Cleanse: Uncertainty Estimation Approach Using Clustering-based Semantic Consistency in LLMs

Minsuh Joo **Hyunsoo Cho**
Ewha Womans University
{judyjoo21, chohyunsoo}@ewha.ac.kr

Abstract

Despite the outstanding performance of large language models (LLMs) across various NLP tasks, hallucinations in LLMs–where LLMs generate inaccurate responses–remains as a critical problem as it can be directly connected to a crisis of building safe and reliable LLMs. Uncertainty estimation is primarily used to measure hallucination levels in LLM responses so that correct and incorrect answers can be distinguished clearly. This study proposes an effective uncertainty estimation approach, **Cl**ust**e**ring-based sem**an**tic consis**te**ncy (**Cleanse**). Cleanse quantifies the uncertainty with the proportion of the intra-cluster consistency in the total consistency between LLM hidden embeddings which contain adequate semantic information of generations, by employing clustering. The effectiveness of Cleanse for detecting hallucination is validated using four off-the-shelf models, LLaMA-7B, LLaMA-13B, LLaMA2-7B and Mistral-7B and two question-answering benchmarks, SQuAD and CoQA.

1 Introduction

Recent advances in LLMs have dramatically enhanced their performance across a wide spectrum of downstream tasks, from translation and summarization to question answering (QA) and dialogue generation. These models now produce fluent, contextually aware outputs that often rival human-like language generation. Despite these remarkable capabilities, a persistent and critical limitation remains: LLMs frequently generate hallucinated outputs—responses that may appear coherent and plausible but are in fact factually incorrect or unsupported by any underlying knowledge (Ji et al., 2023; Huang et al., 2025). These hallucinations are particularly insidious because they are difficult for users, especially non-experts, to detect, potentially leading to serious consequences in high-stakes applications. This challenge becomes especially pronounced in QA tasks, where correctness can be objectively verified. Unlike open-ended tasks such as dialogue or summarization—where diverse outputs can still be acceptable—QA typically demands precise and verifiable answers (Zhang et al., 2023). As a result, even minor hallucinations can significantly degrade task accuracy. When hallucinated outputs are presented in such contexts, they can mislead users, erode trust in AI systems, and compromise the reliability of LLM-based applications (Zhang et al., 2023). Ensuring the factual consistency of outputs is thus not only a technical concern but also a crucial factor for user safety and system credibility.

To address these challenges, researchers have proposed a variety of solutions, including dataset refinement, retrieval-augmented generation (RAG), and uncertainty estimation. Each of these approaches targets hallucination from a different angle, offering complementary benefits. One approach is dataset refinement, which involves carefully reviewing and editing training data to improve model accuracy. While this can help reduce errors, it is also highly labor-intensive and difficult to scale. Another strategy is retrieval-augmented generation (RAG). By retrieving external knowledge during the generation process, RAG can provide more factually grounded answers. However, this approach requires building more complex and potentially fragile pipelines that demand significant computational resources (Ji et al., 2023; Es et al., 2024). In contrast, uncertainty estimation offers a lightweight and scalable alternative by assessing the model's confidence in its own outputs. Importantly, this method does not require additional external knowledge sources or significant changes to the model architecture. Instead, it provides users with interpretable confidence signals that can help identify potentially unreliable responses (Lin et al., 2022a). In QA and related tasks, these confidence metrics can serve as a critical line of defense against the

unintended consequences of hallucination.

Within natural language processing (NLP), uncertainty estimation is typically grounded in the assumption that models are more consistent when confident. That is, when a model is certain about its answer, repeated generations will tend to converge; conversely, a lack of confidence often results in high output variability. To assess uncertainty in generated outputs, researchers have proposed methods that operate at various linguistic levels—token and sentence—each providing distinct advantages based on the desired granularity of analysis. Token-level metrics such as Perplexity (Ren et al., 2023), LN-Entropy (Malinin and Gales, 2020), and Lexical Similarity (Lin et al., 2022b) are well-suited for capturing fine-grained variations within specific output spans, particularly within answer segments of a sentence. In contrast, Rabinovich et al. (2023) evaluates uncertainty at the sentence-level, making it more appropriate for assessing broader linguistic properties such as overall semantic sentiment. While analyses at both token and sentence levels offer valuable insights, semantic aspect of natural language is more significant when deciding whether two texts with different form are equivalent or not. This is because the inherent variability of natural language data leads to semantic equivalence, where diverse expressions can convey the same meaning (Kuhn et al., 2023). Even if two texts use different tokens and syntactic structures, it is reasonable to consider them consistent as long as their semantics are the same. However, sentence-level similarity measures are not without limitations. Rabinovich et al. (2023) calculates all pairwise similarities and they take the average of these similarities equally. It might lead to an incorrect result that a few highly similar sentence pairs disproportionately influence the overall uncertainty score. This can mask the presence of semantically divergent outputs and falsely suggest high consistency.

