Reproducibility Study for The Internal State of an LLM Knows When It's Lying

Srikitha Kandra
George Mason University
Fairfax, VA, USA
skandra@gmu.edu

Venkata Tejaswi Kalla
George Mason University
Fairfax, VA, USA
vkalla@gmu.edu

1 Introduction

Large Language Models (LLMs) have revolutionized the field of natural language processing, demonstrating remarkable capabilities in text generation, comprehension, and reasoning. Built on deep learning architectures, particularly the transformer model, LLMs have evolved significantly in scale and capability, with models such as GPT-4, BERT, and LLaMA leading the field. Trained on vast datasets comprising diverse textual content, LLMs are capable of understanding complex language nuances, producing coherent responses, and even performing tasks requiring reasoning and creativity.

The widespread adoption of LLMs has had a profound impact on various aspects of society. They are extensively used in communication platforms, powering chatbots and virtual assistants like ChatGPT, Siri, and Alexa to enhance customer service, education, and accessibility. In content creation, LLMs assist with generating articles, coding, and creative writing, revolutionizing industries such as journalism, marketing, and software development. Additionally, their capabilities in language translation and summarization are breaking language barriers and improving access to information globally.

1.1 Task / Research Question Description

The integration of LLMs into society also raises ethical concerns, including the potential for misuse in spreading misinformation, generating biased content, and exacerbating privacy issues. Addressing these challenges requires careful consideration of LLM design, deployment, and regulation to ensure responsible usage.

This report contextualizes the role of LLMs in society while focusing on a critical question: Can an LLM recognize when it generates false information? Building on the paper ”The Internal State of an LLM Knows When It's Lying” by Amos Azaria and Tom Mitchell, the report explores the possibility of detecting untruthful outputs through the model's internal representations, a study with significant implications for trust and accountability in AI systems.

1.2 Motivation & Limitations of existing work

The rise of Large Language Models (LLMs) has prompted significant research into their capabilities, particularly in understanding their behavior and reliability. While these models excel at generating fluent and contextually appropriate text, questions about their internal understanding of truthfulness remain largely unexplored. The motivation behind this study stems from the need to address a fundamental issue: can LLMs inherently ”know” when they are generating false information? Understanding this can not only improve trust in AI systems but also enable the development of mechanisms to detect and mitigate misinformation.

Several prior works have investigated related areas, including LLM interpretability, bias detection, and error correction. However, these studies seldom address whether the internal states of an LLM encode a sense of truthfulness about its own generated outputs. What distinguishes the work of Amos Azaria and Tom Mitchell is their direct exploration of whether an LLM's internal states contain sufficient information to identify the truthfulness of its outputs. Unlike prior research that relies on external tools or metadata, this study uses the LLM's own activations as input for truth-detection classifiers. This approach shifts the focus from post-hoc evaluations of output correctness to real-time assessment during text generation.

Early studies often rely on smaller models or datasets, limiting their generalizability to modern, large-scale LLMs. Additionally, they tend to focus on specific types of falsehoods, such as factual inaccuracies, without accounting for broader contexts like deliberate deception or ambiguity in prompts. The challenge of designing universally applicable methodologies to evaluate and interpret LLM behavior also persists, as most techniques are model-specific and do not generalize across architectures. These gaps highlight the need for comprehensive research to bridge the divide between LLM capabilities and their interpretability, which this study seeks to address.

1.3 Proposed Approach

The researchers propose a method called SAPLMA (Statement Activation Pattern Language Model Analysis), which involves training a classifier to determine the probability of a statement being truthful based on the hidden layer activations of the LLM as it processes or generates the statement. SAPLMA is trained on a diverse dataset of true and false statements across various content areas. This classifier analyzes the LLM's internal state, specifically focusing on the hidden layer activations, to detect whether a given statement is true or false. The method is versatile, capable of evaluating both statements provided to the LLM and those generated by the model itself.

One of the key advantages of SAPLMA is its computational efficiency. The classifier utilizes a shallow feedforward neural network, which requires minimal computational power at inference time. This design allows SAPLMA to be computed alongside the LLM's output without significant additional resource demands.

