{'title': 'PoLLMgraph: Unraveling Hallucinations in Large Language Models via State Transition Dynamics',

'authors': [{'name': 'Derui Zhu',

'affiliation': 'Technical University of Munich'},

{'name': 'Dingfan Chen',

'affiliation': 'CISPA Helmholtz Center for Information Security'},

{'name': 'Qing Li', 'affiliation': 'University of Stavanger'},

{'name': 'Zongxiong Chen', 'affiliation': 'Fraunhofer FOKUS'},

{'name': 'Lei Ma',

'affiliation': 'The University of Tokyo, University of Alberta'},

{'name': 'Jens Grossklags', 'affiliation': 'Technical University of Munich'},

{'name': 'Mario Fritz',

'affiliation': 'CISPA Helmholtz Center for Information Security'}],

'abstract': "Despite tremendous advancements in large language models (LLMs) over recent years, a notably urgent challenge for their practical deployment is the phenomenon of 'hallucination', where the model fabricates facts and produces non-factual statements. In response, we propose PoLLMgraph —a Polygraph for LLMs—as an effective model-based white-box detection and forecasting approach. PoLLMgraph distinctly differs from the large body of existing research that concentrates on addressing such challenges through black-box evaluations. In particular, we demonstrate that hallucination can be effectively detected by analyzing the LLM’s internal state transition dynamics during generation via tractable probabilistic models. Experimental results on various open-source LLMs confirm the efficacy of PoLLMgraph, outperforming state-of-the-art methods by a considerable margin, evidenced by over 20% improvement in AUC-ROC on common benchmarking datasets like TruthfulQA.",

'key\_findings': 'Experimental results on various open-source LLMs confirm the efficacy of PoLLMgraph, outperforming state-of-the-art methods by a considerable margin, evidenced by over 20% improvement in AUC-ROC on common benchmarking datasets like TruthfulQA. Our research introduces a novel white-box detection approach for hallucination in LLMs, leveraging internal state transition dynamics and temporal information to model hallucination probability. This method significantly improves hallucination detection performance across various setups and model architectures, even in weakly supervised contexts with less than 100 training samples. Recent studies have shown that hallucination rectification can be partially achieved by self-critique prompting, modifying internal representations, or steering generation towards the most probable factually correct samples in the activation space. This work advances the state of hallucination detection by incorporating temporal information and modeling the entire trajectory of the latent state transitions during LLM generation. Our proposed methods surpass previous state-of-the-art techniques by a noticeable margin, evidenced by an increase of over 0.2 in the detection AUC-ROC on the TruthfulQA dataset and around 0.1 on the HaluEval dataset. A general trend can be identified that white-box methods typically outperform gray-box and black-box approaches in terms of detection effectiveness. All of our proposed variants consistently exhibit superior performance when compared to other white-box approaches due to the integration of temporal information through the analysis of state transition dynamics, which is inherently suited to modelling stateful systems such as LLMs. The inclusion of additional latent state abstractions via HMM enhances the modelling capabilities, leading to improved detection effectiveness. Correct predictions typically align with human intuition, indicating the potential of our approach for interpretability analysis of LLMs. While we observe a trend suggesting that utilizing more annotated data improves the effectiveness of our approach across different reference dataset sizes. An increase in detection effectiveness with more abstraction states, likely due to improved modeling capacity and expressive power. GMM notably outperforms K-means when the number of clusters exceeds 150. An observable improvement in detection effectiveness corresponds with retaining more PCA components during down-projection, with performance plateauing at around 1024 PCA dimensions.',

'limitation\_of\_sota': 'Much of the existing research on addressing hallucinations in LLMs relies on either black-box or gray-box settings, identifying hallucinations via output text or associated confidence scores, or relies on extensive external fact-checking knowledge bases, which has proven to be substantially inadequate. Existing studies on LLM hallucination detection primarily rely on the representation of a single token and do not account for the temporal dynamics in state transitions, which can lead to less effective detection and correction of hallucinations. Existing approaches to hallucination detection in LLMs primarily focus on static analysis of the generated text or internal representations without considering the temporal dynamics of the generation process. Existing methods may not fully leverage temporal information or effectively model state transitions, limiting their ability to accurately detect hallucinations in LLMs.',

