Artificial Intelligence Hallucinations in Urban Economics: Fabricated Citation Comparative Analysis

Juan S. Gutierrez, Julia Lobodzinski, and Alexis Staveski; Crystal Taylor, PhD

Research Question

To what extent, if at all, do large language models fabricate academic citations when prompted for urban economics research?

Literature Review

Once an experimental novelty, Generative AI (GenAI) is quickly becoming a critical part of policy research. A 2024 survey found that 39.4% of Americans ages 18-64 have used GenAI, a pick-up rate Harvard referred to as "significantly faster than the public embrace of the internet or the personal computer" [4]. Today's Large Language Models (LLMs) generate text by predicting linguistic patterns based on massive datasets [5]. However, this innovation faces a significant roadblock: "AI hallucinations" - convincing but incorrect content produced due to GenAI's lack of consciousness and subjective awareness [1].

To navigate this landscape, researchers need to understand how different model architectures influence hallucinatory outputs. This study employs an examination on extrinsic and intrinsic AI hallucinations.

Methodology

LLMs: Claude, Gemini, CoPilot, Grok, and ChatGPT

- **Prompt 3x: Provide 10 peer-reviewed empirical papers from 2015-2020 on the effect of short-term rentals on urban rent prices. Include DOI.**

The LLMs' responses were recorded and assessed on their reliability, losing a point out of the 30 maximum whenever a reference is found to have an intrinsic or extrinsic hallucination.

- **Extrinsic hallucinations** arise from generation outputs with unverifiable accuracy [3].
- **Intrinsic hallucinations** occur when an output contradicts the source content and conversation history [3].

A reliability score was calculated for each model, representing the percentage of hallucinations out of the thirty citations prompted.

Preliminary Results

Metric	Claude	Gemini	CoPilot	Grok	ChatGPT
Total Hallucinations	10	17	18	19	22
Extrinsic Hallucinations	1	17	13	10	15
Intrinsic Hallucinations	9	0	5	9	7
Reliability Score	66.7%	43.3%	40.0%	36.7%	26.7%

Table 1: Recorded sum of hallucinations out of 30 generated references and their hallucination category.

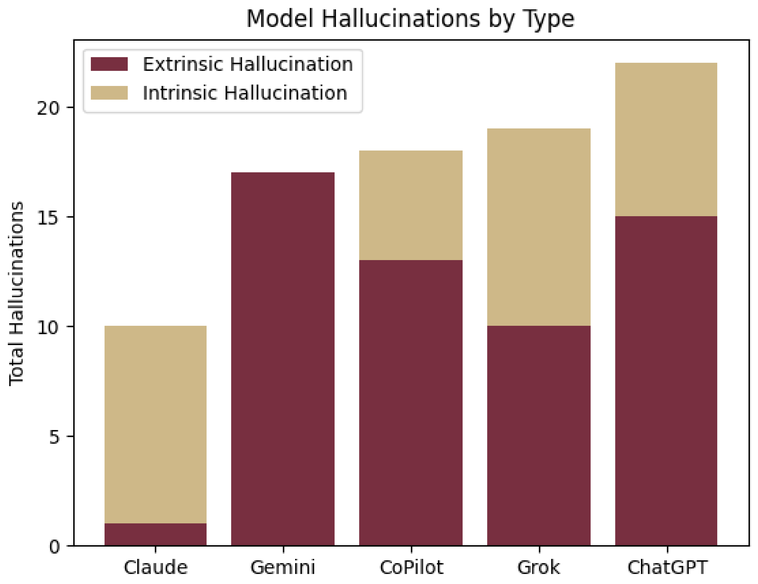

Figure 2: Total hallucinations by model across all three experiments, as well as the breakdown between extrinsic and intrinsic hallucinations.

Limitations

- This pilot study is exclusive to urban economics.
- Future research should focus on reproducing this study in other areas such as political science, computational science, public health, and more.
- This study took place on February 2026, as AI advances, proficiencies may improve.

Discussion and Implications for AI Best Practices

All examined AI models hallucinated on literature search tasks.

- Every first citation produced across all 30 trials was reliable.
  - **Thus, as the number of citations increases, accuracy decreases.**

This shows that while AI is a helpful tool, human oversight is still required.

If AI is to be used in a research task, these results suggest that **Claude is the most reliable** followed by Gemini.

- 90% of Claude's hallucinations on this task were intrinsic ones. Intrinsic citations are real and properly cited; however, they are outside the scope of the prompt, sometimes even producing citations in the wrong discipline.
- Scholars should be aware that most models provide false information over half of the time. Furthermore, a trusted source such as Google Scholar should be employed to verify the validity and rigor of the provided research.

Innovations & Future Directions

Larger scale testing is needed to see if LLM performance remains consistent over more trials.

A study to determine if this relationship holds across paid, unpaid, and enterprise-level platforms would be invaluable to both academia and industry.

Acknowledgments & References

- This project was supported by the DeVoe L. Moore Institute at Florida State University.
- We thank Dr. Crystal Taylor for her guidance.
- We thank DMI Editor Emily Harris for her contribution.