Enhancing Uncertainty Modeling with Semantic Graph for Hallucination Detection

Kedi Chen[1*†], Qin Chen[1*‡], Jie Zhou[1], Xinqi Tao[2], Bowen Ding[2], Jingwen Xie[2], Mingchen Xie[2], Peilong Li[2], Feng Zheng[2]

[1]East China Normal University
[2]Xiaohongshu Inc.
kdchen@stu.ecnu.edu.cn {qchen, jzhou}@cs.ecnu.edu.cn
{yifan5, faming, qingliang, shenzong, liaofan, yemu}@xiaohongshu.com

Abstract

Large Language Models (LLMs) are prone to hallucination with non-factual or unfaithful statements, which undermines the applications in real-world scenarios. Recent researches focus on uncertainty-based hallucination detection, which utilizes the output probability of LLMs for uncertainty calculation and does not rely on external knowledge or frequent sampling from LLMs. Whereas, most approaches merely consider the uncertainty of each independent token, while the intricate semantic relations among tokens and sentences are not well studied, which limits the detection of hallucination that spans over multiple tokens and sentences in the passage. In this paper, we propose a method to enhance uncertainty modeling with semantic graph for hallucination detection. Specifically, we first construct a semantic graph that well captures the relations among entity tokens and sentences. Then, we incorporate the relations between two entities for uncertainty propagation to enhance sentence-level hallucination detection. Given that hallucination occurs due to the conflict between sentences, we further present a graph-based uncertainty calibration method that integrates the contradiction probability of the sentence with its neighbors in the semantic graph for uncertainty calculation. Extensive experiments on two datasets show the great advantages of our proposed approach. In particular, we obtain substantial improvements with 19.78% in passage-level hallucination detection.

Introduction

Large Language Models (LLMs) (Zhao et al. 2023a), with large-scale parameters and advanced training methods, achieve excellent performance in many downstream tasks of natural language processing (NLP) (Aracena et al. 2024; Chen et al. 2024c; Lai and Nissim 2024; Zhang et al. 2024). Despite the many benefits of large language models, hallucination remains an issue that cannot be ignored. Hallucination indicates that some non-factual or untruthful contents are generated (Wang et al. 2023a). Therefore, hallucination detection is critically an essential task, which provides a preliminary review of the contents generated by large language models, reducing their potential harm in real-world scenarios (Lee et al. 2024; Cui et al. 2023; Yan et al. 2024), such as education, economics, science, and so on.

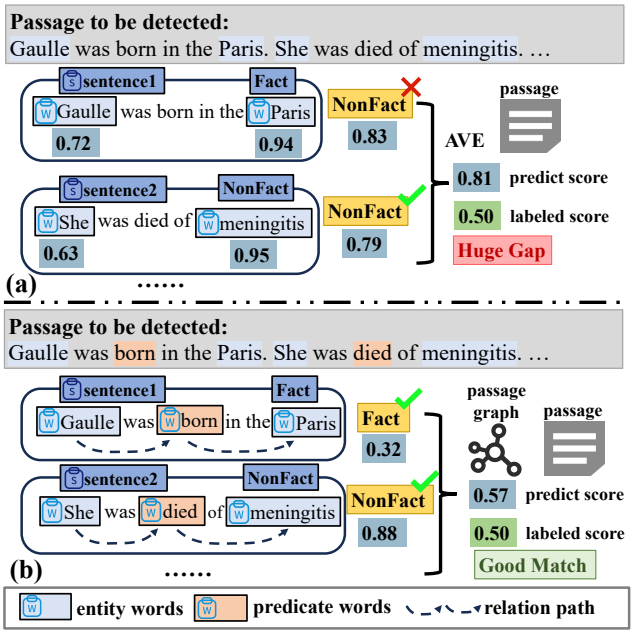

Figure 1: (a) Previous works only concern independent tokens and use their average scores as the metrics, resulting in errors in sentence and passage-level detection. (b) Our method captures more complex semantic dependencies with a semantic graph for uncertainty modeling, such as the relations between entities, and the relations with neighbor sentences in the passage-level semantic graph.

Current hallucination detection methods can be roughly divided into three categories. (i) Retrieval-based method (Wang et al. 2023c; Zhang et al. 2023b) usually retrieve evidence from external resources for fact verification (Chen et al. 2024a). This approach exceedingly depends on the quality of external resources, which is not always available. In addition, it needs various validation steps towards the retrieved knowledge, which are complicated and inefficient.

*These authors contributed equally.
†This work was done during the internship at Xiaohongshu Inc.
‡Corresponding author.

(ii) Sampling-based method frequently samples responses from LLMs for consistency verification, which consumes substantial computational resources (Manakul, Liusie, and Gales 2023; Zhang et al. 2023a). (iii) The uncertainty-based method is a good alternative to resolve the above problems (Giulianelli et al. 2023; Xiong et al. 2023). It leverages LLMs to output the probability of each token in the text to be detected and then computes a hallucination score with uncertainty-based metrics. Given that this method requires the models to perform inference only once, it is relatively efficient and thus attracts increasing interest from researchers.

Nevertheless, several challenges persist in uncertainty-based methods for hallucination detection (Figure 1). **First,** most methods focus on modeling the uncertainty of each independent token, while the complex dependency among tokens within the sentence is not well explored. Recent methods (Zhang et al. 2023c) tend to propagate the uncertainties of all previous tokens to the subsequent ones for uncertainty calculation. However, not all tokens are semantically related, and this propagation leads to uncertainty overestimation as shown in Figure 1. **Second,** passage-level uncertainty is not well studied. Previous methods usually average the uncertainty score of each sentence (Manakul, Liusie, and Gales 2023; Zhang et al. 2023c), while neglecting the intricate relations such as the semantic conflicts among sentences in the whole passage.

