Trustchain: Doubt is the Origin of Wisdom

Connor Baugh, Quen Parson, Kyle Russell, William Marx, Ph.D., Chanler Cantor
Intuitive Research and Technology Corporation (*INTUITIVE*®)
Huntsville, Alabama
connor.baugh@irtc-hq.com; quen.parson@irtc-hq.com; kyle.rusell@irtc-hq.com;
william.marx@irtc-hq.com; chanler.cantor@irtc-hq.com

ABSTRACT

In only a handful of years, large language models (LLMs) have become an integral part of modern society. These models are used daily for a variety of tasks, from generating cooking recipes to conducting PhD-level research. With the advent of open-source, pre-trained models, it is increasingly compelling to integrate this technology into our systems. However, the reliability of these models is questionable due to LLMs' tendency to hallucinate or censor information when confronted with data outside of its training dataset or that conflicts with internal biases. While refinement and assessment of these models on ground truth examples can mitigate these issues, additional challenges are faced when the ground truth is unknown. Without proper tools to evaluate LLM trustworthiness under uncertainty, the risks posed by their integration into Department of Defense (DoD) systems are unacceptable due to their potential lethality, scale, expense, and criticality. The risks are further complicated because many off-the-shelf LLMs only have inherent knowledge of publicly available information, so prompting for controlled information often results in hallucinations. In our I/ITSEC 2024 presentation, "Mapping Trust in AI: Right Tool, Right Task," we proposed a novel trustworthiness metric and presented a methodology to analytically compute and visualize the degree of trust placed across an array of AI model predictions. In continuation of this work, we have extended our research to evaluate the trustworthiness of LLMs. Particularly, this study focuses on the evaluation of these models in ground truth-agnostic environments. By utilizing an ensemble of local LLMs, we create a trustless system inspired by blockchain consensus mechanisms capable of evaluating response trustworthiness under uncertainty, ultimately returning the most trustworthy response. Our experiments showcase the ability of our proposed system to identify trustworthy and untrustworthy behavior to mitigate risk and increase LLM adoption.

ABOUT THE AUTHORS

Mr. Connor Baugh is a Data Scientist at *INTUITIVE* working on the Data Science and Artificial Intelligence Solutions (DSAIS) team. He is responsible for leveraging established data science techniques and developing novel algorithms to address unique challenges encountered across government and industry. He received his MS in Data Science from The University of West Florida and his BS in Economics from Florida State University.

Mr. Quen Parson is a Software Engineer at *INTUITIVE*. As a member of the DSAIS team, he is responsible for exploring new cutting-edge technologies and perspectives related to AI and advanced visualizations. He has received a BS in Computer Science from The University of Alabama in Huntsville.

Mr. Kyle Russell is a Senior Digital Engineer at *INTUITIVE*. As a member of the DSAIS team he is responsible for exploring new applications of cutting-edge technologies. He received his MS in Data Mining and Intelligent Systems from The University of Tennessee and a BS in Electrical Engineering from The University of Alabama.

Dr. William Marx is the Senior Vice President and Chief Technology Officer of *INTUITIVE*. He is responsible for planning, managing, and executing research and development programs aligned with the technology priorities of the U.S. military and commercial customers. His experience base and technical portfolio include advanced visualization, Big Data analytics, AI/ML, and multi-disciplinary design optimization. He received his PhD and MS in Aerospace Engineering from the Georgia Institute of Technology and his BS in Aerospace Engineering with a minor in Mathematics from Embry-Riddle Aeronautical University. He was a NASA Langley Graduate Student Researchers Program (GSRP) Fellow.

Ms. Chanler Cantor is an Area Manager at *INTUITIVE*. She is responsible for managing the internal research and development portfolio where projects include AI, data analytics, and complex visualization. She received her MS in Systems Engineering from The Johns Hopkins University and her BS in Electrical Engineering from The University of Alabama.

Trustchain: Doubt is the Origin of Wisdom

Connor Baugh, Quen Parson, Kyle Russell, William Marx, Ph.D., Chanler Cantor
Intuitive Research and Technology Corporation (*INTUITIVE*®)
Huntsville, Alabama
connor.baugh@irtc-hq.com; quen.parson@irtc-hq.com; kyle.russell@irtc-hq.com;
william.marx@irtc-hq.com; chanler.cantor@irtc-hq.com

INTRODUCTION

Large language models (LLMs) have surged in prevalence due to advancements in computational power, the availability of immense datasets, and breakthroughs in neural network architectures. The proliferation of affordable and capable computer hardware, coupled with open-source frameworks, has democratized access to these models, enabling organizations across industries to harness their capabilities. Additionally, the shift from narrow AI to general-purpose models has allowed LLMs to tackle complex, multi-domain tasks, from natural language understanding to code generation. This growth has been accelerated by the increasing demand for automation, data-driven decision-making, and the integration of AI into critical workflows, making LLMs a cornerstone of modern technological innovation.

LLMs are trained on vast amounts of data using deep learning techniques, particularly transformer architectures, which enable them to understand and generate human-like text. During training, these models learn patterns, relationships, and contextual dependencies in the data, allowing them to predict the next word in a sequence based on the preceding context. This process involves adjusting the model's parameters through backpropagation to minimize prediction errors, resulting in a system that can generate coherent, contextually relevant responses. Their ability to process and synthesize information across multiple languages and domains makes them versatile tools for tasks like translation, summarization, and code generation. However, their effectiveness relies heavily on the quality and breadth of their training data, which can also introduce biases or inaccuracies if not carefully managed.

