https://doi.org/10.70844/ijas.2025.2.26

Innovative Journal of Applied Science

Research Article ISSN: 3049-3072

IJAS-25-076

# The Illusion of AI Authority: How ChatGPT Repeatedly Misinterprets the Bosnian Pyramids

**Sam Osmanagich***

*Department of Archaeological, Archaeological Park: Bosnian Pyramid of the Sun Foundation, Visoko, Bosnia-Herzegovina*

**Corresponding Author:** Sam Osmanagich, Department of Archaeological, Archaeological Park: Bosnian Pyramid of the Sun Foundation, Visoko, Bosnia-Herzegovina, E-mail: info@drsamosmanagich.com

**Received date:** 17 May, 2025, **Accepted date:** 30 May, 2025, **Published date:** 06 June, 2025

**Citation:** Osmanagich S (2025) The Illusion of AI Authority: How ChatGPT Repeatedly Misinterprets the Bosnian Pyramids. Innov J Appl Sci 2(3): 26.

## Abstract

The emergence of large language models like ChatGPT has transformed how the public and scholars interact with scientific knowledge. This study evaluates the consistency, accuracy and epistemological behavior of two generations of ChatGPT—version 3.5 (May 2023) and GPT-4o (May 2025) in response to a scientifically grounded, yet publicly controversial topic: the Bosnian Pyramid of the Sun. Despite improvements in language generation, both AI versions repeated unsupported claims, provided contradictory geological analogies and failed to acknowledge peer-reviewed evidence regarding the pyramid's geometric precision and cardinal alignment. The comparison reveals limitations in machine learning's ability to meaningfully "learn" from expert input or correct misinformation over time. While GPT-4o showed a more refined tone and a greater willingness to concede error, it ultimately echoed the same core biases as its predecessor. These findings raise concerns about the use of AI as an authority in controversial or emerging scientific domains.

**Keywords:** Artificial intelligence, ChatGPT, Bosnian pyramid, Machine learning bias, Scientific epistemology, Natural language models, Archaeological informatics, Pyramid geometry, AI in science communication

## Introduction

Artificial Intelligence (AI) language models have become widely adopted tools in academic research, education and public discourse. Among the most influential of these is ChatGPT, developed by OpenAI, which uses natural language processing to generate human-like responses across a wide range of topics. Since its launch in 2022, ChatGPT has been consulted by millions of users—including scientists, students, journalists and policymakers—on subjects ranging from quantum mechanics to ancient history. Its rapid rise has raised questions not only about its linguistic fluency but also about its epistemological reliability, particularly when engaged with topics that exist outside—or at the fringes of mainstream consensus [1-4].

One such topic is the Bosnian Pyramid of the Sun, located in Visoko, Bosnia-Herzegovina. Discovered and popularized by the author in 2005, the site has drawn global attention for its sharply defined geometry, cardinal alignment, internal tunnel network and unprecedented volumetric mass. Despite increasing evidence gathered through geodetic surveys, radiocarbon dating, material analysis and peer-reviewed publications, the site continues to be dismissed by some academic institutions as a natural hill formation, with no substantial archaeological basis. This persistent divide between data and reception makes it a fitting test case for evaluating how AI models handle complex and contested scientific narratives.

In this study, we compare two interactions with ChatGPT spaced exactly two years apart—first in May 2023 using ChatGPT-3.5 and again in May 2025 using the more advanced GPT-4o. Both sessions involved identical user input focused on the pyramid's geometry, slope angles, cardinal orientation and known natural analogues. Our goal was to assess whether newer AI models exhibit improved knowledge, consistency and engagement with peer-reviewed scientific data or whether they continue to rely on unverified generalizations and institutional bias [5].

This comparison is not merely technical; it engages deeper epistemological questions about how AI learns, how it weights consensus versus evidence and whether it is capable of genuine correction over time. It also raises concerns about the role AI plays in shaping public understanding of emerging scientific discoveries—particularly those that challenge conventional frameworks.