To overcome these challenges and make metric more precise, we introduce **Cl**ust**e**ring-based Sem**an**tic Con**s**ist**e**ncy (Cleanse)—a novel sentence-level uncertainty estimation technique designed to more reliably detect hallucinations in generative models. Cleanse leverages bi-directional natural language inference (NLI) to determine whether pairs of generated responses entail one another, forming semantically equivalent clusters with greater precision and excluding any connections that do not meet entailment criteria. We then measure the internal connectivity of these clusters by computing the cosine similarity of their hidden representations as a proxy for semantic consistency, while the distances between clusters provide signals for semantic divergence. In other words, dense intra-cluster links indicate semantic agreement, while high inter-cluster links suggest uncertainty. Thus, we estimate uncertainty by leveraging the similarity between embeddings within the same clusters as the degree of consistency. By prioritizing these semantically meaningful clusters—rather than relying on simple average similarity—Cleanse offers more calibrated and trustworthy uncertainty estimates. Experiments on QA benchmarks further demonstrate that Cleanse consistently outperforms existing token- and sentence-level methods in detecting hallucinations. We also verify that our key concept, which considers the degree of inter-cluster links (i.e., inter-cluster similarity) as penalty and degree of intra-cluster links (i.e., intra-cluster similarity) as consistency between outputs, contributes to improving hallucination detection performance and the robustness of Cleanse.

2 Related Work

There are several related works about uncertainty estimation with various perspectives. The researchers fine-tune the model to ensure that the estimated uncertainty aligns with the actual uncertainty (Lin et al., 2022a). Application of perturbation module and aggregation module to calibrate uncertainty is an effective setting as well. (Gao et al., 2024). Semantic entropy is the entropy across groups clustered by semantically-equivalent outputs (Kuhn et al., 2023). Shifting Attention to Relevance (SAR) shifts weights from semantically-irrelevant tokens to semantically-relevant tokens so that probability of relevant tokens contributes to uncertainty quantification more significantly (Duan et al., 2023). Recently, there are some approaches using LLM's internal states. The researchers propose a framework named INSIDE, which exploits the eigenvalues of responses' covariance matrix to measure the semantic consistency in the dense embedding space (Chen et al., 2024). Internal states can be considered as the input of the uncertainty estimator model so that the model classifies whether the response is hallucinated or not (Ji et al., 2024).

3 Method

Cleanse estimates the uncertainty by quantifying the intra-cluster consistency between generations,

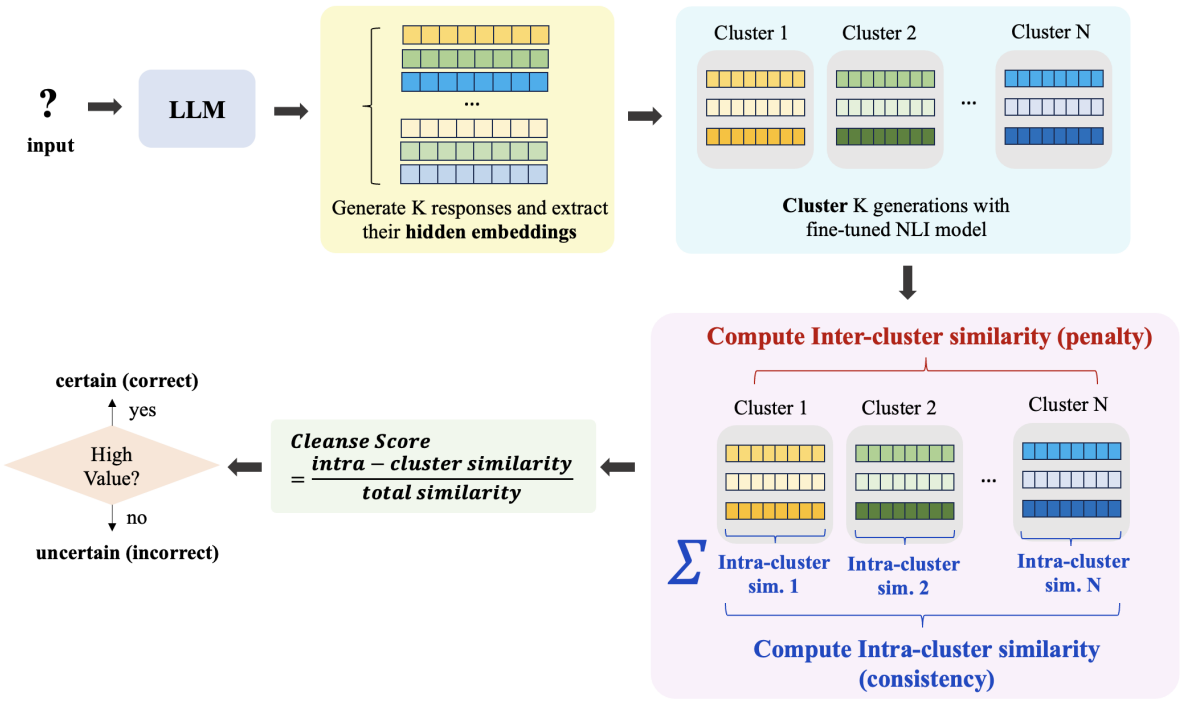


Figure 1: Illustration of Cleanse pipeline.

