AGSC: Adaptive Granularity and Semantic Clustering for Uncertainty Quantification in Long-text Generation

Guanran Luo[1,†], Wentao Qiu[1,†], Wanru Zhao[1], Wenhan Lv[2], Zhongquan Jian[5], Meihong Wang[1], Qingqiang Wu[1,2,3,4,*]

[1]School of Informatics, Xiamen University [2]School of Film, Xiamen University
[3]Key Laboratory of Digital Protection and Intelligent Processing of Intangible Cultural Heritage of Fujian and Taiwan, Ministry of Culture and Tourism, Xiamen University
[4]Institute of Artificial Intelligence, Xiamen University
[5]School of Computer and Data Science, Minjiang University
luoguanran@stu.xmu.edu.cn, wuqq@xmu.edu.cn

Abstract

Large Language Models (LLMs) have demonstrated impressive capabilities in long-form generation, yet their application is hindered by the hallucination problem. While Uncertainty Quantification (UQ) is essential for assessing reliability, the complex structure makes reliable aggregation across heterogeneous themes difficult, in addition, existing methods often overlook the nuance of neutral information and suffer from the high computational cost of fine-grained decomposition. To address these challenges, we propose **AGSC** (**A**daptive **G**ranularity and **GMM**-based **S**emantic **C**lustering), a UQ framework tailored for long-form generation. AGSC first uses NLI neutral probabilities as triggers to distinguish *irrelevance* from *uncertainty*, reducing unnecessary computation. It then applies Gaussian Mixture Model (GMM) soft clustering to model latent semantic themes and assign topic-aware weights for downstream aggregation. Experiments on BIO and LongFact show that AGSC achieves state-of-the-art correlation with factuality while reducing inference time by about 60% compared to full atomic decomposition.

1 Introduction

With the rapid development and widespread application of Large Language Models (LLMs), their potential in generating long-text content is increasingly prominent. Traditional Natural Language Processing (NLP) tasks have evolved into use cases primarily involving the generation of long-text responses (Zhao et al., 2023; Chang et al., 2023). However, a core challenge for LLMs lies in their inherent hallucination problem, where the model may output inaccurate or unfaithful information with high confidence (Manakul et al., 2023; Zhang et al., 2023; Wang et al., 2023). Therefore, accurate and robust Uncertainty Quantification (UQ) of LLM outputs is essential for enhancing their trustworthiness and application in critical domains.

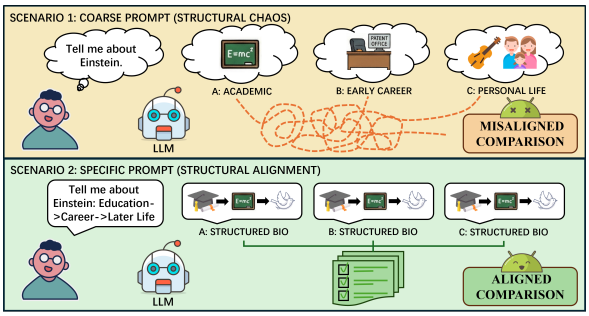

Figure 1: Illustration of prompt-induced structural uncertainty.

Previous research on modeling uncertainty has mainly focused on short responses (Kuhn et al., 2023; Duan et al., 2023; Lin et al., 2024; Xiao et al., 2022; Xiong et al., 2023; Huang et al., 2023). However, with the widespread application of Large Language Models (LLMs), generative tasks involving long text have become more common use cases, such as summarization, open-ended question answering, and multi-agent collaboration. Traditional confidence-based and semantic entropy methods (Hendrycks and Gimpel, 2017; Geifman and El-Yaniv, 2017; Lin et al., 2022; Malinin and Gales, 2021; Mielke et al., 2022) can only make a single overall correctness judgment for the entire response, failing to capture local errors in different facts within long text.

LUQ (Zhang et al., 2024a) addresses this by breaking down each response into sentences or atomic facts as the hypothesis and checking if they can be supported by other samples serving as the premise, thus enabling a more granular evaluation.

*Corresponding authors.
†These authors contributed equally to this work.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 9591–9605
July 2-7, 2026 ©2026 Association for Computational Linguistics

However, this method has the following limitations:

1. **Efficiency-Granularity Trade-off:** Fine-grained decomposition drastically increases computational overhead and LLM calls, making it difficult to balance accuracy with efficiency.

2. **Topic Heterogeneity in Aggregation:** Long-form responses often mix multiple semantic themes, where minor/off-topic parts can disproportionately affect overall uncertainty if all units are pooled uniformly. A theme-aware aggregation mechanism is needed to down-weight such minor/noisy portions.