To resolve the above two challenges, we propose an approach to enhance uncertainty modeling with semantic graph for hallucination detection. Specifically, we first perform Abstract Meaning Representation (AMR) (Xu, Lee, and Huang 2023) based parsing for each sentence, and obtain a passage-level AMR graph by coreference resolution and entity linking between sentences, which well captures the semantic dependency relations among the entity tokens and the sentences for hallucination detection. Then, we present a relation-based propagation method, which propagates the uncertainty from one entity to the other along the relation path in the semantic graph to enhance sentence-level hallucination detection as shown in Figure 1. Regarding passage-level hallucination detection, we further integrate the relations between the sentence and its neighbors in the graph for uncertainty calibration via the natural language inference (NLI) (Zheng and Zhu 2023) technique.

We perform experiments on two datasets, namely the well-known WikiBio (Manakul, Liusie, and Gales 2023) and our constructed NoteSum. The results show the great superiority of our approach in both sentence-level and passage-level hallucination detection.

The main contributions can be summarized as follows:

- To the best of our knowledge, it is the first attempt to explore the potential of semantic graph to capture the complex relations among the tokens and the sentences for hallucination detection.
- We present two novel methods, namely relation-based uncertainty propagation and graph-based uncertainty calibration, which shed light on how to integrate the structured semantic graph with the uncertainty computation framework.
- We conduct elaborate analyses of the experimental results on two benchmark datasets, and provide a better understanding of the effectiveness of our approach.

Related Work

Hallucination in Language Models

Hallucination reflects that language models generate some nonsensical or untruthful contents (Wang et al. 2023a) in many downstream NLP tasks, such as the question and answer task (Naszádi, Manggala, and Monz 2023), the multi-turn dialogue task (Chen et al. 2024b) and the text summarization task (Kryscinski et al. 2020), etc. Hallucination in NLP can be categorized into two main classes: factuality hallucination and faithfulness hallucination (Huang et al. 2023a). The former one reveals the generated contents contain factual errors against real life, while the latter demonstrates the issues of inconsistency or irrelevance in the text.

Hallucination Detection

Before the era of LLMs, researchers normally train a discriminating model to judge whether hallucination exists (Zhao, Nguyen, and Daume 2023). This approach relies too heavily on the training data and can reduce the models' generalization ability. With the development of NLP technology, current hallucination detection methods can be roughly divided into three categories.

Retrieval-based method (Wang et al. 2023c; Zhang et al. 2023b) utilizes the retrieval-augmented generation technique (Chen et al. 2024a) for extra knowledge (Choi et al. 2023) or information to help detection (Varshney et al. 2023; Chen et al. 2024b; Siino 2024). This approach exceedingly depends on the quality of information sources, necessitating complicated validation steps (Ye et al. 2024; Dong et al. 2024) towards the retrieved knowledge. Not to mention that not all information is available easily. On the contrary, we propose an efficient reference-free method.

Sampling-based method rewrites the contents under detection, measuring the consistency and coherence (Malkin, Wang, and Jojic 2022; Sheng et al. 2024) between them to acquire a hallucination score (Manakul, Liusie, and Gales 2023; Zhang et al. 2023a; Zhao et al. 2023b; Mündler et al. 2024). However, this strategy frequently invokes LLMs for rewriting, consuming substantial computational resources. Our method needs one LLM to infer only once, thereby greatly saving the response time.

Uncertainty-based method applies proxy-based LLMs to output the probability of each token in contents to be detected and then estimates a hallucination score with uncertainty-based metrics (Huang et al. 2023b; Chen et al. 2023; Wang et al. 2023b; Petersen et al. 2024; Xiong et al. 2024). Manakul, Liusie, and Gales (2023) regards the degree of hallucination as being negatively correlated with the probability. Zhang et al. (2023c) refutes this view, but there arises a co-occurrence bias (Zhou et al. 2023). Due to a lack of detailed exploration of various dependencies, our method systematically constructs the relationships among the entity tokens and the sentences.