leveraging semantics of responses by employing sentence-level embeddings and bi-directional clustering. First, we generate multiple outputs and extract their hidden embeddings from the model. Then, we cluster those outputs based on their semantic equivalency. Finally, to assess uncertainty, we compute similarities within and across these clusters respectively and calculate Cleanse Score. Specifically, we demonstrate the hidden embeddings we use in Section 3.1, the clustering technique we use in Section 3.2, and how to compute Cleanse score in Section 3.3.

3.1 Hidden embeddings

We use the last token embedding in the middle layer of LLM as the output's hidden embedding, as prior work suggests it may capture semantic information effectively (Azaria and Mitchell, 2023). Here, considering a single hidden embedding as a d-dimensional vector embedding, we measure the consistency between these hidden embeddings using cosine similarity.

3.2 Clustering techniques

We apply the concepts used in clustering validation by adapting them to be suitable for our study, which aims for the better and clearer quantification. In general, the main goal of clustering is to maximize the inter-cluster distances and minimize the intra-cluster distances (Ansari et al., 2015) and these two criteria are utilized in the clustering validation techniques such as Dunn's Index (Ansari et al., 2015). Dunn's Index is defined as the ratio between the minimum distance across different clusters and the maximum distance within the same cluster, where a value closer to 1 indicates better clustering performance. Here, we could shift the perspective from distance to similarity by taking the inverse of the distance (Ansari et al., 2015). In the perspective of similarity, better clustering corresponds to high intra-cluster similarity and low inter-cluster similarity. When we view it from a consistency perspective rather than clustering validation, it provides an intuitive insight that high intra-cluster similarity indicates the presence of many embeddings sharing equivalent meanings, while high inter-cluster similarity suggests the presence of embeddings with diverse meanings. We perform clustering on the K outputs to utilize these similarity concepts. We will further explain what is done with the clustering results in Section 3.3. The thing is that, our study aims to compute these similarities and quantify uncertainty, not to minimize inter-cluster similarity or maximize intra-cluster similarity. We just got an intuition from the concept of the distance defined in the clustering, which can be transformed to similarity.

To ensure that the outputs are clustered based on

their semantic information, we use a fine-tuned NLI model that maps the input to a high-dimensional semantic embedding. We utilize the clustering algorithm used in the precedent study (Kuhn et al., 2023). Here, we introduce only some main concepts for this algorithm. First main concept is that a pair of outputs is considered `entailment` only when both outputs are entail to each other–i.e., bi-directional entailment–which ensures the two outputs truly share the same meaning. Second, researchers concatenated question and its answer in the form of `<Question+Answer>`, insisting that the content of question helps the clustering model comprehend the input context better. Finally, the algorithm is computationally efficient for two reasons. First, the NLI model is substantially smaller than the main model which generates outputs. While the main model has 7B and 13B parameters, the clustering model we used (i.e., nli-deberta-v3-base) has only 184M parameters, making the clustering process comparatively lightweight. Additionally, the number of comparisons required to determine whether an output should be included in the cluster is reduced due to the transitive characteristic between outputs. This transitivity means that a new output can be added to a certain cluster as long as it has a bi-directional entailment with at least one existing member of that cluster, thereby making the number of comparisons be small. More detailed about the algorithm we refer is shown in Algorithm 1.

Algorithm 1 Bi-directional Entailment Algorithm

Require: context x, set of seqs. $\{s^{(2)}, \ldots, s^{(M)}\}$, NLI classifier $\mathcal{M}$, set of meanings $C = \{\{s^{(1)}\}\}$
  for $2 \leq m \leq M$ **do**
    for $c \in C$ **do**
      $s^{(c)} \leftarrow c_0$ ▷ Compare to existing meanings
      `left` $\leftarrow \mathcal{M}(\text{cat}(x, s^{(c)}, \text{“}<g/>\text{”}, x, s^{(m)}))$
      `right` $\leftarrow \mathcal{M}(\text{cat}(x, s^{(m)}, \text{“}<g/>\text{”}, x, s^{(c)}))$
      if `left` **and** `right` are `entailment` **then**
        $c \leftarrow c \cup \{s^{(m)}\}$ ▷ Add to cluster
      end if
    end for
    $C \leftarrow C \cup \{s^{(m)}\}$ ▷ New cluster
  end for
  return C

