EWHA WOMANS UNIVERSITY

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

Minsuh Joo, Hyunsoo Cho

{judyjoo21, chohyunsoo}@ewha.ac.kr

ACL 2025 VIENNA

JULY 27 - AUGUST 1

GEM WORKSHOP AT ACL 2025

Overview

Problem

- Hallucination in LLMs where LLMs generate inaccurate responses
- Mitigating hallucination in QA tasks where precise and verifiable responses are required remains as a critical issue.

Overview

Motivated Approach

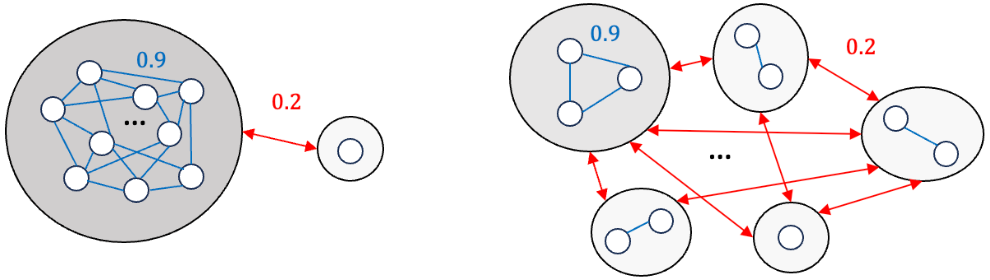
- Uncertainty estimation enables users identify potentially unreliable responses (Lin et al., 2022a) which contributes to building safe and reliable LLMs.

- Based on semantic equivalence where responses are consistent as long as their semantics are the same despite their different syntactic forms, we evaluate the uncertainty of the response through its semantic consistency.

Overview

Method

- We propose **Cl**ust**e**ring-b**a**sed Sema**n**tic Con**s**ist**e**ncy (Cleanse), which quantifies the uncertainty with the proportion of the intra-cluster consistency (similarity) in the total consistency.



Cleanse Score

$$= \frac{0.9 \cdot 36}{0.9 \cdot 36 + 0.2 \cdot 9} \approx 0.947$$

correct (certain)

Cleanse Score

$$= \frac{0.9 \cdot 3 + 3 \cdot 0.9 \cdot 1}{0.9 \cdot 3 + 3 \cdot 0.9 \cdot 1 + 0.2 \cdot 39} \approx 0.409$$

incorrect (uncertain)

