CSS: Contrastive Semantic Similarity for Uncertainty Quantification of LLMs

Shuang Ao, Stefan Reuger, Advaith Siddharthan
Knowledge Media Institute (KMi),
The Open University, UK

Motivation

- Uncertainty Quantification (UQ) in LLMs remains open challenges.
- Information consistency reflects uncertainty of LLMs generations.
- Previous studies employ NLI logits to represent semantic similarity for text-pairs. However, NLI logits represents the probability of classes rather than feature information for text-pairs.
- A better method to extract insightful semantic similarity features between text-pairs is required.

Contributions

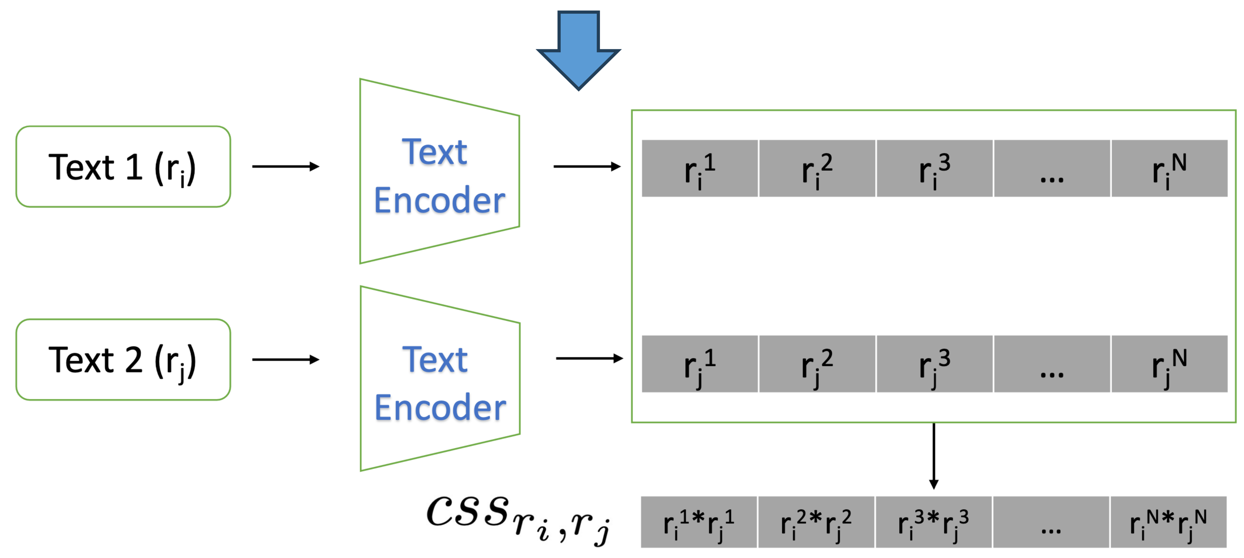
- We introduce novel Contrastive Semantic Similarity (CSS) to extract semantic relations between text-pairs, for estimate uncertainty in generations of LLMs.
- We modify the CLIP text encoder to obtain text-text pairs semantic similarities, then employ spectral clustering for UQ.
- With extensive experiments in three benchmark QA datasets, our proposed methods outperform SOTA techniques.
- Our ablation studies show: (i) our proposed CSS contains more semantic information than NLI logits; (ii) CSS obtained from our text-text encoder of CLIP is superior to regular language models; (iii) our CSS enhances selective natural language generation (NLG).

Proposed Methods

Given one question, the LLM generate m responses (m = 20).

Text-1, Text-2, Text-3, … … , Text-20

Extracting CSS features for text-pairs.



Graph Laplacian for clustering generations with semantic similarities

Symmetric weighted adjacency matrix: W^{css}

Uncertainty with Degree Matrix: $U_{\mathrm{Deg}}^{css} = \mathrm{trace}(m - D^{css})/m^2$

CSS graph Laplacian: $L^{css} := D^{css} - W^{css}$

Ascending order eigenvalues: $\lambda_1^{css} \leq \lambda_2^{css} \leq \ldots \leq \lambda_n^{css}$

Ascending order eigenvectors: $v_1^{css}, v_2^{css} \ldots, v_n^{css}$

Uncertainty with Eigenvalues: $U_{\mathrm{Eig}}^{css} = \sum_{k=1}^{m} \max(0, 1 - \lambda_k^{css})$

Uncertainty with Eccentricity: $U_{\mathrm{Ecc}}^{css} = \left\| \left[\mathbf{e}_1^{\mathbf{css}\prime\top}, \ldots, \mathbf{e}_m^{\mathbf{css}\prime\top} \right] \right\|_2$

Obtain 3 types of uncertainty scores for each generation and m generations, namely degree matrix (Deg), Eigenvalue (EigV) and Eccentricity (Ecc).

- Each generation is assigned with uncertainty score;
- For the accuracy of generations, we utilize Rouge-L > 0.3 and GPT score > 0.7 as the criteria for correctness;
- With the uncertainty score as the threshold, we calculate the AUROC and AUARC score, the higher the better.

★ Given one question, if all generations of LLM fall into similar semantic clusters, it indicate consistent information and lower uncertainty.

★ This consistency suggests that the LLM is well-trained on the concept and can be considered trustworthy.

