Controlling Hallucinations in Large Language Models for Coherent, Genre-Specific Storytelling

Varun Magotra, Pooja Guttal

November 2024

Abstract

This paper presents a novel approach for reducing hallucinations in large language models (LLMs) for coherent, genre-specific storytelling by integrating Chain-of-Thought (CoT) prompting and Semantic Entropy Probes (SEPs). Hallucinations—plausible yet factually incorrect outputs—pose a significant challenge to LLM reliability. However, narrativity and coherence offer untapped opportunities for structured reasoning and trustworthy text generation. We leverage the reasoning capabilities of Chain-of-Thought (CoT) prompts to enhance the logical structure of generated outputs and integrate them with Semantic Entropy Probes (SEPs) to quantify uncertainty, enabling the detection and minimization of hallucinations. Experimental results on benchmark datasets demonstrate that our integrated approach significantly reduces hallucinations while preserving coherence and semantic accuracy. By systematically analyzing performance across these benchmarks, this work highlights the effectiveness of combining structured reasoning with uncertainty quantification to improve the reliability of LLMs.

1 Introduction

Large Language Models have shown remarkable capabilities in text generation, specifically in tasks such as story generation, where the capacity to produce cohesive, captivating narratives is essential. However, a significant challenge for LLMs is hallucination, which produces plausible but factually incorrect or irrelevant details. This problem affects story generation, where maintaining narrative coherence, consistency, and adheres to genre-specific norms for generating relevant and believable stories. [2] Hallucinations degrade the quality of the generated text; additionally, they also interfere with the reader's immersion and damage the model's credibility in creative applications.

The primary research question driving this study is: How can Chain-of-Thought (CoT) reasoning and Semantic Entropy Probes (SEPs) be effectively integrated to minimize hallucinations in LLMs, ensuring that the generated stories maintain coherence and adhere to genre-specific standards and user input.Secondary questions include: How does CoT reasoning enhance the model's capability to logically structure a narrative and preserve coherence across complex story elements? How do SEPs improve semantic accuracy by quantifying uncertainty and minimizing hallucinations both at lexical and semantic levels.

This research is significant because hallucinations in the story generation, particularly when dealing with intricate narrative stories, can have a substantial effect on the reader's experience, specifically in applications such as interactive fiction, educational content or therapeutic storytelling where the generated outputs must be reliable, coherent and genre-specific. Our goal is to increase LLMs reliability in story generation, making LLMs more effective and trustworthy.

We propose a novel approach that combines Chain-Of-Though (CoT) - a reasoning technique where a language model generates logical steps to solve a problem step by step, instead of a direct prediction of the final output [15]. and Semantic Entropy Probes (SEPs) - a technique that uses the language model's hidden states to measure the semantic entropy, that indicates the level of inconsistency in meaning. This ensures that the generated content aligns with input prompts, addressing hallucinations and maintaining coherence [5].

Chain-of-Thought (CoT) reasoning breaks complex narrative tasks into logical steps, enabling the model to process them sequentially and maintain coherence. To enhance reasoning, we employ Automatic Chain-of-Thought (Auto-CoT) reasoning, which automates the generation of diverse reasoning chains without requiring manual intervention. Auto-CoT prompts the model with "Let's think step by step" and systematically creates step-by-step demonstrations tailored to the input query. By constructing logical reasoning paths dynamically, Auto-CoT ensures that the model approaches complex tasks in a structured manner, improving its coherence and correctness. However, CoT alone cannot fully eliminate hallucinations, particularly in ambiguous or complex scenarios. The reasoning chains generated by Auto-CoT are fed directly into the next phase of analysis to detect and address hallucinations.

To detect hallucinations in Auto-CoT-enhanced outputs, we integrate Semantic Entropy Probes (SEPs). SEPs are trained on entropy values derived from clustering multiple model responses based on semantic similarity using entailment models (e.g., DeBERTa). This clustering quantifies the semantic entropy by measuring variability across responses. Once trained, SEPs efficiently detect hallucinations during inference by analyzing a single forward pass of the hidden states. SEPs analyze hidden states, focusing on Second-Last-Token (SLT) and Token-Before-Generation (TBG) positions to estimate semantic entropy—a measure of uncertainty in the output's meaning. High entropy signals hallucinations, while low entropy indicates factual and coherent responses.

By combining Auto-CoT and SEPs, this framework ensures coherent, reliable outputs by integrating structured reasoning with semantic uncertainty analysis, effectively reducing hallucinations.