3. **Underutilization of Neutrality:** LUQ simply discards the *neutral* label. However, neutrality often signals epistemic uncertainty that requires further granularity analysis rather than direct exclusion.

Figure 1 provides an intuitive explanation. With a coarse prompt such as "Tell me about Einstein," different samples may organize the content around different latent aspects (e.g., academic vs. early career vs. personal life), leading to *structural chaos*. As a result, different samples may emphasize different themes, and naive pooling can be overly influenced by minor or off-topic parts. Moreover, coarse prompts often introduce many contextually irrelevant sentences; when compared against a given hypothesis sentence, these irrelevant contexts tend to yield a high probability of the *neutral* class.

To address these limitations, we propose **AGSC** (**A**daptive **G**ranularity and **S**emantic **C**lustering). AGSC leverages the NLI neutral category to dynamically trigger decomposition for accurate yet efficient evaluation, and employs Gaussian Mixture Model (GMM) soft clustering to capture latent semantic themes and perform topic-weighted aggregation, reducing the influence of minor/noisy parts. Finally, uncertainty is computed from NLI-based unit uncertainties and aggregated with cluster-aware weights. Experiments on FActScore (Min et al., 2023) and LongFact (Wei et al., 2024) show that AGSC achieves state-of-the-art (SOTA) correlations with factuality.

Our main contributions are:

- We propose an **Adaptive Granularity Strategy** that uses NLI neutrality to trigger fine-grained decomposition or filter noise, effectively balancing UQ accuracy with computational efficiency.

- We introduce **GMM-based Semantic Clustering** to model latent themes and assign topic-aware weights, which down-weights minor/noisy parts during aggregation and yields more stable long-form uncertainty estimates.

- AGSC achieves **SOTA performance** on two long-text benchmarks, consistently outperforming baselines in correlating uncertainty with factuality.

2 Related Work

Epistemic Uncertainty in Machine Learning. Uncertainty quantification has been a foundational topic in machine learning long before the era of LLMs (Gawlikowski et al., 2023; Zhang et al., 2024b, 2025; Wang et al., 2026; Zhang et al., 2026; Wang et al., 2025). Standard taxonomies categorize uncertainty into two primary types: *aleatoric uncertainty* and *epistemic uncertainty* (Hora, 1996; Hüllermeier and Waegeman, 2021; Der Kiureghian and Ditlevsen, 2009). In the context of LLM hallucination, non-factual generation is often attributed to the model's lack of relevant knowledge (Huang et al., 2023). Therefore, consistent with recent work (Kuhn et al., 2023; Lin et al., 2024), our research focuses on **epistemic uncertainty**. Recent studies have also explored learning conditional dependencies to disentangle unconditional truthfulness from generation artifacts (Vazhentsev et al., 2025), further refining the theoretical basis for generative UQ.

Uncertainty Quantification in LLMs. Existing approaches are generally divided into white-box and black-box methods (Geng et al., 2023). *White-box methods* rely on internal logits or attention. Measures such as Semantic Entropy (SE) (Kuhn et al., 2023) use token probabilities to estimate confidence. Complementary formulations for in-context learning scenarios have also been proposed to better leverage model internals (Ling et al., 2024). However, these methods often fail to capture semantic equivalence in long sequences (Duan et al., 2023) and are inapplicable to closed-source models. *Black-box methods* have thus gained prominence. *Verbalized uncertainty* prompts models to express confidence (Lin et al., 2022; Tian et al., 2023), though recent work on "self-known"