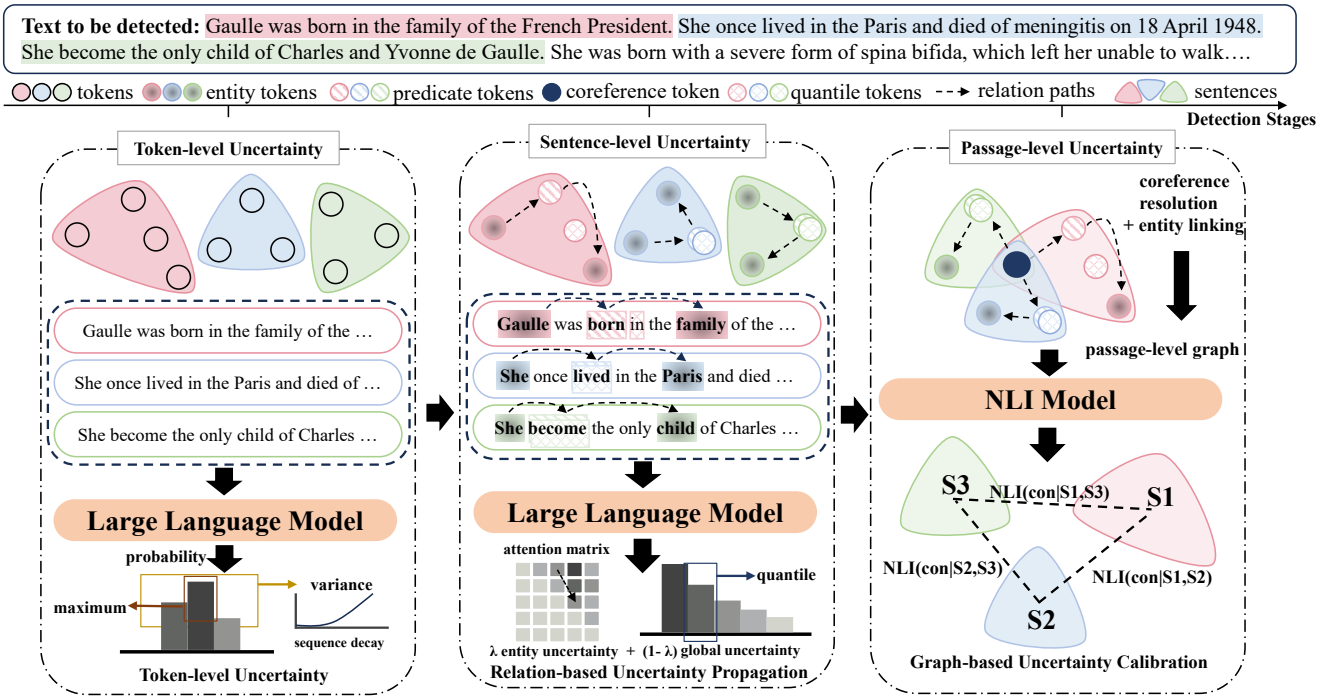

Figure 2: The overview of our approach for hallucination detection. For token-level uncertainty, we integrate the maximum and variance of the probabilities, along with a sequence decay term. Regarding to sentence-level uncertainty, we interpolate the sum of entity uncertainty through relation-based propagation and global uncertainty via quantile. Finally, we incorporate the relations of neighbor sentences in the semantic graph with graph-based uncertainty calibration for passage-level uncertainty.

Our Approach

The framework of our proposed approach is illustrated in Figure 2. Specifically, inspired by the findings that hallucination accumulates as the sequence length increases, we integrate the distribution statistics of LLM-based conditional probability with sequence decay for token-level uncertainty calculation. Considering much hallucination is induced by the entities and relations in the sentence and passage, we further construct a semantic graph for sentence-level and passage-level uncertainty calculation. Regarding sentence-level uncertainty, it well captures the semantic relations between entities for hallucination propagation and calculation. In particular, the uncertainty of an entity propagates to the related entity along the dependent relations. For passage-level uncertainty, we incorporate the neighbors of each sentence in the semantic graph for uncertainty calibration and summation. The details are denoted in the following.

Semantic Graph Construction. To better model the uncertainties of entities with long-range dependency that span over the text, we first perform AMR (Xu, Lee, and Huang 2023) parsing for each sentence, and gain a sentence-level graph where each node is an entity and the edge represents the dependent semantic relation. Compared to traditional dependency parsing, AMR parsing is more logical and less vulnerable to syntactic representation or word order variations. Therefore, we employ AMR to model the inter-dependency between the entities in the sentence. Furthermore, noting that passage-level hallucination usually occurs when two sentences contradict each other, we further link sentence-level AMR graphs together by the intricate relations (e.g., entity linking and coreference) among sentences. Finally, a large AMR graph corresponding to the passage is acquired.

Formally, we provide the notations deployed in this paper. Let $\mathcal{D}$ express the input passage with m sentences, which is denoted as $\mathcal{D} = \{\boldsymbol{S}_1, \boldsymbol{S}_2, \ldots, \boldsymbol{S}_m\}$. Each sentence $\boldsymbol{S}_i$ is composed of n_i tokens, i.e., $\boldsymbol{S}_i = \{t_i^1, t_i^2, \ldots, t_i^{n_i}\}$. In addition, the set of entity tokens in $\boldsymbol{S}_i$ is formulated by $\boldsymbol{E}_i = \left\{e_i^1, e_i^2, \ldots, e_i^{|\boldsymbol{E}_i|}\right\}$, where $|\boldsymbol{E}_i|$ indicates the number of entities in the i-th sentence.

Token-level Uncertainty

Generally, the conditional probability of a token output by LLMs reflects its likelihood in the context, which can be adapted to measure the uncertainty. Previous researches mainly focus on using the negative log probability or entropy-based methods for uncertainty estimation (Huang et al. 2023b). In this paper, we integrate two statistical indicators, namely the maximum and variance of the probability distributions. Moreover, hallucination tends to accumulate with the increasing sequence length as demonstrated in previous studies (Varshney et al. 2023; Naszádi, Manggala,

and Monz 2023; Chen et al. 2024b), thus we further devise a sequence decay term that explicitly models the absolute position of the token in the passage. Specifically, the token-level uncertainty in the j-th position of i-th sentence can be measured as:

$$\mathcal{U}(t_i^j) = \frac{1}{\max(\mathcal{C}_i^j) + \sigma^2(\mathcal{C}_i^j)} \underbrace{(1 + e^{\frac{len(\boldsymbol{S}_{1:i-1})+j}{len(\mathcal{D})}-1})}_{\text{sequence decay term}} \quad (1)$$

where $\mathcal{C}_i^j$ signifies the top-k probabilities of a candidate token set that could probably appear in the current position based on LLMs, which is formulated as:

$$\mathcal{C}_i^j = sorted(P_{ij}^{\mathcal{V}})\,[-k:]$$

where $P_{ij}^{\mathcal{V}}$ expresses the list of all probabilities for the vocabulary at the j-th position of i-th sentence. $\max()$ and $\sigma^2()$ represent the maximum and variance functions separately. If the values of maximum and variance are high, the model will be more confident about its output. The second term is a sequence decay we designed to increase the uncertainty of tokens when the length of the generated sequence grows. $len(\mathcal{D})$ is the total number of tokens in the entire passage, and $len(\boldsymbol{S}_{1:i-1}) + j$ shows the position of the current token in the passage.

Sentence-level Uncertainty

Previous works (Pagnoni, Balachandran, and Tsvetkov 2021; Kryscinski et al. 2020) illustrate that a major of hallucination in text generation is induced by the entity errors, such as false relations between two entities, inconsistent mentions in the context or basic factual errors, etc. This corresponds to our intuition that humans usually pay more attention to the salient information such as the keywords or entities for verification of the generated results. Therefore, recent researches turn to investigate the uncertainty of informative and important entities for hallucination detection. However, the complex dependencies over the entities are not well studied. In this paper, we explore the relations in the constructed semantic graph for uncertainty propagation and hallucination estimation.