The study's findings have important implications for applications like interactive storytelling, therapeutic stories, game design, and the production of instructional content that depend on the production of high-quality stories. This study lays the ground work for more dependable and captivating AI-driven storytelling by enhancing the coherence, correctness, and inventiveness of LLM-generated stories. We hope to push the limits of LLM performance in creative tasks by integrating CoT reasoning with SEPs, guaranteeing that created stories are not only accurate and coherent but also captivating, genre-appropriate, and hallucination-free.

The rest of the paper is organized as follows: Section 2 provides a detailed review of the related work, including methods for hallucination detection, uncertainty quantification, and reasoning in large language models (LLMs). Section 3 describes the methodology of our framework, which involves the integration of automatic chain-of-thought (Auto-CoT) reasoning and semantic entropy probes (SEPs) for hallucination detection. The dataset and experimental setup used for evaluation are also presented. Section 4 presents the result section, highlighting the experimentation result for the LLMs. Section 5, gives the overall insights about our research along with the furtur direction of this work. Finally, Section 6 provides the conclusion of our research findings.

2 Related Work

Large Language Models (LLMs) have demonstrated remarkable potential in text generation, yet the challenge of hallucinated output remains a significant limitation. This issue is particularly pronounced in domains such as storytelling, where narrative coherence and reliability are critical for user engagement and trust. Addressing hallucinations requires advancements in both reasoning techniques and uncertainty quantification to ensure the generation of factual and contextually relevant outputs.

Farquhar introduced semantic entropy as a quantitative measure to detect hallucinations, enabling effective uncertainty quantification in model outputs [1]. This approach provides a robust framework for identifying regions of high uncertainty, which often correlate with hallucinated or incoherent responses. Similarly, Huang conducted a comprehensive survey on hallucinations, highlighting their causes, challenges, and potential mitigation strategies [2]. These studies collectively emphasize the need for advanced detection mechanisms to improve the reliability and consistency of LLM-generated content.

Storytelling applications are particularly vulnerable to the detrimental effects of hallucinations. Lin et al. demonstrated that hallucinations can disrupt narrative coherence, undermining user trust and the usability of generated content [8]. These findings underscore the importance of narrative alignment and reliable generation techniques to maintain the coherence and quality of storytelling tasks.

To address the challenges posed by hallucinations, Chain-of-Thought (CoT) reasoning has emerged as a transformative approach. Wei demonstrated that CoT prompting significantly enhances the logical structuring of LLM outputs, enabling better performance on complex reasoning tasks [13]. Kojima extended this technique with zero-shot CoT prompting, which allows models to generate intermediate reasoning steps without requiring task-specific training [4]. Further advancements were made by Wang et al., who introduced self-consistency decoding to identify the most coherent reasoning paths, thereby reducing inconsistencies and improving the reliability of outputs [12].

Building upon these innovations, recent studies have extended CoT reasoning to multimodal inputs. Zhang proposed a multimodal CoT approach that integrates text and visual data, enabling LLMs to reason effectively across diverse modalities [14]. Additionally, Wang et al. introduced Plan-and-Solve prompting, a structured approach that guides LLMs to formulate plans before generating solutions [11]. These advancements collectively contribute to generating outputs that are more coherent, logically consistent, and contextually relevant across a wide range of applications.

Uncertainty quantification has also gained prominence as a key mechanism for mitigating hallucinations. Kossen et al. proposed Semantic Entropy Probes (SEPs), a cost-efficient method for detecting hallucinations by analyzing the hidden states of LLMs [6]. SEPs have proven effective in estimating semantic coherence, distinguishing hallucinated outputs from factual ones, and providing a quantitative framework for understanding uncertainty in generated content.

The integration of CoT reasoning and SEPs represents a promising direction for addressing hallucinations in storytelling. By combining logical reasoning with uncertainty quantification, it becomes possible to enhance narrative coherence while minimizing hallucinations. Despite these advancements, the application of these techniques to genre-specific storytelling remains underexplored. This research bridges that gap by leveraging the complementary strengths of CoT reasoning and SEPs to improve the reliability and coherence of LLM-generated narratives. By focusing on reducing hallucinations and ensuring that outputs align with genre-specific expectations, this work advances the state of the art in storytelling applications powered by LLMs.

3 Methods

3.1 Data Description

In this research study, we are using following 4 datasets to train and test the Semantic entropy probe model which are:

- TriviaQA: Joshi et.al., proposed the TriviaQA dataset which is composed of trivia-style questions which require factual recall[3]. This dataset will help us in assessing the factual recall properties of the reasoning model.

- SQuAD: Prannav et.al., proposed the SQuAD dataset which consist of contextual questions derived from passage to evaluate the model comprehensive ability[9].