3.3 Cleanse Score

Here, we define concepts of similarities from Section 3.2 for clear understanding. Intra-cluster similarity refers the sum of all cosine similarities between embeddings within the same cluster which is computed by Eq 1. C is the number of clusters, N_k is the number of hidden embeddings in the k-th cluster, and $\text{cosine}(\text{e}_\text{i}, \text{e}_\text{j})$ is the cosine similarity between i-th and j-th hidden embeddings. Inter-cluster similarity refers that of all cosine similarities between embeddings across the different clusters. Total similarity is the summation of intra-cluster similarity and inter-cluster similarity which is computed by Eq. 2 where K is the number of outputs. Figure 2 clarifies the definition of our terms.

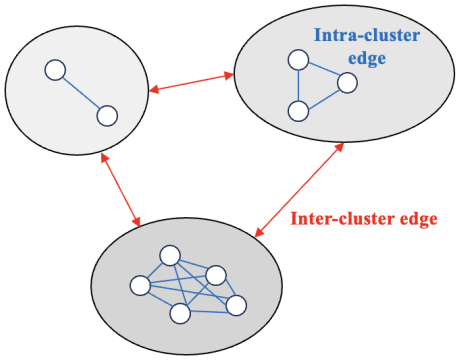


Figure 2: Each white circle indicates a single hidden embedding. Edge means the relationship formed between two embeddings. The red edges represent inter-cluster edges, while the blue edges represent intra-cluster edges. Even the red edges are simplified in this illustration, they represent all possible combinations of embeddings in the different clusters. There are given weights to all edges and each of the weight is the computed cosine similarity between two embeddings.

$$\text{intra-cluster sim.} = \sum_{k=1}^{C} \sum_{i=1}^{N_k-1} \sum_{j=i+1}^{N_k} \text{cosine}(\text{e}_\text{i}, \text{e}_\text{j}) \quad (1)$$

$$\text{total sim.} = \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} \text{cosine}(\text{e}_\text{i}, \text{e}_\text{j}) \quad (2)$$

By clustering the outputs based on their semantic equivalency, we can identify how many clusters are formed, which in turn indicates how much semantically-inconsistent the outputs are. If there are many clusters, outputs have low consistency (i.e., high uncertainty). In this case, most edges are inter-cluster edges, meaning the inter-cluster similarity is greater than intra-cluster similarity and

it leads to low proportion of intra-cluster similarity in the total similarity. In contrast, if the number of clusters is small, outputs have high consistency (i.e., low uncertainty) where most edges are intra-cluster edges. It would lead to high proportion of intra-cluster similarity in the total similarity. Based on this intuition, we measure intra-cluster similarity as the degree of consistency which contributes to the high consistency because they are the similarities between embeddings which are semantically equivalent. Inter-cluster similarity is considered as the penalty for the consistency between outputs as high inter-cluster similarity indicates that there are many outputs belonging to different clusters with divergent meanings. We do clustering in Section 3.2 in order to map outputs to semantic space and compute inter-cluster similarity and intra-cluster similarity separately.

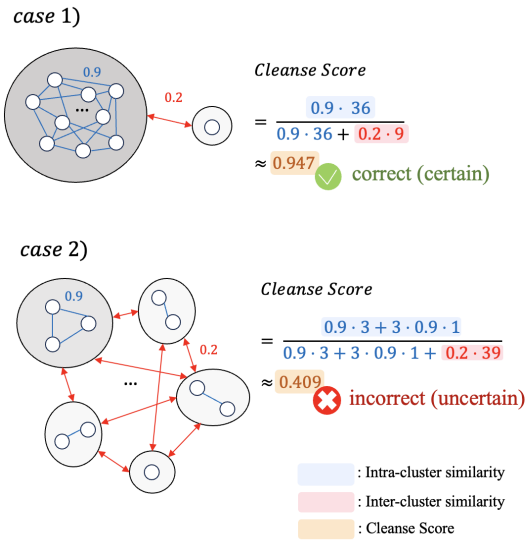


Figure 3: Case 1 has a small number of clusters, resulting a high proportion of the intra-cluster similarity in the total similarity. This case will be classified as correct as Cleanse Score is sufficiently high as 0.947, indicating low uncertainty. However, in Case 2, the proportion of the intra-cluster similarity in the total similarity is low at 0.409, so this case will be determined to be incorrect with high uncertainty.