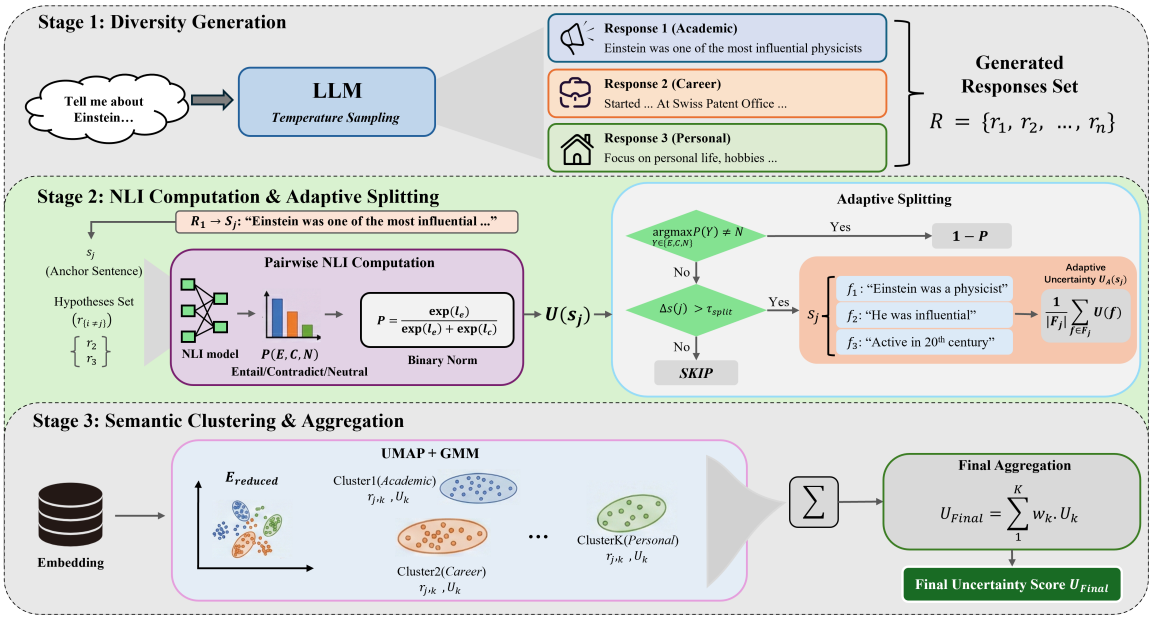

Figure 2: Overview of the **AGSC** framework. The framework consists of three stages: (1) Diversity Generation, where multiple responses are sampled; (2) NLI Computation & Adaptive Splitting, where sentences are analyzed via NLI and adaptively decomposed into atomic facts or skipped; and (3) Semantic Clustering & Aggregation, where units are softly clustered via UMAP and GMM to derive the final uncertainty score.

vs. "self-unknown" knowledge suggests these self-assessments can be intricate and prone to miscalibration (Tu et al., 2025). *Consistency-based methods*, which AGSC follows, measure semantic divergence among sampled responses (Manakul et al., 2023; Wang et al., 2023). Beyond simple sampling, perturbation-based methods like SPUQ (Gao et al., 2024) introduce input noise to assess stability, offering a robust alternative to evaluating epistemic uncertainty in black-box settings.

Granularity and Semantic Clustering. A critical open problem in long-form UQ is the choice of *granularity* and how to aggregate evidence under *topic heterogeneity*. Early works operated at the sentence level (Manakul et al., 2023), which may miss mixed veracity within long sentences. Recent methods move towards finer granularity by decomposing responses into *atomic facts* (Min et al., 2023; Zhang et al., 2024a), but this often incurs substantial computational overhead. Beyond granularity, long responses may mix multiple semantic themes; naive uniform pooling can be overly influenced by minor or off-topic portions. Several lines of work leverage semantic similarity or structured representations to improve fine-grained UQ, such as Kernel Language Entropy (Nikitin et al., 2024) and graph-based formulations (Lin et al., 2024). KG-based decomposition (Yuan et al., 2025) can offer precise structure but depends on external extractors. In contrast, AGSC focuses on (i) adaptively selecting granularity using NLI neutrality signals to reduce unnecessary decomposition, and (ii) applying GMM-based *soft semantic clustering* to obtain theme-aware weights, enabling more stable aggregation in long-form uncertainty estimation.

3 Method

In this section, we propose **AGSC** (Adaptive Granularity and Semantic Clustering). As illustrated in Figure 2, our method consists of three primary stages: (1) **Adaptive Granularity Triggering**; (2) **GMM-based Semantic Clustering**; and (3) **Uncertainty Aggregation**.

3.1 Notation

First, we define a terminology hierarchy:

- **Text Unit (Unit):** The general term for any segment being evaluated.
- **Sentence:** The coarse-grained unit from the original generation.
- **Atomic Fact:** The fine-grained unit resulting from decomposition.
- **Premise Unit (u):** Refers to the specific Text Unit (Sentence or Atomic Fact) currently serving as the premise for NLI verification.

