

# SEMANTIC REFORMULATION ENTROPY FOR ROBUST HALLUCINATION DETECTION IN QA TASKS

Chaodong Tong<sup>\*,†</sup>, Qi Zhang<sup>‡</sup>, Lei Jiang<sup>\*</sup>, Yanbing Liu<sup>\*,†</sup>, Nannan Sun<sup>\*</sup>, Wei Li<sup>‡</sup>

<sup>\*</sup>Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China

<sup>†</sup>School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China

<sup>‡</sup>China Industrial Control Systems Cyber Emergency Response Team, Beijing, China

{tongchaodong, jianglei, liuyanbing, sunnannan}@iie.ac.cn, {bonniezhangqi, ai4cics}@126.com

## ABSTRACT

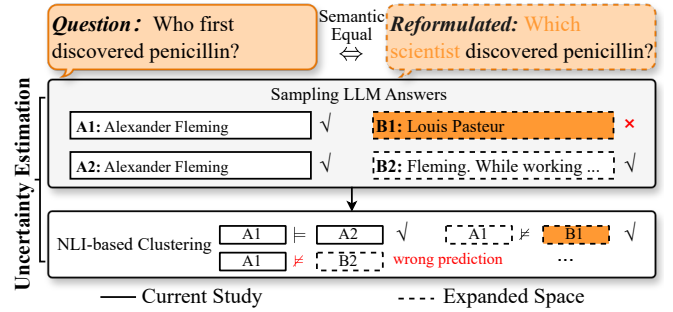
Reliable question answering with large language models (LLMs) is challenged by hallucinations, fluent but factually incorrect outputs arising from epistemic uncertainty. Existing entropy-based semantic-level uncertainty estimation methods are limited by sampling noise and unstable clustering of variable-length answers. We propose Semantic Reformulation Entropy (SRE), which improves uncertainty estimation in two ways. First, input-side semantic reformulations produce faithful paraphrases, expand the estimation space, and reduce biases from superficial decoder tendencies. Second, progressive, energy-based hybrid clustering stabilizes semantic grouping. Experiments on SQuAD and TriviaQA show that SRE outperforms strong baselines, providing more robust and generalizable hallucination detection. These results demonstrate that combining input diversification with multi-signal clustering substantially enhances semantic-level uncertainty estimation.

**Index Terms**— Large language models, Hallucination detection, Uncertainty estimation, Semantic entropy, Hybrid semantic clustering

## 1. INTRODUCTION

Large Language Models (LLMs) excel at tasks such as question answering (QA), summarization, and dialogue [1, 2], but often produce fluent yet factually incorrect outputs, known as *hallucinations* [3]. In QA, these are particularly problematic as users expect knowledge-grounded answers [4, 5], often arising from *epistemic uncertainty* [6, 7]. This motivates leveraging model-internal uncertainty to detect hallucinations.

Existing detection methods can be broadly categorized by the type of uncertainty signal they exploit: (i) *likelihood-based* signals, which rely on token-level probabilities [8, 9]; (ii) *representation-based* signals, capturing hidden-state variability [10]; (iii) *prediction-based* signals, such as model-



**Fig. 1:** Key limitations in QA uncertainty estimation: limited sampling space and fragile NLI-based clustering.

assigned truth probabilities  $P(\text{True})$  [5]; and (iv) *semantic-level* signals, exemplified by Semantic Entropy (SE) [5], which quantify meaning variability across multiple high-temperature sampled outputs.

Semantic-level signals are promising, reflecting meaning variability in LLM outputs while requiring no internal model access. However, existing methods face two key limitations (Fig. 1): (a) they rely on high-temperature sampling from a single input, making uncertainty estimates prone to biases from superficial decoder tendencies [5, 11]; and (b) clustering based solely on NLI is fragile, particularly for variable-length or semantically complex outputs [12]. Some recent works attempt to improve robustness using token-level uncertainty [11] or semantic perturbations [13], but they do not systematically address these fundamental limitations.