Experiments

- Benchmark QA datasets: TriviaQA, CoQA, Natural Questions (NQ).
- LLMs: LLaMA, OPT, GPT (API).
- Each LLM generate sampled response for each dataset.
- Compared SOTA techniques: Semantically Distinct Answers (NumSem)[1], Lexical Similarity (LexiSim)[2], Graph Laplacian with NLI logits (L-GL)[3], Semantic Entropy (SE)[1], P(true)[4].
- White-box (WB method) require the access of predicted probability, which is not available for GPT generated responses.

Results

- With the evaluation of AUROC, our method (CSS) outperform SOTA methods across all LLaMA and OPT generations. Additionally, L-GL achieves superior results with GPT generations..
- With the evaluation of AUARC, our method obtain better results than other techniques, indicating improvement in uncertainty quantification for LLMs.

Results of AUROC with Rouge-L score as the correctness criterion.

Dataset		TriviaQA			CoQA			NQ		
Model		LLaMA	OPT	GPT	LLaMA	OPT	GPT	LLaMA	OPT	GPT
	NumSem	75.06	68.56	68.20	57.76	57.60	51.69	55.59	59.20	61.13
	LexiSim	77.63	76.48	81.13	75.72	76.40	68.70	76.72	73.90	71.65
L-GL	EigV	84.35	82.88	**83.40**	77.95	75.70	78.65	72.59	73.88	80.88
	Ecc	83.66	83.91	82.50	77.26	74.81	77.39	74.44	76.02	79.82
	Deg	84.52	83.36	82.93	77.53	75.85	78.76	74.01	74.75	**81.31**
WB	SE	74.39	81.54	–	74.55	71.25	–	69.50	74.61	–
	P(true)	55.12	41.64	–	55.14	52.67	–	52.52	47.92	–
Ours (CSS)	CSS-EigV	85.52	85.37	82.27	**78.78**	**77.19**	80.04	**76.08**	**77.08**	79.28
	CSS-Ecc	85.17	84.97	81.57	78.40	76.70	**80.40**	75.76	76.53	79.91
	CSS-Deg	**85.63**	**85.82**	81.77	78.68	76.95	79.12	75.81	77.25	80.01

Results of AUARC with Rouge-L score as the correctness criterion.

Dataset		TriviaQA			CoQA			NQ		
Model		LLaMA	OPT	GPT	LLaMA	OPT	GPT	LLaMA	OPT	GPT
	Acc	57.57	25.60	81.07	55.96	51.99	66.38	19.32	9.10	39.83
	Oracle	89.60	54.30	97.91	85.10	78.56	93.00	42.35	24.15	75.58
	NumSem	73.25	33.76	81.07	64.31	57.29	67.81	20.85	10.58	45.97
	LexiSim	78.98	46.72	87.47	79.09	**73.15**	80.39	35.74	17.77	58.01
L-GL	EigV	80.67	48.70	92.32	79.54	71.96	84.34	33.58	14.72	62.13
	Ecc	80.20	48.83	92.01	78.92	70.96	83.94	34.26	17.41	61.80
	Deg	80.71	49.00	92.24	79.12	71.83	84.22	34.23	17.49	62.42
WB	SE	74.09	47.90	–	77.65	67.46	–	28.97	16.62	–
	P(true)	61.85	20.93	–	61.75	58.32	–	20.19	8.27	–
Ours (CSS)	CSS-EigV	81.47	49.85	92.70	**81.92**	72.13	87.26	**36.80**	18.10	64.83
	CSS-Ecc	81.29	49.60	93.07	80.83	71.36	**87.34**	36.62	18.19	**65.04**
	CSS-Deg	**81.55**	**50.08**	**93.18**	81.17	73.18	87.02	36.67	**18.34**	64.87

Ablation Study

Compare Rouge-L and METEOR.

Evaluation Metric	Method	AUARC	AUROC
Rouge-L	L-GL	80.20	83.66
	Ours	81.29	85.17
METEOR	L-GL	80.32	83.79
	Ours	81.35	85.22

Compare CLIP and BERT.

Model	AUARC	AUROC
BERT	83.78	86.24
DeBERTa	83.62	86.53
Sentence-BERT	83.72	87.02
CLIP	**84.32**	**87.19**

Compare NLI logits and feature map.

		AUARC		AUROC	
		LLaMA	OPT	LLaMA	OPT
L-GL	EigV	83.52	50.54	84.90	86.09
	Ecc	83.64	50.42	86.43	86.86
	Deg	84.61	51.06	84.21	86.60
F-GL	EigV	83.54	50.48	84.95	85.92
	Ecc	83.62	51.62	86.53	86.95
	Deg	84.65	51.36	84.16	87.12

- METEOR and Rouge-L shows similar trend;
- Feature maps contains more insightful semantic information than NLI logits;
- Our method better extract contrastive features than BERT language encoder .

Conclusion

- We design Contrastive Semantic Similarity to extract semantic features between text-pairs, to enhancing UQ in LLMs;
- Our method has shown superiority to SOTA techniques via extensive experiments and ablation studies;
- Future work will focus on uncertainty calibration techniques and explore their applications in various domains.

[1] Kuhn, L., Gal, Y. and Farquhar, S., Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation. ICLR 2023.
[2] Lin, C.Y., 2004, July. Rouge: A package for automatic evaluation of summaries. In Text summarization branches out.
[3] Lin, Z., Trivedi, S. and Sun, J., Generating with Confidence: Uncertainty Quantification for Black-box Large Language Models. TMLR 2024.
[4] Kadavath, S., et.al, 2022. Language models (mostly) know what they know. arXiv.

Email: shuang.ao@open.ac.uk

Paper GitHub