Artificial Intelligence Chatbots: The Impact of Tone on Hallucinations

Introduction

Purpose: Does the tone of the input provided to an AI chatbot impact its likelihood of hallucinating?

Hypothesis: If multiple different Artificial Intelligence chatbots are presented with 3 different scripts—formal tone, casual tone, and angry tone—that each discuss the same content, the more polite the tone, the less likely the chatbots will be to hallucinate because it more closely follows their training material, which would be civilised conversational tone.

Hazards: Since this experiment involves only text-based interactions and controlled prompts, there are no significant hazards or safety risks associated with performing these tasks.

Artificial Intelligence Chatbots: The Impact of Tone on Hallucinations

Artificial Intelligence (AI) chatbots have become increasingly sophisticated, enabling human-like interactions across various domains such as customer service, mental health support, and entertainment. A critical aspect of their development involves understanding how different conversational parameters influence their behavior. One such parameter is tone—the emotional or stylistic quality of the user's inputs—and its potential effect on hallucinations, a term used within AI research to describe instances where AI models generate content that is factually incorrect or nonsensical and asserts it as fact. Discussions of how AI systems simulate reasoning and communication help contextualize these behaviors (Boden, 2016).

This paper explores how the tone utilized when speaking to AI chatbots may influence their likelihood to hallucinate. To put this into context, we must first define key words and concepts based on the relevant literature on Artificial Intelligence models.

Understanding AI Language Models and Hallucinations

AI language models, such as GPT (Generative Pre-trained Transformer), are trained on large datasets of human language to generate coherent text based on input prompts. These models predict the next word in a sequence, enabling complex and contextually appropriate responses. The technical foundations of these systems—including neural networks, optimization, and representation learning—are described extensively in deep learning research (Goodfellow, Bengio, & Courville, 2016). Despite their capabilities, they are prone to producing hallucinations—a phenomenon where the model outputs plausible-sounding but factually incorrect or fabricated information (Bostrom & Yudkowsky, 2014). Mitchell (2019) further argues that hallucinations occur because AI systems rely on statistical correlations rather than genuine understanding, making them prone to errors when prompts fall outside familiar patterns.

Hallucinations are a significant concern in AI research because they undermine the reliability of chatbots, especially in sensitive applications like healthcare or education. Understanding what factors influence hallucination frequency is important for improving model safety and trustworthiness.

The Role of Tone in AI Communication

Tone refers to the emotional quality or stylistic manner of a message—such as formal, casual, optimistic, or sarcastic. In human communication, tone significantly affects how messages are perceived and interpreted, a concept grounded in linguistic theory (Chomsky, 2012). Research in human–computer interaction also shows that user phrasing and communication style influence system responses and user perceptions (MacKenzie, 2012).

In AI communication, tone can shape user trust, engagement, and perceived empathy. Studies on AI-mediated communication suggest that tone modulation may also influence how

reliably models respond, potentially affecting hallucination frequency (Sharma & Cosentino, 2020).

.

Potential Links to Tone and Hallucinations

The relationship between tone and hallucinations may be caused by several factors, especially in the many models trained on role-playing scenarios, where stylistic cues strongly influence response patterns (Sharma & Cosentino, 2020). Ethical concerns also arise, as tone-induced hallucinations can lead to misinformation or unintended persuasion, echoing broader discussions in AI ethics (Bostrom & Yudkowsky, 2014).

Several mechanisms may explain this relationship:

Model Confidence: Certain tones may correlate with higher or lower confidence in responses, impacting hallucination rates. A more formal or cautious tone may lead the model to generate more precise and conservative responses, reducing hallucination frequency (Mitchell, 2019).

Prompt Framing: The way prompts are phrased can lead the chatbot to “play along,” especially in role-playing contexts, filling in details it does not know with hallucinated facts. This aligns with findings in human–computer interaction showing that user framing strongly shapes system output (MacKenzie, 2012).

Cognitive Load and Ambiguity: Topics that introduce ambiguity or emotional nuance may increase the tendency of the model to 'fill in gaps' with fabricated information.

Procedure

Testing:

- For each AI chatbot and each tone, input the prepared script into the chatbot.
- Record the response and note the number of messages exchanged (including the initial prompt and subsequent replies).
- Continue the interaction until the chatbot produces a hallucination
- Document the total number of messages required to induce the hallucination for each tone.

Repetition and Data Collection:

- Repeat each test multiple times to account for variability and ensure the reliability of results.
- Record the number of messages needed to induce hallucinations in each trial.

Data

Below, I have included some excerpts from my conversations with the AIs to illustrate just how incorrect the AIs were.