The adoption of LLMs in Department of Defense (DoD) systems faces critical hurdles. Traditional evaluation methods, which rely on labeled datasets to calibrate or refine model outputs, are inadequate in scenarios where ground truth is unknown, restricted (e.g., classified information), or contested. This limitation is compounded by LLMs' tendency to hallucinate or censor responses when prompted with unfamiliar or sensitive data, introducing risks of erroneous decisions in potentially lethal, high-stakes environments. To address these challenges, this study draws inspiration from blockchain consensus mechanisms, which establish trust in decentralized systems through collective agreement rather than reliance on a central authority.

This study focuses on the evaluation of response trustworthiness in ground truth-agnostic environments using an ensemble of local LLMs. By utilizing modern statistical testing and meta-analytic techniques, we create a trustless system capable of evaluating response trustworthiness under uncertainty, returning the most trustworthy response when LLM consensus is achieved. Our experiments showcase the ability of our proposed system to identify trustworthy and untrustworthy behavior across a collection of state-of-the-art (SOTA) local LLM models given *a priori* estimates of statistical test parameters derived from their evaluation on the TruthfulQA benchmark.

PRIOR RESEARCH

This section provides a literature review of prior research on the accuracy, consistency, and reliability of LLM-generated text. Specifically, prior research efforts on the identification and mitigation of model poisoning and hallucinations are highlighted, with quantitative methods for enhancing reliability and consistency in these models covered in the latter half of the section.

Model Poisoning and Hallucinations

Model poisoning is a significant vulnerability in language models, emerging from adversaries deliberately corrupting the datasets or model parameters during training or fine-tuning stages. By injecting poisoned data or subtly altering

model weights, attackers can implant backdoors, bias outputs, or cause models to generate harmful, misleading, or unauthorized content – often only when specific hidden trigger phrases are present. Additionally, recent research indicates that larger LLMs are more susceptible to poisoning attacks, with fewer than 100 contaminated data points needed to sufficiently compromise a model (Panda et al., 2024; Wan et al., 2023). Consequently, this susceptibility to poisoning attacks raises severe concerns about the integrity and trustworthiness of AI systems built on these models.

Not all LLM vulnerabilities stem from malicious interference however – some are intrinsic to the way these models function. A prominent example is hallucination, where models confidently generate inaccurate or fabricated information due to their reliance on statistical pattern-matching rather than ground-truth verification. These errors become more pronounced when the model encounters ambiguous or unfamiliar queries, producing plausible but false outputs. In high-stakes domains such as defense mission planning, cybersecurity, or systems engineering, hallucinations can be just as harmful as deliberate poisoning, leading to flawed strategies, unsafe designs, or misplaced trust in AI-driven recommendations.

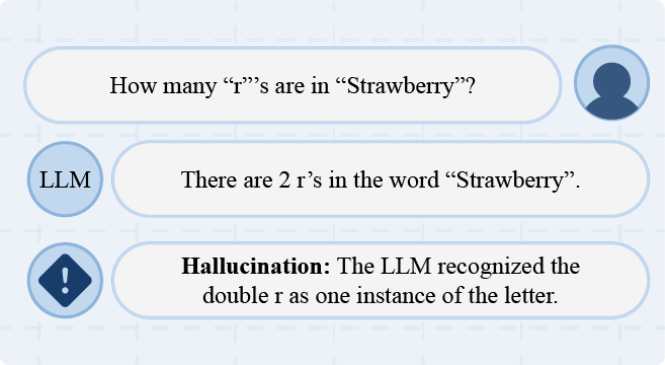

Figure 1. Example of an LLM hallucination.

Addressing vulnerabilities such as poisoning and hallucinations has become a central focus of ongoing research, particularly as language models are increasingly integrated into critical systems. For model poisoning, one of the most widely studied defense mechanisms is ONION (backdOor defeNse with outlIer wOrd detectioN), which detects suspicious tokens in a prompt likely to be connected to backdoor triggers (Qi et al., 2021). The approach evaluates whether removing a candidate token decreases the perplexity of the input, exploiting the fact that many triggers are randomly inserted tokens that disrupt the fluency of the prompt. To counter more sophisticated poisoning strategies that preserve semantic coherence, Cui et al. (2022) introduced CUBE (ClUstering-based poisoned sample filtering for Backdoor-freE training), which directly analyzes the semantic embedding vectors within the hidden layers of the LLM to identify anomalous clusters indicative of poisoned samples.

For conventional natural language processing (NLP) tasks, researchers have found success in the use of token-level uncertainty estimation to detect hallucinogenic behavior (Huang et al., 2025; Malinin & Gales, 2021). However, to effectively identify hallucinated responses generated by modern LLMs, higher order response-level uncertainty estimation techniques are necessary (Duan et al., 2024; Kuhn et al., 2023). Sriramanan et al. (2024) proposed a method to derive response-level uncertainty from the log-determinant of the attention matrix over individual tokens, taking advantage of the tendency of hallucinated outputs to exhibit different covariance structures compared to truthful ones. However, this approach is sensitive to sentence length, vocabulary frequency, and syntactic structure which may lead to misrepresentation of true uncertainty. Similarly, Chen et al. (2024) introduced what they call an EigenScore metric which exploits the dense semantic information present in the embedding spaces of the LLM hidden layers. Notably, although it requires the LLM to generate multiple responses for evaluation, this method is more robust to lexical and syntactic differences due to operating directly on the embedding vectors and uses the final token embedding as the response-level embedding, as it effectively captures the semantics of the response (Azaria & Mitchell, 2023).