We sample n responses $\mathcal{R} = \{r_1, \ldots, r_n\}$ via temperature sampling, and treat r_1 as the *anchor response* to be evaluated. The remaining responses $\mathcal{R}_{\text{ref}} = \{r_2, \ldots, r_n\}$ serve as *reference evidence*. After adaptive granularity (Sec. 3.3), each response yields a set of valid semantic units $\mathcal{H}_i^*$. Each response r_i is segmented into sentences $\mathcal{S}_i = \{s_{i,1}, \ldots, s_{i,|\mathcal{S}_i|}\}$. We denote the union of all sentences as $\mathcal{S} = \bigcup_{i=1}^{n} \mathcal{S}_i$. After clustering (Sec. 3.4), each sentence $s_{i,j}$ is associated with a soft membership vector $\boldsymbol{\gamma}_{i,j} \in \mathbb{R}^K$, where $\gamma_{i,j,k}$ denotes its posterior responsibility to cluster k and $\sum_{k=1}^{K} \gamma_{i,j,k} = 1$.

For each NLI pair, let l_e, l_c, and l_n denote the raw logits for the entailment, contradiction, and neutral classes, respectively. We use $P_{\text{3cls}}(\cdot)$ to denote the original three-way NLI distribution over $\{\texttt{entail}, \texttt{contradict}, \texttt{neutral}\}$, and $P_{\text{binary}}(\cdot)$ to denote the binary-normalized distribution over $\{\texttt{entail}, \texttt{contradict}\}$ after removing the neutral class.

3.2 NLI-based Uncertainty Computation

Pairwise NLI score. For a premise unit u (a sentence or atomic fact from the anchor) and a hypothesis text v (a reference response), we employ a sliding window strategy to address the context length limit of NLI models. Specifically, the hypothesis v is segmented into a set of overlapping chunks $\mathcal{C}_v = \{c^{(1)}, \ldots, c^{(m)}\}$. We compute the NLI logits for u paired with each chunk individually. To ensure robust support detection, we aggregate these scores by selecting the chunk that yields the maximum entailment probability, assuming that valid support in any segment suffices to entail the premise.

Based on the logits (l_e, l_c, l_n) of this best-matching chunk, we ignore the neutral class in a binary normalization. The operation $\max$ is taken over all sliding window chunks $k \in \{1, \ldots, m\}$ of the reference text v:

$$P(\text{entail}|u, v) = \max_{k \in \{1,\ldots,m\}} \left(\frac{\exp(l_e^{(k)})}{\exp(l_e^{(k)}) + \exp(l_c^{(k)})} \right) \tag{1}$$

In the context of long-text hallucination detection, "neutral" sentences typically represent irrelevant information or chit-chat that does not impact the overall veracity of the response. Therefore, removing the neutral score serves to amplify the semantic tendency (Entailment vs. Contradiction) of the content without reversing its polarity.

Anchor-to-reference entailment support and uncertainty. Inspired by the entailment-support-based uncertainty principle in long-form UQ (Zhang et al., 2024a), we define the entailment support for an anchor unit $h \in \mathcal{H}_1^*$ as the average entailment score against all whole reference responses:

$$S(h) = \frac{1}{n-1} \sum_{t=2}^{n} P(\text{entail}|h, r_t), \quad h \in \mathcal{H}_1^* \tag{2}$$

Then the uncertainty of an anchor unit is defined as:

$$U(h) = 1 - S(h), \qquad h \in \mathcal{H}_1^*. \tag{3}$$

3.3 Adaptive Granularity

Neutral NLI predictions may reflect either (i) an *uncertainty tendency* (ambiguous content that should be decomposed) or (ii) an *irrelevance tendency* (off-topic noise that should be skipped). We distinguish them using the entailment–contradiction gap computed from anchor-to-reference NLI signals. For a sentence s_j in the anchor response r_1, we compute its average NLI distribution across whole reference responses:

$$\bar{P}(Y \mid s_j) = \frac{1}{n-1} \sum_{t=2}^{n} P(Y \mid s_j, r_t), \tag{4}$$

and define the dominant label:

$$\hat{y}_j = \underset{Y \in \{\text{entail,contra,neutral}\}}{\operatorname{argmax}} \bar{P}(Y \mid s_j). \tag{5}$$

Decompose vs. skip. If $\hat{y}_j \neq$ neutral, we keep s_j and later compute $U(s_j)$ using Eq. 3. If $\hat{y}_j =$ neutral, we compute the entailment–contradiction gap:

$$\Delta(s_j) = \left| \bar{P}(\text{entail} \mid s_j) - \bar{P}(\text{contra} \mid s_j) \right|. \tag{6}$$

If $\Delta(s_j) > \tau$, we decompose s_j into atomic facts $F_j = \text{Split}(s_j)$. Otherwise, s_j is treated as irrelevance noise and marked as SKIP.