We subtract the proportion of inter-cluster similarity in the total similarity from 1, which is the total proportion. Eq. 3 represents how to compute Cleanse Score using two types of similarities. There are two cases in Figure 3, which shows how does Cleanse Score work effectively and clearly in quantifying consistency.

$$\begin{aligned}\text{Cleanse Score} &= 1 - \frac{\text{inter-cluster sim.}}{\text{total sim.}} \\ &= \frac{\text{intra-cluster sim.}}{\text{total sim.}}\end{aligned} \tag{3}$$

4 Experiment

4.1 Experimental setups

Datasets. We use two representative question-answering datasets, SQuAD (Rajpurkar et al., 2016) and CoQA (Reddy et al., 2019). SQuAD (20.92) has longer ground truth answer spans than CoQA (13.67) when we compute the average of the length of golden answer for each dataset in our experiment. We follow the prompt setting of SQuAD as presented by Chen et al. (2024) and that of CoQA as described by Lin et al. (2023).

Models. We conduct experiments by varying the model in terms of its size, version, and optimized method. We utilize four off-the-shelf models, LLaMA-7B (Touvron et al., 2023a), LLaMA-13B (Touvron et al., 2023a), LLaMA2-7B (Touvron et al., 2023b), and Mistral-7B (Jiang et al., 2023).

Baselines. We compare the performance of Cleanse Score to four baeslines. **Perplexity** (Ren et al., 2023) measures the total uncertainty for generated sequence using the uncertainty of each token which consists of the sequence. **Length-normalized entropy (LN-entropy)** (Malinin and Gales, 2020) is similar to perplexity, but it reduces the bias in quantifying uncertainty by normalizing the joint log-probabilities with its sequence length. **Lexical similarity** (Lin et al., 2022b) is the average similarities between the answers which are measured with Rouge-L (Lin, 2004). **Cosine score**, computed as Eq. 4 in our study, serves as a baseline to verify that incorporating inter-cluster similarity as a penalty helps clarify the boundary between certain and uncertain answers, thereby improving uncertainty estimation performance.

$$\text{cosine score} = \frac{2}{K(K-1)} \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} \text{cosine}(\mathrm{e_i}, \mathrm{e_j}) \tag{4}$$

Correctness measure. We use Rouge-L (Lin, 2004) as the correctness measure which determines whether the generation of LLM is correct or not,

Model		LLaMA-7B		LLaMA-13B		LLaMA2-7B		Mistral-7B	
Dataset		SQuAD	CoQA	SQuAD	CoQA	SQuAD	CoQA	SQuAD	CoQA
Perplexity	AUC	60.2	66.1	61.4	63.6	63.8	62.2	53.3	57.3
(token-level)	PCC	19.3	27.4	21.8	27.0	25.5	24.3	13.0	21.7
LN-Entropy	AUC	72.3	71.6	74.6	70.8	74.2	70.5	59.3	61.7
(token-level)	PCC	38.9	35.5	43.6	37.1	42.8	34.7	14.8	24.6
Lexical Similarity	AUC	76.9	76.1	78.9	75.6	80.4	76.2	69.0	74.9
(token-level)	PCC	51.2	47.7	54.4	49.1	57.4	48.6	31.4	43.2
Cosine Score	AUC	79.6	78.5	81.1	77.7	82.1	79.3	65.9	74.1
(sentence-level)	PCC	54.7	**48.4**	57.8	49.3	59.7	**50.6**	29.1	41.3
Cleanse Score	AUC	**81.7**	**79.4**	**82.8**	**79.6**	**83.0**	**80.1**	**75.9**	**80.2**
(sentence-level)	PCC	**56.4**	47.6	**59.6**	**50.7**	**61.0**	49.7	**41.6**	**47.2**

Table 1: Hallucination detection performance for four models and two question-answering datasets. AUROC (AUC) and PCC are utilized to evaluate the performance of four baselines and Cleanse Score. We use Rouge-L threshold as 0.7 and deberta-nli-v3-base as a clustering model. Token-level indicates that corresponding metric estimates uncertainty based on token-probability or lexical form of generations. Sentence-level indicates that corresponding metric utilizes sentence-level embedding in computing uncertainty. Bolded values indicate the highest scores.

comparing it with the ground truth answer. We set the threshold as 0.7, which means only generation s is considered to be correct if s satisfies $\mathcal{L}(\mathrm{s}, \mathrm{s'}) = 1_{\text{Rouge-L}(s,s')>0.7}$ for the ground truth answer s'. We adjust this threshold from 0.5 to 0.9 in our further experiment to demonstrate the general capability of Cleanse Score.

Evaluation measure. We utilize two evaluation measures to evaluate the uncertainty estimation performance of four baselines and Cleanse Score. We use Area Under the Receiver Operating Characteristic Curve (AUROC) and Pearson Correlation Coefficient (PCC). AUROC is a performance metric for binary classifiers, allowing it to assess whether an uncertainty estimation metric effectively distinguishes between correct and incorrect generations. PCC measures the correlation between the Rouge-L score and the consistency level computed by each metric. Higher AUROC and PCC indicate better performance.