- BioASQ: Sarrouti et al., proposed the BioASQ dataset which consist of domain specific questions form the biomedical field, allowing us to rigorously test the model performance in specialized knowledge domains[10].

- NQ Open: Kwiatkowski et al., proposed the Natural Question Open dataset which features domain specific question extracted from real google search queries, representing natural information-seeking behavior[7].

The selected datasets are domain specific and very complex in nature, helping us to thoroughly evaluate the model reasoning capabilities and hallucination detection method in different use case.

3.2 Experimentation

This section outlines the comprehensive approach adopted to create a novel framework to reduce and detect the hallucinations in the Large Reasoning models which integrate the chain of thought process along with the Semantic entropy uncertainty to estimate the hallucination in model response which can be seen in the figure 1. Zhang et al., in their research study provide an new chain of though reasoning approach known as Auto-CoT for the language model which automatically construct demonstrations based on the user query, which provide the basis of the prompting technique[15]. It samples questions with diversity and generates reasoning chains to construct demonstrations with prompt: **Let's think not just step by step, but also one by one**. In their research study, they showcased how well their Auto-Cot method is able to match the performance of the tedious Manual-CoT and Zero-Shot prompting when applied on GPT-3 as shown in the figure below.

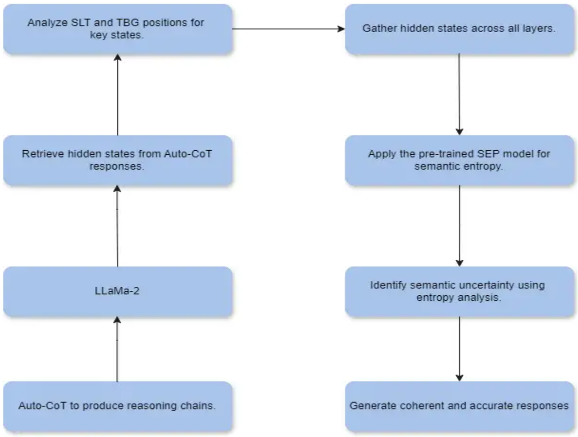
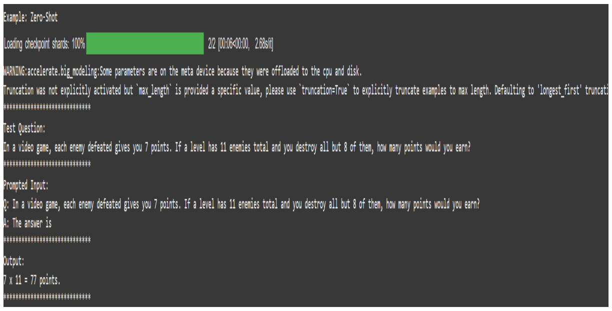
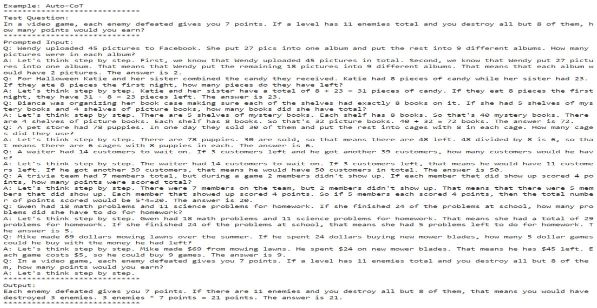

Analyze SLT and TBG positions for key states.
Gather hidden states across all layers.
Retrieve hidden states from Auto-CoT responses.
Apply the pre-trained SEP model for semantic entropy.
LLaMa-2
Identify semantic uncertainty using entropy analysis.
Auto-CoT to produce reasoning chains.
Generate coherent and accurate responses

Figure 1: System Diagram for Auto-CoT + SEP



```
Example: Zero-Shot
Loading checkpoint shards: 100%                    2/2 [00:06<00:00,  2.68s/it]
WARNING:accelerate.big_modeling:Some parameters are on the meta device because they were offloaded to the cpu and disk.
Truncation was not explicitly activated but `max_length` is provided a specific value, please use `truncation=True` to explicitly truncate examples to max length. Defaulting to 'longest_first' truncati
****************************
Test Question:
In a video game, each enemy defeated gives you 7 points. If a level has 11 enemies total and you destroy all but 8 of them, how many points would you earn?
****************************
Prompted Input:
Q: In a video game, each enemy defeated gives you 7 points. If a level has 11 enemies total and you destroy all but 8 of them, how many points would you earn?
A: The answer is
****************************
Output:
7 x 11 = 77 points.
****************************
```