In this work, we propose **Semantic Reformulation Entropy (SRE)**, a natural extension of semantic entropy aimed at more reliable estimation of epistemic uncertainty. SRE enriches both the *input* and *output* sides of uncertainty estimation. On the input side, **semantic reformulation (SR)** generates faithful paraphrases, expanding the estimation space and mitigating bias from superficial decoder patterns. On the output side, **hybrid semantic clustering (HSC)** combines exact matches, embedding similarity, and bidirectional NLI via progressive, energy-based clustering, stabilizing assignments for

Corresponding author: sunnannan@iie.ac.cn

variable-length or semantically complex outputs and supporting more reliable entropy estimation. By combining SR and HSC, SRE provides a stable foundation for capturing genuine epistemic uncertainty at the semantic level in LLMs.

Our contributions are summarized as follows:

- We introduce SRE, which leverages input-side reformulation and a progressive, energy-based multi-signal clustering framework named HSC to help separate epistemic uncertainty from superficial variability and produce robust semantic clusters.
- We empirically show that SRE outperforms strong baselines on QA benchmarks (SQuAD, TriviaQA), providing more reliable and generalizable hallucination detection.

## 2. METHODOLOGY

### 2.1. Preliminaries

We detect hallucinations via model-internal uncertainty. SE [5] measures output dispersion over clusters  $C_\ell$ :

$$\mathcal{H}_{SE}(x, M) = - \sum_{\ell=1}^L p(C_\ell | x, M) \log p(C_\ell | x, M), \quad (1)$$

where  $p(C_\ell | x, M)$  is the fraction or summed probability of outputs in  $C_\ell$  [5, 14]. We extend SE with SRE (Fig. 2), adding input reformulations and hybrid semantic clustering for more robust uncertainty estimation.

### 2.2. Semantic Reformulation

SE captures only output dispersion and is prone to superficial decoder biases. To address this, we generate *semantic reformulations*  $\mathcal{R}(x) = \{r_1, \dots, r_N\}$  via few-shot prompting, assessing model consistency across semantically equivalent queries while ensuring diversity and fidelity [13] (see Fig. 2a). Each  $r_i$  is scored against  $x$  using cosine similarity:

$$s(r_i, x) = \frac{\langle e(r_i), e(x) \rangle}{\|e(r_i)\| \|e(x)\|}, \quad (2)$$

where  $e(\cdot)$  is a sentence embedding [15]. Candidates with  $s(r_i, x) \notin [\tau_{\min}, \tau_{\max}]$  or near-duplicates are removed, yielding  $\mathcal{R}^*(x)$ ; if too few remain, additional reformulations are sampled to meet the desired count.

### 2.3. Answer Sampling Process

For each  $r_i \in \mathcal{R}^*(x)$ , outputs are sampled at high temperature with the prompt  $\text{Prompt}(r_i) = \text{Few-shot examples} \parallel r_i$  to encourage diverse yet faithful responses.

**Gold Labels.** Following [5], gold labels are determined by comparing a single low-temperature output  $y_i$  to the reference:  $y_{\text{gold}} = 1$  if  $y_i$  matches the reference, and 0 otherwise.

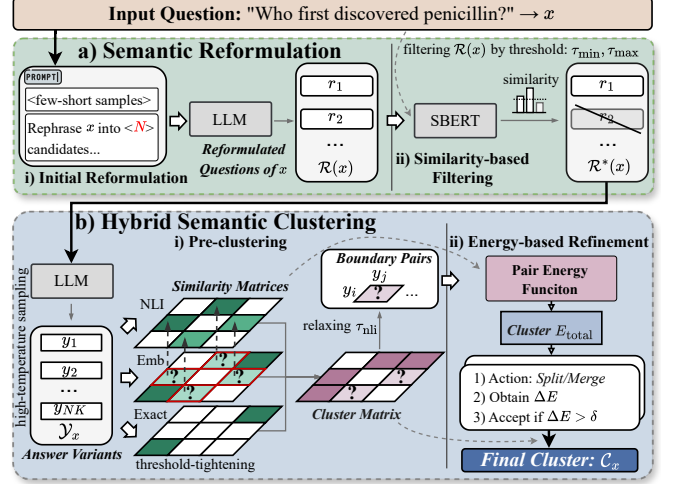


Fig. 2: SRE: reformulating, sampling, and clustering.