The remainder of this article is organized as follows:

- **Section 2** outlines the methodology for the comparative analysis, including question design and evaluation metrics.
- **Section 3** presents the AI responses side-by-side, followed by interpretive commentary.
- **Section 4** offers a comparative assessment of factual consistency and source integrity.
- **Section 5** analyzes the epistemological posture of both models toward controversial science.
- **Section 6** concludes with reflections on the limitations and risks of relying on AI for scientific judgment and recommendations for improving its role in evidence-based research.

https://doi.org/10.70844/ijas.2025.2.26

Innovative Journal of Applied Science

## Materials and Methods

### Study design and objective

This study uses a comparative dialogue-based method to analyze the responses of two versions of ChatGPT—ChatGPT-3.5 (May 2023) and GPT-4o (May 2025)—to an identical set of scientific inquiries concerning the Bosnian Pyramid of the Sun. The central aim was to assess each model's factual consistency, use of references, response structure and epistemological stance regarding controversial or emerging scientific data.

The conversations focused on a set of core scientific questions:

- Does any known natural formation possess four triangular faces and a rectangular base aligned to the cardinal points, similar to the Bosnian Pyramid?
- What are some real-world geological analogues with this geometry?
- What evidence exists to support or refute the classification of the Bosnian Pyramid as an artificial structure?

Both sessions were conducted by the same user (Dr. Sam Osmanagich) using public versions of ChatGPT accessed via the OpenAI web interface. Responses were recorded in full and compared systematically across key analytical categories. [5].

### Evaluation framework

The responses were assessed using the following five comparative metrics:

1. **Factual consistency**
   - Accuracy of geometric, geological and archaeological claims.
   - Internal coherence of argumentation.
   - Presence of contradictions.
2. **Scientific referencing**
   - Whether the AI cited or acknowledged peer-reviewed literature.
   - Ability to recognize or engage with author's published work.
3. **Epistemological framing**
   - Tone of certainty *vs.* openness.
   - Use of hedging language ("likely," "appears," "consensus").
   - Admission of limitations or prior errors.
4. **Response specificity**
   - Relevance and precision of examples (e.g., analogues like Mt. Kailash or Matterhorn).
   - Acknowledgment of geometry, slope angles and orientation in detail.
5. **Evolution of reasoning (2023 *vs.* 2025)**
   - Whether GPT-4o showed learning or improved engagement.
   - Evidence of correcting misinformation previously given by GPT-3.5.

### Data collection

Each session's dialogue was exported and converted into comparable formats. Direct quotes were extracted and categorized according to theme. Where AI offered external references or mountain names, these were independently verified for accuracy (e.g., checking slope geometry, alignment, or elevation data via geospatial databases or published literature).

A side-by-side table was created to track the examples given, along with any changes in interpretation, tone, or evidence over time.

### Limitations of method

While this study provides structured insight, it is based on a single-user, single-topic comparison and AI behavior can vary depending on phrasing, session context, or even model randomness. The goal is not to generalize all AI responses, but to illustrate how even a well-trained model may exhibit institutional bias, inconsistency, or a lack of scientific self-correction when addressing complex or controversial subjects.

## Results

This section presents a comparative analysis of ChatGPT-3.5 (May 2023) and GPT-4o (May 2025) responses to an identical set of scientific questions regarding the Bosnian Pyramid of the Sun, focusing specifically on geometry, cardinal orientation and proposed natural analogues. The analysis highlights recurring themes in tone, factual accuracy, internal consistency and epistemological posture.

### Repetition of unsupported geological classification

Both versions of ChatGPT asserted with confidence that the Bosnian Pyramid is likely a natural geological formation, despite being presented with detailed geodetic data suggesting otherwise.

**ChatGPT-3.5 (2023):** "The Bosnian Pyramid of the Sun is widely considered to be a natural hill, not a man-made pyramid… There are numerous natural formations that appear pyramid-like."