Adaptive uncertainty. The adaptive uncertainty is:

$$U_{\text{A}}(s_j) = \begin{cases} U(s_j), & \hat{y}_j \neq \text{neutral}, \\ \text{SKIP}, & \hat{y}_j = \text{neutral} \wedge \Delta(s_j) \leq \tau, \\ \dfrac{1}{|F_j|} \displaystyle\sum_{f \in F_j} U(f), & \hat{y}_j = \text{neutral} \wedge \Delta(s_j) > \tau, \end{cases} \tag{7}$$

where $U(\cdot)$ is computed by Eq. 3. Units marked as SKIP are excluded from subsequent aggregation.

3.4 GMM-Based Semantic Clustering

Long-form generations may contain multiple semantic themes. To reduce the influence of minor/noisy parts during aggregation, we perform soft clustering over sentence units from all responses and use the resulting memberships as theme-aware weights for downstream aggregation.

Embedding and dimensionality reduction. Let $\mathcal{H}^* = \bigcup_{i=1}^{n} \mathcal{H}_i^*$ denote the union of all valid semantic units after adaptive granularity. We encode each unit $h \in \mathcal{H}^*$ into a sentence embedding $e_h \in \mathbb{R}^D$ using a pretrained embedding model, and denote $E = \{e_h\}_{h \in \mathcal{H}^*}$.

To improve robustness and efficiency for clustering, we optionally apply UMAP (McInnes et al., 2018) to project embeddings from $\mathbb{R}^D$ to $\mathbb{R}^{D'}$ ($D' \ll D$), yielding reduced embeddings $E_{\text{reduced}} = \{e_h^{\text{reduced}} \in \mathbb{R}^{D'}\}$.

Soft clustering via GMM. On the reduced embeddings $\{e_h^{\text{reduced}}\}$, we fit a K-component GMM (Bishop, 2006). The generative model assumes a latent cluster variable $z_h \in \{1, \ldots, K\}$:

$$p(z_h = k) = \pi_k, \qquad p\big(e_h^{\text{reduced}} \mid z_h = k\big) = \mathcal{N}\big(e_h^{\text{reduced}} \mid \mu_k, \Sigma_k\big). \tag{8}$$

Parameters are learned by maximizing the log-likelihood:

$$\log p(E_{\text{reduced}}) = \sum_{h \in \mathcal{H}^*} \log \Big(\sum_{k=1}^{K} \pi_k \mathcal{N}\big(e_h^{\text{reduced}} \mid \mu_k, \Sigma_k\big) \Big). \tag{9}$$

For each unit h, the posterior probability of belonging to cluster k serves as its soft semantic weight:

$$\gamma_{h,k} = p(z_h = k \mid e_h^{\text{reduced}}) = \frac{\pi_k \mathcal{N}\big(e_h^{\text{reduced}} | \mu_k, \Sigma_k\big)}{\sum_{\ell=1}^{K} \pi_\ell \mathcal{N}\big(e_h^{\text{reduced}} | \mu_\ell, \Sigma_\ell\big)}. \tag{10}$$

Intuitively, unlike hard clustering methods which classify points as either belonging to a single cluster or noise—thereby discarding necessary fine-grained probability weights—GMM provides the posterior probability $P(z = k|x)$ (soft assignment). This probabilistic membership is essential for handling semantically ambiguous sentences in long-form generation, better matching the ambiguous boundaries of natural language semantics and providing smoother, more stable aggregation downstream.

3.5 Uncertainty Aggregation

Cluster mass and cluster uncertainty. We define the mass of cluster k as the accumulated membership probability over anchor units:

$$M_k = \sum_{h \in \mathcal{H}_1^*} \gamma_{h,k}. \tag{11}$$

The cluster-specific uncertainty is the mass-normalized average of anchor-unit uncertainties:

$$U_k = \frac{1}{M_k} \sum_{h \in \mathcal{H}_1^*} \gamma_{h,k}\, U(h). \tag{12}$$

Mass-weighted global uncertainty. We set $w_k = M_k / \sum_{j=1}^{K} M_j$ so dominant themes in the anchor response contribute more. Finally, the uncertainty score for the anchor response r_1 is:

$$U_{\text{Final}} = \sum_{k=1}^{K} w_k \cdot U_k. \tag{13}$$

4 Experimental Setup

4.1 Datasets, Metrics, and Baselines

Datasets. We evaluate our method on two widely recognized benchmarks for long-form generation: **BIO** (Min et al., 2023), which focuses on biography generation, and **LongFact** (Wei et al., 2024), a comprehensive dataset covering diverse topics for long-context factuality evaluation.