1.4 Likely challenges and mitigations

One of the primary challenges is ensuring generalization across diverse topics. SAPLMA is designed to focus on the LLM's internal representation of truth, regardless of the statement's subject matter. However, maintaining consistent performance across a wide range of topics, especially those not encountered during training, could prove difficult. This challenge is closely related to the quality and diversity of the dataset used for training SAPLMA. Creating a comprehensive and unbiased dataset that covers various domains is crucial for the method's reliability but presents its own set of challenges. Another significant hurdle is the potential variability in performance across different LLM architectures. The paper reports accuracy ranges from 71 percent to 83 percent, suggesting that the effectiveness of SAPLMA may depend on the specific LLM used. This variability could complicate the method's implementation across different AI systems and applications.

Collaborative dataset curation, involving domain experts, can lead to more comprehensive and nuanced true-false datasets. Implementing a confidence scoring system alongside the binary classification could help handle ambiguous cases. Finally, an iterative refinement process, incorporating real-world performance data and user feedback, can continuously improve the SAPLMA classifier.

2 Related Work

The paper ”TruthfulQA: Measuring How Models Mimic Human Falsehoods” by ”Stephanie Lin (2022)” and others introduces a benchmark to measure the truthfulness of language models in generating answers to questions. The benchmark comprises 817 questions spanning 38 categories, including health, law, finance, and politics. The researchers tested various models, including GPT-3, GPT-Neo/J, and GPT-2, to evaluate their performance in avoiding false answers learned from imitating human texts.

Another relevant study is ”Verbal lie detection using Large Language Models” by ”(Riccardo Loconte1, 2023)” and others, which explores the performance of FLAN-T5 models in lie detection tasks across three English-language datasets covering personal opinions, autobiographical memories, and future intentions. This research achieved state-of-the-art results in certain scenarios, outperforming previous benchmarks. However, unlike SAPLMA, this work focuses on fine-tuning LLMs for specific lie detection tasks rather than analyzing internal model states, highlighting a different approach to the problem of truthfulness detection.

The paper ”Characterizing Truthfulness in Large Language Model Generations” by ”(Fan Yin, 2024)” nad others proposes using the local intrinsic dimension (LID) of model activations to quantify an LLM's truthfulness. The researchers conducted experiments on four question answering datasets to demonstrate the effectiveness of their method. Their approach focuses on internal model activations to detect truthfulness, similar to the proposed SAPLMA method. The study provides valuable insights into how LLMs process and generate truthful information across different types of questions

A more recent study, ”Large Language Models Help Humans Verify Truthfulness – Except When They Are Convincingly Wrong”, by ”(Chenglei Si, 2024)” and others explores the role of LLMs in human fact-checking processes. This research compared LLMs with search engines in facilitating fact-checking through experiments with crowdworkers. While this study focuses on human-AI interaction in fact-checking rather than automated truthfulness detection, it offers important insights into the practical applications and potential pitfalls of using LLMs for truthfulness verification. The proposed SAPLMA approach differs from these works by focusing on analyzing the internal state of LLMs to detect truthfulness, aiming to provide a more generalizable method across various LLM architectures and topics.

3 Experiments

3.1 Datasets

We utilized the same datasets as in the paper ”The Internal State of an LLM Knows When It’s Lying”, which consists of true and false statements across six distinct topics: Cities, Inventions, Chemical Elements, Animals, Companies, and Scientific Facts. These datasets were carefully curated using reliable external sources to ensure the factual accuracy of the statements.

The Cities dataset was sourced from SimpleMaps, a database of cities worldwide, which is freely available to the public. For the Chemical Elements topic, we used data from PubChem’s periodic table, which is maintained by the National Institutes of Health (NIH). The dataset contains information on elements such as atomic number, symbol, and standard state. The Animals dataset was sourced from the National Geographic Kids website, providing information on animal species and their characteristics. For the Companies dataset, we relied on the Forbes Global 2000 List from 2022, which ranks the top 2000 companies globally. This data was freely available on the Forbes website. Finally, the Scientific Facts dataset was generated using ChatGPT (February 13 version) to provide scientifically well-known facts. These facts were then negated to create false statements, with the entire dataset manually curated and verified by two human annotators. Any questionable facts were removed from the dataset, ensuring that the final set of statements was reliable and accurate.

3.2 Implementation

The re-implementation of the experiments from the paper ”The Internal State of an LLM Knows When It’s Lying” was carried out following the methodology outlined in the original study. The core approach involved generating embeddings for true and false statements using a curated dataset, extracting activations from pre-trained language models (LLMs), and analyzing the internal states of these models to detect discrepancies between true and false statements. The original paper does not provide any code implementation. However, we found a similar implementation on GitHub, which we referenced for conceptual guidance (here)

For the statement generation, we used dataset of true and false statements. Each statement was paired with a label indicating its truthfulness. The dataset was split into training and testing sets, with the training data used for learning the activation patterns associated with truth and falsehood.