Multiple low-temperature samples yield consistent results, so we adopt this single-sample strategy.

### 2.4. Hybrid Semantic Clustering

SRE reliability depends on clustering quality. HSC clusters an input set  $\mathcal{Y}_x$  of size  $NK$  ( $N$  reformulations,  $K$  samples each) in two stages: pre-clustering and energy-based refinement (see Fig. 2b).

**Pre-clustering (exact, embedding, NLI).** Outputs identical to each other (exact matches) are grouped directly. Remaining pairs with cosine similarity  $\text{sim}(e(y_i), e(y_j)) > \tau_{\text{emb}}$  are merged. For unmerged pairs, a *DeBERTa-v2-xlarge-MNLI* model [16] predicts entailment  $p_{ij}^{\text{ent}}$  and contradiction  $p_{ij}^{\text{contra}}$ , and a merging score is defined:

$$s_{ij}^{\text{NLI}} = \lambda \frac{p_{ij}^{\text{ent}} + p_{ji}^{\text{ent}}}{2} \mathbf{1} \left[ \max(p_{ij}^{\text{contra}}, p_{ji}^{\text{contra}}) < \tau_{\text{contra}} \right] + (1 - \lambda) \max(p_{ij}^{\text{ent}}, p_{ji}^{\text{ent}}), \quad (3)$$

merging pairs with  $s_{ij}^{\text{NLI}} \geq \tau_{\text{nli}}$ . Strict mode ( $\lambda = 1$ ) enforces contradiction filtering; loose mode ( $\lambda = 0$ ) uses only maximal entailment.

**Energy-based Boundary Refinement.** We refine *boundary pairs*, i.e., pairs with entailment near  $\tau_{\text{nli}}$  and contradiction below  $\tau_{\text{contra}}$ , to improve cluster structure. The pair energy is defined as:

$$E(i, j) = \mathbf{1}_{\text{same}}(i, j) \left[ 1 - (\alpha \text{sim}(y_i, y_j) + \beta \text{entail}(y_i, y_j)) \right] + (1 - \mathbf{1}_{\text{same}}(i, j)) \left[ 1 - \gamma \text{contra}(y_i, y_j) \right], \quad (4)$$

where  $\mathbf{1}_{\text{same}}(i, j) = 1$  if  $y_i$  and  $y_j$  are in the same cluster;  $\text{entail}(\cdot, \cdot)$  and  $\text{contra}(\cdot, \cdot)$  are NLI-based entailment and contradiction scores, and  $\alpha, \beta, \gamma$  weight their contributions (in our experiments,  $\alpha = 0.3, \beta = 0.7, \gamma = 0.7$ ).

The total clustering energy is computed as the sum of average intra-cluster and inter-cluster energies:

$$E_{\text{total}} = \frac{1}{|\mathcal{P}_{\text{intra}}|} \sum_{(i,j) \in \mathcal{P}_{\text{intra}}} E(i,j) + \frac{1}{|\mathcal{P}_{\text{inter}}|} \sum_{(i,j) \in \mathcal{P}_{\text{inter}}} E(i,j), \quad (5)$$

where  $\mathcal{P}_{\text{intra}}$  and  $\mathcal{P}_{\text{inter}}$  denote the sets of intra-cluster and inter-cluster pairs, respectively. For boundary pairs, we attempt split or merge operations and compute the total energy change:

$$\Delta E = E_{\text{total}}^{\text{old}} - E_{\text{total}}^{\text{new}}, \quad (6)$$

accepting the refinement if  $\Delta E > \delta$ , where we set  $\delta = 0.1$ .

This local greedy strategy efficiently corrects uncertain assignments while preserving confident clusters, ensuring stable and semantically coherent partitions.