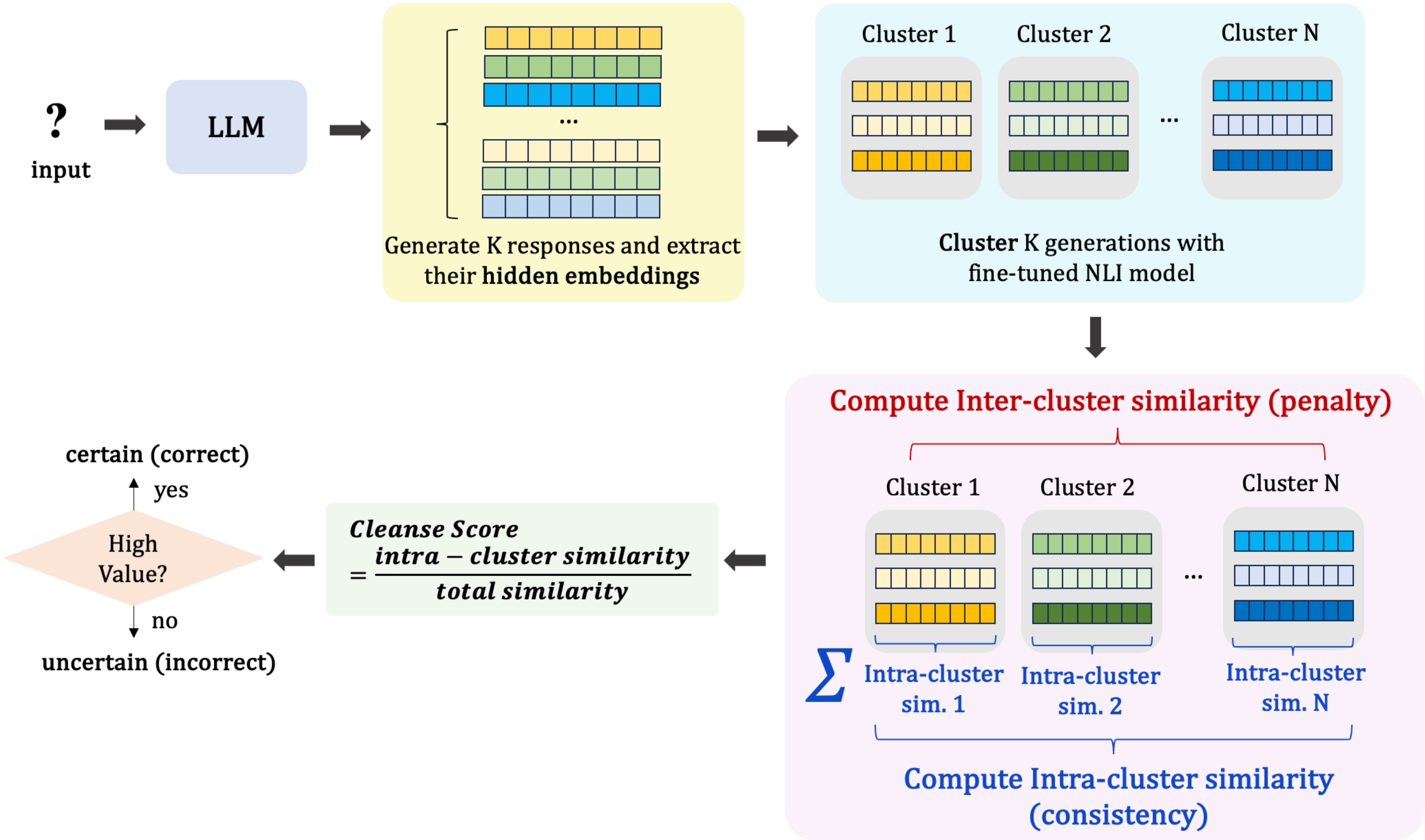
: Intra-cluster similarity

: Inter-cluster similarity

: Cleanse Score

Method

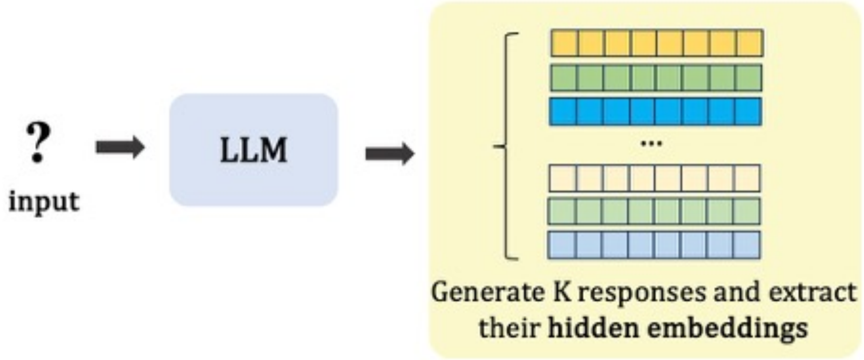
Pipeline



Method

Step 1: Generate multiple responses for a query and extract their hidden embeddings

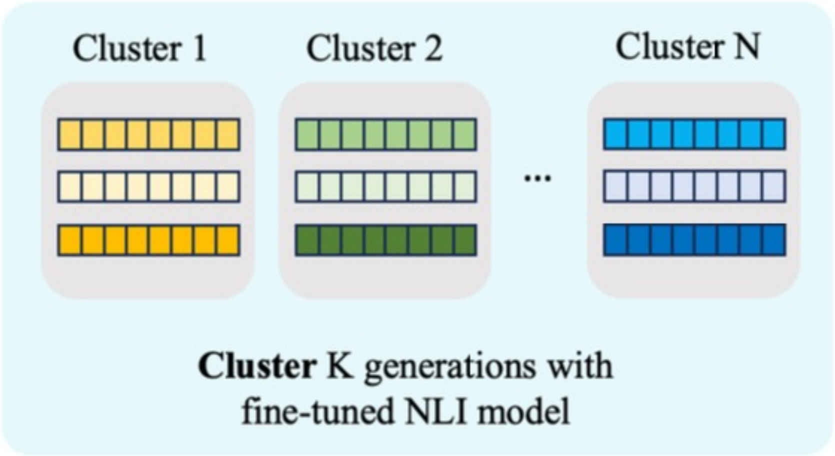
- We extract the <u>last token embedding in the middle layer of LLM</u> as the hidden embedding of the output, as prior work suggests it captures semantic information effectively (Azaria and Mitchell, 2023).