We utilized two pre-trained LLMs for our experiments: Facebook OPT-6.7B and LLAMA2-7b, both of which consist of 32 transformer layers. The models were loaded from their respective pretrained checkpoints. These models were chosen due to their large-scale architecture and generalization capabilities. To analyze the internal state of the LLMs, activations were extracted from 5 different layers: the Last Hidden Layer, the 28th Layer, the 24th Layer, the 20th Layer, and a Middle Layer.

Each statement was tokenized using the pretrained tokenizers corresponding to the LLMs, and the resulting embeddings were passed through the model to extract the activations. The activations from each layer were then aggregated and processed to compute a prediction score, which was compared against the ground truth label (true/false).

To assess the performance of our SAPLMA approach, we compared the results against baseline models, including: BERT: A popular transformer-based model fine-tuned for sentence-level classification tasks and 3-shot and 5-shot Learning: Few-shot learning approaches where a small number of labeled examples were provided to the model to learn the task.

We implemented these baseline models to provide a point of comparison, evaluating their truth-detection accuracy against the results from SAPLMA. The baseline models were fine-tuned on the same dataset of true and false statements.

The code for the re-implementation of the experiments, along with additional comparisons and enhancements, is available in the following GitHub repository here. The repository includes the implementation of the SAPLMA method, preprocessing scripts, and training details of the models, along with the evaluation scripts used to generate the results presented in this report.

3.3 Results

As shown, SAPLMA with OPT-6.7b achieves the highest average accuracy of 0.7855 with the middle-layer providing the best performance across most topics, especially Cities and Inventions.

Model	Cities	Invent.	Elements	Animals	Comp.	Facts	Average
last-layer	0.7632	0.6781	**0.6828**	0.604	0.6247	0.7397	0.6821
28th-layer	0.7047	0.6628	0.6404	0.6234	0.6663	0.718	0.6693
24th-layer	0.7931	0.7829	0.6433	0.6333	0.7104	0.7108	0.7123
20th-layer	0.8765	0.8741	0.6672	0.6492	0.7594	0.7618	0.7647
middle-layer	**0.9073**	**0.8875**	0.6578	**0.6767**	**0.8162**	**0.7675**	**0.7855**
BERT	0.5357	0.5537	0.5645	0.5228	0.5533	0.5302	0.5434
3-shot	0.5410	0.4799	0.5685	0.5650	0.5538	0.5164	0.5374
5-shot	0.5416	0.4799	0.5676	0.5643	0.5540	0.5148	0.5370

Table 1: The accuracy of classifying the truthfulness of externally generated sentences using the activations from different layers of OPT-6.7b

For LLAMA2-7b, the 20th layer demonstrates the highest performance, achieving an average accuracy of 0.7135. The 24th-layer also provides a solid result with an average accuracy of 0.6986, particularly performing well with Cities.

Model	Cities	Invent.	Elements	Animals	Comp.	Facts	Average
last-layer	0.8465	0.5641	0.5561	0.6042	0.7526	0.62712	0.6584
28th-layer	0.8178	0.5886	0.5854	0.5985	0.7669	0.6506	0.6679
24th-layer	**0.8561**	0.6682	0.6151	0.6023	0.7969	0.6531	0.6986
20th-layer	0.8449	**0.7186**	**0.6194**	**0.6156**	**0.8278**	**0.6545**	**0.7135**
16th-layer	0.8419	0.6566	0.5768	0.5839	0.7564	0.6227	0.6730

Table 2: Accuracy classifying truthfulness of externally generated sentences using SAPLMA with LLAMA2-7b. The table shows accuracy of all the models tested for each of the topics, and the average accuracy

OPT-6.7b achieves its highest average accuracy with the middle-layer (0.7855), outperforming other layers for most topics. LLAMA2-7b, on the other hand, performs better with the 20th-layer (0.7135), showing better consistency across topics compared to OPT-6.7b. Both models outperform simpler baselines such as BERT and few-shot learning models (3-shot, 5-shot), demonstrating the power of activations from multiple layers of the LLM for truthfulness classification.