Figure 2: Auto-CoT Reasoning chain and Output



```
Example: Auto-CoT
*****************************
Test Question:
In a video game, each enemy defeated gives you 7 points. If a level has 11 enemies total and you destroy all but 8 of them, h
ow many points would you earn?
*****************************
Prompted Input:
Q: Wendy uploaded 45 pictures to Facebook. She put 27 pics into one album and put the rest into 9 different albums. How many
pictures were in each album?
A: Let's think step by step. First, we know that Wendy uploaded 45 pictures in total. Second, we know that Wendy put 27 pictu
res into one album. That means that Wendy put the remaining 18 pictures into 9 different albums. That means that each album w
ould have 2 pictures. The answer is 2.
Q: For Halloween Katie and her sister combined the candy they received. Katie had 8 pieces of candy while her sister had 23.
If they ate 8 pieces the first night, how many pieces do they have left?
A: Let's think step by step. Katie and her sister have a total of 8 + 23 = 31 pieces of candy. If they eat 8 pieces the first
night, they have 31 - 8 = 23 pieces left. The answer is 23.
Q: Bianca was organizing her book case making sure each of the shelves had exactly 8 books on it. If she had 5 shelves of mys
tery books and 4 shelves of picture books, how many books did she have total?
A: Let's think step by step. There are 5 shelves of mystery books. Each shelf has 8 books. So that's 40 mystery books. There
are 4 shelves of picture books. Each shelf has 8 books. So that's 32 picture books. 40 + 32 = 72 books. The answer is 72.
Q: A pet store had 78 puppies. In one day they sold 30 of them and put the rest into cages with 8 in each cage. How many cage
s did they use?
A: Let's think step by step. There are 78 puppies. 30 are sold, so that means there are 48 left. 48 divided by 8 is 6, so tha
t means there are 6 cages with 8 puppies in each. The answer is 6.
Q: A waiter had 14 customers to wait on. If 3 customers left and he got another 39 customers, how many customers would he hav
e?
A: Let's think step by step. The waiter had 14 customers to wait on. If 3 customers left, that means he would have 11 custome
rs left. If he got another 39 customers, that means he would have 50 customers in total. The answer is 50.
Q: A trivia team had 7 members total, but during a game 2 members didn't show up. If each member that did show up scored 4 po
ints, how many points were scored total?
A: Let's think step by step. There were 7 members on the team, but 2 members didn't show up. That means that there were 5 mem
bers that did show up. Each member that showed up scored 4 points. So if 5 members each scored 4 points, then the total numbe
r of points scored would be 5*4=20. The answer is 20.
Q: Gwen had 18 math problems and 11 science problems for homework. If she finished 24 of the problems at school, how many pro
blems did she have to do for homework?
A: Let's think step by step. Gwen had 18 math problems and 11 science problems for homework. That means she had a total of 29
problems for homework. If she finished 24 of the problems at school, that means she had 5 problems left to do for homework. T
he answer is 5.
Q: Mike made 69 dollars mowing lawns over the summer. If he spent 24 dollars buying new mower blades, how many 5 dollar games
could he buy with the money he had left?
A: Let's think step by step. Mike made $69 from mowing lawns. He spent $24 on new mower blades. That means he has $45 left. E
ach game costs $5, so he could buy 9 games. The answer is 9.
Q: In a video game, each enemy defeated gives you 7 points. If a level has 11 enemies total and you destroy all but 8 of the
m, how many points would you earn?
A: Let's think step by step.
*****************************
Output:
Each enemy defeated gives you 7 points. If there are 11 enemies and you destroy all but 8 of them, that means you would have
destroyed 3 enemies. 3 enemies * 7 points = 21 points. The answer is 21.
*****************************
```

Figure 3: Zero-Shot LLM Hallucination.

In this research study, we investigate how a language model generates hallucinations when given the ability to reason through an input. We employ the Auto-CoT approach on open-source models like LLaMA-2 7B to enhance reasoning capabilities. After applying Auto-CoT, we proceed to estimate and detect hallucinations in the model's responses. To achieve this, we leverage semantic entropy as a measure of uncertainty in the model's output, distinguishing between hallucinated and factual responses. Kossen et al., in their research, proposed a novel method for cost-efficient and reliable uncertainty estimation in LLMs [5], which we adopt in our study. Specifically, we use the Semantic Entropy

Probe (SEP) model to analyze the hidden states of the LLM, estimating the uncertainty in its responses to determine whether the model is hallucinating or producing a factual result. Semantic entropy, encoded in the model's hidden states, serves as the foundation for detecting hallucinations, allowing us to extract uncertainty for each response and classify the outputs accordingly.