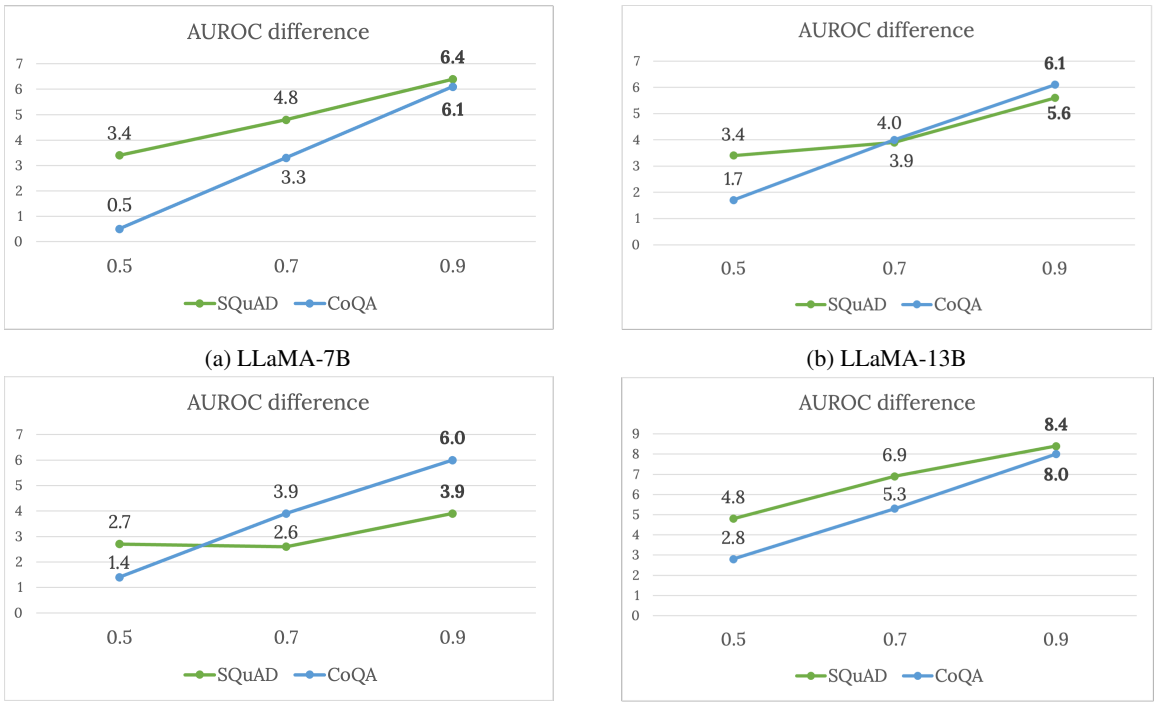
4.2 Main results

Effectiveness of Cleanse. As shown in Table 1, Cleanse Score outperforms all four baselines across LLaMA models and Mistral-7B on the SQuAD and CoQA datasets when evaluated using AUROC and PCC. Cleanse Score consistently achieves the highest AUROC, with a particularly large margin in the Mistral-7B settings. In the Mistral-7B model, Cleanse Score surpasses lexical similarity–the second highest performing baseline in Mistral-7B–by 6.9% in SQuAD and 5.3% in CoQA. There is a tendency that the performance of Cleanse Score improves in LLaMA-13B and LLaMA2-7B than LLaMA-7B and Mistral-7B.

On average, cosine score and Cleanse Score, which both leverage sentence-level embeddings, show better performance than the baselines based on token-probability or lexical similarity. This result supports our discussion in the previous section, demonstrating that prioritizing semantic aspect over lexical aspect is a reasonable approach in determining consistency between texts.

Additionally, in Table 1, Cleanse Score outperforms cosine score in all cases when evaluated with AUROC and in most cases when evaluated with PCC. Through this result, we demonstrate that our core intuition—clustering multiple outputs and using the inter-cluster similarity as a penalty term—successfully enhances uncertainty detection performance when applied to Cleanse Score. Interpreting intra-cluster similarity and inter-cluster similarity as the degree of consistency and inconsistency respectively enables us to filter hallucinated cases better than simply by averaging total similarities.