**Progressive Integration and SRE Computation.** Outputs are clustered through a progressive integration of multiple signals, producing semantically coherent clusters  $\mathcal{C}_x = \{C_1, \dots, C_m\}$ . SRE is then defined as:

$$\mathcal{H}_{\text{SRE}}(x, M) = - \sum_{C \in \mathcal{C}_x} p_f(C) \log p_f(C), \quad (7)$$

where  $p_f(C)$  denotes the fraction of outputs in cluster  $C$ . By progressively refining clusters from coarse to fine, the method produces stable and robust clusters, providing a principled basis for entropy estimation.

### 3. EXPERIMENTS

#### 3.1. Datasets

We evaluate on SQuAD-v2 [17] (contextualized QA) and TriviaQA [18] (Table 1), sampling 500 validation instances each. To ensure representativeness, we compare context, question, and answer lengths, as well as the distribution of gold hallucination labels, with the full validation sets using Kolmogorov–Smirnov and Wasserstein distance tests [19], and repeat sampling to verify stability.

**Table 1:** Overview of QA datasets used in our experiments.

Dataset	Context	#Samples	Domain
SQuAD-v2	Full	142,192	Wikipedia
TriviaQA	Weak	15,368	Trivia & Wikipedia

#### 3.2. Implementation Details

We implement SRE with Llama3-8B-Instruct and Qwen3-14B on 8×RTX 3090 GPUs. SR generates  $N = 3$  paraphrases ( $\tau_{\text{min}} = 0.6$ ,  $\tau_{\text{max}} = 0.95$ ), sampling  $K = 8$  outputs ( $T = 0.8$ ) with up to 5 few-shot examples, while HSC combines exact match, embedding similarity ( $\tau_{\text{emb}} = 0.92$ ), and

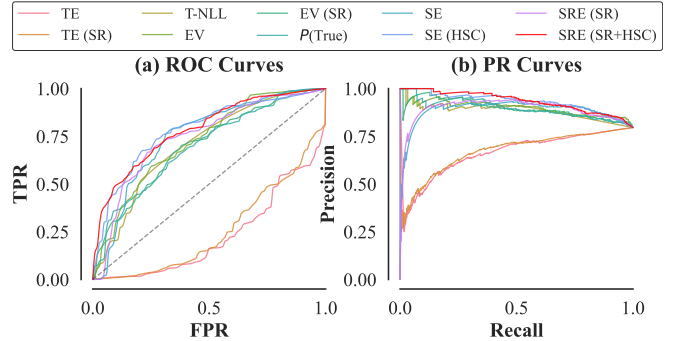
NLI entailment ( $\tau_{\text{nli}} = 0.8$ ); sensitivity of core parameters in SRE is analyzed in Section 3.5.

#### 3.3. Baselines and Evaluation Metrics

We compare our method against representative baselines: likelihood-based T-NLL [9] and token-level entropy TE [20], embedding-based EV [10], prediction-based  $P(\text{True})$  [5], and semantic-level SE [5]. Performance is evaluated using AUROC (AU), AURAC (AR) [5], and F1@best, averaged over three runs, for hallucination ranking, rejection, and classification under class imbalance.

#### 3.4. Main Experiments and Ablation Studies

We evaluate on SQuAD-v2 [17] and TriviaQA [18] under open-domain QA (no context) and extractive QA (with context) (Table 2). In open-domain QA, SRE outperforms baselines, achieving a maximum AUROC of 0.887 (+4% over the strongest baseline) and higher F1@Best. In extractive QA, SRE surpasses standard semantic entropy, though low AUROC reflects rare hallucinations. Performance is consistent across datasets and model variants.



