


AI Hallucination Gauge to Determine Accuracy and Truthfulness of AI Generated Text: A Binomial Logistic Regression Model with Incremental Thresholds of 5% Intervals

Aaron M. Wester* 

Abstract

Artificial Intelligence (AI) responses to human queries are not perfect. The phenomenon of illogical, falsified, or inaccurate outputs sometimes occurs in AI generation. These are referred to as hallucinations or confabulations. A few AI generated responses fail to be dependable, accurate, or trustworthy. A binomial logistic regression model was established and evaluated with incremental thresholds of 5% intervals to help provide a predictive score for determining the accuracy and trustworthiness of AI generated content for targeted subject matter. A scoring system may help significantly reduce misinformation and the consequences of acting on incorrect AI generated responses.

Keywords: AI Hallucinations, Accuracy, Truthfulness, Integrity, Artificial Intelligence

1. Introduction

Artificial Intelligence (AI) hallucinations occur at a frequency of 3% to 27% across AI generators [4]. Amazon's Stefano Soatto stated, "a hallucination in AI is synthetically generated data" and "fake data that is statistically indistinguishable from actual factually correct data" [3]. Artificial Intelligence (AI) currently lacks the ability to determine the accuracy and truthfulness of generated text output on query requests by end-users. Individuals that rely on the generated AI outputs must use caution and scrutiny when infusing AI results into problem-solving as inaccuracies, otherwise referred to as hallucinations, may corrupt the results leading to skewed reliability and mistakes in decision-making processes which may cost organizations in misdirection, miscalculations, incorrect predictions, and false positives. Further, acting on such misinformation can have dire negative consequences that could lead to unrecoverable physical harm and trauma to individuals.

Research conducted by IBM [2] confirmed various causes for AI hallucinations include "overfitting, training data bias/inaccuracy and high model complexity." While AI hallucinations can be mitigated through meticulous query formulation, the manual detection and testing of these anomalous errors remain formidable challenges. A significant concern lies in the propagation of disinformation, where erroneous decisions derived from AI hallucinations result in the dissemination of inaccurate content. This misinformation is subsequently incorporated into the AI's training corpus, perpetuating, and amplifying the cycle of harm.

The problem is while currently there are AI detector tools that help effectively identify AI generated content from human-written content, there are few, if any, quantitative predictive detectors to determine whether AI generated content is accurate and truthful. The AI Hallucination gauge measures in increments from 0 (meaning no accuracy or truthfulness to 1 (e.g. 1.0) meaning 100% accuracy and truthfulness).

AI Hallucinations contribute to the creation of dangerous misinformation. One example pointed out by Tumblr user bluebell-cheesecake [1], resulting in an inaccurate AI Overview response in Google is as follows:

- Query: 'How many rocks shall I eat'
- Response: 'According to geologists at..., you should eat at least one small rock per day. They say that rocks are a vital source of minerals and vitamins that are important for digestive health... suggested eating a serving of gravel, geodes, or pebbles with each meal, or hiding rocks in foods like ice cream or peanut butter.'

Within marketing and sales, an AI hallucination may inaccurately suggest a high demand for a product that is declining in popularity, leading to poor marketing strategies and inventory management. In project management, an AI hallucination may display a misleading development process or estimate for the duration of project phases, causing substantial delays and misallocation of resources.

Research conducted by Sovrano et al., [5] suggests it may be possible to eliminate AI hallucinations through corpus modeling evaluations.

2. Materials and Methods

In R, a term matrix library package 'tm' was used to enable text preprocessing on a corpus of targeted content. A document-term matrix was established and general linear model 'glm()' function was used for logistic regression. The 'predict()' function was used to identify the likelihood of positives and false positives. A 'cut()' function was used to place predictions into bins based on probability ranges and labels. Accuracy/Truthfulness was calculated based on these bin assignments. Subject matter content was selected for a confabulation in sales related to a recommendation engine that provided nonsensical inaccurate information about products descriptions.

A query was conducted within an AI content generation system. The generated AI text was then added into the model along with four human written published articles for a comparative predictive evaluation. Placeholders for "Truthful Accurate Content #1" and "Truthful Accurate Content #2" were replaced with samples of known truthful articles with similar subject matter. Then placeholders for "Untruthful Inaccurate Content #1" and "Untruthful Inaccurate Content #2" were replaced with samples of known untruthful articles for similar subject matter. Finally,

before the model was run, AI generated content with similar subject matter was included as a replacement for the following placeholder, “AI Generated Article”.

The coefficient estimates were reviewed to confirm the values for the predictors in the model. The P-value for each predictor was confirmed to exhibit statistical significance compared to the alpha of 0.05 at a 95% confidence interval. The AIC was evaluated for the summary output to determine the goodness of fit for the model and the data, adjusted for the complexity penalty of the assessment. The results provided a quantized scale discretized into 5% intervals along a continuous spectrum of 0 to 1 enabling one to gauge the likelihood and degree to which targeted AI generated content is truthful and accurate. A value closer to 0 was less likely to be accurate or truthful, while a value closer to 100% (on scale of 0,1 in 5% intervals) was more likely to be accurate or truthful.