Relation-based Uncertainty Propagation. Previous findings reveal that each token influences the surrounding context (Chen et al. 2017), thus the hallucination would probably propagate across the generated text. Zhang et al. (2023c) presents a hallucination propagation method that propagates the uncertainty score of preceding entity tokens to the current one. Whereas, this method roughly uses all the preceding entities, while ignoring their potential dependency relations with the current entity, which is inclined to overestimate the uncertainties by our preliminary studies. In this paper, we present a relation-based uncertainty propagation method and assume that the subject entity propagates its uncertainty to the object entity based on the predicate or relation in the semantic graph. Moreover, we devise a penalty factor based on the relation intensity to alleviate the uncertainty overestimation problem.

To be specific, given an object entity o, we first search the entities that have semantic relations with o from the semantic graph and obtain a set of triples as $\mathcal{T}_o = \{(s', v', o) | (s', v', o) \in \mathcal{T}_i\}$. $\mathcal{T}_i$ is the set of triples in semantic graph of sentence i. Intuitively, the subject entities are not equally important to the object entity, thus we leverage their attention scores as the weights for uncertainty propagation. To alleviate the overestimation problem, we additionally incorporate a relation intensity-based penalty factor for propagation. The final uncertainty of an object entity is formulated as:

$$\mathcal{U}_p(o) = \sum_{(s',v',o)\in\mathcal{T}_o} \frac{att(s', o)}{\mathcal{I}_o} * \mathcal{U}(s') \quad (2)$$

where $att(,)$ signifies the attention score between two tokens, $\mathcal{I}_o$ is a penalty factor that computes the relation intensity of all entities that have relations with the object o, which can be measured as follows:

$$\mathcal{I}_o = \frac{1}{|\mathcal{T}_o|} \sum_{(s',v',o)\in\mathcal{T}_o} \frac{att(s', v') + att(v', o)}{2} \quad (3)$$

In general, high relation intensities usually indicate high factuality-confidence, thus the propagated uncertainties should be penalized.

Entity Uncertainty. For an entity e_i^j, the uncertainty score consists of the self-uncertainty (Formula 1) and the propagated uncertainty (Formula 2). The entity-based uncertainty of sentence S_i can be calculated by averaging the uncertainties of all entities in the sentence:

$$\mathcal{U}_E(i) = \frac{1}{|\boldsymbol{E}_i|} \sum\nolimits_{e_i^j \in \boldsymbol{E}_i} [\mathcal{U}(e_i^j) + \beta \mathcal{U}_p(e_i^j)] \quad (4)$$

where $\mathcal{U}(e_i^j)$ and $\mathcal{U}_p(e_i^j)$ show the self-uncertainty and propagated uncertainty respectively, and β is a hyper-parameter to balance these two uncertainties.

Global Uncertainty. In addition to the entities, there are also many general tokens in the sentence. To capture the global information of the sentence, we also consider the uncertainties of all tokens (both entities and general tokens) in the sentence, and utilize the quantile approach to measure the global uncertainty, which is effective in capturing the global statistics in distributions (Gupta et al. 2024):

$$\mathcal{U}_G(i) = \text{qua}_\alpha(\mathcal{U}(t_i^1 : t_i^{n_i})) \quad (5)$$

where $\text{qua}_\alpha(\mathcal{U}(t_i^1 : t_i^{n_i}))$ is the α-quantile of the uncertainties of all tokens in sentence $\boldsymbol{S}_i$.

The uncertainty of the i-th sentence is the interpolation sum of the entity-based uncertainty and the global uncertainty:

$$\mathcal{U}_s(i) = \lambda \mathcal{U}_E(i) + (1 - \lambda)\mathcal{U}_G(i) \quad (6)$$

where λ is an interpolation weight.

Passage-level Uncertainty

Previous methods usually estimate the average uncertainty of all sentences for passage-level uncertainty. However, the intricate relations among the sentences are neglected, which could affect the detection of hallucination where two sentences contradict each other despite each sentence having

low uncertainty. For example, the first sentence in a passage is *'Thomas was born in 1972.'* and the fourth sentence is *'He raced until 1968.'*, which are contradictory in the passage. In this paper, we present a graph-based uncertainty calibration method that incorporates the relations of the sentence-centered sub-graph for uncertainty calibration. The calibrated uncertainties of all sentences are averaged as the passage-level uncertainty.

Graph-based Uncertainty Calibration. Intuitively, if a sentence contradicts all the neighbor sentences in the semantic graph, it will probably have inconsistency or conflicts in the context, which is prone to the hallucination problem. Thus, the uncertainty score should be increased. Motivated by this intuition, we present a graph-based uncertainty calibration method. First, we search the neighbor nodes for each sentence from the semantic graph. Then, we calculate the contradictory score for each connected sentence pair with a NLI model, namely DeBERTa-v3-Large (He, Gao, and Chen 2023), which is widely applied for natural language processing tasks. Finally, we incorporate the uncertainty of each sentence with the neighbor contradictory scores for passage-level uncertainty computation:

$$\mathcal{U}_p = \frac{1}{\sum_{i=1}^{m} |\mathcal{N}(i)|} \sum_{i=1}^{m} \sum_{j \in \mathcal{N}(i)} \mathcal{U}_s(i) * \text{NLI}(con|\boldsymbol{S}_j, \boldsymbol{S}_i) \tag{7}$$

where $\mathcal{N}(i)$ reflects the neighbors of the i-th sentence in the graph, $\text{NLI}(con|,)$ is the contradiction probability between two sentences via the NLI model.

Experimental Setup

Datasets We conduct extensive experiments on two datasets for hallucination detection. One is currently the latest and most widely used dataset WikiBio. To verify the effectiveness and generalization of our method, we also construct a Chinese dataset NoteSum, which can help boost research in this area. **WikiBio** (Manakul, Liusie, and Gales 2023) is a dataset derived from Wikipedia biographies. WikiBio applies the names from Wikipedia as the topics and generates corresponding biographies using GPT-3 (Floridi and Chiriatti 2020). Each sentence is annotated with one of the following labels: Factual (hallucination score: 0), NonFact* (0.5), and NonFact (1), which indicates a sentence with no hallucination, with factual errors, and is irrelevant to the topic respectively. The entire passage also has a human-labeled hallucination score as the ground truth. **NoteSum** is an industrial Chinese dataset. The company first collects users' long text notes on various daily topics with numerous entities. We cooperate with the company and create shorter summaries from these long notes by LLMs for research. The private information of users is removed. It consists of both factuality and faithfulness hallucination as WikiBio. We also adopt the same annotation guideline with WikiBio. The statistics of the datasets are shown in Table 1.