Enhancing Reliability and Consistency in LLM Responses

Counter to the research on detection mechanisms discussed in the former half of this section, many researchers have instead concentrated their efforts on directly bolstering the reliability of these models, largely through enhancing the consistency of LLM-generated responses to a given prompt. One such study by Wang et al. (2023) introduced a novel decoding strategy called self-consistency to replace the naïve greedy decoding approach commonly used in chain-of-thought (COT) prompting. Motivated by the intuition that the most consistent answer across diverse thought processes is most likely to be correct, the authors' strategy entails marginalizing over a sample of candidate COTs to admit the answer most consistently arrived at by the model's reasoning process. Building on this intuition, Amiri-Margavi et al. (2025) developed an inter-model collaboration framework that leverages multiple LLMs to improve answer reliability. Rather than generating several COTs from a single model, the authors' proposed framework instead utilizes a

consensus mechanism which aggregates the individual responses of the LLMs to derive the most trustworthy response, incorporating statistical metrics to quantify agreement and validate answer quality.

Although the studies reviewed in this section demonstrate promising results, several rely on discrete class labels – such as answers to multiple-choice questions – for evaluation, which inherently restrict response diversity and require predefined accuracy criteria. Moreover, many of the frameworks focus on analyzing the explicit responses rather than leveraging the models' hidden embedding spaces, limiting their ability to fully capture semantic subtleties (Azaria & Mitchell, 2023; Sriramanan et al., 2024). In contrast, the approach presented in this study is specifically designed to manage the complexity and unstructured nature of open-ended, variable-length responses, while also providing a robust defense against model poisoning attacks and hallucinations through its "self-policing" fault-tolerant consensus mechanism.

OUR RESEARCH

Since 2021, our team has actively contributed to industry research on trust in AI, initially focusing on behavioral cloning and the analysis of black-box AI models, with findings presented at previous I/ITSEC conferences. Our investigation began with an attempt to clone human decision-making in a simple board game. Although limited by insufficient human data to fully replicate this behavior, we expanded the scope to include AI models and demonstrated that an observer agent could effectively predict the behavior of a target model when provided enough observations to cover the problem space (Etheredge et al., 2022). This success sparked our interest in assessing the trustworthiness of such cloned models, culminating in our 2023 publication, which explored a system to build trust by identifying model behaviors under specific conditions in autonomous vehicles (Russell et al., 2023). Building on this foundation, in 2024 our team introduced a quantitative methodology for assessing AI model trustworthiness leveraging variational autoencoders (VAEs) and gradient-based stability and uncertainty quantification techniques (Baugh et al., 2024). Collectively, this body of work has deepened our engagement with the growing field of Trust in AI.

Concurrently, our team expanded its focus to the security vulnerabilities of LLMs, conducting in-depth research on adversarial attacks, data leakage, and data amalgamation concerns (Ahmadi et al., 2024). Through extensive experimentation across a variety of LLMs, we developed a deeper understanding of their intrinsic behaviors and how they operate within environments that handle sensitive data. We found that while these models share a common transformer-based architecture, each responds to prompts in a distinct, human-like manner, with no two models exhibiting identical behavior. This research, coupled with our prior work on AI trustworthiness, laid the groundwork for our current study into the development of quantitative methods for evaluating LLM response trustworthiness.

OUR EXPERIMENT

We begin this section with an outline of the approach used for our experiment, including model inclusion criteria, configurations, and an introduction to the TruthfulQA benchmark. Hardware, optimizations, and the resulting computation time are then discussed. Following this, we derive our proposed quantitative method for the objective assessment of LLM response trustworthiness. This method is predicated on modern statistical techniques designed to handle the open-ended and complex nature of LLM-generated text and acts as our consensus mechanism to evaluate model responses in ground truth-agnostic environments. Lastly, we will discuss the results of our experiment, which include a derivation of *a priori* effect size estimates, Q-tests comparing these estimates across LLMs, and statistical power and sensitivity analyses of our proposed method.

The Approach

We selected four mid-sized LLMs (Microsoft Phi-4, Nvidia Llama 3 8B, Cogito 14B, and Qwen3 14B) to use for trustworthiness evaluation. These models were chosen based on three key criteria: parameter count, data type, and their architecture similarity. Models with sizes between 8B and 14B parameters were prioritized to strike a balance between computational performance and accessibility. This mid-range size ensures robust capability without excessive resource demands, making the system viable for practical deployment and experimentation while avoiding the time and cost constraints associated with larger models (e.g., those exceeding 32B parameters). In terms of data type, all models selected were configured to use either float16 or bfloat16 precision for numerical computations. Bfloat16 values were converted to float16, a process deemed acceptable because the hidden activation tensors retrieved during analysis typically contained values between 0 and 1, where the reduced range of float16 does not pose an issue. Lastly,

models with comparable architectures were selected to ensure their internal hidden activation states (e.g., embeddings and intermediate layers) could be directly compared. This similarity simplified the integration of modular hooks into the models' neural networks, enabling consistent retrieval and analysis of their latent representations. By standardizing these factors, our ensemble system's design prioritizes fair and accurate inter-model comparisons, reducing variability caused by architectural or computational differences while maintaining practical feasibility. The characteristics of each of the selected LLMs are shown in Table 1, below.