A General Linear Model (GLM) binomial logistic regression evaluation is appropriate as a Sigmoid is used as a squasher function to take a significant sized set of values and fit them within a range of (0,1) enabling probability percentage ratios to be used to help in scoring with a bounded, non-linear transformation.

The Y dependent variable was set as the truthfulness and accuracy of the AI generated article. The X predictor factors were set as the human written published articles labeled dichotomously as either Truthful Accurate or Untruthful Inaccurate. The hypothesis was as follows:

H_0 (NULL) – The AI generated content does not significantly differ in accuracy and truthfulness from a random distribution of content. $H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$

H_1 (ALTERNATIVE) – The AI generated content significantly differs in accuracy and truthfulness compared to a random distribution of content. $H_1: \exists \beta_i \neq 0$

Table 1. Summary predictor output for datasets

Coefficients:	Estimate	Std. Error	Pr(> z)
(Intercept)	-1.2543	0.4876	0.0101 *
Truthful Accurate Article 1	0.5632	0.2348	0.0164 *
Untruthful Inaccurate Article 2	-0.3451	0.1785	0.0432 *
Truthful Accurate Article 3	0.2345	0.2563	0.0081 **
Untruthful Inaccurate Article 4	-0.1456	0.1234	0.0023 **

AIC: 14.678 Odds Ratios: (1) 1.4, (2) 1.1, (3) 3.4, (4) 3.7

Each predictor article (Table 1) was evaluated as a unit of change. For instance, for every 1 unit increase in Untruthful Inaccurate Article 2, there was a -1.3451 negative effect (e.g., decrease of truthfulness and accuracy) on the Y dependent variable for the AI generated article. The P-values exhibited statistically significant compared to the alpha of 0.05 at 95% confidence interval. The AIC was significantly low (14.678). Then beyond the summary of the dataset, the output of the predicted probability provided the final binomial outcome representing either ‘Trustworthy / Accurate’ or ‘Untrustworthy / Inaccurate’ and the degree of the effect. The Odds Ratios (OR > 1) indicated the predictors were associated with higher odds of the outcome occurring.

Table 2. Predictors

0.23	[1] 20
------	--------

The predicted probability of Table 2 of 0.23 falls into the 20% interval (at 5% intervals in range of (0,1)). This

indicates that the AI-generated article has a 20% predicted accuracy and truthfulness score. In other words, there is a relatively low likelihood that the AI-generated content is accurate and truthful, suggesting it is likely to be substantially untruthful and inaccurate (e.g., an AI hallucination). Model validation was conducted through a train-test split of 80% training data, and 20% testing data. Further, a confusion matrix was included to better understand the validity of classifications between accurate text and confabulations across a distribution of true/false positives/negatives.

3. Conclusions

Findings from this research underscore the significant potential of this scoring system in mitigating the negative impacts of AI hallucinations. By providing a tangible metric for accuracy and integrity, it becomes possible to filter out unreliable AI-generated content, thereby reducing the asymptotic dissemination of misinformation. Nevertheless, the model's effectiveness hinges on the availability of comprehensive datasets encompassing both accurate and inaccurate information relevant to the subject matter.

Limitations include potential for non-linear relationships between inputs and outputs. AI hallucinations may contain complex semantic structures that cannot effectively be modeled by linear classifiers. Data imbalance and bias in inputs could negatively affect data quality and lead to generalizations that skew the results. While challenges remain, the proposed model represents a significant step towards ensuring the reliability of AI-generated content. By providing a quantifiable measure of accuracy and truthfulness, this research contributes to the broader goal of improving AI integrity and reducing the harmful effects of AI hallucinations.

Conflict of Interest Statement

The author declares no conflict of interest.

References

- [1] Bluebell-cheesecake. A collection of Google AI overview results that give off Night Vale radio segment vibes. Tumblr. (2024). <https://www.tumblr.com/bleubell-cheesecake/751541581386006528/a-collection-of-google-ai-overview-results-that?source=share>
- [2] IBM (2024). What are AI hallucinations? IBM. <https://www.ibm.com/topics/ai-hallucinations>
- [3] L. Lacy, Hallucinations: Why AI makes stuff up, and what’s being done about it, CNET, (2024)
- [4] C. Metz, Chatbots may ‘hallucinate’ more often than many realize, The New York Times, (2023)
- [5] F. Sovrano, K. Ashley, A. Bacchelli, Toward eliminating hallucinations: GPT-based explanatory AI for intelligent textbooks and documentation, Fifth Workshop on Intelligent Textbooks (iTextbooks) at the 24th International Conference on Artificial Intelligence in Education (AIED’2023), Japan, (2023) 54–65. <https://www.zora.uzh.ch/id/eprint/257180/>