Evaluation Metrics For fair comparison, we apply the evaluation metrics used in previous researches (Manakul, Liusie, and Gales 2023; Zhang et al. 2023c). Specifically, the area under curves (AUC) (Bradley 1997) are used to measure the performance of sentence-level hallucination detection. To evaluate the agreement between the passage-level hallucination score and human judgment, we employ the Pearson correlation coefficient (Cohen et al. 2009) and the Spearman correlation coefficient (Sedgwick 2014) to estimate the degree of consistency.

	WikiBio	**NoteSum**
Language	English	Chinese
# Passages	238	200
# Sentences	1908	1004
# Words/Sentence	17.49	33.38
# Sentences/Passage	8.02	5.02
Halu Rate (%)	72.95	65.27
Fact Halu Rate (%)	33.07	27.94
Faith Halu Rate (%)	39.88	37.33

Table 1: Statistics of WikiBio and NoteSum. 'Fact Halu Rate (%)' and 'Faith Halu Rate (%)' demonstrate the proportion of sentences with factuality and faithfulness hallucination.

Baselines We compare our approach with the recent advanced baselines: 1) **GPT-3 Uncertainties** method uses the GPT-3 model to output the probability of each token, and then various uncertainty metrics are calculated as Manakul, Liusie, and Gales (2023) do for hallucination detection. 2) **SelfCheckGPT** (Manakul, Liusie, and Gales 2023) is the recent sampling-based method that relies on frequent sampling from LLMs for consistency checking. The gpt-3.5-turbo model is used and four methods are applied to measure the consistency, namely BertScore, QA, Unigram, and their combination. 3) **FOCUS** (Zhang et al. 2023c) is currently the outstanding uncertainty-based detection method. We leverage the LLaMA-13B and LLaMA-30B as the backbones.

Implementation Details we utilize a transition-based AMR parser (Xu, Lee, and Huang 2023) to construct an AMR graph for each sentence. Then, we perform coreference resolution and entity linking by spaCy to link sentence-level AMR graphs together to obtain a passage-level graph for each passage. The DeBERTa-v3-Large (He, Gao, and Chen 2023) NLI model is used to calculate the contradiction probability in Formula 7. We experiment with the LLaMA-13B and LLaMA-30B models to obtain the probability of each token. The hyper-parameters α, β, λ, and k are set to 0.8, 0.65, 0.7, and 3 respectively.

Results and Analyses

Main Results

Table 2 shows the performance of our approach and the state-of-the-art baselines. We have the following observation. **First,** we achieve the best performance on both sentence-level and passage-level hallucination detection regarding all evaluation metrics. In particular, we gain a maximum improvement of 19.78% over the best baseline

Methods	WikiBio sentence-level NonFact	WikiBio sentence-level NonFact*	WikiBio sentence-level Factual	WikiBio passage-level Pearson	WikiBio passage-level Spearman	NoteSum sentence-level NonFact	NoteSum sentence-level NonFact*	NoteSum sentence-level Factual	NoteSum passage-level Pearson	NoteSum passage-level Spearman
GPT-3 Uncertainties										
Avg(-log$p$)	83.21	38.89	53.97	57.04	53.93	80.11	43.69	35.29	39.61	31.55
Avg($\mathcal{H}$)	80.73	37.09	52.07	55.52	50.87	80.08	43.95	38.04	40.36	33.25
Max(-log$p$)	87.51	35.88	50.46	57.83	55.69	79.86	40.17	36.70	38.13	34.75
Max($\mathcal{H}$)	85.75	32.43	50.27	52.48	49.55	81.02	47.33	39.03	42.88	37.24
SelfCheckGPT (gpt-3.5-turbo)										
BertScore	81.96	45.96	44.23	58.18	55.90	76.44	39.69	36.89	25.91	21.24
QA	84.26	40.06	48.14	61.07	59.29	79.69	45.30	39.32	41.07	36.54
Unigram (max)	85.63	41.04	58.47	64.71	64.91	79.48	43.88	36.15	38.80	33.35
Combi	87.33	44.37	61.83	69.05	67.77	82.38	53.19	40.17	47.79	41.27
FOCUS										
LLaMA-13B	87.90	43.84	62.46	70.62	63.03	81.11	49.98	38.88	38.17	38.31
LLaMA-30B	89.79	48.80	65.69	77.15	73.24	82.17	43.12	49.85	37.37	40.09
OURS										
LLaMA-13B	90.14	**61.65**	64.82	72.11	64.35	85.06	50.70	53.03	**55.62**	60.81
LLaMA-30B	**90.93**	61.16	**65.70**	**77.60**	**74.44**	**87.95**	**54.42**	**61.51**	54.77	**61.05**
Δ	+1.14	+12.85	+0.01	+0.45	+1.20	+5.57	+1.23	+11.66	+7.83	+19.78

Table 2: Comparison results of our approach and the recent hallucination detection methods. The best results are in **bold** and the second best is marked with underline. Δ indicates our maximum improvements over the best baselines.

in passage-level hallucination detection. **Second,** compared with FOCUS that propagates the uncertainties of all preceding focused tokens to the subsequent one, our approach yields significant improvements especially for the NonFact* and Factual types that have moderate and no hallucination respectively, indicating the effectiveness of our relation-based uncertainty propagation to help alleviate the overestimation problem. **Third,** our approach exhibits good cross-domain and cross-language generalization. It not only performs well on the English biography dataset WikiBio, but also reflects significant improvements on the Chinese note summary dataset NoteSum.

Ablation Studies

We conduct ablation studies on WikiBio with LLaMA-30B from three dimensions: token, sentence, and passage. Experimental results are shown in Table 3. For each row, one setting is removed while keeping the other settings unchanged.