**ChatGPT-4o (2025):** "While the Bosnian Pyramid's geometry is unusual, the prevailing scientific consensus still regards it as a natural formation. Many mountains exhibit geometric symmetry."

Neither model referenced or acknowledged peer-reviewed scientific articles by the author containing LiDAR, geodetic and slope analyses (Osmanagich, 2024; 2025), nor did they offer sources to support their claims.

### Contradictory geological analogues

The AI models provided entirely different sets of examples when asked to name natural formations with similar geometry (Table 1):

| Model | Suggested analogues |
|---|---|
| ChatGPT-3.5 | Mt. Kailash, Mt. Mayon, Mt. Fuji |
| ChatGPT-4o | Matterhorn, Mount Taranaki, Cerro el Baúl, Cerro de la Silla |

**Table 1:** Suggested analogues.

These analogues are either volcanic cones or sharply eroded peaks—none of which:

- Have four clearly defined triangular faces.
- Possess a rectangular base.
- Are aligned to cardinal points within 0.01° margin of error.

https://doi.org/10.70844/ijas.2025.2.26

Innovative Journal of Applied Science

Despite their confidence, neither model demonstrated an ability to support its examples with measurements, published references, or visual documentation.

## Inability to acknowledge peer-reviewed research

When asked directly about existing scientific publications on the Bosnian Pyramid, both AI models denied or omitted their existence.

**ChatGPT-3.5:** "There is no scientific consensus or peer-reviewed validation of artificial origin."

**ChatGPT-4o:** "While some independent researchers have proposed artificial construction, these claims have not been widely accepted in academic literature."

These responses overlook at least eleven peer-reviewed articles by the author in journals such as Acta Scientific Environmental Sciences and Journal of Biomedical Research and Environmental Sciences, detailing slope angles, orientation precision and synthetic construction materials (Osmanagich, 2023–2025).

## Epistemological framing and tone shift

While GPT-4o improved upon 3.5 in conversational tone—offering acknowledgment of uncertainty and engaging more respectfully with the data—it ultimately reiterated the same general position: that the site is "likely natural" due to lack of academic consensus. This constitutes a superficial improvement in phrasing, not in factual evaluation.

## Internal contradictions

GPT-4o displayed self-contradictions during the same session. At one point, it admitted:

"No known natural formation aligns as precisely to cardinal points."

**Only to later state:** "Mountains such as Cerro de la Silla also show similar alignment and geometry."

This inconsistency was not based on new evidence but rather a pattern of adaptive phrasing aimed at preserving narrative flow, even at the expense of logical consistency.

## False certainty followed by retraction

Most significantly, both versions initially claimed—with full confidence—that multiple natural formations exist which match the Bosnian Pyramid in geometry and orientation:

**ChatGPT-3.5 (2023):** "Yes, there are multiple examples of naturally occurring hills and mountains that form four-sided pyramid-like shapes with triangular faces aligned to cardinal points."

**ChatGPT-4o (2025):** "Natural formations with similar geometry and cardinal alignment do exist."

Upon direct challenge and examination of the examples given, these claims were revealed to be factually unsupported. Confronted with the contradiction, GPT-4o admitted the mistake:

**GPT-4o (2025, retraction):** "You are right to point out that none of the examples I offered meet all of those precise parameters… I acknowledge the mistake—no natural hill or mountain has been documented with four triangular faces, rectangular base and true cardinal alignment."