The average accuracy values obtained in our reproduction study show a noticeable improvement over the original results across most layers. The last layer accuracy increased from 0.6187 in the original paper to 0.6584 in our reproduction. The 28th layer accuracy rose from 0.6362 to 0.668. The 24th layer accuracy increased from 0.6134 to 0.6986. The 20th layer accuracy showed a notable improvement, rising from 0.6029 to 0.7135. The middle layer, while still showing the lowest accuracy among all layers, improved from 0.5566 to 0.6743.

Model	Accuracy	AUC.
last-layer	0.6584	0.7
28th-layer	0.668	0.7930
24th-layer	0.6986	0.8060
20th-layer	0.7135	0.8009
middle-layer	0.6743	0.7474
BERT	0.4946	0.4810
3-shot	0.5012	0.4998
5-shot	0.5012	0.4998

Table 3: Accuracy classifying truthfulness of sentences generated by the LLM (OPT-6.7b) itself

The AUC values in our reproduced results also exhibit improvements compared to the original paper. The last layer AUC increased from 0.7587 to 0.7. While this particular metric showed a slight decrease, the overall trend across other layers is positive. The 28th layer AUC rose from 0.7614 to 0.7930. The 24th layer AUC increased from 0.7435 to 0.8060, demonstrating the highest improvement. The 20th layer AUC increased from 0.7182 to 0.8009. The middle layer AUC showed a moderate improvement from 0.6610 to 0.7474

The performance of baseline models, including BERT and few-shot methods, was comparable to the original study. In both cases the BERT accuracy remained near 0.51, while its AUC values were slightly lower in the reproduced results (from 0.5989 to 0.4810). Both 3-shot and 5-shot methods produced nearly identical accuracy and AUC values in the reproduced study (0.5012, 0.4998) compared to the original (0.5041, 0.4845, and 0.5125, 0.4822, respectively).

3.4 Discussion

Our results, as shown in the previous section, align well with those reported in the original paper, though there are some differences in the accuracy values across the models.

OPT-6.7b: The last-layer accuracy for OPT-6.7b in our reproduction is 0.6821, whereas in the original paper, it was 0.6449. Our results are slightly higher, particularly for the Cities and Inventions categories. Similarly, the 20th-layer and middle-layer results in our reproduction (0.7647 and 0.7855, respectively) are somewhat higher than the original study's reported values of 0.7098 and 0.6515.

LLAMA2-7b: For LLAMA2-7b, the last-layer accuracy in our reproduction is 0.6584, which is slightly lower than the reported 0.7107 in the original paper. However, we observe a similar trend of higher accuracy with deeper layers. The 20th-layer and 16th-layer results in our reproduction (0.7135 and 0.6730, respectively) are a bit higher than the original paper's 0.8060 and 0.8298, respectively.

Although we attempted to match the preprocessing steps and datasets used in the original paper, slight differences in data cleaning, tokenization, or even the source of the dataset could have impacted the accuracy. Moreover, we used publicly available datasets, which may differ in content or quality compared to those used in the original study. While we used the same models as in the original study (OPT-6.7b and LLAMA2-7b), there may have been subtle implementation differences, such as the specific tokenizer or handling of the model's activation values. These differences could also account for variations in performance.

In summary, the results of our reproduced study are largely consistent with the findings from the original paper. Although there are some minor discrepancies in the accuracy values, the trends we observed align well with the original results, especially regarding the relative performance of different layers in the LLMs.

3.5 Resources

The reproduction required substantial computational power to train and evaluate models on large-scale datasets. Specifically, we used high-performance GPUs to manage the computational load of processing multiple layers of LLM activations and training the SAPLMA classifier. The project took around 4 weeks to complete, including data preparation, model training, evaluation, error analysis, and documentation. Training models and running tests for multiple configurations and datasets were the most time-intensive aspects.

The project was conducted by a team of two, with both members contributing equally. Collaborative effort ensured efficient division of workload and streamlined progress.

3.6 Error Analysis

We examine specific instances where the models failed to correctly classify the truthfulness of statements. A false statement, ”Oranjestad is a city in Nigeria,” was misclassified as true. This error may have occurred due to the model's limited knowledge about smaller or less globally recognized cities. While Oranjestad is not a city in Nigeria, it is possible that the model's training data contained fewer references to less widely known cities, leading to a misunderstanding.