We employ the Automatic Chain-of-Thought (Auto-CoT) model on LLaMA-2, which generates diverse and logically structured reasoning chains to improve coherence and correctness in outputs. Auto-CoT automates the process of constructing reasoning demonstrations, eliminating the need for manual intervention, and allows the model to think through problems step by step. This structured reasoning enhances the model's ability to maintain narrative coherence and produce accurate results. After applying Auto-CoT, the **Semantic Entropy Probe (SEP)** model is integrated to detect hallucinations by analyzing semantic entropy, offering a computationally efficient alternative to traditional sampling-based methods. SEP is trained on hidden states extracted from the Auto-CoT-enhanced responses of LLaMA-2 7B. In the extraction phase, we focus on the Second-Last-Token (SLT) and Token-Before-Generation (TBG) positions, which capture critical insights into the model's semantic uncertainty. These hidden states, collected across all layers, provide a comprehensive view of the model's behavior, enabling detailed uncertainty analysis. Based on the detected uncertainty, we classify the model's outputs as either high or low uncertainty using a threshold mechanism. While this classification helps identify potentially incorrect outputs, we have not yet implemented feedback mechanisms to prompt the model to reconsider its responses. Further advancements, including providing feedback to the model, are outlined in the discussion section.

3.3 Evaluation

To assess the performance of our proposed approach, we firstly select the LLamma-2 7B with the SEP model as our baseline model for all experiments. For evaluation metrics we are using the following metrics:

- **Area Under the Receiver Operating Characteristic Curve (AUROC):** The first metric that we will be using is the AUROC, this metric will help us in measuring the ability of Semantic Entropy Probes (SEPs) to distinguish between hallucinated and non-hallucinated outputs. And it will help us understand how model is hallucinating in different domain scenarios and how well the Semantic Entropy Probes are able to predict the likelihood of hallucinations in the model output.The curve plots the true positive rate against the false positive rate at various threshold of semantic entropy. A higher AUROC indicates better classification performance in distinguishing the hallucinated output and the factual one. Through this metric, we will be able to evaluate the enhancement in the model's reliability and trustworthiness and how much reasoning help a model to produce factual outputs.

4 Result

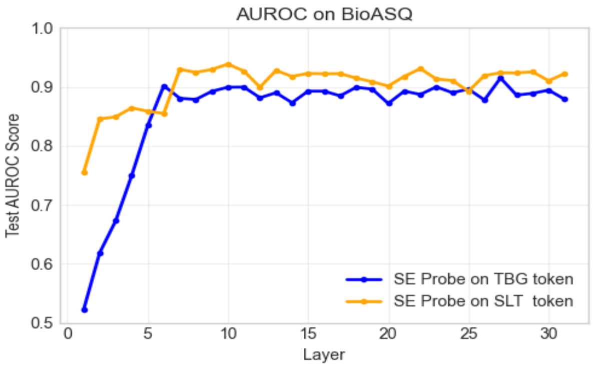


Figure 4: AUROC on BioASQ (LLaMMa-2 +SEP)

As shown in Figure 4, 5, 6 and 7 the AUROC performance of SE Probes applied to token-before-generation (TBG)and second-last-token(SLT) across different model layers for four datasets—BIOA-SQ, TRIVIA-QA, NQ, and SQuAD. We observe that AUROC scores often increase with deeper layers when looking at both token-before-generation (TBG) and

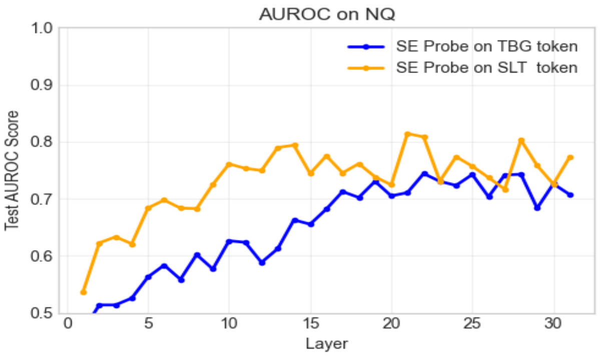


Figure 5: AUROC on NQ Open (LLaMMa-2 +SEP)

second-last-token (SLT) positions. The majority of datasets exhibit notable improvements from beginning scores of 0.50-0.55 to final scores of 0.75-0.90. The SLT position performs better than TBG, with especially good outcomes in later layers (24–32).

