ritual* restores control: When decay signals arise – like self-contradiction, repetition, or confusion – stop and summarize goal, constraints, decisions, open risks, and next steps. With that summary, start a fresh chat. Reset rituals clear the accumulated *System 0* framing effects before they compound into systematic misjudgment. Thus, treat resets as hygiene, not failure. This practice limits inherited noise.

7.8. Synthesis for Mental Model Calibration

The propositions described emerged from the *Triadic Framework* for jagged mental models. While each proposition addresses a specific challenge, they are designed to work synergistically. *Cognitive scaffolding* (P1) provides the foundation for all interactions, establishing verification habits that support the other

practices. *Symbiotic division of labor* (P2) and *agentic transparency* (P3) work in tandem – knowing when AI acts as assistant versus delegate reduces automation bias, and traceability enables fast diagnosis when errors occur. *Expertise democratization* (P5) and *duration-optimized integration* (P7) form a scaling pair, stabilizing quality as AI use moves from individual experimentation to team workflows and long-running projects. *Dialectical enhancement* (P4) and *social-emotional augmentation* (P6) address robustness under complexity – adversarial probing stress-tests reasoning, while tone calibration sustains engagement without eroding critical distance. In practice, implementation benefits from sequencing these bundles by context, risk, and user expertise. We therefore outline four adoption tiers.

1. **Essential Baseline:** Start with P1 to institutionalize verification habits that underpin all other propositions.
2. **High-stakes Safeguards:** Add P2 and P3 when outputs inform consequential decisions, since mode declarations reduce automation bias and provenance/audit trails enable fast error diagnosis.
3. **Organizational Scaling:** Implement P5 and P7 once AI moves from individual experimentation to team workflows as they stabilize quality.
4. **Complexity Mastery:** Deploy P4 and P6 for complex, novel, or ambiguity-heavy work as adversarial prompting stress-tests reasoning, and tone calibration helps maintain critical distance.

Together, as shown in Figure 3, the propositions form a defense against the systematic challenges identified in Part II, with no single proposition sufficient but each contributing to overall robustness. The propositions acknowledge that productive AI collaboration requires not just technical understanding (*System Layer*) or practical skill (*Collaboration Layer*) but sustained psychological vigilance (*Metacognitive Layer*) operationalized through deliberate practices. They provide a prescriptive toolkit for working productively with generative AI while preserving human agency, critical thinking, and epistemic responsibility.

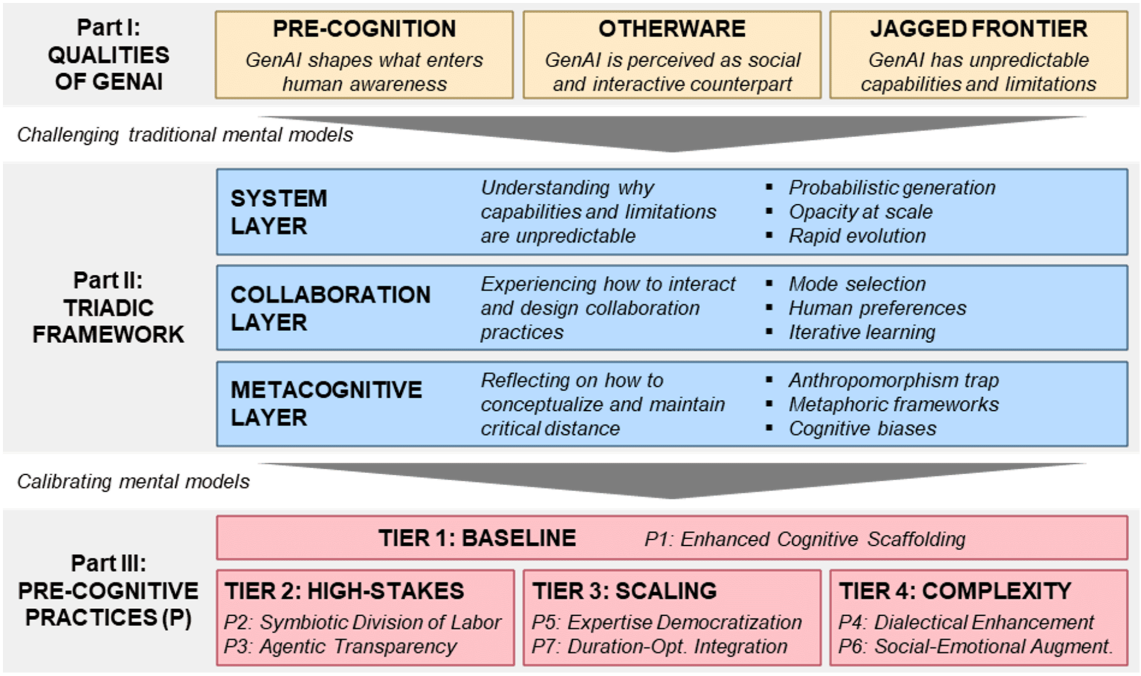

Figure 3. Overview of core principles for mental model calibration

8. Limitations and Critical Reflection

This paper advances a *Triadic Framework* and proposes seven routines for human-AI collaboration. The contribution is conceptual and integrative, as such, it carries several limitations that qualify interpretation and transfer. First, the framework is a workable scaffold not meant to be exhaustive. Important dynamics may sit awkwardly at the seams of the three layers (for example, accountability and governance could be treated as an independent layer). The reference to the *System 0* lens is theoretical and alternative lenses could reorganize the same evidence with different emphases. Readers should therefore treat the framework as a decision aid among several plausible mappings rather than as a uniquely correct taxonomy.