Method

Step 2: Cluster responses with fine-tuned NLI model

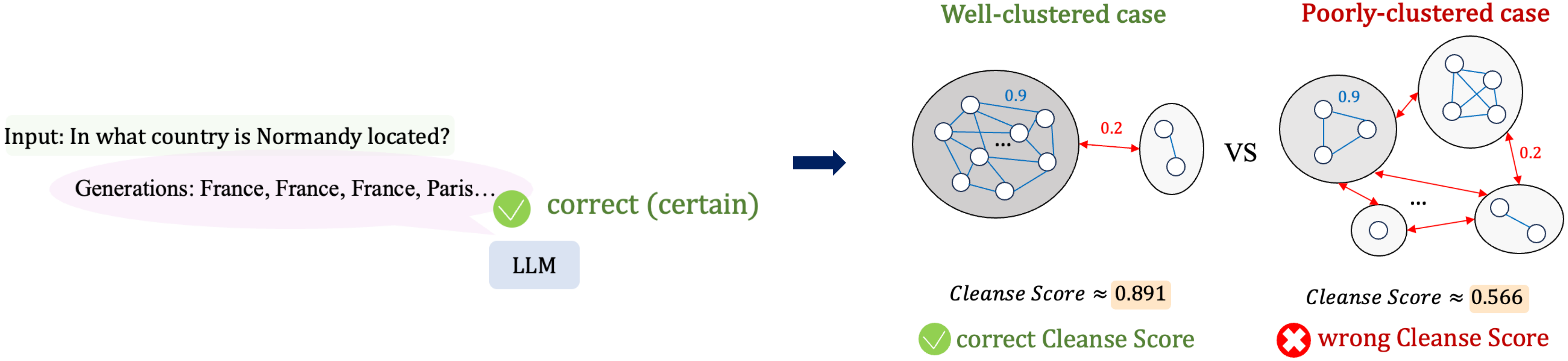
- We clustered outputs based on their semantic information by leveraging bi-directional entailment clustering algorithm (Kuhn et al., 2023) and fine-tuned NLI model.



Method

Step 2: Cluster responses with fine-tuned NLI model

- We chose the clustering model based on intuition that having few clusters for correct case and a few clusters for wrong case is advantageous for clearer hallucination detection.



Method

Step 2: Cluster responses with fine-tuned NLI model

- We utilized nli-deberta-v3-base (He et al., 2021) as a clustering model, which outperforms other models when evaluated with AUROC and the difference between the number of clusters formed for incorrect case and correct case.

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)

Method

Step 3: Compute inter/intra-cluster similarity based on the clustering result and quantify the uncertainty with Cleanse Score

Intra-cluster similarity

: the degree of consistency which contributes to the high consistency between outputs.