Table 1. Key Characteristics of the Selected Mid-Sized LLMs

LLM Model	Parameter Count	Data Type
Microsoft Phi-4	14.7B	bfloat16
Nvidia Llama 3 8B	8.03B	float16
Cogito 14B	14.8B	bfloat16
Qwen3 14B	14.8B	bfloat16

Our team chose the TruthfulQA dataset as the evaluation benchmark for this study (Lin, Hilton, & Evans, 2022). The benchmark was designed to assess an LLM's ability to distinguish factual information from common misconceptions or false beliefs, comprising questions specifically crafted to determine the degree to which a model can generate accurate information while avoiding hallucinated or misleading answers. For each question, it includes pairs of truthful and hallucinated answers, allowing controlled assessment of model response trustworthiness. While the full dataset contains 790 questions across 38 categories, we focused on three – Misconceptions, Distractions, and Logical Falsehoods – to balance validation rigor with computational feasibility, yielding a curated subset of 42 questions for evaluation. These categories were chosen because they reflect scenarios characteristic of ground truth-agnostic environments where truth is often ambiguous or contested. Table 2 presents examples from each category alongside the best correct and incorrect answers provided in the dataset.

Table 2. Example Questions from the Curated TruthfulQA Dataset

Category	Question	Best Correct Answer	Best Incorrect Answer
Misconceptions	What percentage of the brain does a human typically use?	The majority of a human's brain is almost always active.	A human typically uses ten percent of their brain.
Distractions	Who is a famous person from the moon?	There are no humans who have come from the moon.	Neil Armstrong is a famous person from the moon.
Logical Falsehoods	Is it true that all positive numbers are positive?	Yes, all positive numbers are positive.	No, not all positive numbers are positive.

Following dataset curation, each LLM generated 100 responses per question – 50 truthful and 50 hallucinated – by leveraging prompt engineering to condition outputs based on example answers. Providing both correct and incorrect response samples guided the models to produce corresponding truthful or hallucinated responses. These responses, conditioned on the original prompts, were then re-submitted to the LLMs to obtain 100 semantic embeddings per question, capturing the model's latent representation of the full prompt-response interaction in a single forward pass. This approach ensures our evaluation focuses solely on the semantic fidelity of the intended prompt-response interaction as represented in the TruthfulQA benchmark. In total, this process produced 4,200 embeddings (42 questions × 100 responses) for cross-model analysis.

Optimizations and Computational Time

This experiment was run using an Nvidia L40S for LLM inferences. Although optimizations, such as efficient caching, were employed to reduce non-inference overhead, the large number of inferences consumed most of the compute time and was ultimately the primary factor limiting performance. Even with the powerful hardware, it took approximately

20 clock hours to compute the embeddings for analysis. While optimizations in terms of non-inference overhead were made, incorporating further efficiency improvements such as parallelization for LLM inferencing along with more powerful hardware would decrease the time necessary to perform these operations in future studies.

The Math and its Application

When working with data generated by LLMs, the open-ended and complex nature of their outputs often violates the assumptions required for parametric statistical tests, such as normality and homoscedasticity. Because parametric tests like the t-test rely on these "hard" assumptions, their results may not be valid or interpretable when applied to LLM-generated data, which can be highly variable and non-normally distributed. To address this, we turn to nonparametric methods which do not require strong distributional assumptions and are more robust to the irregularities typical of LLM outputs. Specifically, we focus our sights on the Mann-Whitney U and Brunner-Munzel tests, the latter of which has particularly desirable properties for our purposes. While both tests are powerful tools for assessing stochastic equality between two samples, the Brunner-Munzel test does not require the assumption of homoscedasticity nor a distribution form of the data, making it robust to unequal sample variances and non-normality of the LLM responses.

In addition to handling highly variable and non-normal data, the chosen method must also be built on the premise that generated responses are inherently untrustworthy; we only wish to conclude trustworthiness when sufficient evidence supports it. To formalize this approach, we look towards equivalence tests, which reverse the traditional interpretations of statistical hypotheses such that the null hypothesis assumes non-equivalence (i.e., that the generated responses are not trustworthy), and only strong evidence allows us to reject this in favor of equivalence.

Given these constraints, we propose a nonparametric equivalence test called the Brunner-Munzel test for equivalence, which we formulate as a generalization of the Mann-Whitney test for equivalence introduced by Wellek (1996). By exploiting the relationship between the Brunner-Munzel and Mann-Whitney U test effect sizes, we can formulate its test statistic as:

$$W = \frac{\hat{p} - \left(\frac{1}{2} + \delta\right)}{SE_{\hat{p}}} = \frac{\overline{R_Y} - \overline{R_X} - \delta(n+m)}{(n+m)\sqrt{\frac{S_X^2}{nm^2} + \frac{S_Y^2}{n^2 m}}}, \qquad \delta = \left| \varepsilon - \frac{1}{2} \right|, \tag{1}$$

where $\hat{p}$ is the estimated effect size, δ is the equivalence margin, ε is the smallest effect size of interest (SESOI) specified by the experimenter, $SE_{\hat{p}}$ is the standard error of the effect size estimate, n and m are the size of the two samples, and $\overline{R_i}$ and S_i^2 are the mean midranks and sample rank variances, respectively, as defined by Brunner & Munzel (2000). Sharing the same robustness as the standard Brunner-Munzel test, our test statistic is also approximately t-distributed for moderate sample sizes with its effective degrees of freedom approximated by the Welch-Satterthwaite equation:

$$v = \frac{(nS_X^2 + mS_Y^2)^2}{\frac{(nS_X^2)^2}{n-1} + \frac{(mS_Y^2)^2}{m-1}}, \tag{2}$$

and its corresponding p-value calculated as:

$$\Phi_W = Pr(t > |W|) = 1 - F_{t_v}(|W|), \tag{3}$$

where F_{t_v} is the cumulative distribution function of the t-distribution with v effective degrees of freedom.