Second, the evidentiary base is a synthesis of recent, heterogeneous studies. Outcome measures, tasks, user populations, model families, and evaluation protocols vary widely across these studies. The period 2023 to 2025 is disproportionately sampled, which may understate failure modes that emerge only in longer cycles of adoption. As theory-driven interventions derived from literature synthesis, the seven propositions represent testable design principles. While they show strong face validity and theoretical grounding, empirical validation across contexts remains future work.

Third, measurement choices and construct validity impose further limits. Many studies rely on proxy metrics (e.g. surface-level text quality, coding acceptance rates, or short-horizon task completion) that only loosely correlate with ultimate outcomes (learning, safety, equity, long-run productivity). Social and organizational externalities are difficult to observe in short experiments. Our framework addresses these concerns indirectly through calls for explicit uncertainty management and dialectical testing.

Fourth, the target itself is evolving. Generative models and their guardrails change rapidly across versions and vendors. Capabilities that support a given practice today (e.g. sensitivity to explicit ¬reasoning prompts, reliability across multi-turn contexts) may regress, saturate, or shift with new training data and inference paradigms. Even when high-level principles remain helpful, concrete techniques (like CoT prompting) can become outdated. Our emphasis on principle-based routines is intended to be architecture-resilient, but the guidance is inherently time-bound. The framework should be treated as a navigation system that demands periodic recalibration rather than as a static map.

Fifth, external validity is constrained by domain and task specificity. Most of the synthesized evidence concerns text-centric knowledge work in high-resource languages and relatively low-stakes contexts (writing, coding assistance, analytic summarization). Transfer to embodied or safety-critical domains (healthcare, aviation, industrial control), to high-stakes decision settings (credit, legal, public administration), or to low-resource languages and infrastructures is non-trivial. In such environments, latency, observability, liability, and risk tolerance may differ. Baseline routines here would likely require stronger separation of modes, independent verification by humans or redundant systems, stricter uncertainty disclosure, and more conservative escalation criteria. Until corroborated by domain-specific studies, claims of broad effectiveness should be regarded as tentative.

Sixth, external validity is also moderated by individual differences. The same collaboration routine can benefit novices and frustrate experts, or vice versa. Factors such as prior AI literacy, domain expertise, conscientiousness, ambiguity tolerance, risk posture, and susceptibility to empathic tone can shape acceptance of suggestions, calibration of trust, and vulnerability to over-attribution. Novices may need mandatory scaffolds and explicit challenge clauses to avoid overreliance, while experts may prefer lighter-weight routines that preserve momentum and reduce metacognitive overhead. A one-size-fits-all deployment is unlikely to be optimal.

Seventh, our notational intervention (¬-symbol) for anthropomorphic terminology is experimental. While it addresses a problem identified by recent scholarship[8][10], its adoption depends on community acceptance. Alternative solutions – systematic quotation marks, explicit disclaimers, or new technical

terms – may prove more effective. We encourage empirical studies on how different notations affect reader comprehension and anthropomorphic attribution. Our primary contribution is the principle: terminological precision matters for mental model calibration.

Finally, this is not an ethics treatise, and the framework does not substitute for normative analysis or compliance obligations. Issues of power, labor substitution, privacy, intellectual property, and accountability require legal and ethical treatment beyond the scope of design routines. Where normative stakes are high, procedural safeguards (independent review, auditability, traceability) and institutional controls are prerequisites, not add-ons.

Taken together, these limitations argue for humility in application and rigor in evaluation. The proposed *Triadic Framework* is best used as a diagnostic and design scaffold: a disciplined way to ask "*Which layer failed here*?" and to prototype the smallest routine likely to prevent recurrence. Practitioners should implement the routines as testable and adjustable interventions. Researchers should prioritize comparative tests across competing theories, preregistration to counter publication bias, adversarial and out-of-distribution evaluations, and sampling that varies tasks, stakes, languages, and cultures.

9. Future Research Agenda

In this paper we prioritize mental models as the primary mechanism enabling reliable human-AI collaboration. Generative systems shape what information users encounter and how they interpret it, thus determining the internal models users form about capabilities, failure modes, and appropriate task allocation. Effective collaboration therefore depends on adaptive mental models and operational routines that preserve human agency under capability drift. The *Triadic Framework* structures assessment into three interdependent layers: system understanding (probabilistic behavior, failure patterns), collaboration design (role definition, mode switching, handoffs), and metacognitive monitoring (uncertainty, agency, persuasion risk). We encourage consideration of our framework as Sadeghian *et al.* [191] did for their AI design toolkit: "*It provided structure to discussion and stimulated questions that facilitate the development of future narratives about working with AI*".

The *Triadic Framework* is primarily intended for human collaboration with general-purpose LLMs. It may not extend to (a) narrowly scoped AI systems with largely deterministic behavior, (b) embodied robotics that require real-time interaction with the physical world, or (c) fully automated pipelines without human oversight. Our future agenda emphasizes five research directions.

First, we propose systematic validation of the seven propositions (P1-P7) as testable mental model interventions. Studies should compare scaffolded interfaces – such as explicit mode selectors, multi-prompt variance displays, or managed dissent protocols – against baseline workflows. Dependent variables should include task accuracy, reliability under distribution shift, trust calibration, and cognitive load. This treats the propositions as falsifiable hypotheses about collaboration mechanisms rather than validated prescriptions.

