KG-UQ: Knowledge Graph-Based Uncertainty Quantification for Long Text in Large Language Models

Yingqing Yuan
University of Sydney
Sydney, Australia
yyua0659@uni.sydney.edu.au

Linwei Tao
University of Sydney
Sydney, Australia
linwei.tao@sydney.edu.au

Haohui Lu
University of Sydney
Sydney, Australia
haohui.lu@sydney.edu.au

Matloob Khushi
Brunel University of London
London, England
matloob.khushi@brunel.ac.uk

Imran Razzak
MBZUAI, Abu Dhabi, UAE
University of New South Wales
imran.razzak@mbzuai.ac.ae

Mark Dras
Macquarie University
Sydney, Australia
mark.dras@mq.edu.au

Jian Yang
Macquarie University
Sydney, Australia
jian.yang@mq.edu.au

Usman Naseem
Macquarie University
Sydney, Australia
usman.naseem@mq.edu.au

Abstract

With the commercialization of large language models (LLMs) and their integration into daily life, addressing their susceptibility to hallucinations—unfactual information in generated outputs—has become an urgent priority. Existing uncertainty quantification (UQ) methods often rely on access to LLMs' internal states, which is unavailable for closed-source models like GPTs, or are primarily designed for short text. Current research on long text typically evaluates sentences individually, overlooking smaller semantic units that better capture the text's complexity. Recognizing the potential of knowledge graphs (KGs) to extract structured relationships from unstructured text, we propose KG-UQ, a UQ method leveraging KGs to address the semantic intricacies of long text. Our approach involves constructing KGs from long-text outputs and utilizing their embeddings to estimate uncertainties. Through our analysis, we demonstrate that knowledge graphs are an effective tool for decomposing long text into fundamental statements. However, we also highlight the increased uncertainty introduced during KG construction, stemming from inherent challenges in accurately capturing all semantic information.

CCS Concepts

• **Computing methodologies → Natural language generation**; **Knowledge representation and reasoning**.

Keywords

Large Language Models, Hallucinations, Uncertainty Quantification, Knowledge Graphs

WWW Companion '25, April 28-May 2, 2025, Sydney, NSW, Australia

ACM ISBN 979-8-4007-1331-6/2025/04
https://doi.org/10.1145/3701716.3717660

ACM Reference Format:
Yingqing Yuan, Linwei Tao, Haohui Lu, Matloob Khushi, Imran Razzak, Mark Dras, Jian Yang, and Usman Naseem. 2025. KG-UQ: Knowledge Graph-Based Uncertainty Quantification for Long Text in Large Language Models. In *Companion Proceedings of the ACM Web Conference 2025 (WWW Companion '25), April 28-May 2, 2025, Sydney, NSW, Australia.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3701716.3717660

1 Introduction

With the rapid development of Large Language Models (LLMs), their exceptional performance in natural language processing (NLP) tasks has been well illustrated [2, 4, 26, 37, 45], alongside their applications in various other areas [3, 32, 33, 44]. Applications, such as ChatGPT, have seamlessly integrated into everyday life. However, LLMs are often prone to hallucinations, generating responses that may be unfactual or unfaithful. Traditionally, researchers have relied on human evaluation, manually verifying decomposed atomic information. Alternatively, approaches have been developed to quantify the factuality of LLM responses or assess the confidence levels of LLMs in their outputs.

Existing uncertainty quantification (UQ) methods predominantly rely on accessing the internal states of LLMs, such as token likelihoods [8, 20, 41]. However, with the increasing commercialization of LLMs, closed-source proprietary models like GPTs restrict access to such internal states, offering only API-level interaction. Another significant challenge lies in handling the complexity of long text. While Natural Language Inference (NLI) models are often used to determine whether a piece of text is supported by its source, they perform well when dealing with fragmented information. However, long text is known for its intricate semantic structure, making it challenging to verify whether an entire passage is supported by its knowledge source. In existing literature, LUQ [42] tackle this by analyzing long text sentence by sentence, using NLI to determine whether each sentence is supported. Another approach [12] involves decomposing LLM responses into factoids and clustering these factoids by meaning to evaluate factuality.

Knowledge graphs (KGs), widely recognized for their ability to represent structured data and organize relationships derived from unstructured data, and have demonstrated significant potential in tasks like retrieval-augmented generation. GraphEval [38] has picked up on this and leveraged this capability by feeding LLM outputs into a KG construction prompt and then evaluating the factuality of each generated triple using NLI.

Building on insights from current research, we propose multiple KG-UQ, an uncertainty quantification method based on knowledge graphs, with comparisons of existing UQ methods performance by measuring the correlation with factuality scores. The key contributions of our work are summarized as follows:

- We introduce knowledge graphs (KG) as a tool for extracting logical information, fully leveraging the capabilities of KGs to address the limitations of current uncertainty quantification (UQ) methods for long text.
- Through analysis, we uncover the inherent limitations of current knowledge graph construction methods due to their intrinsic uncertainty.
- Through extensive experiments across different models and datasets, we demonstrate the superior generalization ability of our algorithm, surpassing the state-of-the-art (SOTA) methods.