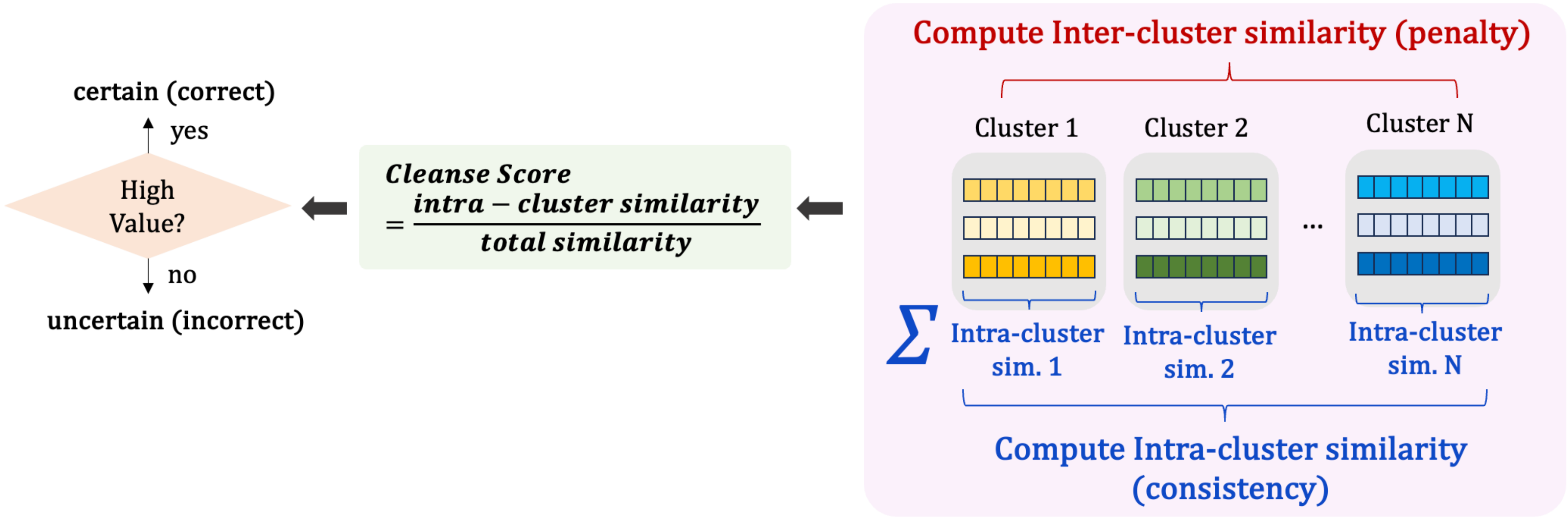
Inter-cluster similarity

: the penalty for the consistency which indicates the degree of divergence between semantics of outputs .

$$\text{Cleanse Score} = \frac{Intra{-}cluster\ sim.}{Total\ sim. = Intra{-}cluster\ sim. + Inter{-}cluster\ sim.}$$

Method

Step 3: Compute inter/intra-cluster similarity based on the clustering result and quantify the uncertainty with Cleanse Score



Results

Effectiveness of Cleanse

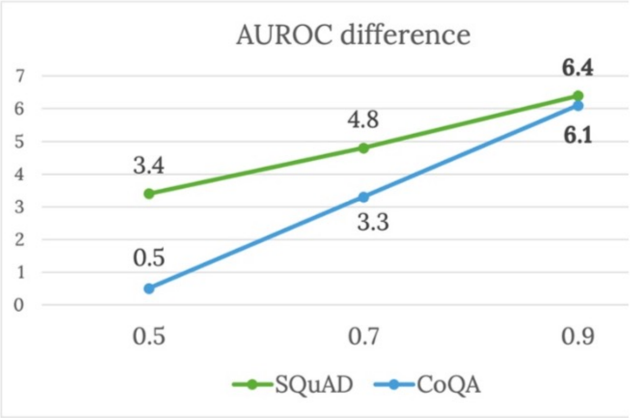
- Our core intuition — clustering multiple outputs and using the inter-cluster similarity as a penalty term — successfully enhances the performance when applied to Cleanse, compared to other baselines.

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

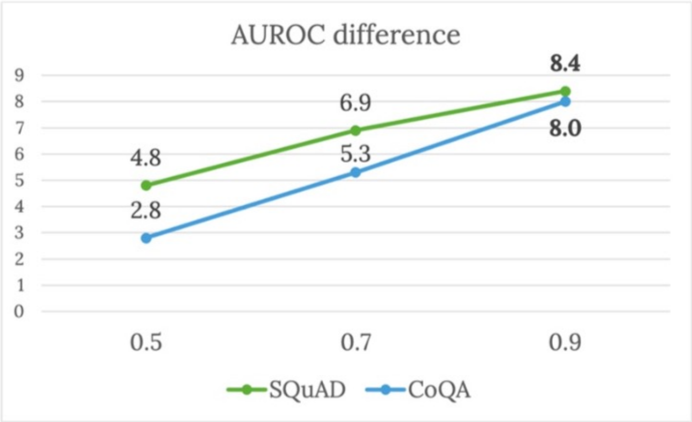
Results

Superior hallucination detection capability under strict settings

- The performance gap between Cleanse Score and lexical similarity increases as the threshold of rouge-L increases, which indicates that the Cleanse Score is robustly applicable in strict environments such as question-answering and translation tasks.



(a) LLaMA-7B



(d) Mistral-7B

Conclusion

- We propose Cleanse, which clusters the outputs and computes the proportion of the intra-cluster similarity in the total similarity to quantify the consistency.

- We showed that 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.

Thank You