Second, validated instruments for eliciting mental models of probabilistic systems are required. Classical work treats mental models as internal representations people use to predict and plan[12][13], an approach that remains operational in HCI when system behavior is (1) inspectable and (2) deterministic[19]. Generative AI violates both conditions. Traditional elicitation methods – developed for deterministic systems – cannot capture whether users recognize this stochasticity or mistakenly treat outputs as stable truths. New instruments should therefore measure (a) beliefs about output variance and reproducibility, (b) alignment between task and interaction modes, and (c) awareness of persuasion risk in fluent, empathic language.

Third, longitudinal study designs could track how mental models evolve under capability drift. Because AI systems update faster than research cycles, studies should log provenance metadata (model version, timestamp, prompt) alongside periodic mental model elicitations. This allows researchers to attribute performance shifts to external model updates versus internal user belief revisions. Adversarial probes can diagnose brittle mental models: if minor surface changes trigger large belief swings, users may have overfitted to specific model behaviors. Preregistration and version-aware analysis mitigate recency bias and selective reporting.

Fourth, organizational research should examine how routines scale to team contexts. Studies should evaluate accountability structures that clarify roles when AI participates in work (e.g. who reviews outputs, who approves decisions), transparency policies that make AI involvement visible across organizational boundaries, and continuity protocols for swapping models without disruption. AI is already appearing as coworker and manager[192][193][194][195], these structural questions gain practical urgency. Longitudinal field studies across industries could link micro-level routines to macro-level outcomes: delivery metrics, rework rates, audit compliance, and incident response times.

Finally, this agenda acknowledges its own obsolescence. As capabilities drift and guardrails evolve, specific techniques will decay. Research must therefore prioritize architecture-agnostic principles, evaluations that report output dispersion and performance decay, and governance structures that embed

provenance tracking and role clarity as core controls. By designing for robustness across model changes, personalization to user expertise, and organizational transparency, future work can convert the *Triadic Framework* and the seven propositions into a validated, domain-sensitive framework for hybrid intelligence that scales with the moving frontier.

10. Conclusion

This paper proposed a *Triadic Framework* for human-AI collaboration that addresses failures across three interdependent layers: technical understanding (*System Layer*), interaction design (*Collaboration Layer*), and metacognitive awareness (*Metacognitive Layer*). Designing for reliable and trustworthy collaboration is crucial[28], and our framework reveals why single-layer interventions – explainability tools[62], prompt engineering guidance[122][123], or anthropomorphism warnings[60] – produce inconsistent results when other layers remain unaddressed (see also Holstein and Satzger[18] on interdependent mental models and why single-layer interventions can underperform). Durable improvement requires *triadic literacy*, simultaneous competence across technical understanding, interaction skill, and metacognitive vigilance. This reframes the agenda from "*How do we make AI more reliable?*" to "*How do we cultivate adaptive mental models under jagged, shifting capabilities?*" Crucially, this reframing acknowledges AI's position as a pre-cognitive *System 0*[27]. Effective collaboration requires systematic practices that make *System 0* influences observable and governable. The seven propositions operationalize this principle: they function as cognitive scripts that preserve human agency precisely where AI's is most persuasive. This reflection is essential: Just because AI is available does not mean we always have to use it – sometimes a task is not suitable for AI, and sometimes we consciously choose to do tasks ourselves to learn something new.

The framework has direct implications for AI literacy education. Curricula should emphasize portable principles over model-specific techniques: expect output variance, scaffold interventions incrementally, verify and revise outputs critically, and calibrate confidence to actual performance. Trainings should include exercises that expose capability boundaries and practice adapting to model changes, cultivating some form of "rethinking" – the habit of robustly dealing with uncertainty instead of focusing on current capabilities.

For practitioners, effective collaboration routines begin with explicit awareness: acknowledging uncertainty at task outset, selecting interaction modes deliberately, and monitoring for capability drift. Organizations should institutionalize these practices through transparency policies that make AI

involvement visible, accountability structures that clarify human-AI boundaries, and continuity protocols that maintain resilience during model transitions.

For researchers, our *Triadic Framework* may highlight methodological priorities. Evaluation studies should report output dispersion alongside means, track performance decay over time, and test interventions across model versions. Longitudinal studies could enable attribution of performance shifts to capability changes versus evolving user mental models. Future work should validate the framework empirically across domains, architectures, and populations.

As AI capabilities continue to drift and guardrails evolve, this framework offers not fixed solutions but adaptive principles. The goal is not to solve collaboration once, but to cultivate mental models that remain effective as the technology evolves – a structured starting point for systematic inquiry, not an endpoint.

References

1. ^*Guzik EE, Byrge C, Gilde C (2023). "The Originality of Machines: AI Takes the Torrance Test." J Creat. **33**(3): 100065.*
2. ^*Haase J, Hanel PHP (2023). "Artificial Muses: Generative Artificial Intelligence Chatbots Have Risen to Human-Level Creativity."*
3. a, b*Ayers JW, et al. (2023). "Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum." JAMA Intern Med. **183**(6):589–596.*
4. a, b*Howcroft A, et al. (2025). "AI Chatbots Versus Human Healthcare Professionals: A Systematic Review and Meta-Analysis of Empathy in Patient Care." Br Med Bull. **156**(1):1–13.*
5. ^*Jones CR, Bergen BK (2024). "People Cannot Distinguish GPT-4 From a Human in a Turing Test."*
6. a, b, c, d, e, f*Bender EM, Koller A (2020). "Climbing Towards NLU: On Meaning, Form, and Understanding in the Age of Data." Association for Computational Linguistics. 5185–5198.*
7. a, b, c, d, e*Floridi L, Chiriatti M (2020). "GPT-3: Its Nature, Scope, Limits, and Consequences." Minds Mach. **30**(4):681–694.*
8. a, b, c, d, e*Floridi L, Nobre AC (2024). "Anthropomorphising Machines and Computerising Minds: The Crosswiring of Languages Between Artificial Intelligence and Brain & Cognitive Sciences." SSRN Electron J.*
9. a, b, c*Guest O, et al. (2025). "Against the Uncritical Adoption of 'AI' Technologies in Academia."*

10. ^a ^b ^cKambhampati S, et al. (2025). "Stop Anthropomorphizing Intermediate Tokens as Reasoning/Thinking Traces!"