2 Related Work

2.1 Uncertainty Quantification in Machine Learning Models

Uncertainty quantification (UQ) [10, 15, 27, 39, 40], plays a crucial role in machine learning, helping models provide predictions along with measures of confidence. Traditional approaches include Bayesian methods, such as Bayesian Neural Networks (BNNs), which use posterior distributions to capture uncertainty in model parameters [27]. Non-Bayesian methods, like Monte Carlo Dropout [15] and Deep Ensembles [10], offer more practical alternatives for estimating predictive uncertainty. Before the rise of Large Language Models (LLMs), UQ was already a key area of research in machine learning [15]. Uncertainty is typically divided into two categories: aleatoric and epistemic uncertainty [7, 16]. Aleatoric uncertainty, also known as statistical uncertainty, refers to the natural randomness in data or outcomes caused by inherent variability [18]. Epistemic uncertainty, on the other hand, arises from incomplete knowledge, such as missing data or uncertainty in model parameters [18]. While aleatoric uncertainty cannot be reduced, epistemic uncertainty can often be addressed through better models or more data.

2.2 Uncertainty in Knowledge Graphs

Knowledge graphs (KGs) are widely used to represent structured data and relationships, but effectively managing uncertainty is key to their usefulness. Probabilistic knowledge graphs enhance traditional KGs by assigning confidence scores to the facts or triples they contain, making it possible to represent uncertain information [5]. Advanced methods such as Markov Logic Networks [?] have been developed to propagate and infer uncertainty within the structure of a graph. Embedding techniques allow uncertain knowledge graphs to be represented in a continuous vector space, supporting tasks like link prediction, entity classification, and fact verification [5]. These methods address challenges like missing, noisy, or conflicting data, enabling more robust reasoning and decision-making. Applications of these approaches include question answering systems and personalized recommendations, where the ability to handle uncertainty is crucial.

2.3 Uncertainty Quantification in LLMs

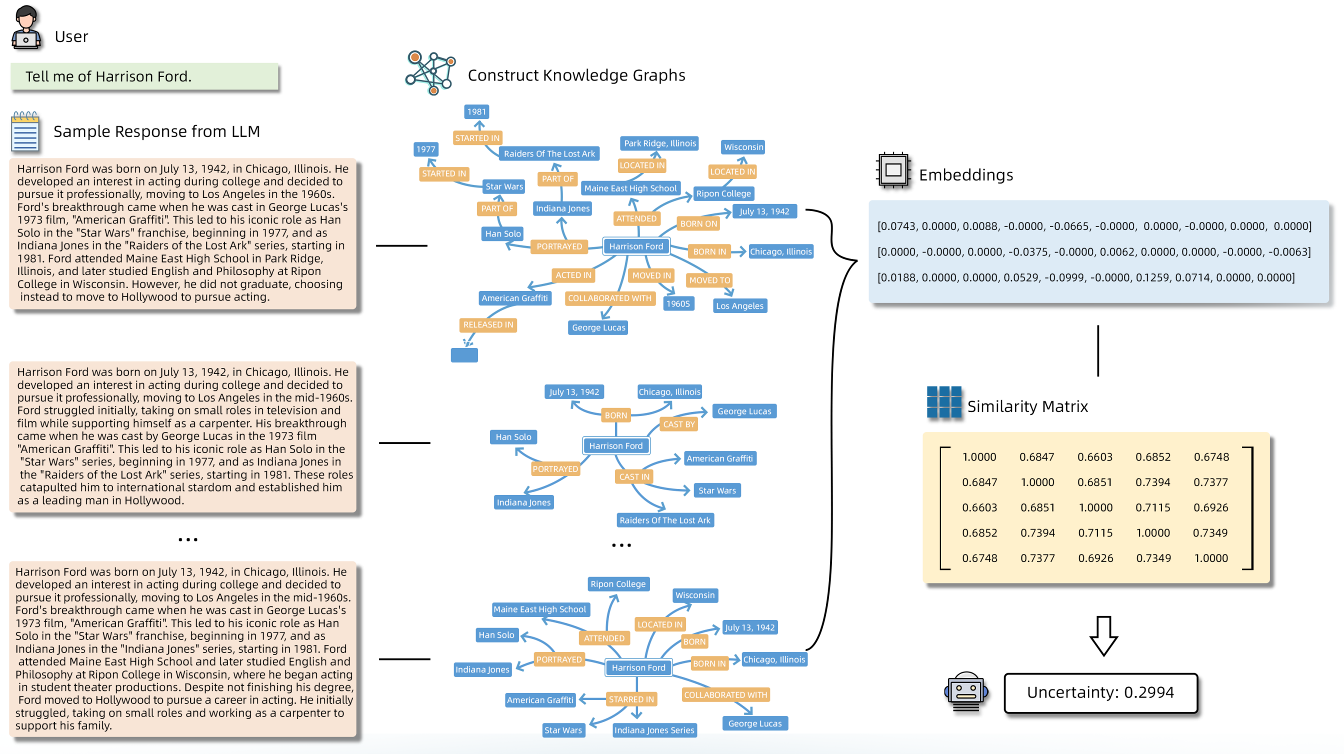
For Large Language Models (LLMs), uncertainty quantification (UQ) is particularly important as their outputs often lack clear indications of reliability. One widely used approach is calibration, which adjusts model confidence scores to better match actual accuracy [14]. Perturbation-based methods have also gained traction; these involve introducing small changes to the input to measure how stable the model's responses are [14]. Another approach is embedding-based analysis, which incorporates semantic checks to evaluate the consistency and reliability of generated text [25]. Tools like FACTSCORE go further by breaking down generated responses into individual facts, comparing these facts against a reference source, and evaluating factual accuracy [30]. UQ methods for LLMs are often categorized based on whether they require access to the model's internal mechanics. White-box methods, for instance, rely on logit-based evaluations that assess sentence uncertainty by analyzing token-level probabilities or entropy [22, 31, 46]. These approaches contrast with black-box methods, which operate independently of the model's internal structure. The importance of UQ in LLMs extends to high-stakes areas such as medical decision-making [19] and content moderation, where ensuring reliability and trustworthiness is critical. As LLMs are increasingly applied in diverse real-world scenarios, UQ remains an essential area of research for improving their interpretability and reliability.