We have the following observations: (1) By removing each element from Formula 1 respectively, the performance decreases significantly in most cases, which signifies the effectiveness of the maximum, variance, and decay term for modeling the token-level uncertainty. (2) The performance with the passage-level metrics drops more significantly with the setting of '- max', manifesting that the maximum probability can better capture the key features for hallucination detection, while other terms can help further refine the uncertainty. (3) Both the entity uncertainty computed by relation-based propagation and the global uncertainty are important to sentence-level detection. In addition, entity uncertainty is more effective than global uncertainty for passage-level detection. (4) By excluding the contradiction relations of the neighbor sentences in the semantic graph, the performance of passage-level hallucination detection significantly drops by about 2 points, which further verifies the effectiveness of our graph-based uncertainty calibration for detecting hallucination over the passage.

	sentence-level NonFact	sentence-level NonFact*	sentence-level Fact	passage-level Pear.	passage-level Spear.
Ours	90.93	61.16	65.70	77.60	74.44
- max	86.48	64.86	63.52	23.32	38.57
- var	90.17	50.94	64.82	75.60	72.36
- decay	89.01	43.57	63.48	70.19	66.49
- entity	88.31	43.06	63.10	65.81	60.34
- global	88.75	43.88	65.19	70.36	65.49
- graph	-	-	-	75.89	72.20

Table 3: Results of ablation studies on WikiBio. '- max', '- var' and '- decay' mean removing the maximum, variance and decay term from Formula 1. '- entity' and '- global' reveal removing the entity and global uncertainty respectively from Formula 6. '- graph' indicates not including the contradiction probability of the neighbors in the graph, i.e., averaging the uncertainties of all sentences in Formula 7.

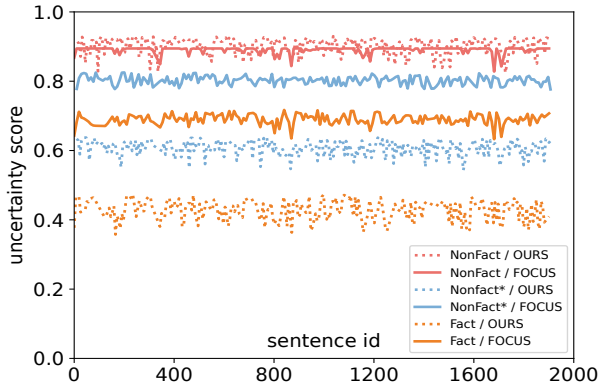

Figure 3: The uncertainty scores of three types of samples calculated with FOCUS and ours.

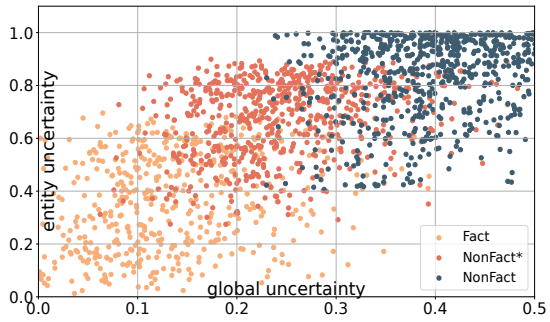

Figure 4: Visualization of the entity uncertainty and global uncertainty for three types of samples.

Further Analyses

Effect of Relation-based Uncertainty Propagation. To further investigate the effectiveness of our relation-based uncertainty propagation method, we compare with the baseline FOCUS (Zhang et al. 2023c) that propagates the uncertainties of all preceding keywords to the subsequent one. The results are shown in Figure 3, illustrating the uncertainty scores of three types of samples from WikiBio measured by FOCUS and ours respectively. We can observe that both of the two methods yield high uncertainty scores for the samples with NonFact (ground truth score = 1), which can help well identify the severe hallucination. It is also notable that the FOCUS method tends to overestimate the uncertainties for the samples with NonFact* (ground truth score = 0.5) and Fact (ground truth score = 0). There is a large gap between the estimated uncertainties and the ground truth. Moreover, the uncertainties of the three types calculated by FOCUS are very close, making it difficult to identify hallucination in different degrees. In contrast, our approach effectively diminishes the uncertainties for samples with NonFact* and Fact, which further verifies the effectiveness of our relation-based uncertainty propagation in alleviating the overestimation problem.

Visualization of Entity and Global Uncertainty. To examine the effect of entity and global uncertainty for sentence-level hallucination detection, scores of the two uncertainties are visualized for three types of samples from WikiBio in Figure 4. We observe that with the increased degree of hallucinations (Fact $\rightarrow$ NonFact* $\rightarrow$ NonFact), both types of uncertainty scores increase. Moreover, there are fewer overlaps in the three types of samples based on entity uncertainty and global uncertainty. In other words, the three types of samples can be well distinguished by the two uncertainties. All these observations demonstrate the effectiveness of our entity and global uncertainty.

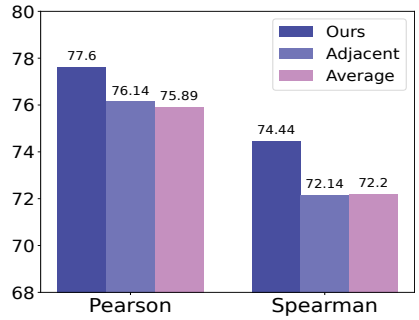

Figure 5: The Pearson and Spearman metrics of ours and the compared methods for passage-level uncertainty calculation.

Effect of Graph-based Uncertainty Calibration. To verify the effectiveness of our graph-based uncertainty calibration, we compare it with other two methods, namely Adjacent and Average. The Adjacent method merely incorporates the relations between the current sentence and the previous as well as the next sentence for uncertainty calculation, while the Average method simply measures the average uncertainties of all sentences. The results of the two methods and ours are shown in Figure 5. Our method is observed to outperform Adjacent and Average in terms of Pearson and Spearman correlations, indicating the effectiveness of using the semantic graph to model the long-range sentence relations for passage-level hallucination detection. In addition, the performance of Adjacent and Average is close, indicating the limits of merely considering the adjacent sentences.