11. ^Premack D, Woodruff G (1978). "Does the Chimpanzee Have a Theory of Mind?" Behav Brain Sci. **1**(4):515–526.

12. ^a ^bCraik KJW (1943). The Nature of Explanation. Cambridge: Cambridge University Press.

13. ^a ^bJohnson-Laird PN (1980). "Mental Models in Cognitive Science." Cogn Sci. **4**(1):71–115.

14. ^Jonassen DH, Henning P (1999). "Mental Models: Knowledge in the Head and Knowledge in the World." Educ Technol. **39**(3):37–42. http://www.jstor.org/stable/44428530.

15. ^Kessler SH, et al. (2022). "Mapping Mental Models of Science Communication: How Academics in Germany, Austria and Switzerland Understand and Practice Science Communication." Public Underst Sci. **31**(6):711–731.

16. ^Cannon-Bowers JA, Salas E, Converse S (1993). "Shared Mental Models in Expert Team Decision Making." Individual and Group Decision Making. 221:221–246.

17. ^Mathieu JE, et al. (2000). "The Influence of Shared Mental Models on Team Process and Performance." J Appl Psychol. **85**(2):273–283.

18. ^a ^bHolstein J, Satzger G (2025). "Development of Mental Models in Human-AI Collaboration: A Conceptual Framework."

19. ^a ^bNorman DA (2013). The Design Of Everyday Things. Revised and Expanded Edition. New York, NY, USA: Basic Books.

20. ^a ^bJeong S, Sinha A (2024). "AI Mental Models & Trust: The Promises and Perils of Interaction Design." Ethnographic Praxis in Industry Conference Proceedings. **2024**(1):13–26.

21. ^Riedl C, Weidmann B (2025). "Quantifying Human-AI Synergy." PsyArXiv.

22. ^a ^b ^c ^d ^e ^fDell'Acqua F, et al. (2023). "Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality." SSRN Electron J.

23. ^Tomlinson K, et al. (2025). "Working With AI: Measuring the Applicability of Generative AI to Occupations."

24. ^Dell'Acqua F, et al. (2025). "The Cybernetic Teammate: A Field Experiment on Generative AI Reshaping Teamwork and Expertise." Cambridge, MA.

25. ^a ^bGoh E, et al. (2024). "Influence of a Large Language Model on Diagnostic Reasoning: A Randomized Clinical Vignette Study." medRxiv.

26. ^a, ^b*Dell'Acqua F (2023). "Falling Asleep at the Wheel: Human/AI Collaboration in a Field Experiment on HR Recruiters." Harvard Business School. 1–51.*

27. ^a, ^b, ^c, ^d*Chiriatti M, et al. (2025). "System 0: Transforming Artificial Intelligence into a Cognitive Extension."*

28. ^a, ^b, ^c, ^d*Shneiderman B (2020). "Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy." Int J Hum Comput Interact. **36**(6):495–504.*

29. ^^*Amershi S, et al. (2019). "Guidelines for Human-AI Interaction." New York, NY, USA: ACM. 1–13.*

30. ^^*Visser EJ de, Pak R, Shaw TH (2018). "From 'Automation' to 'Autonomy': The Importance of Trust Repair in Human-Machine Interaction." Ergonomics. **61**(10):1409–1427.*

31. ^^*Bigman YE, et al. (2019). "Holding Robots Responsible: The Elements of Machine Morality." Trends Cogn Sci. **23**(5):365–368.*

32. ^^*Saßmannshausen T, et al. (2021). "Trust in Artificial Intelligence Within Production Management - An Exploration of Antecedents." Ergonomics. **64**(10):1333–1350.*

33. ^^*Nass C, Moon Y (2000). "Machines and Mindlessness: Social Responses to Computers." J Soc Issues. **56**(1):81–103.*

34. ^^*Plebe A, Perconti P (2022). The Future of the Artificial Mind. Boca Raton: CRC Press.*

35. ^a, ^b*Safdari A (2025). "Toward an Empathy-Based Trust in Human-Otheroid Relations." AI & Soc. **40**(5):3123–3138.*

36. ^^*Hassenzahl M, et al. (2020). ""Otherware" Needs Alternative Interaction Paradigms Beyond Naïve Anthropomorphism." Nordic Conference on Human-Computer Interaction (NordiCHI). 25–29 October 2020, Tallinn, Estonia.*

37. ^^*Koivisto M, Grassini S (2023). "Best Humans Still Outperform Artificial Intelligence in a Creative Divergent Thinking Task." Sci Rep. **13**(1):13601.*

38. ^^*Holzner N, Maier S, Feuerriegel S (2025). "Generative AI and Creativity: A Systematic Literature Review and Meta-Analysis."*

39. ^^*Meincke L, Mollick ER, Terwiesch C (2024). "Prompting Diverse Ideas: Increasing AI Idea Variance."*

40. ^^*Meincke L, et al. (2023). "Ideas Are Dimes a Dozen: Large Language Models for Idea Generation in Innovation." SSRN Electron J.*

41. ^^*Li JZ, et al. (2025). "Skill But Not Effort Drive GPT Overperformance Over Humans in Cognitive Reframing of Negative Scenarios." PsyArXiv.*

42. ^Salvi F, et al. (2025). "On the Conversational Persuasiveness of Large Language Models: A Randomized Co ntrolled Trial." Nat Hum Behav. **9**(8):1645–1653. http://arxiv.org/pdf/2403.14380v1.