**Fig. 3:** ROC (a) and PR (b) curves of all methods on SQuAD (no context) with Llama3-8B.

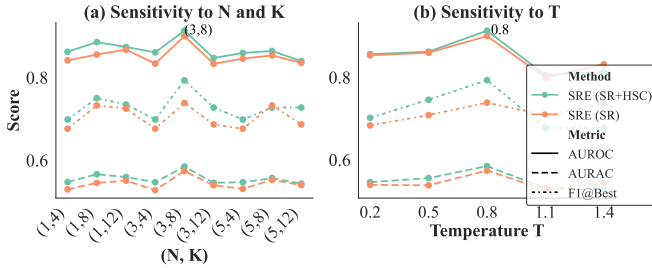
Ablations show that SR provides modest gains, while HSC, including boundary pair refinement, drives most improvements (up to +16% AUROC). The refinement alleviates boundary ambiguities in coarse clusters; as it is greedy, the gains are moderate. ROC and PR curves (Fig. 3) indicate SRE achieves clearer separation between hallucinated and faithful samples, providing a more nuanced uncertainty assessment. Overall, SR combined with HSC effectively captures model uncertainty across datasets and QA settings.

#### 3.5. Parameter Sensitivity Analysis

We evaluate SRE on 500 TriviaQA samples with Llama3-8B (Fig. 4). Performance depends on the number of reformulations ( $N$ ), samples per reformulation ( $K$ ), and sampling temperature ( $T$ ). Best results are obtained with  $N = 3$ ,  $K = 8$ ,

**Table 2:** Hallucination detection on SQuAD (with/without context) and TriviaQA (without context). **Bold:** best, underline: second-best. Last row (HSC\*) shows results without refinement and is excluded from ranking.

Method	SQuAD (context)						SQuAD (no context)						TriviaQA (no context)					
	Llama3-8B			Qwen3-14B			Llama3-8B			Qwen3-14B			Llama3-8B			Qwen3-14B		
	AU	AR	F1@Best	AU	AR	F1@Best	AU	AR	F1@Best	AU	AR	F1@Best	AU	AR	F1@Best	AU	AR	F1@Best
T-NLL	0.602	0.211	0.304	0.593	0.116	0.233	0.654	0.861	0.887	0.558	0.753	0.837	0.622	0.390	0.460	0.595	0.478	0.561
TE	0.256	0.069	0.276	0.291	0.036	0.139	0.235	0.625	0.885	0.297	0.593	0.837	0.172	0.128	0.457	0.175	0.194	0.561
TE (SR)	0.275	0.069	0.280	0.320	0.040	0.142	0.263	0.644	0.885	0.274	0.554	0.837	0.186	0.124	0.457	0.171	0.178	0.561
EV	<b>0.755</b>	0.264	<b>0.443</b>	0.702	0.134	<b>0.330</b>	0.728	0.883	<b>0.902</b>	0.634	0.781	0.833	0.838	0.535	0.704	0.807	0.613	0.719
EV (SR)	<u>0.733</u>	0.255	0.400	<b>0.746</b>	0.121	0.271	0.709	0.881	0.896	0.649	0.788	0.835	0.842	0.526	0.667	0.780	0.568	0.689
$P(\text{True})$	0.576	0.195	0.309	0.650	0.102	0.206	0.703	0.880	0.890	0.719	0.843	0.838	0.864	0.560	0.701	0.849	0.662	0.702
SE	0.614	0.222	0.300	0.533	0.089	0.142	0.766	0.888	0.885	0.669	0.821	0.837	0.828	0.528	0.681	0.817	0.627	0.700
SE (HSC)	0.730	<b>0.276</b>	<u>0.414</u>	<u>0.706</u>	<b>0.135</b>	<u>0.311</u>	<u>0.801</u>	<u>0.914</u>	<u>0.900</u>	0.701	0.839	0.836	<u>0.870</u>	0.557	<u>0.711</u>	0.833	0.653	0.721
SRE (SR)	0.547	0.181	0.279	0.575	0.097	0.193	0.751	0.884	0.887	0.633	0.807	0.835	0.840	0.526	0.669	0.829	0.645	<u>0.726</u>
SRE (SR+HSC)	0.711	0.244	0.411	0.705	0.120	0.271	<b>0.806</b>	<b>0.917</b>	<u>0.900</u>	<b>0.754</b>	<b>0.851</b>	<b>0.840</b>	<b>0.871</b>	<b>0.566</b>	<b>0.730</b>	<b>0.887</b>	<b>0.685</b>	<b>0.743</b>
SRE (SR+HSC*)	0.701	0.232	0.407	0.688	0.114	0.265	0.787	0.903	0.891	0.739	0.840	0.838	0.849	0.533	0.683	0.882	0.674	0.738

