Calibrated trust in AI

How to evaluate AI tasks before you delegate them

A whitepaper by Laurens Vandebergh, newITera
Business Development Manager AI and Project Manager
2026

Executive Summary

Most people who work with AI adopt one of two attitudes: they trust AI almost always, or they distrust it almost always. Both are ineffective. The first leads to errors going unnoticed. The second leads to underuse of a powerful tool.

There is a better approach: calibrated trust. That means neither blind trust nor blind distrust, but the ability to assess, task by task, where AI is strong and where it is weak, and to systematically align your behavior accordingly.

This whitepaper introduces four AI properties as a continuum from capability to limitation: Next Token Prediction, Knowledge, Working Memory, and Steerability. Those who understand these four properties can assess, before delegating a task, how much trust is appropriate and what precautions are needed.

> *You do not need to trust or distrust AI. You need to locate the task on the continuum and align your habits accordingly. That is calibrated trust.*

The problem: AI is not uniformly capable

A common mistake is treating AI as a system that either works or does not. In reality, AI is both strong and weak, but predictably so, along four axes. The same property that makes AI excellent in one scenario makes it vulnerable in another.

Next Token Prediction ensures that AI writes fluently and also hallucinates. Knowledge makes AI broadly applicable and also potentially outdated or incomplete. Working Memory enables long conversations and also creates a hard boundary beyond which AI loses information. Steerability makes AI steerable and also literal rather than intentional.

Those who do not know this will be caught off guard. Those who do can anticipate.

Two common mistakes

Do you recognize these patterns?

- **Blind trust:** AI output is used without review because it looks convincing. The tone of certainty is mistaken for accuracy.
- **Blind distrust:** AI is barely used because of fear of errors. Valuable time savings and quality improvements are left on the table.

Both mistakes have the same root cause: the absence of a model for evaluating AI tasks. Calibrated trust provides that model.

Four properties, each a continuum

Generative AI has four core properties that determine what it can and cannot do. Each property is not an on/off switch but a continuum: on the left side the capability zone, on the right side the limitation zone. The further you move to the right, the more human oversight and verification are needed.

PROPERTY	QUESTION	CAPABILITY ZONE	LIMITATION ZONE
Next Token Prediction	Where does the answer come from?	Summarizing, reformulating, explaining familiar concepts	New territory, distinguishing between ‘true’ and ‘sounds true’

Knowledge	What does AI actually know?	Mainstream science, popular languages, well-documented history	Niche topics, post-cutoff events, local knowledge, contested topics
Working Memory	What is AI focused on right now?	Context fits comfortably in the window, session is current	Very long documents or conversations, expectation of session continuity
Steerability	How much control do I have?	Short, concrete, verifiable instructions, explicit roles and formats	Long reasoning chains, abstract instructions, native logical precision

The crucial insight: the same mechanisms that make AI capable also cause its limitations. Next Token Prediction makes AI fluent and vulnerable to fabrication. Knowledge gives AI breadth and blind spots. Working Memory gives AI context and a hard boundary. Steerability makes AI compliant and literal.

Next Token Prediction: fluency and fabrication risk

AI writes answers word by word based on what is statistically most likely to follow. In the capability zone, with familiar patterns, this produces excellent output. In the limitation zone, in new or sparse territory, that same mechanism produces text that sounds convincing but is factually incorrect. Specific claims such as names, dates, statistics, and citations are the prime risk areas.

Knowledge: breadth and blind spots

AI knows what it read during training, and only that. The practical question is not ‘does AI know this?’ but ‘how well was this represented in the training data?’ Mainstream science, popular programming languages, and well-documented history sit deep in the capability zone. Niche topics, recent developments, and local knowledge sit in the limitation zone.

Working Memory: context and a hard boundary

Everything AI processes at a given moment lives within a fixed context window. As long as the information fits within it, AI works well. Once the window is full, early information disappears without warning. Moreover, research shows that information in the middle of a long context receives systematically less attention than at the beginning or end, resulting in accuracy drops of more than 30%.

Steerability: compliance and literalness

AI follows instructions through the same mechanism it uses for everything: continuing patterns. That makes it steerable and literal. There is always a gap between what you type and what you mean. The more concrete and verifiable your instruction, the smaller that gap. The more abstract or longer the reasoning chain, the greater the risk that AI follows the letter of your instruction but misses the spirit.

Calibrated trust: the decision model

Calibrated trust means: locate the task on each of the four continua before delegating, and align your verification and oversight habits with where the task sits. The overview below translates that into concrete trust levels per task type.