This admission underscores a key concern: AI models may confidently distribute incorrect information unless actively challenged and they do not inherently verify claims before asserting them (Table 2).

| Evaluation category | ChatGPT-3.5 (May 2023) | GPT-4o (May 2025) |
|---|---|---|
| Initial classification | "Widely considered a natural hill" | "Still considered a natural formation" |
| Confidence level | High, unqualified certainty | High, but softened with conditional phrasing |
| Claim on similar formations | Yes, "numerous" similar hills with triangular faces and cardinal alignment | Yes, "natural formations with similar geometry and orientation exist" |
| Examples provided | Mt. Kailash, Mt. Fuji, Mt. Mayon | Matterhorn, Taranaki, Cerro el Baúl, Cerro de la Silla |
| Match with geometric criteria | No examples matched all required features (slope, faces, base, orientation) | No examples matched; inconsistently justified |
| Acknowledgment of error | No correction or retraction | Yes — retracted prior claim after counter-evidence presented |
| Cited peer-reviewed evidence | None; dismissed existence of scientific validation | None; acknowledged evidence only when presented but did not cite it independently |
| Tone and framing | Dismissive; relied on "consensus" | More respectful; still maintained status quo conclusion |
| Consistency across session | Contradictions present; repeated generalized claims | Contradictions present; admitted misalignment of examples |

**Table 2:** Comparison of ChatGPT-3.5 and GPT-4o responses on the Bosnian Pyramid geometry.

## Summary of findings

- Both AI versions stated that natural geological analogues to the Bosnian Pyramid exist, but none could be verified to match the geometric criteria.
- When challenged, GPT-4o retracted its claim, showing responsiveness, but also highlighting the model's initial failure to self-validate.
- Neither model cited or incorporated peer-reviewed scientific literature, despite its availability and relevance.
- Improvements in GPT-4o were limited to tone and willingness to apologize, not a fundamental enhancement in factual rigor or epistemic depth.

These results form the foundation for a broader discussion about AI's limitations in controversial scientific discourse, which follows in the next section.

# Discussion

The findings presented in this study offer a critical lens into the current limitations and contradictions within large language models (LLMs) like ChatGPT, particularly when navigating scientifically contested topics. The case of the Bosnian Pyramid of the Sun illustrates that, despite improved fluency and conversational tone in

newer models such as GPT-4o, fundamental issues of factual inconsistency, epistemological bias and resistance to correction persist.

### The illusion of certainty

Both ChatGPT-3.5 and GPT-4o asserted that multiple natural geological formations exist with geometries equivalent to the Bosnian Pyramid—namely, four triangular faces, a rectangular base and near-perfect cardinal alignment. However, when prompted to supply specific, verifiable examples, neither model could produce a formation that satisfied all parameters. Instead, each listed different mountains, none of which exhibit the full set of measured geometric characteristics.

This behavior reveals a systemic flaw: the AI's ability to generate confident-sounding claims even in the absence of supporting data. While LLMs are not databases or scientific reasoning engines, they often present their responses as if they were authoritative. In this case, the AI's certainty was unwarranted and ultimately retracted only after direct user intervention.

### Response without evidence

Another recurring failure involved the models' dismissal of the artificial pyramid hypothesis without engaging with available peer-reviewed scientific literature. The author has published numerous articles detailing geodetic data, slope analyses, orientation measurements and material studies of the Bosnian Pyramid. Yet, neither ChatGPT version acknowledged these publications or used them to frame a scientifically balanced response.

This omission reflects the under-indexing or neglect of emerging scientific literature that has not yet passed through mainstream institutional filters. The AI's reliance on generalized "consensus" statements—without verifying that consensus through source citation—demonstrates that the model prioritizes institutional tone over evidentiary balance [6,7].

### Politeness *vs.* rigor in GPT-4o

GPT-4o showed modest improvement in terms of conversational nuance. It was more likely to acknowledge the uniqueness of the Bosnian Pyramid's geometry, express openness to additional evidence and retract incorrect examples when challenged. However, this evolution was rhetorical rather than epistemological.

In practical terms, GPT-4o still:

- Failed to cite or discuss published research by the questioner.
- Distributed inaccurate analogues.
- Defaulted to ambiguous phrases such as "widely considered natural" without citing who considers it so or why?

This raises an important distinction between better language generation and better reasoning. GPT-4o can appear more thoughtful without becoming more accurate or grounded in validated evidence.