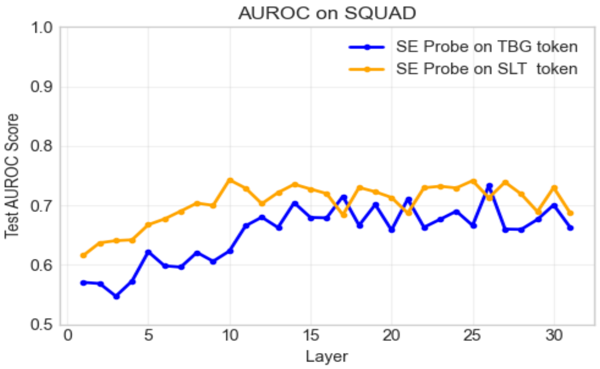


Figure 6: AUROC on SQUAD (LLaMMa-2 +SEP)

- **Performance Across Datasets:** BIOASQ exhibits the highest AUROC scores, with both tokens often achieving scores above 0.9 in deeper layers, indicating good semantic alignment.TRIVIA-QA, NQ, and SQuAD exhibit comparable performance across various datasets, with final AUROC scores stabilizing between 0.75 and 0.85 and exhibiting comparatively similar trends.

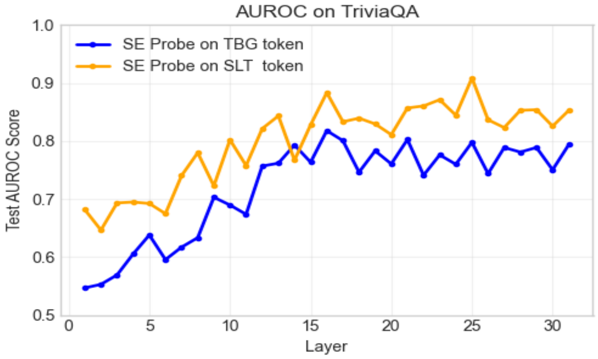


Figure 7: AUROC on triviaQA (LLaMMa-2 +SEP)

Although later layers consistently yield more dependable results, this variance raises the possibility that the best layer selection for hallucination detection may vary depending on the dataset. We emphasize SLT positions from levels 24-32 in our implementation because the SLT position notably exhibits more consistent performance patterns, especially in the last third of the model's layers. These results represents our baseline model. Further, we will be evaluating our framework and estimate the extent of reduction for hallucination compared to the baseline model.

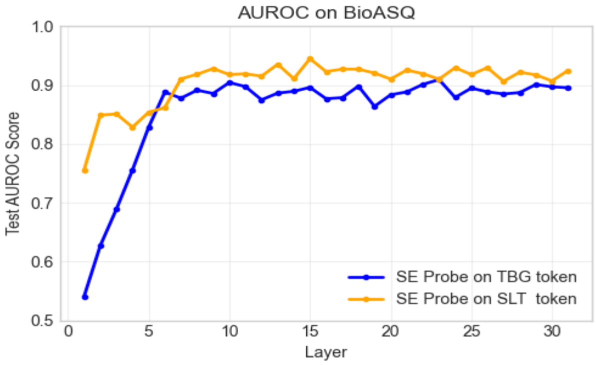


Figure 8: AUROC on BioASQ (Auto-CoT + SE).

As shown in Figure 8, 9, 10 and 11 the AUROC scores of Auto-CoT LLaMa 2-7b + SE Probes

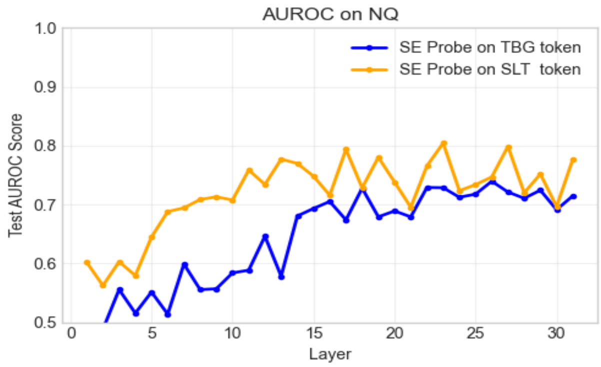


Figure 9: AUROC on NQ Open (Auto-CoT + SE).

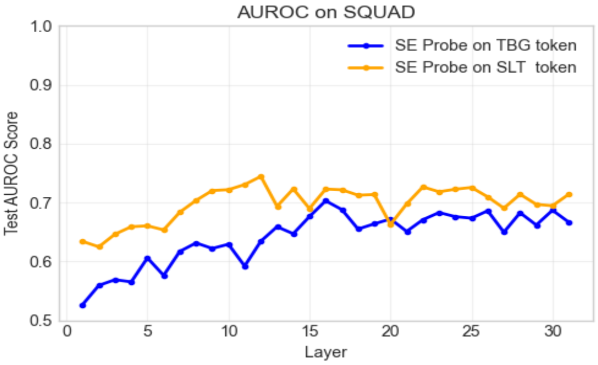


Figure 10: AUROC on SQUAD (Auto-CoT + SE).