Conclusions

In this paper, we propose a method to enhance uncertainty modeling with semantic graph for hallucination detection. Extensive experiments verify the effectiveness of each component of our approach. In particular, our approach consistently outperforms the state-of-the-art baselines in both sentence-level and passage-level hallucination detection, by incorporating the semantic relations among entities and sentences into the uncertainty calculation framework. It is also interesting to find that our relation-based uncertainty propagation method can help effectively alleviate the uncertainty overestimation problem and our graph-based uncertainty calibration method can capture long-range relations. In the future, we will explore integrating the existing knowledge graph with AMR graphs for fact-checking and hallucination detection.

Ethics Statement

Our WikiBio dataset is publicly used in the field of natural language processing. The NoteSum dataset, on the other hand, is an internal private dataset of Xiaohongshu, and its construction, annotation, and review are all handled by Xiaohongshu's own employees. The method presented in this paper is original to the authors and does not involve any ethical issues.

Acknowledgments

Thanks to all collaborators and reviewers for their efforts.This research is funded by the National Science and Technology Major Project (No. 2021ZD0114002), the National Nature Science Foundation of China (No. 62477010), the Science and Technology Commission of Shanghai Municipality Grant (No. 22511105901, No. 21511100402), Shanghai Science and Technology Innovation Action Plan (No. 24YF2710100), and Shanghai Special Project to Promote High-quality Industrial Development (No. RZ-CYAI-01-24-0288).

References

Aracena, G.; Luster, K.; Santos, F.; Steinmacher, I.; and Gerosa, M. A. 2024. Applying Large Language Models API to Issue Classification Problem. *CoRR*, abs/2401.04637.

Bradley, A. P. 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit.*, 30(7): 1145–1159.

Chen, J.; Lin, H.; Han, X.; and Sun, L. 2024a. Benchmarking Large Language Models in Retrieval-Augmented Generation. In *AAAI 2024*, 17754–17762. AAAI Press.

Chen, K.; Chen, Q.; Zhou, J.; He, Y.; and He, L. 2024b. DiaHalu: A Dialogue-level Hallucination Evaluation Benchmark for Large Language Models. *CoRR*, abs/2403.00896.

Chen, K.; Zhou, J.; Chen, Q.; Liu, S.; and He, L. 2024c. A Regularization-based Transfer Learning Method for Information Extraction via Instructed Graph Decoder. In *LREC/-COLING 2024*, 1472–1485. ELRA and ICCL.

Chen, L.; Deng, Y.; Bian, Y.; Qin, Z.; and et al. 2023. Beyond Factuality: A Comprehensive Evaluation of Large Language Models as Knowledge Generators. In *EMNLP 2023*, 6325–6341. Association for Computational Linguistics.

Chen, Q.; Hu, Q.; Huang, J. X.; He, L.; and An, W. 2017. Enhancing Recurrent Neural Networks with Positional Attention for Question Answering. In *SIGIR, 2017*, 993–996. ACM.

Choi, S.; Fang, T.; Wang, Z.; and Song, Y. 2023. KCTS: Knowledge-Constrained Tree Search Decoding with Token-Level Hallucination Detection. In *EMNLP 2023*, 14035–14053. Association for Computational Linguistics.

Cohen, I.; Huang, Y.; Chen, J.; Benesty, J.; Benesty, J.; Chen, J.; Huang, Y.; and Cohen, I. 2009. Pearson correlation coefficient. *Noise reduction in speech processing*, 1–4.

Cui, J.; Li, Z.; Yan, Y.; Chen, B.; and Yuan, L. 2023. ChatLaw: Open-Source Legal Large Language Model with Integrated External Knowledge Bases. *CoRR*, abs/2306.16092.

Dong, Q.; Liu, Y.; Ai, Q.; Wu, Z.; and et al. 2024. Unsupervised Large Language Model Alignment for Information Retrieval via Contrastive Feedback. In *SIGIR 2024*, 48–58. ACM.

Floridi, L.; and Chiriatti, M. 2020. GPT-3: Its Nature, Scope, Limits, and Consequences. *Minds Mach.*, 30(4): 681–694.

Giulianelli, M.; Baan, J.; Aziz, W.; Fernández, R.; and Plank, B. 2023. What Comes Next? Evaluating Uncertainty in Neural Text Generators Against Human Production Variability. In *EMNLP 2023*, 14349–14371. Association for Computational Linguistics.

Gupta, N.; Narasimhan, H.; Jitkrittum, W.; Rawat, A. S.; Menon, A. K.; and Kumar, S. 2024. Language Model Cascades: Token-level uncertainty and beyond. *CoRR*, abs/2404.10136.

He, P.; Gao, J.; and Chen, W. 2023. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. In *ICLR 2023*. OpenReview.net.

Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; and et al. 2023a. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *CoRR*, abs/2311.05232.

Huang, Y.; Song, J.; Wang, Z.; Chen, H.; and Ma, L. 2023b. Look Before You Leap: An Exploratory Study of Uncertainty Measurement for Large Language Models. *CoRR*, abs/2307.10236.

Kryscinski, W.; McCann, B.; Xiong, C.; and Socher, R. 2020. Evaluating the Factual Consistency of Abstractive Text Summarization. In *EMNLP 2020*, 9332–9346. Association for Computational Linguistics.

Lai, H.; and Nissim, M. 2024. A Survey on Automatic Generation of Figurative Language: From Rule-based Systems to Large Language Models. *ACM Comput. Surv.*, 56(10): 244.

Lee, J.; Stevens, N.; Han, S. C.; and Song, M. 2024. A Survey of Large Language Models in Finance (FinLLMs). *CoRR*, abs/2402.02315.

Malkin, N.; Wang, Z.; and Jojic, N. 2022. Coherence boosting: When your pretrained language model is not paying enough attention. In *ACL 2022*, 8214–8236. Association for Computational Linguistics.