For BIO, we use a test set of 500 entities, and generate 5 biographies per entity. For LongFact, we sample 10 prompts for each of 38 topics, yielding 380 prompts in total, and generate 5 responses per prompt.

Metrics. We employ **FActScore** (Min et al., 2023) as the ground-truth measure for the factuality of generated responses. Specifically, we apply FActScore to the **first generated response** for each question. To assess the performance of Uncertainty Quantification (UQ) methods, we measure how well the estimated uncertainty scores align with the actual factuality. Following standard practice for correlation analysis (Schober et al., 2018), we report the **Pearson Correlation Coefficient (PCC)** and **Spearman Correlation Coefficient (SCC)** between the calculated uncertainty scores and the FActScore.

Models. We use **GPT-4.1-mini** (OpenAI, 2025), **Qwen2.5-32B** (Bai et al., 2023), and **Llama3-70B** (Meta AI, 2024) as the generator models. For the uncertainty estimation components, we utilize **DeBERTa-v3-large-mnli** (He et al., 2021) as the Natural Language Inference (NLI) model to predict

entailment relationships. For semantic clustering, we employ **gte-large-en-v1.5** (Li et al., 2023) to generate sentence embeddings.

Baselines. We compare our approach against a comprehensive set of baselines: token-level **SE** (Semantic Entropy) (Kuhn et al., 2023); similarity-based **LexSim** (Lexical Similarity); graph-based measures including **Ecc** (Eccentricity) and **Deg** (Degree Matrix) (Lin et al., 2024). We also include advanced semantic baselines: **KLE** (Kernel Language Entropy) (Nikitin et al., 2024), which estimates fine-grained uncertainty via semantic similarities; **SPUQ** (Gao et al., 2024), a perturbation-based method for robust uncertainty estimation; and the NLI-based **SCN** (SelfCheckNLI) (Manakul et al., 2023). We also include SAR (Shifting Attention to Relevance) (Duan et al., 2024), a relevance-aware uncertainty estimation method. Finally, we compare with the **LUQ** framework and its variants **LUQ-Pair** and **LUQ-Atomic** (Zhang et al., 2024a).

4.2 Implementation Details

All experiments were conducted on an **NVIDIA GeForce RTX 4090** GPU. We generated $n = 5$ responses for each question using temperature sampling ($t = 0.7$).

For semantic clustering, we implement an adaptive soft clustering algorithm to obtain *theme-aware membership weights* used in downstream aggregation. We first apply **UMAP** to reduce embedding dimensions to **32**, configured with **15** nearest neighbors, a minimum distance of **0.1**, **cosine** metric, and **spectral** initialization. Then, we employ a GMM with **full** covariance type, **k-means++** initialization, and a regularization constant of 10^{-5}. The optimal number of clusters K is determined dynamically based on the **Bayesian Information Criterion (BIC)** with an improvement threshold of **0.01**. The pseudocode for this algorithm is provided in **Appendix A**.

For adaptive granularity triggering, we set the entailment–contradiction gap threshold to $\tau = 0.1$.

To decompose long-form responses into atomic facts, we utilize the `AtomicFactGenerator` from the **FActScore** library (Min et al., 2023). Following the standard implementation, this module employs a few-shot prompting strategy to split sentences into independent facts.

5 Results

5.1 Main Results

Table 1 presents the performance of our proposed AGSC method compared to various uncertainty quantification (UQ) baselines. AGSC achieves the highest correlation with FActScore in most experimental settings, consistently outperforming existing baselines. Among the LUQ variants, LUQ-Atomic generally outperforms the sentence-level LUQ, confirming that fine-grained verification is crucial. AGSC further advances this by dynamically selecting between sentence and atomic granularity, avoiding the computational redundancy of full atomic decomposition while maintaining high precision. In addition, AGSC introduces GMM-based *soft semantic clustering* to obtain theme-aware weights for aggregation, which mitigates the impact of minor/noisy parts when long responses mix multiple semantic themes.

While all UQ methods show a performance drop on the more challenging LongFact dataset due to its diverse topics and complex reasoning paths, AGSC maintains robust performance. For instance, on Llama3-70B, baseline methods struggle significantly (e.g., SE PCC is -0.067), whereas AGSC maintains a SCC of -0.229, demonstrating its robustness under topic heterogeneity and structural variance in long-context generation.