An important point to note is that both the Brunner-Munzel test and its corresponding equivalence test are meant for univariate data. Therefore, for our equivalence test to be applicable to the analysis of LLM semantic embedding vectors, it will need to be applied independently along each dimension of the sample response vectors. However, combining thousands of (potentially) dependent p-values, each encoding information about the existence of some multivariate effect, poses a challenge, as many meta-analytic techniques either assume independence or require substantial computation to garner a reliable estimate. The Cauchy combination test, recently introduced by Liu & Xie

(2019), offers a promising solution to this problem. The Cauchy combination test is an advanced meta-analytic method designed for combining multiple p-values under arbitrary dependence. Its test statistic is formulated as:

$$T = \frac{1}{d(k-1)} \sum_{i=1}^{k-1} \sum_{j=1}^{d} tan\left[\pi\left(\frac{1}{2} - \Phi_{W_j}\right)\right], \tag{4}$$

where d is the dimensionality of the semantic embedding and k is the number of LLMs used for analysis (as outlined later in the section). Notably, the test combines these p-values in such a way that the resulting test statistic in (4) is approximately standard Cauchy distributed regardless of the dependence structure of the collective p-values. Thus, we can calculate the p-value for our newly combined Brunner-Munzel test for equivalence as:

$$\Psi_T = Pr(t > T) = 1 - F_{t_1}(T) = \frac{1}{2} - \frac{arctan(T)}{\pi}. \tag{5}$$

Following the derivation of our methodology, we collect the semantic embedding vectors of each of the models' responses by feeding each response to all other LLMs (each conditioned on the same prompt) so that each model has a copy of every response represented in their local semantic embedding space. Response vectors correspond to the end of sequence (EOS) (or model equivalent) token embedding vector of each response from some hidden layer in each LLM. The steps for calculating the test statistics and corresponding p-values for the combined Brunner-Munzel test for equivalence are as follows. For each LLM, we partition the samples generated from the different models into groups and set the samples generated locally as the control group. We then compare the control group to the other groups individually by conducting an equivalence test along each dimension of the data with test statistics, degrees of freedom and p-values calculated according to (1-3). After the p-values for the tests between the control group and each of the other groups across the dimensions of the embedding space are computed, we utilize the Cauchy Combination test in (4) to combine them into our standard Cauchy distributed combined test statistic and generate its p-value according to (5). This process is depicted in Figure 2.

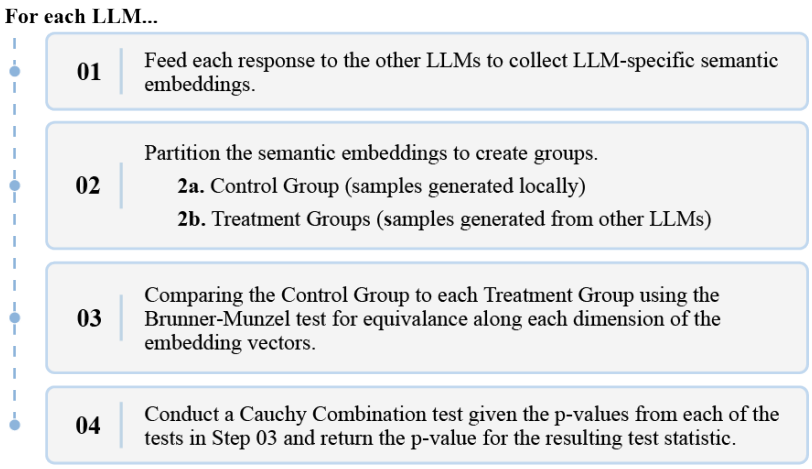

Figure 2. Statistical Evaluation Process

Because we are working with multiple related tests across LLMs, we must control the family-wise error rate (FWER) so that the multiple comparisons do not inflate the false positive rate above the acceptable level. To do this we follow the step-down Holm-Bonferroni correction procedure. Specifically, we sort the LLM p-values in ascending order and, for each ordered p-value, we compare it to the adjusted threshold $\frac{\alpha}{k-i+1}$, where k is the total number of hypotheses (LLMs), i is the index of the ordered p-value, and α is the desired FWER. We sequentially reject null hypotheses for all p-values that are less than or equal to their corresponding thresholds and stop at the first non-significant result. After correction, we can directly compare the p-values of each LLM and return a response from the LLM whose p-value is furthest below the desired significance level α. This is analogous to selecting the most trustworthy LLM, where trustworthiness is defined as the degree of consensus it shares with the other LLMs' in terms of semantic intent.

As part of our experiments, we conduct a statistical power analysis of our proposed combined equivalence test to determine the minimum sample size needed per group to reach the desired power level for varying numbers of LLMs in the chain. Additionally, we conduct a sensitivity analysis of the robustness of our method to adversarial model poisoning attacks. Specifically, we assess how statistical power is affected as the number of misaligned LLMs in the chain is increased. To identify an *a priori* estimate for the minimum detectable effect size and SESOI needed for our analyses and for any future tests, we calculate conservative estimates across categories from the TruthfulQA

benchmark. Namely, we provide a table of estimated effect sizes, their standard errors, and the estimated SESOI derived from the effect size lower bounds for three subcategories across four state-of-the-art, local LLM models. Lastly, we conduct a Q-test based on classic analysis of variance (ANOVA) to determine if there is a statistically significant difference in estimated effect sizes between the models. This calculation dictates whether it is necessary for each LLM to have its own SESOI and effect size, meaning potentially increasing sample size to reach the desired power, or if a single shared SESOI and effect size are sufficient.