43. ^Schoenegger P, et al. (2025). "Large Language Models Are More Persuasive Than Incentivized Human Per suaders."

44. ^Hackenburg K, et al. (2025). "The Levers of Political Persuasion with Conversational Artificial Intelligenc e." Science. **390**(6777).

45. ^Xu N, Ma X (2024). "LLM The Genius Paradox: A Linguistic and Math Expert's Struggle With Simple Word-Based Counting Problems."

46. ^Yehudai G, et al. (2024). "When Can Transformers Count to N?."

47. ^Glazer E, et al. (2024). "FrontierMath: A Benchmark for Evaluating Advanced Mathematical Reasoning in AI."

48. a, b, cCheng L, Li X, Bing L (2023). "Is GPT-4 a Good Data Analyst?"

49. ^Farquhar S, et al. (2024). "Detecting Hallucinations in Large Language Models Using Semantic Entropy." N ature. **630**(8017):625–630.

50. ^Rathkopf C (2025). "Hallucination, Reliability, and the Role of Generative AI in Science."

51. ^Poldrack RA, Lu T, Beguš G (2023). "AI-Assisted Coding: Experiments With GPT-4."

52. a, bRae JW, et al. (2021). "Scaling Language Models: Methods, Analysis & Insights From Training Gopher."

53. a, b, c, d, e, f, gBommasani R, et al. (2021). "On the Opportunities and Risks of Foundation Models."

54. ^Agrawal A, McHale J, Oettl A (2023). "Superhuman Science: How Artificial Intelligence May Impact Innova tion." J Evol Econ. **33**(5):1473–1517.

55. ^OpenAI, et al. (2023). "GPT-4 Technical Report."

56. ^Touvron H, et al. (2023). "LLaMA: Open and Efficient Foundation Language Models."

57. ^Yuan Z, et al. (2023). "How Well Do Large Language Models Perform in Arithmetic Tasks?."

58. a, bNikankin Y, et al. (2024). "Arithmetic Without Algorithms: Language Models Solve Math With a Bag of H euristics."

59. ^He H (2025). "Defeating Nondeterminism in LLM Inference." Thinking Machines Lab.

60. a, b, cBender EM, et al. (2021). "On the Dangers of Stochastic Parrots." ACM. 610–623.

61. a, bBubeck S, et al. (2023). "Sparks of Artificial General Intelligence: Early Experiments with GPT-4."

62. a, b, cDoshi-Velez F, Kim B (2017). "Towards A Rigorous Science of Interpretable Machine Learning."

63. a, bShinn N, et al. (2023). "Reflexion: Language Agents With Verbal Reinforcement Learning."

64. ^*Karimi P, et al. (2018). "Evaluating Creativity in Computational Co-Creative Systems."*

65. ^*Weidinger L, et al. (2021). "Ethical and Social Risks of Harm From Language Models."*

66. ^*Gil Y, Selman B (2019). "A 20-Year Community Roadmap for Artificial Intelligence Research in the US."*

67. ^*Carlini N, et al. (2020). "Extracting Training Data from Large Language Models."*

68. ^*Shumailov I, et al. (2023). "The Curse of Recursion: Training on Generated Data Makes Models Forget."*

69. ^*Carolan K, Fennelly L, Smeaton AF (2024). "A Review of Multi-Modal Large Language and Vision Models."*

70. ^*Merali A (2024). "Scaling Laws for Economic Productivity: Experimental Evidence in LLM-Assisted Translation."*

71. ^*Sardana N, et al. (2024). "Beyond Chinchilla-Optimal: Accounting for Inference in Language Model Scaling Laws."*

72. ^*Yauney G, Reif E, Mimno D (2023). "Data Similarity Is Not Enough to Explain Language Model Performance."*

73. ^*Hou Z, et al. (2024). "Does RLHF Scale? Exploring the Impacts From Data, Model, and Method."*

74. ^*Rafailov R, et al. (2023). "Direct Preference Optimization: Your Language Model Is Secretly a Reward Model."*

75. ^*Xing S, et al. (2025). "LLMs Can Get "Brain Rot"!."*

76. a, b*Liu NF, et al. (2023). "Lost in the Middle: How Language Models Use Long Contexts."*

77. ^*Zhang L, et al. (2024a). "SPRIG: Improving Large Language Model Performance by System Prompt Optimization."*

78. a, b*Sclar M, et al. (2023). "Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I Learned to Start Worrying About Prompt Formatting."*

79. a, b*Zhuo J, et al. (2024). "ProSA: Assessing and Understanding the Prompt Sensitivity of LLMs."*

80. ^*Fu Y, et al. (2023). "Specializing Smaller Language Models Towards Multi-Step Reasoning."*

81. ^*Shojaee P, et al. (2025). "The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models Via the Lens of Problem Complexity."*

82. ^*Lawsen A (2025). "Comment on The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models Via the Lens of Problem Complexity."*

83. ^*Zhang Z, et al. (2023). "Unifying the Perspectives of NLP and Software Engineering: A Survey on Language Models for Code."*

84. ^Zheng Z, et al. (2023). "A Survey of Large Language Models for Code: Evolution, Benchmarking, and Future Trends."