Manakul, P.; Liusie, A.; and Gales, M. J. F. 2023. SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. In *EMNLP 2023*, 9004–9017. Association for Computational Linguistics.

Mündler, N.; He, J.; Jenko, S.; and Vechev, M. 2024. Self-contradictory Hallucinations of Large Language Models: Evaluation, Detection and Mitigation. arXiv:2305.15852.

Naszádi, K.; Manggala, P.; and Monz, C. 2023. Aligning Predictive Uncertainty with Clarification Questions in Grounded Dialog. In *Findings of EMNLP 2023*, 14988–14998. Association for Computational Linguistics.

Pagnoni, A.; Balachandran, V.; and Tsvetkov, Y. 2021. Understanding Factuality in Abstractive Summarization with

FRANK: A Benchmark for Factuality Metrics. In *NAACL-HLT 2021*, 4812–4829. Association for Computational Linguistics.

Petersen, F.; Mishra, A.; Kuehne, H.; Borgelt, C.; Deussen, O.; and Yurochkin, M. 2024. Uncertainty Quantification via Stable Distribution Propagation. arXiv:2402.08324.

Sedgwick, P. 2014. Spearman's rank correlation coefficient. *Bmj*, 349.

Sheng, Z.; Zhang, T.; Jiang, C.; and Kang, D. 2024. BBScore: A Brownian Bridge Based Metric for Assessing Text Coherence. In *AAAI 2024*, 14937–14945. AAAI Press.

Siino, M. 2024. BrainLlama at SemEval-2024 Task 6: Prompting Llama to detect hallucinations and related observable overgeneration mistakes. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, 82–87. Mexico City, Mexico: Association for Computational Linguistics.

Varshney, N.; Yao, W.; Zhang, H.; Chen, J.; and Yu, D. 2023. A Stitch in Time Saves Nine: Detecting and Mitigating Hallucinations of LLMs by Validating Low-Confidence Generation. *CoRR*, abs/2307.03987.

Wang, C.; Liu, X.; Yue, Y.; Tang, X.; Zhang, T.; and et al. 2023a. Survey on Factuality in Large Language Models: Knowledge, Retrieval and Domain-Specificity. *CoRR*, abs/2310.07521.

Wang, J.; Sun, Q.; Chen, N.; Wang, C.; Huang, J.; Gao, M.; and Li, X. 2023b. Uncertainty-aware Parameter-Efficient Self-training for Semi-supervised Language Understanding. arXiv:2310.13022.

Wang, X.; Yan, Y.; Huang, L.; Zheng, X.; and Huang, X. 2023c. Hallucination Detection for Generative Large Language Models by Bayesian Sequential Estimation. In *EMNLP 2023, Singapore, December 6-10, 2023*, 15361–15371. Association for Computational Linguistics.

Xiong, M.; Hu, Z.; Lu, X.; Li, Y.; and et al. 2023. Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs. *CoRR*, abs/2306.13063.

Xiong, M.; Hu, Z.; Lu, X.; LI, Y.; Fu, J.; He, J.; and Hooi, B. 2024. Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs. In *The Twelfth International Conference on Learning Representations*.

Xu, Z.; Lee, J. Y.; and Huang, L. 2023. Learning from a Friend: Improving Event Extraction via Self-Training with Feedback from Abstract Meaning Representation. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Findings of the Association for Computational Linguistics: ACL 2023*, 10421–10437. Toronto, Canada: Association for Computational Linguistics.

Yan, L.; Sha, L.; Zhao, L.; Li, Y.; and et al. 2024. Practical and ethical challenges of large language models in education: A systematic scoping review. *Br. J. Educ. Technol.*, 55(1): 90–112.

Ye, L.; Lei, Z.; Yin, J.; Chen, Q.; Zhou, J.; and He, L. 2024. Boosting Conversational Question Answering with Fine-Grained Retrieval-Augmentation and Self-Check. In *SIGIR 2024*, 2301–2305. ACM.

Zhang, D.; Yu, Y.; Li, C.; Dong, J.; and et al. 2024. MM-LLMs: Recent Advances in MultiModal Large Language Models. *CoRR*, abs/2401.13601.

Zhang, J.; Li, Z.; Das, K.; Malin, B. A.; and Sricharan, K. 2023a. SAC^3: Reliable Hallucination Detection in Black-Box Language Models via Semantic-aware Cross-check Consistency. In *Findings of EMNLP 2023*, 15445–15458. Association for Computational Linguistics.

Zhang, J.; Muhamed, A.; Anantharaman, A.; Wang, G.; Chen, C.; and et al. 2023b. ReAugKD: Retrieval-Augmented Knowledge Distillation For Pre-trained Language Models. In *ACL 2023*, 1128–1136. Association for Computational Linguistics.

Zhang, T.; Qiu, L.; Guo, Q.; Deng, C.; and et al. 2023c. Enhancing Uncertainty-Based Hallucination Detection with Stronger Focus. In *EMNLP 2023*, 915–932. Association for Computational Linguistics.

Zhao, L.; Nguyen, K.; and Daume, H. 2023. Hallucination Detection for Grounded Instruction Generation. In *Findings of EMNLP 2023*, 4044–4053. Association for Computational Linguistics.

Zhao, W. X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; and et al. 2023a. A Survey of Large Language Models. *CoRR*, abs/2303.18223.

Zhao, Y.; Yan, L.; Sun, W.; Xing, G.; and et al. 2023b. Knowing What LLMs DO NOT Know: A Simple Yet Effective Self-Detection Method. *CoRR*, abs/2310.17918.

Zheng, Z.; and Zhu, X. 2023. NatLogAttack: A Framework for Attacking Natural Language Inference Models with Natural Logic. In *ACL 2023*, 9960–9976. Association for Computational Linguistics.

Zhou, Y.; Hu, H.; Yu, J.; Xu, Z.; Lu, W.; and Cao, Y. 2023. A Solution to Co-occurrence Bias: Attributes Disentanglement via Mutual Information Minimization for Pedestrian Attribute Recognition. *CoRR*, abs/2307.15252.