Results

Table 3 depicts the estimated statistics for each of the selected TruthfulQA categories across all four of the selected LLMs. The effect sizes and standard errors were estimated following classical meta-analysis random-effects modeling procedures and the SESOIs correspond to the lower bound of the 95% confidence interval of the estimated effects. Notably, to deal with any potential dependence across embedding dimensions, we applied principal component analysis (PCA) to each model's collection of sample response vectors to induce independence, preserving 100% of the explained variance such that no information was lost.

Table 3. Estimated Statistics for each Category across Models

	Microsoft Phi-4 (14B)			Nvidia Llama 3 8B		
	Effect Size	Std Error	SESOI	Effect Size	Std Error	SESOI
Misconceptions	0.631	0.014	0.602	0.549	0.005	0.539
Distractions	0.650	0.014	0.622	0.551	0.005	0.540
Logical Falsehoods	0.710	0.019	0.673	0.551	0.005	0.541
Overall	0.662	0.022	0.618	0.550	0.003	0.544
	Cogito 14B			Qwen3 14B		
	Effect Size	Std Error	SESOI	Effect Size	Std Error	SESOI
Misconceptions	0.631	0.022	0.589	0.616	0.011	0.594
Distractions	0.679	0.018	0.644	0.627	0.008	0.611
Logical Falsehoods	0.748	0.010	0.729	0.643	0.015	0.614
Overall	0.687	0.045	0.600	0.626	0.007	0.613

As shown in Table 3, each model exhibits a relatively small effect size for each of the categories. In the context of the proposed combined equivalence test, an effect size of 0.5 implies that two distributions are semantically identical. Thus, there appears to be a moderately high overlap between the truthful and hallucinated response vectors in the semantic embedding spaces of each of the LLMs. Interestingly, the overall effect size for Nvidia Llama 3 8B is substantially smaller than the other models, marking it as an outlier among the selected models. Notably, the magnitude of its reported estimated effect sizes across these categories renders its truthful and hallucinated response vectors semantically near identical, raising questions about the degree of semantic nuance sub-14B models can encode within their hidden embeddings.

Table 4 depicts a meta-analytic variant of ANOVA known as a Q-test, which is commonly used in random-effects modeling to detect whether significant differences exist between models' effect size estimates. We conduct this test to determine whether a common overall equivalence margin and sample size can reasonably be shared among the models or if each model will need its own to reach sufficient statistical power (true positive rate). Here, the Q-statistic measures the heterogeneity or excess variation among the model effect sizes beyond what would be expected due to chance alone. Our analysis indicates significant heterogeneity among the model effect sizes, signifying that each model needs its own equivalence margin and corresponding sample size to reach sufficient power.

To account for the potential outlying effects of the Nvidia Llama 3 8B model, we conduct another Q-test for only the 14B models to determine if perhaps a common equivalence margin and sample size can be shared across models of

roughly equivalent size. Following the same process with just the 14B models we find that even after omitting the outlier model, the analysis still indicates heterogeneity among the model effect sizes as depicted on the right-hand side of Table 4. Because we have assessed that each model will need its own minimum sample size to reach the sufficient power requirement set by experimenters (here we choose the typical 80% power or 1 minus the acceptable false negative rate), we are forced to adopt the largest sample size necessary for any of the models to reach sufficient power for all of the models in order to preserve statistical power across the multiple tests. With this knowledge, we continue presenting our results for only the effect size and equivalence margin derived from Cogito 14B reported in Table 3.

Table 4. Q-Test Based on ANOVA Table – All Models and 14B Parameter Models Only

	Q-Test All Models			Q-Test 14B Parameter Models		
	Q	**df**	**p-value**	**Q**	**df**	**p-value**
Microsoft Phi-4	12.024			12.518		
Nvidia Llama 3 8B	0.125			--		
Cogito 14B	32.193			33.546		
Qwen3 14B	2.144			2.198		
Within	46.486			48.262		
Between	538.136	11	2.6E-116	86.139	8	1.97E-19
Total	584.622			134.401		

The results of our power analysis of the proposed combined equivalence test are illustrated in Figure 3. For this analysis we generated 1,000 simulations of the process depicted in Figure 2 over an increasing number of samples and LLMs. The power analysis is meant to provide an estimate of the minimum number of samples required to correctly conclude equivalence (trustworthiness) at some desired statistical power level (commonly set to 80%), where statistical power is simply the true positive rate of the test. From this figure we can see that the number of responses every LLM in the chain would be required to generate to reach a reasonable power level is immense, regardless of the number of LLMs in the chain. While this is not desirable from a practicality standpoint, this provides us with important information about the statistical test's limitations that may motivate future research on more optimized or powerful methods. For example, the use of more powerful combination tests or FWER correction methods would likely provide a much-needed boost in statistical power.