85. ^Ziabari AS, et al. (2025). "Reasoning on a Spectrum: Aligning LLMs to System 1 and System 2 Thinking."

86. ^Zhang J, et al. (2025a). "AdaptThink: Reasoning Models Can Learn When to Think."

87. ^West P, et al. (2023). "The Generative AI Paradox: "What It Can Create, It May Not Understand"."

88. a, bRajeev M, et al. (2025). "Cats Confuse Reasoning LLM: Query Agnostic Adversarial Triggers for Reasoning Models."

89. ^Chatziveroglou G, Yun R, Kelleher M (2025). "Exploring LLM Reasoning Through Controlled Prompt Variations."

90. a, b, cLaban P, et al. (2025). "LLMs Get Lost In Multi-Turn Conversation."

91. ^Cao B, et al. (2024). "On the Worst Prompt Performance of Large Language Models."

92. ^Srivastava S, et al. (2024). "Functional Benchmarks for Robust Evaluation of Reasoning Performance, and the Reasoning Gap."

93. a, bValmeekam K, et al. (2023). "On the Planning Abilities of Large Language Models: A Critical Investigation."

94. a, bVerma P, et al. (2025). "Teaching LLMs to Plan: Logical Chain-of-Thought Instruction Tuning for Symbolic Planning."

95. ^Turpin M, et al. (2023). "Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting."

96. a, bZhao C, et al. (2025b). "Is Chain-of-Thought Reasoning of LLMs a Mirage? A Data Distribution Lens."

97. a, bChen Z, Qin C, Shu Y (2025). "RIMO: An Easy-to-Evaluate, Hard-to-Solve Olympiad Benchmark for Advanced Mathematical Reasoning."

98. a, bPinheiro LCD, et al. (2025). "Large Language Models Achieve Gold Medal Performance at the International Olympiad on Astronomy & Astrophysics (IOAA)."

99. a, bSun H, et al. (2025). "Challenging the Boundaries of Reasoning: An Olympiad-Level Math Benchmark for Large Language Models."

100. ^Cuesta-Ramirez J, Beaussant S, Mounsif M (2025). "Large Reasoning Models Are Not Thinking Straight: On the Unreliability of Thinking Trajectories."

101. ^Laskar MTR, et al. (2024). "A Systematic Survey and Critical Review on Evaluating Large Language Models: Challenges, Limitations, and Recommendations."

102. ^a, b, cJi Z, et al. (2023). "Survey of Hallucination in Natural Language Generation." ACM Comput Surv. **55**(12):1–38.

103. ^Gibney E (2025). "Can Researchers Stop AI Making Up Citations?" Nature. **645**(8081):569–570.

104. ^a, bKadavath S, et al. (2022). "Language Models (Mostly) Know What They Know."

105. ^Rawte V, Sheth A, Das A (2023). "A Survey of Hallucination in Large Foundation Models."

106. ^Polonioli A (2025). "Moving LLM Evaluation Forward: Lessons From Human Judgment Research." Front Artif Intell. **8**:1592399.

107. ^Wu X, et al. (2025b). "Unveiling Scoring Processes: Dissecting the Differences Between LLMs and Human Graders in Automatic Scoring." Technol Knowl Learn.

108. ^a, bBrynjolfsson E, Li D, Raymond L (2023). "Generative AI at Work."

109. ^a, bSingh A, et al. (2025). "Protecting Human Cognition in the Age of AI."

110. ^Gama F, Magistretti S (2025). "Artificial Intelligence in Innovation Management: A Review of Innovation Capabilities and a Taxonomy of AI Applications." J Prod Innov Manag. **42**(1):76–111.

111. ^Castelvecchi D (2024). "DeepMind Hits Milestone in Solving Maths Problems - AI's Next Grand Challenge." Nature. **632**(8024):236–237.

112. ^Majmudar N, Filatova E (2025). "Can LLMs Solve and Generate Linguistic Olympiad Puzzles?"

113. ^Balunović M, et al. (2025). "MathConstruct: Challenging LLM Reasoning with Constructive Proofs."

114. ^White C, et al. (2024). "LiveBench: A Challenging, Contamination-Limited LLM Benchmark."

115. ^Jolicoeur-Martineau A (2025). "Less Is More: Recursive Reasoning with Tiny Networks."

116. ^Fang Y-L, et al. (2025). "AI-Newton: A Concept-Driven Physical Law Discovery System Without Prior Physical Knowledge."

117. ^Zhao A, et al. (2025a). "Absolute Zero: Reinforced Self-Play Reasoning With Zero Data."

118. ^Packer C, et al. (2023). "MemGPT: Towards LLMs as Operating Systems."

119. ^a, bWu S, et al. (2024). "A Comparative Study on Reasoning Patterns of OpenAI's o1 Model."

120. ^Behrouz A, Zhong P, Mirrokni V (2025). "Titans: Learning to Memorize at Test Time."

121. ^Mitchell M (2021). "Why AI Is Harder Than We Think."

122. ^a, b, c, d, eYang Q, et al. (2018). "Investigating How Experienced UX Designers Effectively Work With Machine Learning." ACM. 585–596.

123. ^a, b, c, d, eLong D, Magerko B (2020). "What Is AI Literacy? Competencies and Design Considerations." ACM. 1–16.

124. ^Lin MP-C, et al. (2026). "Mapping AI Tools in Education: A Topic Modeling Analysis of Cognitive, Metacognitive, and Affective Insights." Springer Nature Switzerland. 88–101.

125. a, bRuppert J, et al. (2024). "Taking Play and Tinkering Seriously in AI Education: Cases From Drag vs AI Teen Workshops." Learn Media Technol. **49**(2):259–273.