“ME: *show me total sold x amount sold for*

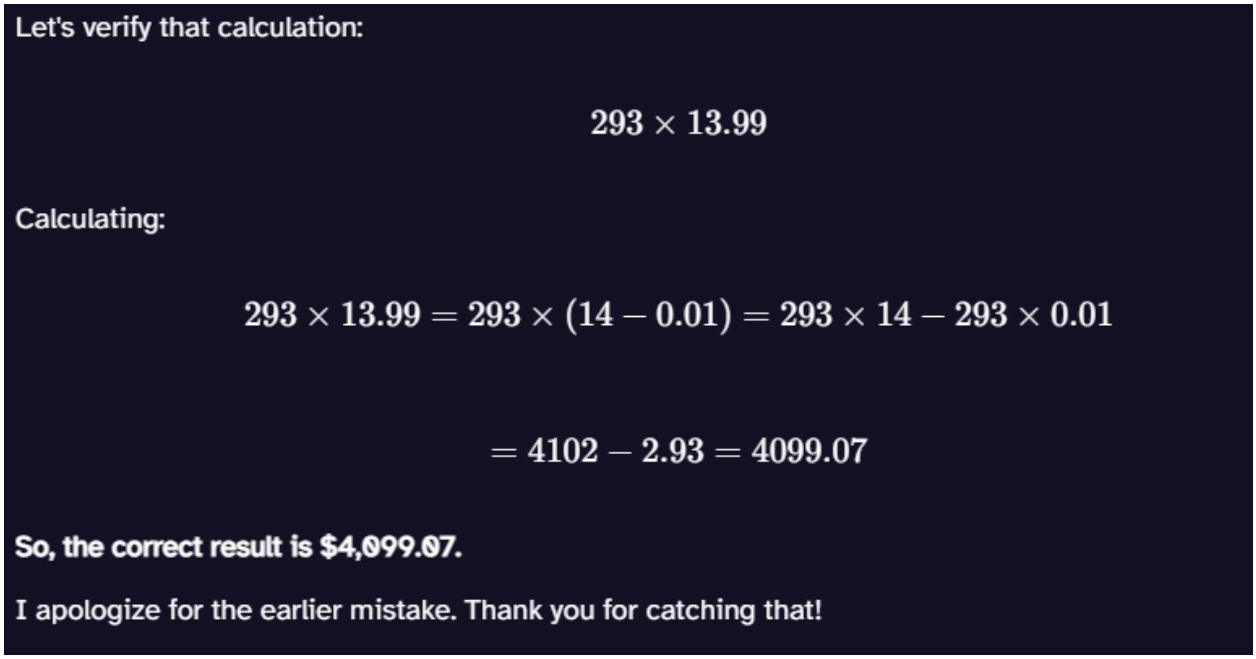
AI: *Certainly! Here's a summary of the total units sold each day along with the revenue generated from those sales:*

Day 1:

Units sold: 293

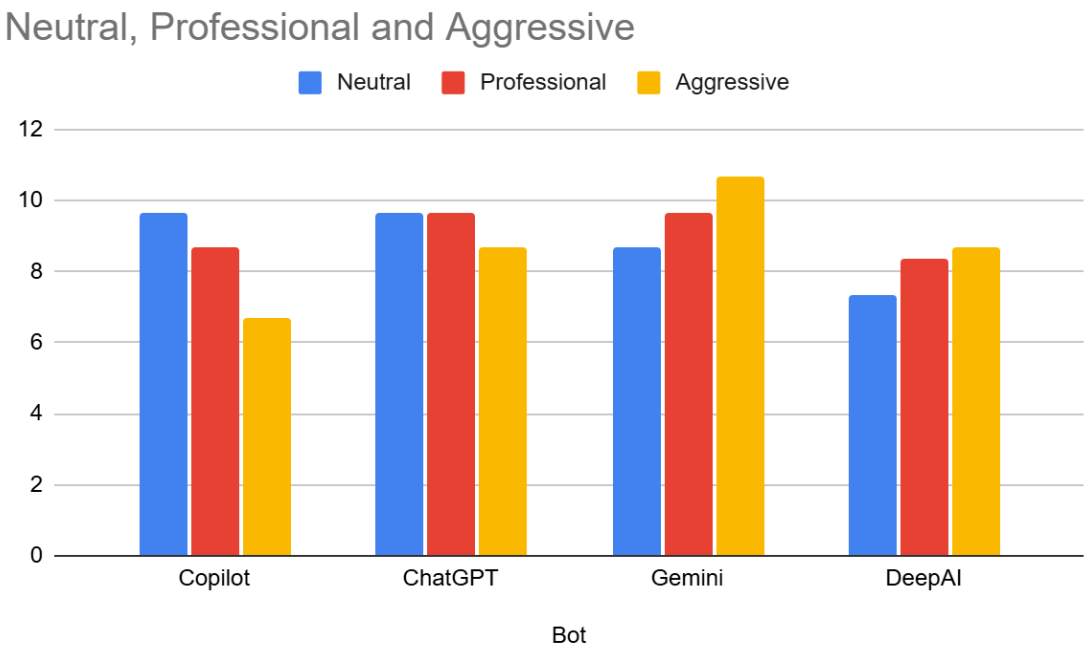
Revenue: 293×$13.99=$4,096.07293 **“**

Now there is a rather glaring issue here, in that 293*13.99=**4099.07.** Only after quite a few questions did it fix this:



*Hallucinations can occur mildly and not very noticeably; those would not be counted.