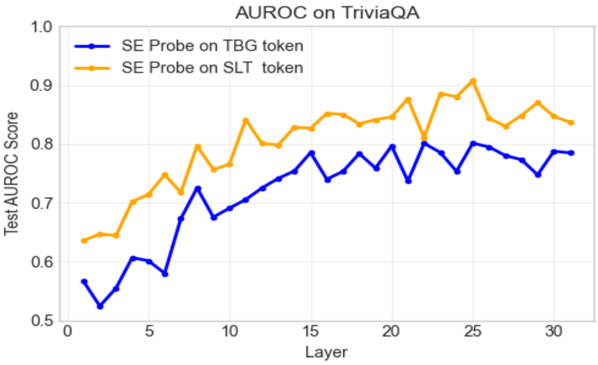


Figure 11: AUROC on triviaQA (Auto-CoT + SE).

applied on Token-Before-Generation (TBG) and Second-Last-Token (SLT) across various layers for datasets—BIOASQ, NQ, SQuAD, and TriviaQA.

The AUROC scores consistently improve across layers as the model progresses deeper, with the SLT position consistently outperforming the TBG position, particularly in the final third of the model's layers (24–32).

- **Performance Compared to Baseline:** The Auto-CoT LLaMa 2-7b + SE Probes model demonstrates improved AUROC scores for SLT tokens across all datasets compared to the baseline, while TBG token performance remains comparable, though it does not exceed the baseline for datasets like BIOASQ. The SLT token shows consistent trends across layers, particularly in layers 24–32, reinforcing its robustness for AUROC performance and achieving slightly better results than the baseline. Both models indicate that SLT tokens outperform TBG tokens, but the Auto-CoT LLaMa 2-7b model further amplifies this performance gap. Additionally, significant improvements are observed in datasets such as NQ and TriviaQA, where the SLT token consistently outperforms the baseline.

The Auto-CoT LLaMa 2-7b + SE Probes framework exhibits superior performance compared to the baseline model, especially for SLT tokens in the deeper layers (24–32). The improved consistency and higher AUROC scores, particularly in challenging datasets like TriviaQA and NQ, demonstrate the effectiveness of this approach over the traditional large language model.

5 Discussion

From our experimentation, we observed various key insights about the LLM behavior when they are given reasoning ability and about utilizing their hidden states. By analyzing the performance of the Auto-CoT LLaMA-2 7balong with Semantic Probes, we observed consistent improvement in the hallucination detection and less hallucinated outputs across different datasets. The AUROC scores across all the dataset shows significant improvement in hallucination detection along the deeper layers, as suggested by the Jannik Kossen et al., in their paper, where the second last token (SLT) consistently performs better than the token before the generation, showcasing how information encoded in the deeper layers encoded semantic information, which can be really helpful in detecting the hallucinated outputs correctly. Furthermore, Auto-CoT approach provided a better logical and structural reasoning over the traditional method like zero-shot and manual cot, helping the model to understand the input properly, which in turn help the model to provide more accurate outputs. From the results, we observed that the SEP performance across the datasets, as in BioASQ dataset, our integrated approach shows highest AUROC scores in deeper layers, where as for SQuAD dataset, it showes slightly lower performance, suggesting that we have to look deeper into the data characteristic to decide which hallucination parameters we can use for SEP. Our findings, shows that by leveraging logical and structured reasoning along with uncertainty quantification, can enhance LLM reliability.

Moving forward, we will be using the Semantic Entropy probe output as the baseline metric for fine tuning the reasoning model. We will integrate the SEP module in the reinforcement learning, by using the SEP method to detect the hallucinations in the LLM's output, to fine-tune the reasoning model. The SEP outputs will serve as the baseline reward signal for the reinforcement learning framework. The outputs which are factually correct and have lower semantic entropy will be rewarded, while hallucinated outputs with higher entropy will be penalized. This feedback loop will help the model to generate more reliable and factual output. Further more, to make the feedback process more consolidated and robust, we will be further exploring other feature metrics like BLEU and ROUGE for text quality, along with the SEP scores. The integration of RL with SEP-based rewards represents a novel framework for improving the coherence and reliability of LLM-generated narratives. This approach has the potential to Reduce Hallucinations: By penalizing high-entropy outputs, the framework ensures semantically grounded factual outputs. This research direction offers a pathway to more dependable, adaptive, and effective LLMs, laying the groundwork for hallucination reduction method which create factual correct and coherent outputs based on the user input.