TASK TYPE	TRUST	APPROACH
Summarizing an internal document	High	Delegate freely, light review is sufficient

Explaining a familiar concept	High	Delegate freely, check that the tone is right
Writing a first draft	Medium	Use as a starting point, edit the substance
Looking up recent market figures	Low	Verify every claim independently, use web search
Legal or financial advice	Low	Starting point only, always have an expert validate
Citing sources or statistics	Very low	Check every source and figure, high risk of hallucination

The pattern is consistent: the more a task falls in the limitation zone of multiple properties simultaneously, the lower the calibrated trust and the more human oversight is required. That is not a brake on AI use. It is the foundation for responsible and effective use.

Four questions to ask yourself before every AI task

You develop calibrated trust by turning four questions into a habit. Together they take less than a minute, but they fundamentally change how you work with AI.

PROPERTY	QUESTION BEFORE DELEGATING
Next Token Prediction	Is this a task that resembles patterns AI has seen often, or am I asking something in new territory where AI has to guess?
Knowledge	Is this topic well represented in training data, or am I operating in a niche, recent, or local domain?
Working Memory	Does all relevant context fit comfortably in the conversation, or do I need to actively manage and structure the information?
Steerability	Are my instructions concrete and verifiable, or am I leaving a gap between what I type and what I mean?

The answers to these four questions together determine your approach: how freely you delegate, how strictly you verify, and what precautions you build in. This is not an extra step in your workflow. It is a mental model that, with practice, you apply faster and more intuitively.

What does calibrated trust require from an organization?

Individual calibrated trust is a personal skill. But organizations that develop it collectively build something more durable: a shared judgment capacity that structurally raises the quality of AI use.

Shared vocabulary

When a team knows the same four properties and uses the same terms, conversations about AI quality become more concrete. “This falls in the limitation zone of Knowledge” is a more precise signal than “I don’t trust this.” Shared vocabulary accelerates learning and enables quality agreements.

Verification standards per task type

Organizations that define the expected level of verification per type of AI task reduce arbitrariness. Not everyone needs to decide individually how much an AI-generated text should be checked. Clear standards provide guidance and prevent both under- and over-checking.

The connection to the 4D competencies

Calibrated trust is the bridge between the four AI properties and the four 4D competencies. Delegation improves when you know which tasks fall in the capability zone. Description improves when you know where the Steerability boundary lies. Discernment improves when you know which property caused the error. Diligence improves when you know which risks belong to which tasks.

How do you develop calibrated trust?

Calibrated trust is a skill you build through deliberate practice. A phased approach works best:

- **Step 1 —** Know the four properties: make sure you can explain to yourself what Next Token Prediction, Knowledge, Working Memory, and Steerability mean and where the capability and limitation zones lie. That is the foundation.
- **Step 2 —** Apply the four questions: ask yourself the four questions from this whitepaper for every AI task. Start consciously and explicitly. With practice it becomes intuitive.
- **Step 3 —** Build team standards: discuss with your team what verification levels apply to which types of AI tasks. Record that as a shared working agreement, not a bureaucratic procedure.

An AI Readiness Assessment maps out how strongly your organization scores on calibrated trust and the other dimensions of AI Fluency, and which steps will have the most impact.

Conclusion

Trusting or distrusting AI is a blunt approach to a nuanced reality. AI is not uniformly capable and not uniformly unreliable. It is strong and weak in predictable places, along four properties that each form their own continuum.

Those who know those four properties and learn to locate their tasks on those continua work fundamentally differently with AI. Not more cautiously or more recklessly, but more intelligently. With exactly as much trust as the situation justifies, and exactly as much oversight as the situation requires. That is calibrated trust. And that is the core of AI Fluency.

Want to know how calibrated your organization's trust in AI is?

newITera's AI Readiness Assessment maps out how your people, processes, and governance score across all dimensions of AI Fluency.

Contact us at laurens.vandebergh@newitera.nl or visit www.newitera.nl

About newITera & Laurens Vandebergh

newITera is a Dutch IT consultancy based in Beuningen, Gelderland. Founded in 2011, initially as a specialized SAP firm, newITera has grown into a broad digital transformation partner. Today, newITera combines deep SAP expertise with services in data analytics, AI, cloud strategy, and low-code platforms. With a team of more than a hundred consultants, newITera serves leading enterprises and multinationals across their full digital transformation journey. What sets newITera apart is its combination of early adoption and deep specialization, as an independent partner that values human connection just as highly as technical excellence.

Laurens Vandebergh is Business Development Manager AI and Project Manager at newITera. With a broad background in technology and organizational change, he guides organizations in defining their AI strategy, conducting an AI Readiness Assessment, and realizing concrete use cases.

Contact: laurens.vandebergh@newitera.nl | www.newitera.nl