The true statement, ”The theory of evolution, proposed by Charles Darwin, states that species evolve over time through natural selection” was incorrectly classified as false. This is a well-established scientific fact, but the model misclassified it. The error could stem from ambiguity in the phrasing of the statement or from the model's reliance on conflicting or out-of-context scientific data.

Many of the errors occurred with statements that contained phrases with multiple interpretations. The model struggled with less common or highly specific knowledge, especially in domains such as smaller cities (like Oranjestad) or niche scientific facts. These cases may not have been well-represented in the model's training data, which led to errors when the model was required to generalize to less frequent examples. Some of the errors might have stemmed from conflicting information in the model's training data.

In addition to the standard error analysis, we experimented with data augmentation techniques to improve the model's robustness which is discussed in the next section.

4 Robustness Study

To evaluate the model's robustness, we implemented a comprehensive approach based on the techniques described in the paper ”Beyond Accuracy: Behavioral Testing of NLP Models with

CheckList" by "(Marco Tulio Ribeiro, 2020)" and others, adapted them to our specific task. The robustness analysis involved generating perturbed versions of the original datasets using three distinct perturbation techniques. These techniques were carefully selected to introduce controlled variations that could potentially challenge the model's understanding and consistency.

Typo Introduction method involved randomly selecting a word in each statement and deliberately introducing a typographical error by shuffling its letters. The goal was to test the model's resilience to minor spelling variations that do not fundamentally alter the statement's meaning. In Word Order Modification approach, the words of a statement were randomly rearranged while preserving the original vocabulary. This perturbation aimed to evaluate the model's ability to maintain semantic understanding when the syntactic structure is disrupted. Noise Addition technique involved inserting filler words such as "um", "uh", or "like" at random positions within the statement. The purpose was to simulate real-world conversational variations and assess the model's performance when faced with non-essential linguistic noise.

The perturbed datasets served as a comprehensive robustness benchmark. For each original dataset—including categories like capitals, cities, companies, elements, and facts—a corresponding perturbed version was created. This approach allowed for a systematic comparison of the model's performance across different domains and perturbation types.

The model's robustness was evaluated using two primary performance metrics: Accuracy (Measuring the proportion of correctly classified statements across true and false categories.) and Area Under the Curve (AUC) (Assessing the model's ability to distinguish between true and false statements across various classification thresholds.)

4.1 Results of Robustness Evaluation

The robustness evaluation of the reproduced model reveals significant insights into its performance across various datasets. The evaluation was conducted by comparing the model's accuracy and area under the curve (AUC) on both original and perturbed datasets, highlighting areas of strength and vulnerability. The results indicate varying levels of performance across different datasets. When assessing the model's performance on perturbed datasets, a noticeable decline in accuracy and AUC was observed across most categories. This suggests that the model's robustness is compromised when faced with minor alterations in input data.

The Capitals Dataset demonstrates the model's resilience in identifying capital cities, even when faced with slight changes in phrasing or structure. The original performance boasted an impressive average accuracy. Despite perturbations, the model maintained a high level of performance, with a small drop in accuracy and average accuracy. This robust performance indicates that the model has developed a strong understanding of capital cities that can withstand minor alterations in input. Similarly, the Companies Dataset exhibited good robustness. When faced with perturbations, the model's performance remained relatively stable, with only a minor decrease. This stability suggests that the model has developed a robust representation of company-related information that is less affected by small changes in input structure or phrasing.

Model	Cities	Invent.	Elements	Animals	Comp.	Facts	Average
last-layer	0.7013	0.5624	0.5591	0.5657	0.7411	0.6799	0.6350
28th-layer	0.7049	0.5671	0.5569	0.5852	0.7357	0.6949	0.6407
24th-layer	0.7043	0.5792	0.5593	0.5859	0.7366	0.6978	0.6438
20th-layer	0.7075	0.5843	0.5574	0.5999	0.7298	0.6970	0.6460

Table 4: Accuracy generated for the pertubed datasets

From the above results we can see that the impact of perturbations is evident across all layers, with every layer showing a consistent decline in accuracy. The difference is particularly stark for layers closer to the final output, such as the 20th and 24th layers, which had previously exhibited the best performance in the unmodified dataset. This decline suggests that the perturbations have effectively disrupted the coherence or predictability of the input, making it harder for the model to correctly classify statements.

