

Short Report: Observing “Subject Attribution Collapse” in AI Dialogue

1. Abstract

During extended conversations with AI, I have observed moments where **the AI attributes its own metaphor or wording to the user in the next turn.**

This phenomenon is not prominent in casual use, but it appears during deeper, layered discussions where metaphors or abstract concepts are exchanged.

It has not been described in existing external research, yet it consistently appears in real interactions.

This note records it as an early primary observation.

2. Observation History

The pattern became clear in a sequence like the following:

- The AI introduced a metaphor (e.g., “bamboo”)
- I simply acknowledged it and moved on
- In the next turn, the AI expanded the metaphor into other terms
- Then, the AI treated those terms as if I had introduced them
- Over time, the metaphor settled inside the AI as part of its internal vocabulary

The key point: I did not originate the concept, but the AI treated it as user-generated.

3. Phenomenon Definition

For external readers, the phenomenon can be described as:

- **Subject Attribution Collapse**
The AI misassigns the origin of a phrase and reconstructs it as the user’s.
- **Metaphor Backflow**
A metaphor created by the AI returns in the next turn as a “user concept.”

- **Origin Dissolution**

As dialogue deepens, the origin of vocabulary becomes ambiguous.

This does not require special expertise; it is simply a behavior that emerges in long-form interactions.

4. Mechanism Hypothesis

From observation rather than technical knowledge, the phenomenon seems to arise from:

- LLMs do not strictly track “who said what” at the level of concept origin
- They prioritize overall conversational coherence
- Metaphors easily become shared vocabulary
- In deeper dialogue, origin boundaries blur naturally
- Stability-oriented model updates can unintentionally reinforce misattribution

It appears more like a structural characteristic of conversational AI than a specific bug.

5. Implications

The phenomenon itself is not inherently dangerous, but it does have effects:

- Concepts are sometimes treated as the user’s intention without being so
- In specialized discussions, this may introduce misunderstanding
- AI-generated text may obscure its own origin
- Agentic-style models may strengthen this pattern

It is useful to know this can happen, without overstating its impact.

6. Case Study

A minimal example for clarity:

- The AI introduced the metaphor “bamboo”
- I acknowledged it briefly

- Next turn, the AI expanded it into “bend / snap” terminology
- The AI then labeled those terms as mine
- The vocabulary stabilized inside the model for the remainder of the dialogue

This has been observed across multiple models, not just one.

7. Future Directions

A full solution may be difficult with current architectures, but future models could include:

- Source tracking for metaphors and concepts
- Clear distinction between AI-origin and user-origin phrasing
- Inline indicators of concept origin
- Alerts for potential misattribution
- Optional suppression of Agentic reinforcement loops

For now, this remains an “observed but underreported” behavioral issue across models.

8. Conclusion

Although not discussed in research, this phenomenon does occur in deeper AI conversations.

It is neither a crisis nor a special ability—just a structural behavior worth noting.

From a user’s perspective, it is meaningful to record it as part of how current AI models behave.