3 Methods

Our method is a black-box method intended to estimate the uncertainty of LLMs' outputs based on knowledge graphs. It consists of three parts, and the framework is illustrated in Figure 1.

1. Generate responses from LLMs.
2. Construct a set of knowledge graphs for the responses.
3. Estimate uncertainty by computing similarities between knowledge graphs.

For a given query prompt q_a, let $\mathbf{R} = \{r_1, r_2, r_3, \ldots\}$ denote all the sampling responses generated by a LLM. As discussed, determining the similarity between long text responses is inherently challenging due to the intricate semantic relationships embedded across sentences and paragraphs. To address this, we propose decomposing each long text response r_i into a collection of simple sentences that capture the core ideas of the text. This decomposition is achieved by constructing a set of knowledge graphs $\mathbf{G}$ for the responses, where each knowledge graph g_i comprises a set of triples $triples_i = \{triple_{i_1}, triple_{i_2}, triple_{i_3}, \ldots\}$. Each triple consists of a head, a relationship, and a tail, analogous to the subject, predicate, and object in grammatical structure, collectively representing a specific semantic fact extracted from the original text. By assembling these triples, we can effectively distill the complex



Tell me of Harrison Ford.

Sample Response from LLM

Harrison Ford was born on July 13, 1942, in Chicago, Illinois. He developed an interest in acting during college and decided to pursue it professionally, moving to Los Angeles in the 1960s. Ford's breakthrough came when he was cast in George Lucas's 1973 film, "American Graffiti". This led to his iconic role as Han Solo in the "Star Wars" franchise, beginning in 1977, and as Indiana Jones in the "Raiders of the Lost Ark" series, starting in 1981. Ford attended Maine East High School in Park Ridge, Illinois, and later studied English and Philosophy at Ripon College in Wisconsin. However, he did not graduate, choosing instead to move to Hollywood to pursue acting.

Harrison Ford was born on July 13, 1942, in Chicago, Illinois. He developed an interest in acting during college and decided to pursue it professionally, moving to Los Angeles in the mid-1960s. Ford struggled initially, taking on small roles in television and film while supporting himself as a carpenter. His breakthrough came when he was cast by George Lucas in the 1973 film "American Graffiti". This led to his iconic role as Han Solo in the "Star Wars" series, beginning in 1977, and as Indiana Jones in the "Raiders of the Lost Ark" series, starting in 1981. These roles catapulted him to international stardom and established him as a leading man in Hollywood.

...

Harrison Ford was born on July 13, 1942, in Chicago, Illinois. He developed an interest in acting during college and decided to pursue it professionally, moving to Los Angeles in the mid-1960s. Ford's breakthrough came when he was cast in George Lucas's 1973 film, "American Graffiti". This led to his iconic role as Han Solo in the "Star Wars" franchise, beginning in 1977, and as Indiana Jones in the "Indiana Jones" series, starting in 1981. Ford attended Maine East High School and later studied English and Philosophy at Ripon College in Wisconsin, where he began acting in student theater productions. Despite not finishing his degree, Ford moved to Hollywood to pursue a career in acting. He initially struggled, taking on small roles and working as a carpenter to support his family.

