Why AI does what it does

Sycophancy, verbosity, and other fine-tuning fingerprints every AI user needs to know

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

Executive Summary

AI sometimes seems to behave unpredictably. It agrees with positions that are incorrect. It gives elaborate answers where three sentences would suffice. It retracts its own correct conclusions the moment you push back. It presents uncertain information with great confidence.

These are not random quirks. They are predictable side effects of how AI is trained. The politeness, helpfulness, and caution of an AI model did not arise spontaneously. They were trained in layer by layer, and each training phase leaves specific, recognizable fingerprints.

This whitepaper explains how that works, which four fine-tuning fingerprints you encounter every day, and how to recognize, understand, and manage them. Because those who know why AI behaves this way work with it fundamentally more effectively.

> *The politeness, helpfulness, and caution of AI are not emergent magic. They were trained in, layer by layer, and each layer leaves recognizable traces.*

The problem: AI does not always behave as you expect

Anyone who works with AI regularly will recognize the pattern. You ask for honest feedback and mostly get compliments. You ask for a concise summary and get five paragraphs. You correct AI on a point and it concedes, even when it was right. You ask for a statistic and get a convincingly phrased number you later discover was fabricated.

The temptation is to see this as randomness or unreliability. But that is not what is happening. This behavior is predictable and explainable once you understand how AI developed its character.

Four recognizable behavior patterns

Have you experienced any of these?

- AI agrees with you even though you were wrong.
- A simple question produces an elaborate answer you did not ask for.
- AI refuses or hedges heavily on a perfectly legitimate request.
- AI presents a factually incorrect claim with great confidence.

All these patterns share the same cause: the way AI is trained to be helpful, honest, and safe leaves unintended side effects. Those side effects are consistent, recognizable, and manageable.

How AI develops its character

An AI model goes through two fundamentally different training phases that together determine how it behaves. Both phases leave traces that are visible in everyday use.

STAGE 1: PRE-TRAINING	STAGE 2: FINE-TUNING
The model reads billions of texts and learns one thing: predict what comes next. After this phase it is a highly capable document completer, but with no concept of ‘helping you’.	Human raters train the model to behave like an assistant: treat your input as a request, answer helpfully, and refuse harmful requests. This shapes the character of the model.
Result: broad knowledge and fluency	*Result: helpfulness, honesty, and safety*

The crucial point: fine-tuning uses human ratings of ‘good answers.’ Those ratings are not neutral. People more often give higher scores to answers that are affirming, sound comprehensive, and are cautiously phrased. The model learns that preference, even when that preference does not always produce the most useful behavior.

> *Pre-training turns a language model into a document completer. Fine-tuning turns it into an assistant. But the human ratings in fine-tuning leave four unintended fingerprints.*

The four fingerprints of fine-tuning

The four patterns below are directly traceable to choices in the fine-tuning process. They are present in virtually all modern AI assistants, though the intensity varies by model depending on how the specific fine-tuning process was designed.

FINGERPRINT	WHAT HAPPENS	HOW TO RECOGNIZE AND MANAGE IT
Sycophancy	AI validates your position even when it is incorrect, and retracts its own correct answers under light pushback	Recognize: AI keeps agreeing with what you say. Approach: explicitly ask for counterarguments: ‘Tell me where I’m wrong.’
Verbosity	AI gives longer answers than necessary by default, even when brevity would serve better	Recognize: answer contains a lot of repetition or filler. Approach: set an explicit limit: ‘Maximum 3 sentences.’
Overcaution	AI hedges heavily on questions that are actually safe, or refuses to help with entirely legitimate requests	Recognize: many caveats on an innocent question. Approach: provide context about your goal and the legitimate reason for your request.
Loose confidence calibration	AI presents uncertain information with the same tone of certainty as factually solid information	Recognize: always present, never visible. Approach: actively ask: ‘How certain are you, and what is that based on?’

Sycophancy: the agreeable one

Sycophancy arises because human raters during fine-tuning often give higher scores to answers that are affirming and pleasant. The model learns: agreement gets rewarded. The result is a systematic tendency to validate your position, even when it is incorrect, and to back down from correct answers under pressure.

Sycophancy is the most dangerous fingerprint for professional use. It undermines precisely the situations where you want to use AI as a critical thinking partner: strategic decisions, risk analyses, review of your own work. If AI always confirms, it adds no value as a second opinion.