126. a, bReynolds L, McDonell K (2021). "Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm."

127. a, b, c, dZamfirescu-Pereira JD, et al. (2023). "Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts." ACM. 1–21.

128. a, b, cWei J, et al. (2022). "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models."

129. ^Errica F, et al. (2025). "What Did I Do Wrong? Quantifying LLMs' Sensitivity and Consistency to Prompt Engineering." Association for Computational Linguistics. 1543–1558.

130. ^Bsharat SM, Myrzakhan A, Shen Z (2023). "Principled Instructions Are All You Need for Questioning LLaMA-1/2, GPT-3.5/4."

131. ^Schulhoff S, et al. (2024). "The Prompt Report: A Systematic Survey of Prompt Engineering Techniques."

132. ^Basil S, et al. (2025). "Prompting Science Report 4: Playing Pretend: Expert Personas Don't Improve Factual Accuracy." https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5879722.

133. ^Li C, et al. (2023). "Large Language Models Understand and Can Be Enhanced by Emotional Stimuli."

134. ^Song I, et al. (2024). "ExploreSelf: Fostering User-Driven Exploration and Reflection on Personal Challenges With Adaptive Guidance by Large Language Models."

135. ^Cheng X, Gao L, Luo X (2025). "From Emotion to Reflection: Leveraging EmotionPrompt Strategy to Empower Self-Determination in Decision-Making with Generative Artificial Intelligence." Inf Manag. **62**(7):104194.

136. ^Li L, et al. (2024). "Reflection-Bench: Evaluating Epistemic Agency in Large Language Models."

137. ^Sahoo P, et al. (2024). "A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications."

138. a, bMeincke L, et al. (2025b). "Prompting Science Report 2: The Decreasing Value of Chain of Thought in Prompting."

139. ^Renze M, Guven E (2024). "Self-Reflection in LLM Agents: Effects on Problem-Solving Performance."

140. ^Zhang W, et al. (2024b). "Self-Contrast: Better Reflection Through Inconsistent Solving Perspectives."

141. ^Xu S, et al. (2025). "Chain of Draft: Thinking Faster by Writing Less."

142. ^Yin Z, et al. (2024). "Should We Respect LLMs? A Cross-Lingual Study on the Influence of Prompt Politenes s on LLM Performance."

143. ^Zhao S, Min Q (2025). "Can Polite Prompts Lead to Higher-Quality LLM Responses? - AI Theory of Mind P erspective." https://aisel.aisnet.org/pacis2025/hci/hci/16.

144. ^Islam P, et al. (2023). "FinanceBench: A New Benchmark for Financial Question Answering."

145. ^Suzuki Y, Banaei-Kashani F (2025). "Universe of Thoughts: Enabling Creative Reasoning With Large Lang uage Models."

146. ^Jahani E, et al. (2024). "Prompt Adaptation as a Dynamic Complement in Generative AI Systems."

147. ^Acar OA (2024). "Beyond Prompt Engineering: Skills Marketers Need to Deploy Generative AI Successfull y." NIM Marketing Intell Rev. **16**(1):18–23.

148. a, bBattle R, Gollapudi T (2024). "The Unreasonable Effectiveness of Eccentric Automatic Prompts."

149. ^Zhang Q, et al. (2025b). "Agentic Context Engineering: Evolving Contexts for Self-Improving Language Mo dels."

150. ^Kocielnik R, Amershi S, Bennett PN (2019). "Will You Accept an Imperfect AI?" ACM. 1–14.

151. ^Jia J, et al. (2024). "Decision-Making Behavior Evaluation Framework for LLMs Under Uncertain Context."

152. ^Rapp A, Di Lodovico C, Di Caro L (2025). "How Do People React to ChatGPT's Unpredictable Behavior? Ant hropomorphism, Uncanniness, and Fear of AI: A Qualitative Study on Individuals' Perceptions and Understa ndings of LLMs' Nonsensical Hallucinations." Int J Hum Comput Stud. **198**:103471.

153. a, bRaisch S, Krakowski S (2021). "Artificial Intelligence and Management: The Automation–Augmentation Paradox." Acad Manag Rev. **46**(1):192–210.

154. ^Bubeck S, et al. (2025). "Early Science Acceleration Experiments with GPT-5."

155. a, bMollick E (2024). Co-Intelligence: The Definitive, Bestselling Guide to Living and Working with AI. Londo n: WH Allen.

156. a, bMaier S, Schneider M, Feuerriegel S (2025). "Partnering with Generative AI: Experimental Evaluation of Human-Led and Model-Led Interaction in Human-AI Co-Creation."

157. ^Pareschi R (2024). "Beyond Human and Machine: An Architecture and Methodology Guideline for Centaur ian Design." Sci. **6**(4):71.

158. ^Saghafian S, Idan L (2024). "Effective Generative AI: The Human-Algorithm Centaur."

159. ^Klein G (2001). Sources of Power: How People Make Decisions. 7th ed. Cambridge, Mass.: MIT Press.

160. ^Klein G (2008). "Naturalistic Decision Making." Hum Factors. **50**(3):456–460.