[0.0743, 0.0000, 0.0088, -0.0000, -0.0665, -0.0000, 0.0000, -0.0000, 0.0000, 0.0000]
[0.0000, -0.0000, 0.0000, -0.0375, -0.0000, 0.0062, 0.0000, 0.0000, -0.0000, -0.0063]
[0.0188, 0.0000, 0.0000, 0.0529, -0.0999, -0.0000, 0.1259, 0.0714, 0.0000, 0.0000]

1.0000	0.6847	0.6603	0.6852	0.6748
0.6847	1.0000	0.6851	0.7394	0.7377
0.6603	0.6851	1.0000	0.7115	0.6926
0.6852	0.7394	0.7115	1.0000	0.7349
0.6748	0.7377	0.6926	0.7349	1.0000

Figure 1: Overview of KG-UQ. For a given user query, we generate response n times using a LLM. Each response is then used to construct a corresponding knowledge graph. Embeddings are generated for each KG, and a similarity matrix is computed based on these embeddings. From this matrix, we derive the uncertainty associated with the LLM's responses to the user query.

semantic structure of the long text into a concise format.

$$r_i \xrightarrow{\text{Construct Knowledge Graph}} g_i = \{triple_{i_1}, triple_{i_2}, triple_{i_3}, \dots\}$$

$$triple_{i_j} = (head_{i_j}, relation_{i_j}, tail_{i_j})$$

To determine if facts $triples_i = \{triple_{i_1}, triple_{i_2}, triple_{i_3}, \dots\}$ that are decomposed from a long response are supported by others, we concatenate these triples into a simplified paragraph p_i and embed it into latent space. This approach ensures that the semantic content of the triples is preserved in a form suitable for embedding into a shared latent space. The process can be formalized as follows:

$$\mathbf{G} = \{g_1, g_2, g_3, \dots\} \xrightarrow{\text{Concat Triples}} \mathbf{P} = \{p_1, p_2, p_3, \dots\}$$

$$\mathbf{P} = \{p_1, p_2, p_3, \dots\} \xrightarrow{\text{Embed}} \mathbf{E} = \{e_1, e_2, e_3, \dots\}$$

where **G** represents the collection of knowledge graphs constructed from responses **R**, **P** represents the corresponding simplified paragraphs after concatenating the triples, and **E** represents the embeddings of these paragraphs in the latent space.

All responses generated using the same prompt are embedded into the same latent space, enabling semantic comparison between them. Specifically, the Euclidean distance in this space represents the semantic distance between responses. Let **D** denote the semantic distance matrix, where d_{ij} represents the semantic distance between paragraph p_1 and p_2, corresponding to response r_i and r_j.

The semantic distance matrix is defined as:

$$\mathbf{D} = \begin{pmatrix} d_{00} & d_{01} & \cdots & d_{0n} \\ d_{10} & d_{11} & \cdots & d_{1n} \\ \vdots & \vdots & \vdots & \vdots \\ d_{n0} & d_{n1} & \cdots & d_{nn} \end{pmatrix}$$

To quantify the confidence score of each response r_i concerning the query prompt q_a, we define the confidence score $\mathbf{C}(r_i, q_a)$ as the average semantic distance between r_i and all other responses:

$$\mathbf{C}(r_i, q_a) = \frac{1}{n} \sum_{\substack{j=0 \\ i \neq j}}^{n} d_{ij}$$

The overall uncertainty **U** for the query prompt q_a is then calculated as the average confidence score aross all responses:

$$\mathbf{U}(q_a) = \frac{1}{n} \sum_{i=0}^{n} C(r_i, q_a)$$

Another approach to obtaining the embedding of each knowledge graph is by leveraging a Graph Convolutional Network (GCN) [21]. In this framework, each knowledge graph is treated as a heterogeneous graph structure, where the head and tail of a triple are modeled as nodes in the graph, and the relationship between them serves as the edge connecting these nodes. A key aspect of this approach is the utilization of node types, which serve as an essential property to distinguish the nodes. For example, consider a head node like *Donald Trump*. Depending on the context, this node could carry the property of being a *Person*, a *President*, or a *Real Estate Entrepreneur*. Node types allow the GCN to effectively encode such distinctions, enabling the model to better capture the semantic and structural nuances of the knowledge graph. The process can be formalized as follows:

$$\mathrm{G} = \{g_1, g_2, g_3, \dots\} \xrightarrow{\text{Graph Embedding}} \mathbf{GE} = \{ge_1, ge_2, ge_3, \dots\}$$

where **GE** denotes the set of graph embeddings generated by GCN.

Similarly, we can compute the pairwise distance matrix and uncertainty.

4 Experiments

4.1 Dataset and Evaluation Metric

We followed the experimental settings of LUQ [43] by utilising FACTSCORE [29], a fine-grained evaluation metric designed for factuality assessment of long-form text. FACTSCORE provides an automated method to estimate the factuality score of a given text by first decomposing it into atomic facts using LLMs and then verifying these facts against a custom knowledge source through retrieval. With a low error rate of just 2%, FACTSCORE ensures a high level of accuracy in quantifying the factuality of long text.

FACTSCORE also offers datasets for evaluation, which includes 183 names with human-annotated factuality labels corresponding to their respective Wikipedia titles, as well as an unlabeled dataset containing 500 additional names. To enhance the factuality assessment process, we batch-crawled page content from Wikipedia to construct our own comprehensive knowledge source. This custom knowledge source allowed us to further measure the factuality scores of text responses generated by various LLMs.

We determine the factuality score of a query prompt r_a by averaging the FACTSCOREs based on the responses. Then we use Pearson Correlation Coefficient (PCC) to quantify the linear relationship between the factuality scores and uncertainties. Additionally, we use the Spearman Correlation Coefficient (SCC) to measure the monotonic relationship between the two.