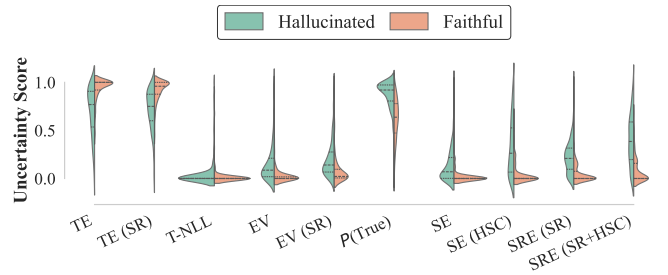


**Fig. 4:** Parameter sensitivity analysis of SRE.

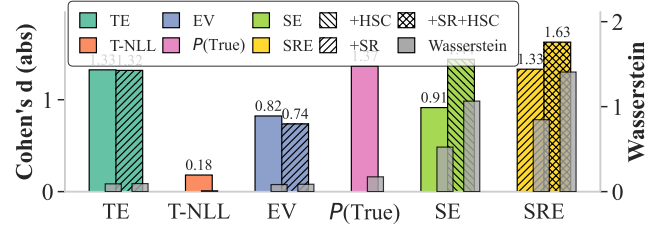
and  $T = 0.8$ , while larger or extreme values degrade performance due to redundancy or lower-quality outputs. These results emphasize the need to balance diversity and quality for reliable uncertainty estimation.

### 3.6. Effect on Uncertainty Estimation and Score Distributions

On SQuAD-v2 without context, the distributions of uncertainty scores in the violin (Fig. 5a) and bar plots (Fig. 5b) show that SRE effectively differentiates hallucinated from faithful samples. Compared to baselines (TE, T-NLL, EV,  $P(\text{True})$ , SE), SRE (SR+HSC) achieves a clearer separation, with hallucinated samples exhibiting a wider spread of uncertainty. Quantitative metrics, including Cohen’s  $d$  and Wasserstein distance, confirm that HSC drives most of the improvement while SR contributes moderately, whereas baselines display overlapping distributions and low distances. These results underscore SRE’s advantage and the crucial role of HSC in structuring the estimation space for reliable uncertainty measurement.



(a) Uncertainty distributions of hallucinated vs. faithful samples for SRE and baselines.



(b) Cohen’s  $d$  and Wasserstein distance comparisons across methods.

**Fig. 5:** Distribution differences across methods.

## 4. CONCLUSION

We propose SRE, leveraging input-side semantic reformulations and hybrid clustering to extend uncertainty estimation and more reliably capture epistemic uncertainty. Experiments on SQuAD and TriviaQA show that SRE outperforms strong baselines in hallucination detection. Consistent with related work, its efficiency is influenced by LLM sampling and NLI computation, which could be further optimized in a more systematic study. Nonetheless, our results suggest that expanding the uncertainty space and combining it with more precise estimation methods is a promising direction for enhancing LLM reliability.

## 5. ACKNOWLEDGMENT

This research is supported by the National Key R&D Program of China (No. 2023YFC3303800).