'proposed\_solution': 'PoLLMgraph proposes a model-based white-box detection and forecasting approach that analyzes the LLM’s internal state transition dynamics during generation via tractable probabilistic models to effectively detect hallucinations. We propose PoLLMgraph, a novel approach that models the hallucination probability based on observed intermediate state representation traces during LLM generation. This method utilizes temporal information in state transition dynamics, offering a closer approximation of the LLM decision-making process and significantly improving detection performance. In our work, we apply commonly used LLM-based judgments for assessing hallucinations and evaluating the detection effectiveness of our approach, due to their reliability and suitability for our setup. The proposed solution involves analyzing the traces of internal state transitions in LLMs with tractable probabilistic models, such as Markov models and hidden Markov models, to detect hallucinations. This approach abstracts the concrete state space into a manageable form using Principal Component Analysis and Gaussian Mixture Models, allowing for the identification of hallucinations through inference on the probabilistic models. The proposed solution involves defining distinct abstract states that represent different clusters of the model’s internal characteristics for training and inference in probabilistic models. After summarizing the internal characteristics of generated texts into traces, standard probabilistic models are used to capture the transitions between these states. Additionally, semantics are bound with hallucination detection through the use of a few annotated reference samples. The proposed solution involves the use of Markov Models (MM) and Hidden Markov Models (HMM) to detect hallucinations in the generation process of Large Language Models (LLMs). The MM captures the autoregressive nature of standard LLM generation, allowing for the modeling of state transitions and the prediction of hallucinations by calculating posterior probabilities. The HMM introduces a layer of analytical depth by accommodating latent variables, which capture unobserved heterogeneity within the state traces, enabling the recognition of various modes in the abstract state space that may induce hallucinations. The effectiveness of detection is commonly evaluated using the AUC-ROC (Area under the ROC Curve), which ranges from 0.5 to 1, with a higher value indicating a more effective detection method. Our key design intuition connects the occurrence of hallucinations to the internal workings of the model, which is particularly relevant when considering practical use cases, where detection is typically conducted by the model owner, who possesses comprehensive knowledge and control over the model. Our approach integrates temporal information through the analysis of state transition dynamics and includes additional latent state abstractions via HMM to enhance modelling capabilities and improve detection effectiveness of hallucinations in LLMs. Investigating the effectiveness of the approach across different reference dataset sizes. PoLLMgraph, a novel method leveraging state transition dynamics within activation patterns to detect hallucination issues in LLMs. It constructs a probabilistic model that intricately captures the characteristics within the LLM’s internal activation spaces, enabling more effective analysis and reasoning of LLM hallucinations.',

'paper\_limitations': "The paper does not specify its limitations. The approach requires manual labeling of reference data for training the probabilistic models, which may limit scalability and adaptability to new domains or languages. The effectiveness of the modeling framework using MM and HMM is acknowledged, with an anticipation of possible future improvements through more advanced designs for the probabilistic models. The model's responses to open-ended questions tend to be categorized as factual, which might arise from the open-ended nature of these responses leading them to be misinterpreted within the context of our model's latent states. Additionally, unusual word combinations trigger hallucination predictions, indicating a potential area for enhancing the language model by incorporating a broader spectrum of less common information into the LLM's training dataset. The hyperparameter settings are crucial in identifying hallucination behavior based on state transition dynamics. Exploring scenarios with a larger degree of distribution shifts could be insightful, especially when the reference and testing data have very different semantics or are limited in scope and when the LLM undergoes extra fine-tuning that causes potential concept shifts in its internal representations. More comprehensive experiments with varied LLM architectures and broader datasets will enhance the validation of the generalizability of PoLLMgraph."}