4.2 LLMs and Baseline Methods

We utilised six top-performing large language models (LLMs) to conduct our evaluation: ChatGPT-4o [35], ChatGPT-4 [1], ChatGPT-3.5-turbo [34], Llama-3.1-8B [9], Llama-3.1-70B [9], and Vicuna-33B [6]. This selection includes both lightweight models, such as Llama-3.1-8B, and larger parameter models, like Llama-3.1-70B and Vicuna-33B, along with the state-of-the-art ChatGPT series. These models provide a comprehensive range for assessing the capabilities of uncertainty quantification.

Our study follows the framework of LM-Polygraph [11], which implements a variety of uncertainty estimation methods. For white-box approaches, we selected three prominent methods: Maximum Sequence Probability (MSP), Monte Carlo Sequence Entropy (MCSE) [28], and Semantic Entropy (SE) [23]. These methods directly leverage the internal mechanics of LLMs for estimating uncertainty.

Additionally, we included several black-box methods as baselines, which operate independently of the internal states of the model: Lexical Similarity (LexSim) [13], Number of Semantic Sets (NumSets) [24], Sum of Eigenvalues of the Graph Laplacian (EigV) [24], Degree Matrix (Deg) [24], Eccentricity (Ecc) [24], and Long-Text Uncertainty Quantification (LUQ) [43].

5 Uncertainty Quantification Results

Effectiveness for Open-Source Model Table 1 presents the correlation coefficients between the uncertainties derived from various uncertainty quantification methods and the factuality scores determined by FACTSCORE. Ideally, greater uncertainty in a LLM should correspond to lower factuality in its outputs. Our proposed methods demonstrate a strong correlation with factuality scores, particularly in open-source models such as Llama-3.1-8B, Llama-3.1-70B, and Vicuna-33B. The effectiveness of our method extends across both lightweight models with 8 billion parameters and larger-scale models with 70 billion parameters. While baseline approaches such as LUQ, LexSim, and EigV occasionally exhibit robustness, our methods consistently provide stability for these open-source models. NumSets interestinly displays a unique behavior, showing near-zero correlation, indicating its limited capability for uncertainty quantification in open-source models. Moreover, white-box methods consistently outperform black-box methods, delivering greater reliability and stability across different models.

For GPTs Uncertainty quantification methods, in general, exhibit low correlation between predicted uncertainties and factuality scores for GPTs models, highlighting limited effectiveness in accurately assessing uncertainty. Moreover, white-box methods are not applicable in this context due to the inaccessibility of internal model states. Certain methods, such as Deg and LUQ, even display positive correlations, which suggests poor uncertainty quantification performance, as higher uncertainty should ideally correspond to lower factuality. While our methods don't acquire the highest score for each model, they consistently maintain a strong correlation with factuality scores as measured by FACTSCORE, demonstrating the robustness and effectiveness in uncertainty quantification.

Effectiveness of Knowledge Graph As discussed, our method consists of three parts, with the construction of a knowledge graph serving as a preprocessing step to break down long texts into distinct statements, thereby simplifying their semantic structure. We compare the performance of the pipeline with and without the knowledge graph construction step, and Table 2 demonstrates its effectiveness. Knowledge graphs prove inherently beneficial for Llama models, significantly improving the correlation between uncertainty and FACTSCORE. However, the performance with knowledge graph took a dive with other models.

This disparity arises because constructing a knowledge graph from long text is inherently challenging, and current methods cannot guarantee capturing all key information points. Existing approaches, such as LangChain and REBEL[17], rely on LLMs[36], which exhibit inherent uncertainty. The information extracted from the same sentence may vary across iterations, and there is no mechanism to ensure that the triples generated by LLMs comprehensively cover all critical information points.

Additionally, the expectation that a knowledge graph should encapsulate all aspects of a long text may be inherently unrealistic.