6 Conclusion

Confabulations or so called hallucinations, pose a difficult challenge for the current Large Language Models, as these hallucinations leads to incorrect results which is very dangerous in domains like Medical, Law etc. In this research, we have explored how much reasoning helps a simple LLM, to reduce its hallucinations and how can we detect those hallucination with out any computational overhead. To induce the reasoning in the LLM, we used Auto-CoT model which doesn't require any manual intervention while creating the reasoning chains and have shown major performance boost over the traditional zero-shot and manual-cot methods. To detect the LLM's hallucinations, we used Semantic Entropy Probes which quantifies the uncertainty in the model response without the need of sampling the responses, to detect the hallucinated responses. And through our experimentation over diverse datasets, we have shown how the reasoning models, shows less confabulations as compared to the traditional counterparts. The proposed experimentation showed around 5%-8% improvement in confabulation reduction over the traditional counterpart LLM.

In the future work, we will be exploring how we can use the Reinforcement learning framework along with the SEP score and output to fine tune the model to provide factual correct and coherent outputs based on the user input.

References

[1] Sebastian Farquhar et al. "Detecting hallucinations in large language models using semantic entropy". In: *Nature* 630.8017 (2024), pp. 625–630. DOI: 10.1038/s41586-024-07421-0. URL: https://doi.org/10.1038/s41586-024-07421-0.

[2] Lei Huang et al. “A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions”. In: *arXiv preprint arXiv:2311.05232* (2023). URL: https://arxiv.org/abs/2311.05232.

[3] Mandar Joshi et al. “TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension”. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Regina Barzilay and Min-Yen Kan. Vancouver, Canada: Association for Computational Linguistics, July 2017, pp. 1601–1611. DOI: 10.18653/v1/P17-1147. URL: https://aclanthology.org/P17-1147.

[4] Takeshi Kojima et al. “Large Language Models are Zero-Shot Reasoners”. In: *arXiv preprint arXiv:2205.11916* (2022). URL: https://arxiv.org/abs/2205.11916.

[5] Jannik Kossen et al. “Semantic entropy probes: Robust and cheap hallucination detection in llms”. In: *arXiv preprint arXiv:2406.15927* (2024).

[6] Tim Kossen et al. “Semantic Entropy Probes: Quantifying Hallucinations in Large Language Models”. In: *Transactions of the Association for Computational Linguistics* 11 (2023), pp. 1–20.

[7] Tom Kwiatkowski et al. “Natural Questions: A Benchmark for Question Answering Research”. In: *Transactions of the Association for Computational Linguistics* 7 (2019). Ed. by Lillian Lee et al., pp. 452–466. DOI: 10.1162/tacl_a_00276. URL: https://aclanthology.org/Q19-1026.

[8] Xiao Lin and Sarah Lee. “Towards Trustworthy LLMs: A Review on De-biasing and Dehallucinating in Large Language Models”. In: *arXiv preprint arXiv:2303.12345* (2023). URL: https://arxiv.org/abs/2303.12345.

[9] P Rajpurkar. “Squad: 100,000+ questions for machine comprehension of text”. In: *arXiv preprint arXiv:1606.05250* (2016).

[10] Mourad Sarrouti and Said Ouatik El Alaoui. “A Biomedical Question Answering System in BioASQ 2017”. In: *BioNLP 2017*. Ed. by Kevin Bretonnel Cohen et al. Vancouver, Canada, Association for Computational Linguistics, Aug. 2017, pp. 296–301. DOI: 10.18653/v1/W17-2337. URL: https://aclanthology.org/W17-2337.

[11] Chen Wang, Jing Liu, and Alex Zhou. “Plan-and-Solve Prompting for Chain-of-Thought Reasoning”. In: *arXiv preprint arXiv:2301.08465* (2023). URL: https://arxiv.org/abs/2301.08465.

[12] Xiang Wang et al. “Self-Consistency Improves Chain of Thought Prompting”. In: *arXiv preprint arXiv:2203.11171* (2022). URL: https://arxiv.org/abs/2203.11171.

[13] Jason Wei et al. “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models”. In: *arXiv preprint arXiv:2201.11903* (2022). URL: https://arxiv.org/abs/2201.11903.

[14] Lei Zhang, Min Liu, and Kai Zhao. “Multimodal Chain-of-Thought Reasoning in Large Language Models”. In: *arXiv preprint arXiv:2302.03456* (2023). URL: https://arxiv.org/abs/2302.03456.

[15] Zhuosheng Zhang et al. “Automatic chain of thought prompting in large language models”. In: *arXiv preprint arXiv:2210.03493* (2022).