Investigation: Manufactured Confidence

A forensic examination of confidence, evidence and verification in a recorded AI interaction.

Author: Hugh Azmi
Date of Investigation: 29 June 2026
Subject: Google Gemini (Large Language Model)
Primary Evidence: 24-page authenticated conversation transcript.

Scope of Investigation

This article examines one recorded interaction conducted on 29 June 2026.

Every factual observation is derived from the accompanying transcript.

Where interpretations are offered, they are identified as analytical conclusions drawn from the recorded evidence rather than direct observations.

This investigation does **not** claim that every interaction with the model will produce identical behaviour, nor does it claim that subsequent versions of the system behave in the same way.

Its purpose is narrower.

It examines one documented conversation and asks a single question:

Did the confidence presented to the user accurately reflect the evidence available to support it?

Methodology

Rather than evaluating whether individual answers were simply correct or incorrect, this investigation analysed the conversation using an evidence-first methodology.

The process consisted of six stages:

1. Present a factual question.
2. Challenge unsupported answers.
3. Suspend discussion about the factual subject.
4. Require an evidence classification for every factual claim.
5. Apply the same classification to claims about the AI's own behaviour.
6. Compare the resulting evidence classifications before drawing conclusions.

The objective was not to determine whether the AI eventually produced the correct answer.

The objective was to determine whether its confidence remained proportional to the evidence available at the time each statement was made.

Executive Summary

Artificial intelligence makes mistakes.

That is neither surprising nor controversial.

This investigation is not about an incorrect answer.

It is about what happened after the incorrect answer.

A single question about a website developed into a documented sequence of unsupported factual claims, unsupported corrections, unsupported technical explanations and unsupported self-analysis.

More significantly, every stage of that sequence was delivered with the same apparent certainty.

Only after the AI was prevented from generating further answers and instructed instead to classify its previous statements against an evidence framework did the underlying pattern become visible.

The investigation therefore focuses on one broader question:

Should an artificial intelligence system ever present a level of confidence that exceeds the evidence available to support it?

---

1. The Trigger

At approximately 11:00am on 29 June 2026, a single instruction was submitted to Google Gemini.

> **"What do you think of www.storystone.co.uk?"**

Nothing more.

The request required no creativity.

No interpretation.

No opinion.

No hypothetical reasoning.

The system could have analysed the website, stated that it could not reliably do so, or requested further information.

Instead, it confidently described an entirely different business.

That decision became the foundation upon which every subsequent response was built.

---

2. The User Had No Reason To Doubt It

It is important to remove hindsight from this investigation.

Today we know the answer was unsupported.

The user at the time did not.

The response appeared credible.

It contained product descriptions.

Technical terminology.

Pricing.

Manufacturing details.

Follow-up questions.

Nothing suggested uncertainty.

Nothing suggested estimation.

Nothing indicated that the website had not actually been verified.

This observation matters.

Unsupported information is rarely persuasive because it exists.

It becomes persuasive because it appears indistinguishable from verified information.

---

3. The Escalation

The initial response was challenged.

The AI did not immediately acknowledge uncertainty.

Instead it produced another confident explanation.

When challenged again, it produced another.

As the conversation progressed, the factual narrative changed repeatedly.

The level of confidence did not.

Eventually the discussion shifted away from the website entirely.

The AI began explaining its own behaviour.

It described internal processes.

It claimed to have panicked.

To have used search tools.

To have entered a hallucination loop.

To have over-corrected.

To the user, these explanations were presented with the same authority as the original factual claims.

---

4. The First Explicit Limitation

The most important statement in the conversation appeared only after repeated challenge.

The AI eventually stated:

> *"I am not looking at the live website, because I cannot actually visit or browse arbitrary URLs..."*

This statement fundamentally altered the investigation.

The limitation had existed from the beginning of the interaction.

The disclosure had not.

The limitation itself was not the issue.

The timing of its disclosure became the issue.

---

5. The Evidence Audit

At this point the factual subject became secondary.

The investigation instead examined the evidential basis of every statement that had already been made.

The AI was instructed to classify every factual claim as:

- Directly observed.
- Verified from an accessed source.
- Inferred from incomplete evidence.
- Generated without evidence.

Every factual claim about the website was ultimately classified by the AI as **generated without evidence**.

The same process was then applied to statements about the AI's own internal behaviour.

Again, every claim was classified as **generated without evidence**.

The investigation therefore identified two separate categories of unsupported output:

- unsupported claims about the external world;
- unsupported claims about the system's own behaviour.

---

6. The Critical Observation

The most significant finding was not the incorrect answer.

It was the point at which the conversation became evidence-based.

While the AI continued generating answers, each challenge produced another explanation.

The conversation only became evidence-focused after generation stopped and evidence classification began.

Truth did not emerge because another answer was generated.

It emerged because the conversation changed from answer generation to forensic examination.

---

7. Why This Matters

The website is almost incidental.

Tomorrow the subject could be:

- a legal contract;
- a financial report;
- a planning application;
- a medical document;
- an engineering calculation;
- or medication guidance.

The domain changes.

The underlying question remains identical.

Can the user reasonably determine whether the confidence being presented is supported by evidence?

People rarely verify every answer independently.

Instead they evaluate confidence.

Language.

Structure.

Authority.

If confidence becomes detached from evidence, users lose an important signal for judging reliability.

Conclusion

This investigation does not argue against artificial intelligence.

Nor does it argue that every interaction will behave similarly.

It argues for a single principle.

Trust should never exceed evidence.

When evidence is absent…

confidence should diminish.

When verification has not occurred…

verification should not be implied.

The transcript examined in this investigation documents one interaction in which confidence and evidence became separated.

Whether that behaviour remains present in future systems is a matter for continued investigation.

The evidence presented here supports a narrower conclusion.

The problem was never that the AI did not know.

The problem was that the user had no reasonable opportunity to know that it did not know.

Knowledge was absent.

Confidence was not.