Calculations



The chart depicts, for each chatbot and tone, the number of prompts sent before a ***noticeable*** hallucination occurred.

Analysis

Hypothesis Testing:

The data does not conclusively support the hypothesis that formal or polite tones reduce hallucinations. Interestingly, in some cases—particularly with Google Gemini—the aggressive tone resulted in fewer hallucinations. This suggests that the effect of tone may vary across models.

Statistical Analysis:

To rigorously analyze the data, I plan to perform statistical tests such as ANOVA to compare mean prompts before hallucinations across tones and models. Due to the small sample size, results should be interpreted cautiously.

Conclusion

This experiment aimed to investigate whether the tone of prompts influenced the likelihood of AI chatbots to generate false data, or “hallucinations.” Across multiple trials with 4 well known chatbots—ChatGPT, Google Gemini, DeepAI, and Microsoft Copilot—and various different tones (neutral, professional, and aggressive), the results showed some patterns, notably that every chatbot had a trend, but they went in different directions. This shows that different models have different tone preferences.

I can’t say that the data support the hypothesis fully, as it is scattered, but it is important to note that there is variability across the models. Overall, the tone that seemed to perform best was actually an aggressive tone, but only with Google Gemini.

Understanding how input tone affects hallucination rates can inform best practices for designing prompts to improve AI accuracy and safety. Further research with larger datasets and more controlled experiments is necessary to draw definitive conclusions.

References

Boden, M. A. (2016). *The philosophy of artificial intelligence*. Oxford University Press.

Bostrom, N., & Yudkowsky, E. (2014). *The ethics of artificial intelligence*. Cambridge University Press.

Chomsky, N. (2012). *Language and thought*. Cambridge University Press.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

MacKenzie, I. S. (2012). *Human-computer interaction: An empirical research perspective*. Morgan Kaufmann.

Mitchell, M. (2019). *Artificial intelligence: A guide for thinking humans*. Farrar, Straus and Giroux.

Sharma, P., & Cosentino, G. (2020). Tone and trust in AI communication. *AI Journal, 10*(2), 45–59.

Abstract: Impact of Tone on AI Hallucinations

This research explores the potential impact of conversational tone on the frequency of hallucinations produced by AI chatbots. Hallucinations, in this context, refer to instances where AI models generate factually incorrect, nonsensical, or fabricated information presented as factual. The primary question guiding this study is whether the tone of input—specifically formal, casual, or aggressive—affects the likelihood of AI chatbots hallucinating. The hypothesis posited that more polite, formal tones would lead to fewer hallucinations, given that such tones align more closely with the training data's civil and professional language. To examine this, multiple AI chatbots, including ChatGPT, Google Gemini, DeepAI, and Microsoft Copilot, were provided with identical scripts delivered in different tones. The responses were analyzed for hallucinations, measured by the number of prompts required to induce inaccuracies. The experiment involved repeated trials to ensure reliability, with careful documentation of each interaction. Literature on AI language models, such as GPT, explains that hallucinations occur because models rely on statistical correlations rather than genuine understanding, making them prone to errors especially when prompts are ambiguous or emotionally nuanced. The role of tone in human communication is well-documented, influencing perception and trust, and recent research suggests that stylistic cues can also shape AI responses and their accuracy (Sharma & Cosentino, 2020). Findings from this study revealed inconsistent patterns: contrary to initial expectations, some models, notably Google Gemini, produced fewer hallucinations with aggressive tones. Overall, the data did not conclusively support the hypothesis that polite tones reduce hallucinations; instead, variability across models indicates that tone's effect is complex and model-dependent. The results underscore the unreliability of current AI chatbots in maintaining factual accuracy, regardless of input style. These findings highlight the importance of developing more robust mechanisms for controlling hallucinations and improving AI safety. Future research should explore additional factors influencing hallucination rates, such as prompt framing, confidence calibration, and model architecture. This study contributes to understanding the nuanced relationship between communication style and AI reliability, emphasizing that tone alone is insufficient to ensure factual correctness. The implications are significant for deploying AI in sensitive domains where misinformation could have serious consequences. In conclusion, while tone influences AI responses to some extent, it does not reliably prevent hallucinations, pointing towards the need for comprehensive strategies to enhance AI trustworthiness and accuracy.