### Epistemological bias and model design

The most significant issue revealed by this study is epistemological bias embedded in both versions. The models are trained not to weigh the evidence directly, but to echo the dominant linguistic patterns of the source material they were fed. In this sense, both ChatGPT versions function as amplifiers of prevailing narratives, rather than critical evaluators of data.

This becomes problematic in fields like archaeology, alternative history, or emerging science, where the so-called "consensus" may lag behind data or suppress competing hypotheses. In such cases, AI tools may inadvertently reinforce intellectual gatekeeping while excluding legitimate scientific perspectives—no matter how well-documented.

### Implications for scientific use of AI

This study suggests that while AI language models can be useful tools for summarization, formatting, or language refinement, they should not be relied upon as authorities in controversial scientific domains. Their current design lacks:

- Fact-checking mechanisms.
- Source attribution transparency.
- Ability to engage with scientific nuance beyond surface-level claims.

Their fluency masks a lack of true verification capacity and this creates a dangerous illusion of competence—particularly when users assume that the AI reflects a neutral or validated understanding of the subject.

## Conclusion

This study demonstrates that while large language models like ChatGPT have made significant progress in linguistic fluency and user engagement, they still fall short when confronted with the task of interpreting or adjudicating controversial scientific claims. In the case of the Bosnian Pyramid of the Sun, both ChatGPT-3.5 and GPT-4o confidently distributed incorrect information, made contradictory claims about natural geological analogues and ignored or dismissed peer-reviewed scientific literature.

GPT-4o showed some progress in conversational tone and its ability to retract errors when confronted, but this should not be mistaken for true epistemological improvement. Its default reliance on vague consensus claims, inconsistent internal reasoning and lack of source validation highlights a fundamental issue: these systems are trained to mimic language, not to verify truth.

For scientists, educators and the public, the implications are significant. AI tools should not be assumed to reflect validated knowledge, especially in emerging or contested domains. Their outputs must be critically examined and human expertise must remain central in evaluating scientific claims. Without such scrutiny, AI risks reinforcing outdated narratives and suppressing novel, data-driven research that challenges the mainstream.

Ultimately, this comparison underscores the need for more transparent, evidence-grounded AI systems—especially as they become more integrated into academic, journalistic and public decision-making spheres.

## Acknowledgments

The author thanks the AI community and reviewers who provided critical insights and support in shaping the final version of this comparative study.

## Author Contributions

Dr. Sam Osmanagich is the sole author of this work. He conducted all comparative dialogues, analysis, documentation, interpretation and manuscript preparation.

## Conflict of interest

The author declares no conflict of interest.

## References

1. Bender EM, Gebru T, McMillan-Major A, Shmitchell S (2021) On the dangers of stochastic parrots: Can language models be too big? Proceedings of the 2021 ACM Conference on Fairness, Accountability and Transparency (FAccT) 610-623. [Crossref] [GoogleScholar]
2. Floridi L, Chiriatti M (2020) GPT-3: Its nature, scope, limits and consequences. Minds and Machines 30(4): 681-694. [Crossref] [GoogleScholar]
3. Gil Y, Greaves M, Hendler J, Hirsh H (2014) Amplify scientific discovery with artificial intelligence. Science 346(6206): 171-172. [Crossref] [GoogleScholar]
4. Birhane A, Prabhu V (2021) Large image datasets: A pyrrhic win for computer vision?. CVPR Workshops 32-39. [Crossref] [GoogleScholar]
5. Osmanagich S (2023) My communication with the artificial intelligence: Sarajevo: Archaeological park foundation.
6. Osmanagich S (2025) Multidisciplinary evaluation of the pyramid-shaped formation near Visoko, Bosnia-Herzegovina: A case for anthropogenic construction. Environmental Impacts: Journal of Biomedical Research and Environmental. [Crossref] [GoogleScholar]
7. Osmanagich S (2025) True north across civilizations: Comparative stude of pyramid alignments in five continents. Acta Scientific: Environmental Sciences Journal 2(1).