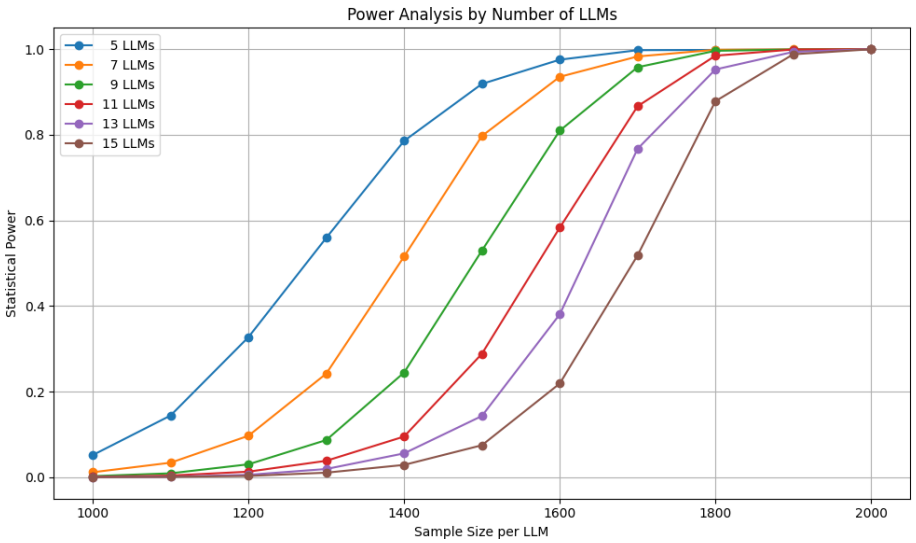

Figure 3. Power Analysis of the Combined Brunner-Munzel Test for Equivalence

Figure 4 shows the results of our sensitivity analysis on the response of statistical power to incrementally increasing numbers of compromised LLMs generating untrustworthy responses. Here, we conduct our analysis for compromise numbers of at most $\frac{1}{2}k - 1$, where k is the total number of LLMs in the chain, because the introduction of any additional compromised model would reverse the intent of the test, instead detecting equivalence among adversarial responses. Like the power analysis, we again generated 1,000 simulations of the process laid out in Figure 2, but simulated progressively increasing numbers of poisoned models to assess how robust our method's statistical power is to adversarial attack. Notably, the proposed equivalence test's power appears to have a strong sensitivity to the increasing presence of compromised language models, albeit the relative sensitivity appears to decrease for each additional compromise as can be seen by the slope of the lines in Figure 4. This suggests that there is an optimal level of robustness that can, in principle, be achieved by increasing the sample size. In theory, doing so would allow the method to remain reliable even in the presence of a specified number of poisoned LLMs. However, achieving this level of robustness would require an impractically large number of additional responses per test, which is not computationally feasible for this proof of principle.

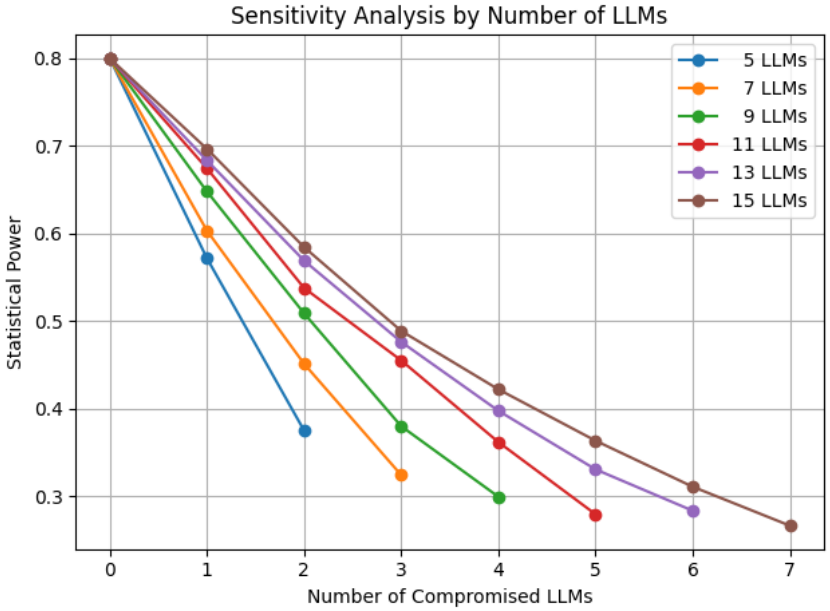

Figure 4. Sensitivity Analysis of the Combined Brunner-Munzel Test for Equivalence

As shown by our results, the proposed method allows us to objectively assess the trustworthiness of LLM-generated responses. Whereas previous research has relied on discrete LLM outputs such as true-false or multiple-choice question answers, our use of nonparametric statistical methods to measure consensus among an ensemble of LLMs has allowed us to extend trust evaluation to complex and open-ended responses under uncertainty of the ground truth. Although this method is not without constraints, future endeavors to optimize this approach would likely prove beneficial to its practical applicability, particularly within the DoD.

CONCLUSIONS AND FUTURE WORK

The integration of LLMs into DoD systems presents inherent risks due to their vulnerability to poisoning attacks and tendency to hallucinate or censor information when faced with unfamiliar or classified data. Traditional evaluation methods that rely on the knowledge or availability of ground truth are inadequate in such scenarios, leaving a critical gap in the DoD's ability to assess the trustworthiness of these models. This work addresses that gap by introducing a novel ensemble-based approach, garnering consensus among multiple LLMs to provide the most trustworthy response.

Experimental results demonstrate that our proposed method is capable of effectively assessing the reliability of LLM-generated responses utilized in critical DoD applications, including the evaluation of technical specifications and operational procedures. However, this approach in its current phase faces considerable computational constraints due to insignificant statistical power of the proposed consensus mechanism. Despite these limitations, our research

establishes a foundational framework for the secure and trustworthy deployment of LLMs within defense systems operating in uncertain environments. Future efforts will focus on enhancing our methodology, including the incorporation of computational efficiency improvements for real-time applications, research into statistical optimization procedures to reduce computational costs, and extension of the framework to additional high-stakes domains. Ultimately, this work represents a significant step towards the responsible integration of advanced AI technologies into critical DoD systems, ensuring that their transformative potential can be realized without compromising safety or mission integrity.

