CX AI LEADER ROUNDTABLE

Confidently Wrong

An executive playbook for knowing how to trust AI

Why GenAI systems produce persuasive but unreliable conclusions + a practical verification architecture.

DAVID LAHEY · DAVIDLAHEY.COM · 15-MINUTE BRIEFING

The most dangerous AI answer isn't obviously wrong.

It's 95% right, beautifully written, and wrong about the **5%** that determines your outcome.

GenAI produces answers. It does not inherently establish truth.

THE COST OF THE 5%

The 5% isn't random. It lands on your biggest decisions.

LOW STAKES

Drafting, brainstorming, summarizing

COST WHEN WRONG

Minutes

The error surfaces as you use it.

›

OPERATIONAL DECISIONS

Root-cause, prioritization, forecasting

COST WHEN WRONG

Quarters of effort

Resources aim at the wrong fix.

›

STRATEGIC DECISIONS

Pricing, renewals, board metrics, M&A

COST WHEN WRONG

ARR + credibility

Committed decisions that compound.

A 5% error rate is not 5% risk. When plausible-but-wrong answers reach judgment-heavy questions, the cost can compound.

THE FAILURE MODE

Coherent, but unverified

Prompt: *“Why did this strategic customer churn?”*

THE SYSTEM CONFIDENTLY CONCLUDES

- Product adoption declined
- The executive sponsor departed
- Support problems remained unresolved
- Competitor activity increased
- Renewal was in 90 days

SOURCE CHECK FINDS

- Adoption remained stable
- Sponsor change was 18 months earlier
- Competitor activity came from one comment
- Renewal was nine months away

The output is coherent. The evidence is not.

WHY IT HAPPENS

Plausibility meets confirmation bias

Plausibility

A coherent answer can be generated without independently establishing the claims inside it.

Sycophancy

Responses can shift toward the assumptions or preferences expressed in the interaction.

WHEN WE PROMPT

“I think we're going to lose this customer. Look at these notes and tell me why.”

INSTEAD OF

“What does the evidence suggest about this account?”

The first prompt supplies a question, a hypothesis, and a preferred conclusion. The model now has an anchor.

AI Sycophancy × Confirmation Bias → Persuasive Error

WHY SMART TEAMS STILL GET TRAPPED

The Confidence Trap

THE MODEL

Plausibility
Looks right

Sycophancy
Agrees with your framing

Lost nuance
Drops the contradictions

THE CONTEXT

Ambiguity
Unclear scope or intent

Missing evidence
Gaps filled by inference

Misplaced emphasis
Attention ≠ importance

US

Benchmark bias
Averages hide tail failures

Confirmation bias
We test what we expect

Fluency bias
Readable feels correct

Eloquence ≠ Evidence. Failure emerges from the model, the context, and us — and the last three we can catch ourselves doing.

Two fixes that feel right — and still fail

BAD FIX #1

Ask AI to check itself

The model still carries the same framing, context, and assumptions. It may defend and polish its original reasoning rather than overturn it.

Self-review can produce consistency without correctness.

BAD FIX #2

Just add more AI models

Better — independent models can expose assumptions and errors. But models trained on overlapping data can share correlated blind spots.

Agreement can be insight — or shared error. From the outside they look identical.

Both can add signal. Neither is the same thing as independent evidence.

What verification actually requires

One model → **answer** Multiple independent models → **challenge** Independent evidence → **verification**

1 PRODUCE

AI Role: Analyst

Turn the answer into explicit claims; separate what is known from what is inferred.

Facts · Assumptions · Inferences

2 ATTACK

AI Role: Adversary

Challenge the claims without inheriting the first model's rationale.

Evidence gaps · Hidden assumptions · Alternatives

3 VERIFY

Role: Verifier

Test material claims against evidence that exists independently of the model.

Sources · Tools · Deterministic checks

4 SYNTHESIZE

Role: Arbiter

Reconcile the competing outputs with the evidence; classify each conclusion.

Reconcile · Classify · Escalate

Model agreement is not evidence. Different vendors ≠ independent reasoning.

LIVE DEMONSTRATION

Multi-Model Adversarial Verification in practice.

1 PRODUCE

AI Role: Analyst

PROMPT

You're an expert in [ROLE]. Help me build a ____ for ____. The primary goal is to [GOAL1]. The secondary goal is to [GOAL2]. Format the output into ____ for a v1. Ask me any clarifying questions.

2 ATTACK

AI Role: Adversary

PROMPT

Attached is v1 of ____ for ____. The primary goal is to ____. The secondary goal is to ____. **Rate it on a scale of 0-10, 10 being perfect to the two goals. Show your reasoning and suggest any needed improvements.** Ask me any clarifying questions.

3 VERIFY

Role: Verifier

PROMPT

Is the attached v2 better or worse than or the v1 we created on the last turn? Show your reasoning and quantify the differences. Also rate v2 on a scale of 0-10, 10 being perfect to the two goals. Show your reasoning and suggest any needed improvements.

4 SYNTHESIZE

Role: Arbiter

PROMPT

I agree/disagree with ____. Update v2 of the document and create a v3 based on ____. Highlight any changes in v3 so they standout and are obvious.

Which model was most confident? Irrelevant. **What determined the conclusion?** Evidence.

MONDAY MORNING

Five rules for high-stakes AI

1. **Separate fact from inference.** Make the model label both — every claim carries its status.
2. **Don't unnecessarily reveal your preferred answer.** Reduce anchoring and sycophancy before the model responds.
3. **Engineer disagreement.** Independent models in different roles — and different vendors don't guarantee independence.
4. **Verify against source evidence.** Models are not evidence for one another.
5. **Scale rigor with consequence.** Bigger decisions get stronger verification.

CONCLUSION

Confidence is not evidence.
Consensus is not evidence.
Independent evidence earns trust.

MULTI-MODEL ADVERSARIAL VERIFICATION

Multiple models create challenge. Independent evidence creates verification.

CONTINUE THE CONVERSATION

David Lahey

Advisor to growth-stage B2B CEOs.

SIX EXITS AS FOUNDER & OPERATOR · AI DISRUPTION & GTM TRANSFORMATION

davidlahey.com · linkedin.com/in/davidlahey