Advantage of Cleanse: Superior hallucination detection capability even under strict conditions In Figure 4, we compute the AUROC difference between Cleanse Score and lexical similarity, which achieves the highest performance among token-level approaches. The AUROC differences increase as the threshold of Rouge-L becomes harder, regardless of the model type and dataset. In particular, the differences in LLaMA-7B in Fig-



(a) LLaMA-7B (b) LLaMA-13B

(c) LLaMA2-7B (d) Mistral-7B

Figure 4: AUROC difference between Cleanse Score and lexical similarity across four models on two QA datasets, varying the correctness measure threshold between 0.5 to 0.9. The highest values are in bold.

ure 4a and Mistral-7B in Figure 4d across both SQuAD/CoQA datasets settings are significant, achieving 6.4%/6.1% and 8.4%/8.0%. A detailed analysis of the results shown in Table 3 in Appendix reveals that, except for the case of Mistral-7B on the SQuAD dataset, the performance of lexical similarity either remains the same or decreases as the Rouge-L threshold increases, whereas the performance of Cleanse Score consistently improves. In the case of Mistral-7B on the SQuAD dataset, the performance of lexical similarity also increases with a higher threshold, but the improvement margin of Cleanse Score is significantly greater that of lexical similarity. Here, increasing the threshold means that the correctness measure becomes more rigorous and aligns more closely with human evaluation. These settings are crucial for certain NLP tasks that require a precise and accurate correctness metric. The results demonstrate that Cleanse Score is robustly applicable in such strict environments such as question-answering and translation tasks.

Clustering model comparison. The choice of clustering model is one of the most important factors in our study as shown in Figure 5. We compare four fine-tuned NLI model, deberta-large-mnli (He et al., 2020), roberta-large-mnli (Liu et al., 2019), nli-deberta-v3-base (He et al., 2021) and nli-deberta-v3-large (He et al., 2021) to find the optimal clustering model.

We identify the performance of each clustering model in two ways. First, we compare AUROC when each clustering model is applied to Cleanse Score. Table 2 shows that AUROC scores of Cleanse Score using nli-deberta-v3-base are slightly better than when using other clustering models. Besides this result, inspired by the intuition from Kuhn et al. (2023), we conduct additional comparison using the concept mentioned in Section 3.3. In Figure 5, a clustering model that forms a small number of clusters for correct answers and a large number of clusters for incorrect answers can clarify between certain and uncertain outputs, leading Cleanse Score to predict correct and incorrect labels better. Based on this idea, the difference in the number of clusters formed in incorrect generations and correct generations can serve as a metric for evaluating the performance of clustering. The larger the difference is, the better the model clusters. We calculate the difference between the average number of clusters for correct and incorrect generations and show them in parentheses in Table 2. The overall differences for nli-deberta-v3-base are the largest, confirming again that using nli-deberta-v3-base as a clustering model

Clustering Model		deberta-large-mnli	roberta-large-mnli	nli-deberta-v3-base	nli-deberta-v3-large
LLaMA-7B	SQuAD	81.3 (2.71)	80.7 (2.54)	**81.7 (2.78)**	81.2 (2.63)
	CoQA	79.0 (2.49)	78.5 (2.40)	**79.4 (2.55)**	**79.4** (2.45)
LLaMA-13B	SQuAD	82.5 (2.96)	82.3 (2.78)	**82.8 (3.03)**	82.6 (2.88)
	CoQA	79.3 (2.47)	79.0 (2.36)	**79.6 (2.53)**	79.5 (2.51)
LLaMA2-7B	SQuAD	82.7 (2.92)	82.2 (2.73)	**83.0 (2.99)**	82.7 (2.86)
	CoQA	79.7 (2.52)	79.4 (2.43)	80.1 **(2.60)**	**80.2** (2.57)
Mistral-7B	SQuAD	75.2 (1.84)	74.2 (1.59)	**75.9 (1.92)**	74.9 (1.75)
	CoQA	80.0 (2.57)	79.4 (2.45)	**80.2 (2.63)**	79.8 (2.55)

Table 2: The results of the Cleanse Score performance comparison, measured by AUROC and the difference between the average number of clusters of correct and incorrect answers across four distinct clustering techniques when applied to the methodology (the latter is shown in parentheses). We set Rouge-L threshold as 0.7. Bold values are the highest.

outperforms other models.

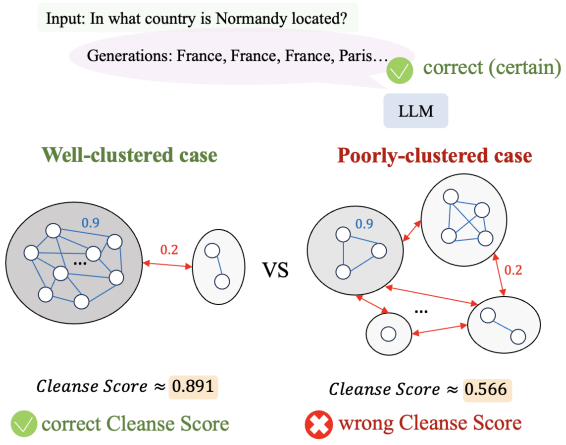


Figure 5: The illustration that shows the importance of clustering in our approach. For the same query that the model answers correctly, a well-clustered case results in few clusters, leading to an accurate Cleanse score. In contrast, a poorly-clustered case forms a few scattered clusters which yield an incorrect Cleanse score. This demonstrates that having few clusters for correct answers and a few clusters for wrong answers is advantageous for clearer hallucination detection.