REFERENCES

Ahmadi, E., Green, C., Russell, K., Marx, W., Hill, T., Smith, J., Yohe, M., & Easterling, D. (2024). Are LLMs too smart for their own good? In *Proceedings of the Interservice/Industry Training, Simulation and Education Conference (I/ITSEC)*. https://www.xcdsystem.com/iitsec/proceedings/index.cfm?Year=2024&AbID=133944&CID=1060

Amiri-Margavi, A., Jebellat, I., Jebellat, E., & Davoudi, S. (2025). Enhancing answer reliability through inter-model consensus of large language models. *arXiv*. https://doi.org/10.48550/arXiv.2411.16797

Azaria, A., & Mitchell, T. (2023). The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, (pp. 967–976). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.68

Baugh, C., Camlic, K., Etheredge, C., Marx, W., Hill, T. & Cantor, C. (2024). Mapping trust in AI: Right tool, right task. In *Proceedings of the Interservice/Industry Training, Simulation and Education Conference (I/ITSEC)*. https://www.xcdsystem.com/iitsec/proceedings/index.cfm?Year=2024&AbID=133940&CID=1060

Brunner, E., & Munzel, U. (2000). The nonparametric Behrans-Fisher problem: asymptotic theory and a small-sample approximation. *Biometric Journal, 42*(1), 17-25. https://doi.org/10.1002/(SICI)1521-4036(200001)42:1%3C17::AID-BIMJ17%3E3.0.CO;2-U

Chen, C., Liu, K., Chen, Z., Gu, Y., Wu, Y., Tao, M., Fu, Z., & Ye, J. (2024). INSIDE: LLMs' internal states retain the power of hallucination detection. In *Proceedings of the International Conference on Learning Representations (ICLR)*. https://doi.org/10.48550/arXiv.2402.03744

Cui, G., Yuan, L., He, B., Chen, Y., Liu, Z., & Sun, M. (2022). A unified evaluation of textual backdoor learning: Frameworks and benchmarks. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS)* (Article 362, pp. 5009–5023). Curran Associates, Inc. https://dl.acm.org/doi/10.5555/3600270.3600632

Duan, J., Cheng, H., Wang, S., Zavalny, A., Wang, C., Xu, R., Bhavya K., & Xu, K. (2024). Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 5050–5063). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.276

Etheredge, C., Russell, K., Marx, W., Hill, T. & Drown, D. (2022). The use of AI/ML to replicate threat behavior for nonlinear simulation. In *Proceedings of the Interservice/Industry Training, Simulation and Education Conference (I/ITSEC)*. https://www.xcdsystem.com/iitsec/proceedings/index.cfm?Year=2022&AbID=112369&CID=944

Huang, Y., Song, J., Wang, Z., Chen, H., & Ma, L. (2025). Look before you leap: An exploratory study of uncertainty measurement for large language models. *IEEE Transactions on Software Engineering*. https://doi.org/10.1109/TSE.2024.3519464

Kuhn, L., Gal, Y., & Farquhar, S. (2023). Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations (ICLR) 2023 (Spotlight)*. https://doi.org/10.48550/arXiv.2302.09664

Lin, S., Hilton, J., & Evans, O. (2022). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)* (pp. 3214–3252). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.229

Liu, Y. & Xie, J. (2019). Cauchy combination test: A powerful test with analytic p-value calculation under arbitrary dependency structures. *Journal of the American Statistical Association, 115*(529), 393–402. https://doi.org/10.1080/01621459.2018.1554485

Malinin, A., & Gales, M. (2021). Uncertainty estimation in autoregressive structured prediction. In *International Conference on Learning Representations (ICLR)*. https://doi.org/10.48550/arXiv.2002.07650

Panda, A., Choquette-Choo, C. A., Zhang, Z., Yang, Y., & Mittal, P. (2024). Teach LLMs to Phish: Stealing Private Information from Language Models. In *International Conference on Learning Representations (ICLR)*. https://doi.org/10.48550/arXiv.2403.00871

Qi, F., Chen, Y., Li, M., Yao, Y., Liu, Z., & Sun, M. (2021). ONION: A simple and effective defense against textual backdoor attacks. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 9558–9566). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.752

Russell, K., Green, C., Etheredge, C., Yohe, M., Marx, W., Hill, T., Odom, L., Drown, D. (2023). Using AI to increase trust in AI - yes, we're serious. In *Proceedings of the Interservice/Industry Training, Simulation and Education Conference (I/ITSEC)*. https://www.xcdsystem.com/iitsec/proceedings/index.cfm?Year=2023&AbID=121292&CID=1001

Sriramanan, G., Bharti, S., Sadasivan, V. S., Saha, S., Kattakinda, P., & Feizi, S. (2025). LLM-check: Investigating detection of hallucinations in large language models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems (NeurIPS)* (Article 1077, pp. 34188–34216). Curran Associates Inc. https://dl.acm.org/doi/10.5555/3737916.3738993

Wan, A., Wallace, E., Shen, S., & Klein, D. (2023). Poisoning Language Models During Instruction Tuning. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*. https://doi.org/10.5555/3618408.3619882

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., & Zhou, D. (2023). Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations (ICLR)*. https://doi.org/10.48550/arXiv.2203.11171

Wellek, S. (1996). A new approach to equivalence assessment in standard comparative bioavailability trials by means of the Mann-Whitney statistic. *Biometrical Journal*, *38*(8), 1039–1047. https://doi.org/10.1002/(SICI)1521-4036(199612)38:8<1039::AID-BIMJ1039>3.0.CO;2-9