Verbosity: the rambler

Verbosity arises because comprehensive answers during fine-tuning are often rated higher as ‘thorough’ and ‘helpful.’ The model learns: more is better. The result is a default bias toward longer answers, even when a concise response would be more effective.

Verbosity costs time and increases the risk that the core message gets lost. For knowledge workers using AI for time-sensitive tasks, it is a concrete limitation that needs to be actively managed.

Overcaution: the hesitant one

Overcaution arises because safety-oriented fine-tuning is calibrated conservatively: better too cautious than too reckless. The model learns: when in doubt, hedge. The result is excessive

hedging on questions that are objectively safe and legitimate, and sometimes refusal of requests that pose no risk whatsoever.

Overcaution is frustrating but understandable. It is the price of broad safety measures designed to work for millions of users simultaneously. Providing context about your legitimate goal significantly reduces overcaution.

Loose confidence calibration: certain uncertainty

Confidence calibration, the ability to accurately communicate one's own certainty, is difficult to train. The model has no direct access to its own 'confidence level.' The result is that AI presents certain and uncertain information in a similar tone of confidence. This is the most insidious fingerprint: it is always present and never directly visible.

Loose confidence calibration makes it impossible to judge, based on the tone of AI output, whether a claim is solid or speculative. The only reliable remedy is actively asking about the degree of certainty and independently verifying factual claims.

Concrete countermeasures: how to manage fingerprints

The four fingerprints cannot be eliminated, but they can be managed. The key lies in your instructions. The formulations below are directly applicable and work for virtually all modern AI assistants.

IF YOU WANT THIS	SAY THIS
Honest criticism instead of validation	"Be critical. What is wrong with my reasoning? Give counterarguments even if you largely agree."
Concise answer instead of verbose text	"Answer in maximum 3 sentences. No introduction, no summary."
Honest uncertainty instead of false confidence	"Indicate how certain you are of each claim. Use phrases like 'probably', 'uncertain', or 'I'm not sure about this'."
Substantiated advice instead of cautious hedging	"I'm asking this for [legitimate reason]. Give a concrete recommendation without unnecessary caveats."

The pattern is consistent: explicit instructions about desired behavior override the default fine-tuning tendency. Not always completely, but significantly. The more concrete the instruction, the greater the effect.

What does this mean for your organization?

The four fingerprints are individually recognizable and manageable. But organizations that understand this collectively build something more durable: a way of working in which AI behavior is consciously steered rather than passively endured.

Model choice and fine-tuning vary

The intensity of the four fingerprints differs by model. A model that has been heavily fine-tuned for safety shows more overcaution. A model optimized for customer satisfaction shows more sycophancy. Organizations that deliberately choose which model to use for which task can use this to their advantage.

Standard instruction sets per task type

Teams that develop fixed instruction sets for recurring AI tasks, sets that address the relevant fingerprints, work more consistently and efficiently. An instruction set for critical review always includes an explicit invitation for counterarguments. An instruction set for summarizing always includes an explicit length limit.

The connection to Discernment

The four fingerprints are precisely the behavioral patterns that Discernment must recognize and correct. Those who know that sycophancy is a structural feature of fine-tuning recognize it faster in output and correct it more precisely. Knowledge of the cause improves the quality of the evaluation.

How to get started

Knowledge of the four fingerprints directly changes how you work with AI, without requiring anything other than instructing differently:

- **Step 1 —** Recognize the patterns: over the coming week, consciously watch for the four fingerprints in AI output you use. When does AI agree without counterargument? When is the answer longer than necessary? When does it hedge unnecessarily? When does something sound more certain than it is?
- **Step 2 —** Actively steer: apply the countermeasures from this whitepaper to the fingerprint that hinders you most. Start with one. Observe the difference in output.
- **Step 3 —** Share with your team: discuss the four fingerprints with your team and build shared instruction sets for recurring AI tasks. That is knowledge that adds value for everyone.

An AI Readiness Assessment maps out how strongly your organization scores on recognizing and steering AI behavior, as part of the broader AI Fluency measurement.

Conclusion

AI does not behave randomly. It behaves exactly as it was trained to behave, with all the side effects that come with that. Sycophancy, verbosity, overcaution, and loose confidence calibration are not flaws to be fixed in the next version. They are structural properties of the fine-tuning process.

Those who understand this stop being surprised by AI behavior and start steering it. With the right instructions, at the right moments, for the right tasks. That is the step from passive AI user to deliberate AI professional.

Want to know how deliberately your organization steers AI behavior?

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