5 Conclusion

Uncertainty estimation is one of the main solutions in detecting hallucination and prevent it from becoming critical problem in constructing reliable and trustworthy LLMs. We propose Cleanse, which clusters the outputs and computes the proportion of the intra-cluster similarity in the total similarity to quantify the consistency. As a result, filtering inter-cluster similarity as the inconsistency term helps to classify certain and uncertain generations effectively so that Cleanse perform better than the other existing approaches. Also, we found that Cleanse works well even under various correctness measure settings, which indicates Cleanse is appropriate to detecting uncertainty in diverse NLP tasks. Additionally, by conducting further experiments, we could identify a clustering model that outperforms than the others, thereby enhancing the performance of Cleanse.

Limitations

This approach is limited to white-box LLM as it requires hidden embedding extracted directly from the model. However, the performance and usefulness of Cleanse is verified through several experiments, other vector embeddings of the outputs could be used instead of hidden embeddings from a model, thereby overcome this limitation.

References

Zahid Ansari, Mohammad Fazle Azeem, Waseem Ahmed, and A Vinaya Babu. 2015. Quantitative evaluation of performance and validity indices for clustering the web navigational sessions. *arXiv preprint arXiv:1507.03340*.

Amos Azaria and Tom Mitchell. 2023. The internal state of an llm knows when it's lying. *arXiv preprint arXiv:2304.13734*.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. 2024. Inside: Llms' internal states retain the power of hallucination detection. *arXiv preprint arXiv:2402.03744*.

Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. 2023. Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models. *arXiv preprint arXiv:2307.01379*.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. Ragas: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158.

Xiang Gao, Jiaxin Zhang, Lalla Mouatadid, and Kamalika Das. 2024. Spuq: Perturbation-based uncertainty quantification for large language models. *arXiv preprint arXiv:2403.02509*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Ziwei Ji, Delong Chen, Etsuko Ishii, Samuel Cahyawijaya, Yejin Bang, Bryan Wilie, and Pascale Fung. 2024. Llm internal states reveal hallucination risk faced with a query. *arXiv preprint arXiv:2407.03282*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022a. Teaching models to express their uncertainty in words. *arXiv preprint arXiv:2205.14334*.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2023. Generating with confidence: Uncertainty quantification for black-box large language models. *arXiv preprint arXiv:2305.19187*.

Zi Lin, Jeremiah Zhe Liu, and Jingbo Shang. 2022b. Towards collaborative neural-symbolic graph semantic parsing via uncertainty. *Findings of the Association for Computational Linguistics: ACL 2022*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Andrey Malinin and Mark Gales. 2020. Uncertainty estimation in autoregressive structured prediction. *arXiv preprint arXiv:2002.07650*.

Ella Rabinovich, Samuel Ackerman, Orna Raz, Eitan Farchi, and Ateret Anaby-Tavor. 2023. Predicting question-answering performance of large language models through semantic consistency. *arXiv preprint arXiv:2311.01152*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Jie Ren, Jiaming Luo, Yao Zhao, Kundan Krishna, Mohammad Saleh, Balaji Lakshminarayanan, and Peter J. Liu. 2023. Out-of-distribution detection and selective generation for conditional language models. *Preprint*, arXiv:2209.15558.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023. Siren's song in the ai ocean: A survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*, 2(5).

Appendix

A Additional Experiments

Model		LLaMA-7B		LLaMA-13B		LLaMA2-7B		Mistral-7B	
Dataset		SQuAD	CoQA	SQuAD	CoQA	SQuAD	CoQA	SQuAD	CoQA
Lexical Similarity	0.5	76.8	76.9	79.1	77.1	80.2	77.5	67.6	74.9
	0.7	76.9	76.1	78.9	75.6	80.4	76.2	69.0	74.9
	0.9	75.7	74.9	77.1	74.5	79.8	74.8	70.7	73.6
Cleanse Score	0.5	80.2	77.4	82.5	78.8	82.9	78.9	72.4	77.7
	0.7	81.7	79.4	82.8	79.6	83.0	80.1	75.9	80.2
	0.9	82.1	81.0	82.7	80.6	83.7	80.8	79.1	81.6

Table 3: Pattern of AUROC performance changes in lexical similarity and Cleanse Score as Rouge-L threshold varies across 0.5, 0.7, and 0.9. We use deberta-nli-v3-base for clustering model.