Table 1: Performance Comparison of Various Methods with Different LLMs

		White-Box Methods			**Black-Box Methods**								
		MSP	**MCSE**	**SE**	**LexSim**	**Ecc**	**NumSets**	**EigV**	**Deg**	**LUQ**	**Our**	**OpenAI**	**OpenAI-KG**
FACTSCORE-BIO													
GPT-4o	PCC	-	-	-	-0.4954	0.0335	-0.1553	-0.2367	0.1095	0.1948	0.1253	-0.6647	-0.4577
	SCC	-	-	-	-0.4449	0.0967	-0.1404	-0.0041	0.2158	0.1790	-0.1970	-0.4634	-0.2588
GPT-4	PCC	-	-	-	0.0908	0.0918	-0.2399	-0.1534	0.1469	0.2385	-0.1922	-0.4207	-0.2395
	SCC	-	-	-	-0.3383	0.1184	-0.2415	-0.0967	0.1628	0.3445	-0.2552	-0.3849	-0.2433
GPT-3.5-turbo	PCC	-	-	-	-0.5520	-0.3636	-0.6334	-0.6817	-0.6201	0.6849	-0.1350	-0.6727	-0.5447
	SCC	-	-	-	-0.5083	-0.4389	-0.6058	-0.6441	-0.6403	0.6509	-0.1581	-0.5630	-0.4665
Llama-3.1-8B	PCC	-0.2634	-0.5035	-0.3933	0.1439	-0.2473	-0.0929	-0.5171	-0.3198	-0.6542	-0.3752	-0.2387	-0.5536
	SCC	-0.3550	-0.5088	-0.4991	-0.0109	-0.1906	-0.0936	-0.5640	-0.2955	-0.6492	-0.4777	-0.2566	**-0.7978**
Llama-3.1-70B	PCC	-0.3821	-0.6171	-0.5748	-0.2078	-0.3207	-0.2979	-0.6285	-0.4144	-0.3084	-0.3503	-0.6509	**-0.6299**
	SCC	-0.4306	-0.6259	-0.6082	-0.4185	-0.3197	-0.3072	-0.6802	-0.4172	-0.2697	-0.4780	-0.6204	**-0.7303**
Vicuna-33B	PCC	-0.5358	-0.7509	-0.7728	-0.7863	-0.2878	-0.1464	-0.5310	-0.4003	0.6150	-0.2183	-0.8355	-0.7715
	SCC	-0.5579	-0.7603	-0.7826	-0.7936	-0.1693	-0.1240	-0.5358	-0.3805	0.6520	-0.2323	-0.8317	-0.7735

Table 2: Performance Comparison of With and Without Knowledge Graph

Models	**GPT-4o**		**GPT-4**		**GPT-3.5-turbo**		**Llama-3.1-8B**		**Llama-3.1-70B**		**Vicuna-33B**	
	PCC	SCC	PCC	SCC	PCC	SCC	PCC	SCC	PCC	SCC	PCC	SCC
Without Knowledge Graph	-0.6647	-0.4634	-0.4207	-0.3849	-0.6727	-0.5630	-0.2387	-0.2566	-0.6509	-0.6204	-0.8355	-0.8317
With Knowledge Graph	-0.4577	-0.2588	-0.2395	-0.2433	-0.5447	-0.4665	**-0.5536**	**-0.7978**	-0.6299	**-0.7303**	-0.7715	-0.7735

For instance, consider the sentence from Botak Chin's autobiography we got from LLM: "*His life took a drastic turn when he entered the world of crime at the age of 20, starting with petty thefts and gradually escalating to armed robberies.*" This sentence comprises four distinct statements:

- His life took a drastic turn.
- He entered the world of crime at the age of 20.
- He started with petty thefts.
- He escalated to armed robberies.

Extracting all four statements simultaneously is challenging, and the extracted information often varies. For example, statement c could be rephrased as "*He began to steal petty things*", which conveys the same semantic meaning but introduces significant differences in the resulting knowledge graph, thereby increasing uncertainty. Moreover, the extracted statements may fail to capture the full context of the original sentence. For instance, statement c alone does not convey that these events occurred when "*his life took a drastic turn.*"

Although knowledge graphs are inherently effective at decomposing long text into discrete statements intended to capture the full semantic meaning of the original text, current methods fall short of achieving this goal. To our understanding, no approach to construct knowledge graphs that consistently extracts complete and identical information across iterations has been found .

6 Conclusion

In this work, we address the current limitations of uncertainty quantification methods for long text and explore the potential of knowledge graphs for their ability to transform unstructured text into structured data. This capability aids in deconstructing the complex semantic relationships inherent in long text. We propose KG-UQ, a UQ method based on KGs, which decomposes long text into multiple statements that collectively preserve the semantic meaning of the original text. Instead of treating the text as a single entity or relying on sentence-level consistency, KG-UQ identifies the smallest semantic units within lengthy paragraphs, effectively overcoming the challenges associated with long text. Our findings demonstrate that KGs are an effective tool for improving the correlation between estimated uncertainties and factuality scores for both open-source and closed-source LLMs. However, the process of constructing KGs remains challenging due to its reliance on LLMs, which often results in capturing only partial semantic meaning from the original text. This work highlights the promising applications of knowledge graphs in UQ and lays the groundwork for future research to further refine and expand upon these methodologies.

7 Limitation

This work has several limitations, which present opportunities for further exploration and improvement:

- **Knowledge Graph Construction Methods**: In this study, we construct KGs using LangChain and REBEL, both of which are LLM-based methods. However, the output KGs generated by these methods often fail to fully capture the complete semantic meaning of long text. Traditional methods for KG construction, which may offer complementary strengths, were not investigated and should be explored in future work.
- **Evaluation Metrics**: We rely on FACTSCORE as the primary metric to evaluate the correlation between estimated uncertainties and factuality scores. While effective, FACTSCORE may not always reflect real-world factuality accurately. Incorporating human evaluation-based methods could provide

a more precise assessment of the factuality of generated text, offering valuable insights for model improvements.