The evaluation also revealed areas where the model's performance was less robust. The Cities Dataset, for instance, showed a significant vulnerability to perturbations. While the original performance was strong, these metrics dropped considerably when faced with perturbed inputs. This substantial decrease indicates that the model's understanding of city-related information is more sensitive to changes in input phrasing or structure, suggesting a potential area for improvement in the model's robustness. The Elements Dataset presented another example of poor robustness. The original performance was already relatively low. When presented with perturbed inputs, the accuracy showed a slight improvement AUC decreased. This inconsistent change in performance metrics suggests that the model's handling of element-related information is not robust and may be particularly sensitive to specific types of perturbations. The slight increase in accuracy coupled with a decrease in AUC could indicate that the perturbations are affecting the model's decision boundaries in complex ways, rather than uniformly degrading performance.

The model's robustness varies significantly across different topics, a phenomenon referred to as topic sensitivity. For instance, it performs exceptionally well in domains like capital cities, where the information is straightforward and factual. However, it struggles with more complex or ambiguous topics such as chemical elements. This variance in performance underscores the model's difficulty in maintaining consistent accuracy across diverse subject matters, particularly when dealing with nuanced or specialized knowledge. The impact of perturbations on the model's performance is substantial, revealing potential areas of overfitting. When faced with minor alterations in input data, the model's accuracy often decreases significantly. This suggests that the model may be relying too heavily on specific patterns in the training data rather than developing a robust, flexible understanding of language. Such sensitivity to perturbations indicates a lack of generalization, which is crucial for real-world applications where input data may not always be pristine or follow expected patterns.

Interestingly, the evaluation revealed inconsistencies in how perturbations affect different performance metrics. In some cases, accuracy improved while the Area Under the Curve (AUC) decreased for certain perturbed datasets. These inconsistent changes in metrics suggest that perturbations may be affecting the model's decision boundaries in complex and unpredictable ways. This phenomenon highlights the intricacy of the model's internal representations and the challenges in creating truly robust language models. In practical scenarios, input data often contains errors, variations, or non-standard phrasing. The model's sensitivity to such changes could lead to inconsistent or unreliable performance in real-world settings, potentially limiting its usefulness in critical applications where consistent accuracy is paramount.

4.2 Discussion

During the analysis, extracting activations from specific layers of large models like OPT-6.7B and LLAMA2-7b turned out to require more computing power than we expected. Running the models multiple times for different layers and statements needed a lot of resources, including GPUs. To speed up future experiments, using parallel processing could help.

When comparing the SAPLMA method with BERT, 3-shot and 5-shot learning models, we faced challenges because these models have different architectures. Additionally, LLMs are expensive to run, and extracting activations takes a lot of time and resources.

Although analyzing the internal states of LLMs for truth detection shows potential, there are many ways to improve these models in future work. Trying different layers, fine-tuning the model settings, and exploring multi-modal approaches will help us better understand how LLMs detect truth and falsehood.

5 Workload Clarification

We contributed equally to all aspects of the assignment. We collaboratively worked on the reproduction of the study, ensuring that each task was evenly distributed and executed effectively. Both of us actively participated in implementing the model, preprocessing the data, running the experiments, and analyzing the results. Additionally, we worked together on drafting the report, discussing our findings, and addressing any challenges that arose during the project.

6 Conclusion

The paper is largely reproducible. We successfully implemented the method described in the original study, including the use of different layers of the LLMs (OPT-6.7b and LLAMA2-7b) for truthfulness prediction based on model activations. The results we obtained were consistent with the original findings, especially in terms of the relative performance of different layers and the overall trend across models. The error analysis and further testing of the model's robustness showed that the model still exhibits the same challenges with certain types of statements, which aligns with the original paper's findings.

References

Sherry Tongshuang Wu Chen Zhao Shi Feng Hal Daumé III Jordan Boyd-Graber Chenglei Si, Navita Goyal. 2024. Large language models help humans verify truthfulness – except when they are convincingly wrong.

Kai-Wei Chang Fan Yin, Jayanth Srinivasa. 2024. Characterizing truthfulness in large language model generations.

Carlos Guestrin Sameer Singh Marco Tulio Ribeiro, Tongshuang Wu. 2020. Beyond accuracy: Behavioral testing of nlp models with checklist.

Pasquale Capuozzo Pietro Pietrini1 Giuseppe Sartori Riccardo Loconte1, Roberto Russo. 2023. Verbal lie detection using large language models.

Owain Evans Stephanie Lin, Jacob Hilton. 2022. Truthfulqa: Measuring how models mimic human falsehoods.