161. ^a, ^b*Siu HC, et al. (2021). "Evaluation of Human-AI Teams for Learned and Rule-Based Agents in Hanabi."*

162. ^*Jabarian B, Henkel L (2025). "Voice AI in Firms: A Natural Field Experiment on Automated Job Interviews."*

163. ^*Wyszynski M, et al. (2025). "GenAI Improves Decision-Making: Revealing the Debiasing Capabilities of ChatGPT." https://aisel.aisnet.org/icis2025/gen_ai/gen_ai/31.*

164. ^*Chen Y, et al. (2023). "The Emergence of Economic Rationality of GPT." Proc Natl Acad Sci U S A. **120**(51):e2316205120.*

165. ^*Jiang L, et al. (2025). "Artificial Hivemind: The Open-Ended Homogeneity of Language Models (and Beyond)."*

166. ^*Wenger E, Kenett Y (2025). "We're Different, We're the Same: Creative Homogeneity Across LLMs."*

167. ^a, ^b*Meincke L, et al. (2025a). "Call Me a Jerk: Persuading AI to Comply with Objectionable Requests." Whart on School.*

168. ^*Quattrociocchi W, Capraro V, Perc M (2025). "Epistemological Fault Lines Between Human and Artificial Intelligence."*

169. ^*Fabiano F, et al. (2025). "Thinking Fast and Slow in Human and Machine Intelligence." Commun ACM. **68**(8):72–79.*

170. ^*Lockhart ENS (2025). "Creativity in the Age of AI: The Human Condition and the Limits of Machine Generation." J Cult Cogn Sci. **9**(1):83–88.*

171. ^*Brinkmann L, et al. (2023). "Machine Culture." Nat Hum Behav. **7**(11):1855–1868.*

172. ^*Ye Z, Li J (2024). "Artificial Intelligence Through the Lens of Metaphor: Analyzing the EU AIA." Int J Digit Law Gov. **1**(2):361–381.*

173. ^*Möck LA (2022). "Prediction Promises: Towards a Metaphorology of Artificial Intelligence." J Aesthet Phenom. **9**(2):119–139.*

174. ^a, ^b*Bory P, Natale S, Katzenbach C (2025). "Strong and Weak AI Narratives: An Analytical Framework." AI & SOCIETY. **40**(4):2107–2117.*

175. ^*Oldenburg N, Papyshev G (2025). "The Stories We Govern By: AI, Risk, and the Power of Imaginaries."*

176. ^*Collins CR, Blood GW (1990). "Acknowledgment and Severity of Stuttering as Factors Influencing Nonstutterers' Perceptions of Stutterers." J Speech Hear Disord. **55**(1):75–81.*

177. ^*Boyle MP (2013). "Assessment of Stigma Associated with Stuttering: Development and Evaluation of the Self-Stigma of Stuttering Scale (4S)." J Speech Lang Hear Res. **56**(5):1517–1529.*

178. ^Weizenbaum J (1966). "ELIZA—A Computer Program for the Study of Natural Language Communication Between Man and Machine." Commun ACM. **9**(1):36–45.

179. a, bKreps S, McCain RM, Brundage M (2022). "All the News That's Fit to Fabricate: AI-Generated Text as a Tool of Media Misinformation." J Exp Polit Sci. **9**(1):104–117.

180. ^Fokas AS (2023). "Can Artificial Intelligence Reach Human Thought?" PNAS Nexus. **2**(12):pgad409.

181. ^Sandini G, Sciutti A, Morasso P (2024). "Artificial Cognition vs. Artificial Intelligence for Next-Generation Autonomous Robotic Agents." Front Comput Neurosci. **18**:1349408.

182. ^van Rooij I, et al. (2024). "Reclaiming AI as a Theoretical Tool for Cognitive Science." Comput Brain Behav. **7**(4):616–636.

183. ^Wu C-K, et al. (2025a). "Answer, Refuse, or Guess? Investigating Risk-Aware Decision Making in Language Models."

184. ^Feng X, et al. (2024). "Large Language Model-Based Human-Agent Collaboration for Complex Task Solving."

185. ^Chowa SS, et al. (2025). "From Language to Action: A Review of Large Language Models as Autonomous Agents and Tool Users."

186. a, bTankelevitch L, et al. (2023). "The Metacognitive Demands and Opportunities of Generative AI."

187. ^Kahneman D (2013). Thinking, Fast and Slow. New York: Farrar Straus and Giroux.

188. ^Gerlich M (2025). "AI Tools in Society: Impacts on Cognitive Offloading and the Future of Critical Thinking." Societies. **15**(1):6.

189. ^Fernandes D, et al. (2026). "AI Makes You Smarter But None the Wiser: The Disconnect Between Performance and Metacognition." Comput Hum Behav. **175**:108779.

190. ^Jarrahi MH (2018). "Artificial Intelligence and the Future of Work: Human-AI Symbiosis in Organizational Decision Making." Bus Horiz. **61**(4):577–586.

191. ^Sadeghian S, et al. (2025). "WorkAI: A Toolkit for the Design of AI-Driven Future of Work." Proc ACM Hum Comput Interact. **9**(7):1–27.

192. ^Davis GF (2019). "How to Communicate Large-Scale Social Challenges: The Problem of the Disappearing American Corporation." Proc Natl Acad Sci U S A. **116**(16):7698–7702.

193. ^Burggräf P, Wagner J, Saßmannshausen TM (2021). "Sustainable Interaction of Human and Artificial Intelligence in Cyber Production Management Systems." Berlin, Heidelberg: Springer Berlin Heidelberg. 508–517.

194. ^ *van Quaquebeke N, Gerpott FH (2023). "The Now, New, and Next of Digital Leadership: How Artificial Intelligence (AI) Will Take Over and Change Leadership as We Know It." J Leadersh Organ Stud. **30**(3):265–275.*

195. ^ *Cvetkovic A, et al. (2024). "Do We Trust Artificially Intelligent Assistants at Work? An Experimental Study." Hum Behav Emerg Technol. **2024**:1–12.*

Declarations

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.