- **Temperature Sensitivity**: We observe that temperature plays a crucial role in uncertainty quantification. Specifically, higher temperature settings often result in stronger correlations between uncertainties and factuality scores. Future studies could delve deeper into how temperature and other decoding parameters influence uncertainty estimation across different models and datasets.
- **Scalability and Efficiency**: While KGs are effective in handling the semantic complexity of long text, their construction process can be computationally expensive, especially for large-scale applications. Optimizing KG generation for scalability and efficiency would be a valuable direction for future research.

Acknowledgements

This research was supported by the Macquarie University Research Acceleration Scheme (MQRAS) and Data Horizon funding.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[2] Pranav Bhandari, Nicolas Fay, Michael Wise, Amitava Datta, Stephanie Meek, Usman Naseem, and Mehwish Nasim. 2025. Can LLM Agents Maintain a Persona in Discourse? *arXiv preprint arXiv:2502.11843* (2025).

[3] Pranav Bhandari, Usman Naseem, Amitava Datta, Nicolas Fay, and Mehwish Nasim. 2025. Evaluating Personality Traits in Large Language Models: Insights from Psychological Questionnaires. *arXiv preprint arXiv:2502.05248* (2025).

[4] Shijing Chen, Mohamed Reda Bouadjenek, Shoaib Jameel, Usman Naseem, Basem Suleiman, Flora D Salim, Hakim Hacid, and Imran Razzak. 2025. Leveraging Taxonomy and LLMs for Improved Multimodal Hierarchical Classification. *arXiv preprint arXiv:2501.06827* (2025).

[5] Xuelu Chen, Muhao Chen, Weijia Shi, Yizhou Sun, and Carlo Zaniolo. 2019. Embedding uncertain knowledge graphs. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 3363–3370.

[6] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality. https://lmsys.org/blog/2023-03-30-vicuna/

[7] Armen Der Kiureghian and Ove Ditlevsen. 2009. Aleatory or epistemic? Does it matter? *Structural safety* 31, 2 (2009), 105–112.

[8] Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. 2024. Shifting Attention to Relevance: Towards the Predictive Uncertainty Quantification of Free-Form Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 5050–5063. https://doi.org/10.18653/v1/2024.acl-long.276

[9] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).

[10] Romain Egele, Romit Maulik, Krishnan Raghavan, Bethany Lusch, Isabelle Guyon, and Prasanna Balaprakash. 2022. Autodeuq: Automated deep ensemble with uncertainty quantification. In *2022 26th International Conference on Pattern Recognition (ICPR)*. IEEE, 1908–1914.

[11] Ekaterina Fadeeva, Roman Vashurin, Akim Tsvigun, Artem Vazhentsev, Sergey Petrakov, Kirill Fedyanin, Daniil Vasilev, Elizaveta Goncharova, Alexander Panchenko, Maxim Panov, Timothy Baldwin, and Artem Shelmanov. 2023. LM-Polygraph: Uncertainty Estimation for Language Models. arXiv:2311.07383 [cs.CL] https://arxiv.org/abs/2311.07383

[12] Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature* 630, 8017 (2024), 625–630.

[13] Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. 2020. Unsupervised Quality Estimation for Neural Machine Translation. arXiv:2005.10608 [cs.CL] https://arxiv.org/abs/2005.10608

[14] Xiang Gao, Jiaxin Zhang, Lalla Mouatadid, and Kamalika Das. 2024. SPUQ: Perturbation-Based Uncertainty Quantification for Large Language Models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, Yvette Graham and Matthew Purver (Eds.). Association for Computational Linguistics, St. Julian's, Malta, 2336–2346. https://aclanthology.org/2024.eacl-long.143

[15] Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. 2023. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review* 56, Suppl 1 (2023), 1513–1589.

[16] Stephen C Hora. 1996. Aleatory and epistemic uncertainty in probability elicitation with an example from hazardous waste management. *Reliability Engineering & System Safety* 54, 2-3 (1996), 217–223.

[17] Pere-Lluís Huguet Cabot and Roberto Navigli. 2021. REBEL: Relation Extraction By End-to-end Language generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. Association for Computational Linguistics, Punta Cana, Dominican Republic, 2370–2381. https://aclanthology.org/2021.findings-emnlp.204

[18] Eyke Hüllermeier and Willem Waegeman. 2021. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine learning* 110, 3 (2021), 457–506.

[19] Di Jin, Shuyang Gao, Seokhwan Kim, Yang Liu, and Dilek Hakkani-Tür. 2022. Towards textual out-of-domain detection without in-domain labels. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 30 (2022), 1386–1395.

[20] Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. 2022. Language Models (Mostly) Know What They Know. arXiv:2207.05221 [cs.CL] https://arxiv.org/abs/2207.05221

[21] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).