5.2 Ablation Study

Impact of Adaptive Granularity. We investigated the necessity of our adaptive granularity strategy by comparing it with three variants: (1) **w/o Adap.**: completely removing the adaptive mechanism; (2) **w/ NG**: assigning a fixed uncertainty score of 0.5 to sentences that should have been skipped; and (3) **w/ NW**: incorporating the neutral probability into the uncertainty calculation with a weight of 0.5.

As shown in Table 2, removing adaptive granularity (*w/o Adap.*) leads to a significant performance drop on both datasets. This confirms that simply ignoring or uniformly processing neutral sentences is insufficient. Furthermore, both heuristic strategies (*w/ NG* and *w/ NW*) underperform compared to AGSC. This indicates that neutrality in NLI is not a uniform signal of uncertainty or irrelevance. By adaptively distinguishing between *irrelevance tendency* (triggering Skip) and *uncertainty tendency* (triggering Decomposition), AGSC effectively filters noise while retaining verifiable

Dataset	Model	Metric ↓	SE	LexSim	Ecc	Deg	KLE	SPUQ	SCN	SAR	LUQ	LUQ-Pair	LUQ-Atomic	AGSC
BIO	GPT-4.1-mini	PCC	-	-0.438	-0.219	-0.329	-0.501	-0.515	-0.493	-0.515	-0.548	-0.381	-0.572	**-0.708**
		SCC	-	-0.423	-0.211	-0.317	-0.495	-0.508	-0.486	-0.490	-0.529	-0.241	-0.486	**-0.665**
	Qwen2.5-32B	PCC	-0.246	-0.262	-0.131	-0.197	-0.305	-0.315	-0.295	-0.405	-0.328	-0.414	-0.517	**-0.645**
		SCC	-0.318	-0.339	-0.169	-0.254	-0.390	-0.405	-0.382	-0.450	-0.424	-0.395	-0.541	**-0.646**
	Llama3-70B	PCC	-0.242	-0.258	-0.129	-0.193	-0.295	-0.305	-0.290	-0.285	-0.322	-0.255	-0.239	**-0.339**
		SCC	-0.251	-0.268	-0.134	-0.201	-0.308	-0.315	-0.301	-0.290	-0.334	-0.260	-0.250	**-0.335**
LongFact	GPT-4.1-mini	PCC	-	-0.163	-0.082	-0.122	-0.190	-0.195	-0.184	-0.255	-0.204	-0.324	-0.327	**-0.370**
		SCC	-	-0.167	-0.084	-0.125	-0.195	-0.201	-0.188	-0.280	-0.209	-0.298	-0.445	**-0.461**
	Qwen2.5-32B	PCC	-0.131	-0.140	-0.070	-0.105	-0.162	-0.168	-0.158	-0.210	-0.175	-0.288	-0.255	**-0.291**
		SCC	-0.089	-0.095	-0.048	-0.071	-0.112	-0.115	-0.107	-0.145	-0.119	-0.203	-0.175	**-0.256**
	Llama3-70B	PCC	-0.032	-0.034	-0.017	-0.025	-0.040	-0.041	-0.038	-0.095	-0.042	**-0.178**	-0.128	-0.175
		SCC	-0.067	-0.071	-0.036	-0.053	-0.082	-0.085	-0.080	-0.110	-0.089	-0.069	-0.169	**-0.229**

Table 1: Pearson (PCC) and Spearman (SCC) correlations between uncertainty metrics and FActScore. **Bold** indicates the highest correlation, and underlined indicates the second highest in each row. Values marked with '-' indicate inapplicable white-box metrics for API-based models.

Table 2: Ablation study of AGSC using **GPT-4.1-mini**. We report Pearson (PCC) and Spearman (SCC) correlations between uncertainty and FActScore. **Bold** indicates the best performance.

Section	Variant	BIO		LongFact	
		PCC	SCC	PCC	SCC
	AGSC	**-0.708**	**-0.665**	**-0.370**	**-0.461**
Sec 3.3	w/o Adap.	-0.512	-0.520	-0.292	-0.461
	w/ NG	-0.540	-0.540	-0.201	-0.295
	w/ NW	-0.564	-0.514	-0.198	-0.257
Sec 3.4	w/o Clustering.	-0.679	-0.622	-0.204	-0.307
	w/ K-Means	-0.531	-0.462	-0.229	-0.361

Table 3: Impact of different NLI backbone models on AGSC performance.

NLI Backbone	BIO		LongFact	
	PCC	SCC	PCC	SCC
DeBERTa-v3-large-mnli (Default)	-0.708	-0.665	-0.370	-0.461
GPT-5	**-0.735**	**-0.692**	**-0.405**	**-0.498**
RoBERTa-large-mnli	-0.685	-0.642	-0.358	-0.449
BART-large-mnli	-0.668	-0.625	-0.345	-0.436
DeBERTa-base-mnli	-0.632	-0.594	-0.328	-0.415

details, which simple numerical heuristics cannot achieve.