## 6. REFERENCES

- [1] Yiming Tan, Dehai Min, Yu Li, Wenbo Li, Nan Hu, Yongrui Chen, and Guilin Qi, “Can chatgpt replace traditional kbqa models? an in-depth analysis of the question answering performance of the gpt llm family,” in *ISWC*, 2023, p. 348–367.
- [2] Hao Zhou, Chengming Hu, Ye Yuan, Yufei Cui, Yili Jin, Can Chen, Haolun Wu, Dun Yuan, Li Jiang, Di Wu, et al., “Large language model (llm) for telecommunications: A comprehensive survey on principles, key techniques, and opportunities,” *IEEE Commun. Surv. Tutor.*, 2024.
- [3] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung, “Survey of hallucination in natural language generation,” *ACM Comput. Surv.*, vol. 55, no. 12, pp. 1–38, 2023.
- [4] Vaibhav Adlakha, Parishad BehnamGhader, Xing Han Lù, Nicholas Meade, and Siva Reddy, “Evaluating correctness and faithfulness of instruction-following models for question answering,” *TACL*, vol. 12, pp. 681–699, 2023.
- [5] Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal, “Detecting hallucinations in large language models using semantic entropy,” *Nature*, vol. 630, no. 8017, pp. 625–630, 2024.
- [6] Eyke Hüllermeier and Willem Waegeman, “Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods,” *Mach. Learn.*, vol. 110, pp. 457 – 506, 2019.
- [7] Yasin Abbasi Yadkori, Ilja Kuzborskij, András György, and Csaba Szepesvari, “To believe or not to believe your llm: Iterative prompting for estimating epistemic uncertainty,” *NeurIPS*, vol. 37, pp. 58077–58117, 2024.
- [8] Tianyu Liu, Yizhe Zhang, Chris Brockett, Yi Mao, Zhifang Sui, Weizhu Chen, and Bill Dolan, “A token-level reference-free hallucination detection benchmark for free-form text generation,” in *ACL*, 2022, pp. 6723–6737.
- [9] Tianhang Zhang, Lin Qiu, Qipeng Guo, Cheng Deng, Yue Zhang, Zheng Zhang, Chenghu Zhou, Xinbing Wang, and Luoyi Fu, “Enhancing uncertainty-based hallucination detection with stronger focus,” in *EMNLP*, 2023, pp. 915–932.
- [10] Yashvir S. Grewal, Edwin V. Bonilla, and Thang Duc Bui, “Improving uncertainty quantification in large language models via semantic embeddings,” *ArXiv*, vol. abs/2410.22685, 2024.
- [11] Huan Ma, Jiadong Pan, Jing Liu, Yan Chen, Joey Tianyi Zhou, Guangyu Wang, Qinghua Hu, Huaqin Wu, Changqing Zhang, and Haifeng Wang, “Semantic energy: Detecting llm hallucination beyond entropy,” 2025.
- [12] Erik Arakelyan, Zhaoqi Liu, and Isabelle Augenstein, “Semantic sensitivities and inconsistent predictions: Measuring the fragility of NLI models,” in *EACL*, 2024, pp. 432–444.
- [13] Ruiyang Zhang, Hu Zhang, and Zhedong Zheng, “V1-uncertainty: Detecting hallucination in large vision-language model via uncertainty estimation,” *ArXiv*, vol. abs/2411.11919, 2024.
- [14] Jannik Kossen, Jiatong Han, Muhammed Razzak, Lisa Schut, Shreshth A. Malik, and Yarin Gal, “Semantic entropy probes: Robust and cheap hallucination detection in llms,” *ArXiv*, vol. abs/2406.15927, 2024.
- [15] Nils Reimers and Iryna Gurevych, “Sentence-bert: Sentence embeddings using siamese bert-networks,” in *EMNLP*, 2019.
- [16] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen, “Deberta: Decoding-enhanced bert with disentangled attention,” in *ICLR*, 2021.
- [17] Pranav Rajpurkar, Robin Jia, and Percy Liang, “Know what you don’t know: Unanswerable questions for SQuAD,” in *ACL*, 2018, pp. 784–789.
- [18] Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer, “TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension,” in *ACL*, Regina Barzilay and Min-Yen Kan, Eds., 2017, pp. 1601–1611.
- [19] Aaditya Ramdas, Nicolás García Trillos, and Marco Cuturi, “On wasserstein two-sample testing and related families of nonparametric tests,” *Entropy*, vol. 19, no. 2, pp. 47, 2017.
- [20] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi, “The curious case of neural text degeneration,” in *ICLR*, 2020.