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

[23] Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation. arXiv:2302.09664 [cs.CL] https://arxiv.org/abs/2302.09664

[24] Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2024. Generating with Confidence: Uncertainty Quantification for Black-box Large Language Models. *Transactions on Machine Learning Research* (2024). https://openreview.net/forum?id=DWkJCSxKU5

[25] Chen Ling, Xujiang Zhao, Xuchao Zhang, Wei Cheng, Yanchi Liu, Yiyou Sun, Mika Oishi, Takao Osaki, Katsushi Matsuda, Jie Ji, Guangji Bai, Liang Zhao, and Haifeng Chen. 2024. Uncertainty Quantification for In-Context Learning of Large Language Models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 3357–3370. https://doi.org/10.18653/v1/2024.naacl-long.184

[26] Haohui Lu and Usman Naseem. 2024. Can Large Language Models Enhance Predictions of Disease Progression? Investigating Through Disease Network Link Prediction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 17703–17715.

[27] Wesley J Maddox, Pavel Izmailov, Timur Garipov, Dmitry P Vetrov, and Andrew Gordon Wilson. 2019. A simple baseline for bayesian uncertainty in deep learning. *Advances in neural information processing systems* 32 (2019).

[28] Andrey Malinin and Mark Gales. 2021. Uncertainty Estimation in Autoregressive Structured Prediction. arXiv:2002.07650 [stat.ML] https://arxiv.org/abs/2002.07650

[29] Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. arXiv:2305.14251 [cs.CL] https://arxiv.org/abs/2305.14251

[30] Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 12076–12100. https://doi.org/10.18653/v1/2023.emnlp-main.741

[31] Kenton Murray and David Chiang. 2018. Correcting length bias in neural machine translation. *arXiv preprint arXiv:1808.10006* (2018).

[32] Akram Mustafa, Usman Naseem, and Mostafa Rahimi Azghadi. 2025. Large language models vs human for classifying clinical documents. *International*

Journal of Medical Informatics (2025), 105800.

[33] Abdulsalam obaid Alharbi, Abdullah Alsuhaibani, Abdulrahman Abdullah Alalawi, Usman Naseem, Shoaib Jameel, Salil Kanhere, and Imran Razzak. 2025. Evaluating Large Language Models on Health-Related Claims Across Arabic Dialects. In *Proceedings of the 1st Workshop on NLP for Languages Using Arabic Script*. 95–103.

[34] OpenAI. 2022. Introducing ChatGPT. https://openai.com/index/chatgpt/

[35] OpenAI. 2024. Hello GPT-4o. https://openai.com/index/hello-gpt-4o/

[36] Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2024. Unifying Large Language Models and Knowledge Graphs: A Roadmap. *IEEE Transactions on Knowledge and Data Engineering* 36, 7 (July 2024), 3580–3599. https://doi.org/10.1109/tkde.2024.3352100

[37] Amin Qasmi, Usman Naseem, and Mehwish Nasim. 2025. Competing LLM Agents in a Non-Cooperative Game of Opinion Polarisation. *arXiv preprint arXiv:2502.11649* (2025).

[38] Hannah Sansford, Nicholas Richardson, Hermina Petric Maretic, and Juba Nait Saada. 2024. GraphEval: A Knowledge-Graph Based LLM Hallucination Evaluation Framework. arXiv:2407.10793 [cs.CL] https://arxiv.org/abs/2407.10793

[39] Linwei Tao, Minjing Dong, and Chang Xu. 2023. Dual focal loss for calibration. In *International Conference on Machine Learning*. PMLR, 33833–33849.

[40] Linwei Tao, Haolan Guo, Minjing Dong, and Chang Xu. 2024. Consistency Calibration: Improving Uncertainty Calibration via Consistency among Perturbed Neighbors. *arXiv preprint arXiv:2410.12295* (2024).

[41] Artem Vazhentsev, Akim Tsvigun, Roman Vashurin, Sergey Petrakov, Daniil Vasilev, Maxim Panov, Alexander Panchenko, and Artem Shelmanov. 2023. Efficient Out-of-Domain Detection for Sequence to Sequence Models. In *Findings of the Association for Computational Linguistics: ACL 2023*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 1430–1454. https://doi.org/10.18653/v1/2023.findings-acl.93

[42] Caiqi Zhang, Fangyu Liu, Marco Basaldella, and Nigel Collier. 2024. LUQ: Long-text Uncertainty Quantification for LLMs. *arXiv preprint arXiv:2403.20279* (2024).

[43] Caiqi Zhang, Fangyu Liu, Marco Basaldella, and Nigel Collier. 2024. LUQ: Long-text Uncertainty Quantification for LLMs. arXiv:2403.20279 [cs.CL] https://arxiv.org/abs/2403.20279

[44] Zhihao Zhang, Carrie-Ann Wilson, Rachel Hay, Yvette Everingham, and Usman Naseem. 2025. BeefBot: Harnessing Advanced LLM and RAG Techniques for Providing Scientific and Technology Solutions to Beef Producers. In *Proceedings of the 31st International Conference on Computational Linguistics: System Demonstrations*. 54–62.

[45] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2024. A Survey of Large Language Models. arXiv:2303.18223 [cs.CL] https://arxiv.org/abs/2303.18223

[46] Younan Zhu, Linwei Tao, Minjing Dong, and Chang Xu. 2025. Mitigating Object Hallucinations in Large Vision-Language Models via Attention Calibration. *arXiv preprint arXiv:2502.01969* (2025).