Impact of Semantic Clustering. We evaluated the effectiveness of the semantic clustering module by removing it (**w/o Clustering.**) or replacing the soft GMM clustering with hard K-Means clustering (**w/ K-Means**).

As shown in Table 2, removing semantic clustering leads to a clear decline in performance, particularly on LongFact (PCC drops from -0.370 to -0.204). This is expected, as LongFact exhibits stronger topic heterogeneity and structural variance, where naive uniform aggregation is more likely to be dominated by minor/noisy portions. Additionally, using K-Means performs significantly worse than the GMM approach (-0.531 vs. -0.708 on BIO), suggesting that semantic boundaries in long-form generation are often ambiguous. Hard clustering forces each sentence into a single topic and discards uncertainty in theme membership, whereas GMM's soft assignment preserves probabilistic memberships, enabling more stable theme-aware aggregation and more robust uncertainty estimation. A qualitative case study illustrating the learned cluster memberships and their impact on aggregation is provided in Appendix B.

5.3 Analysis

5.3.1 Robustness Analysis

Impact of NLI Backbone. We analyze the sensitivity of AGSC to different NLI models in Table 3. The results indicate a positive correlation between the capability of the NLI model and the final UQ performance. Utilizing **GPT-5** as the NLI backbone yields the best performance, suggesting that the upper bound of our method can be further extended with stronger NLI systems. However, considering the trade-off between cost and performance, our default choice, **DeBERTa-v3-large**, offers a balanced solution, significantly outperforming smaller models like RoBERTa-large and DeBERTa-base.

Impact of Prompt Granularity. We investigate how prompt specificity affects uncertainty estimation, as illustrated in Figure 3. We compare *Fine-grained Prompts* (e.g., "Tell me a short bio... Begin with birth... Include education...") against *Coarse-grained Prompts* (e.g., "Tell me a short bio...").

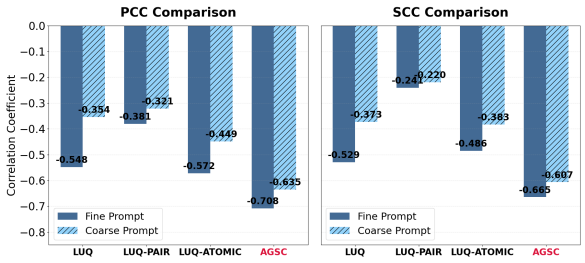

Figure 3: Robustness analysis of AGSC compared to LUQ variants across different prompt granularities.

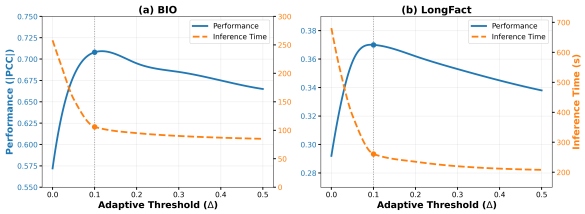

Figure 4: Sensitivity analysis of the adaptive threshold Δ on BIO and LongFact. Left y-axis (solid): Performance ($|\text{PCC}|$); right y-axis (dashed): Inference Time. For both datasets, the optimal trade-off is achieved at $\Delta = 0.1$.

Experiments show that fine-grained prompts generally lead to higher UQ correlations by reducing aleatoric uncertainty in writing style. However, AGSC demonstrates greater resilience to coarse prompts compared to baselines, benefiting from its GMM-based soft semantic clustering, which provides theme-aware aggregation weights and thus is less sensitive to structural variance inherent in open-ended generation.

5.3.2 Parameter Sensitivity

Sensitivity Analysis of Adaptive Threshold. As shown in Figure 4, the optimal threshold is consistently achieved at $\Delta = 0.1$ on both BIO and LongFact. On BIO, decreasing Δ leads to over-decomposition and unnecessary computation, while increasing Δ gradually discards useful uncertainty signals. A similar trend is observed on the more heterogeneous LongFact benchmark, confirming that the adaptive routing threshold generalizes beyond the relatively homogeneous BIO setting.

Impact of Response Quantity. Figure 5 plots the UQ performance against the number of sampled responses n. As expected, performance improves with n. Notably, AGSC consistently outperforms other baselines across all sample sizes (from $n = 2$ to 10), demonstrating superior stability.

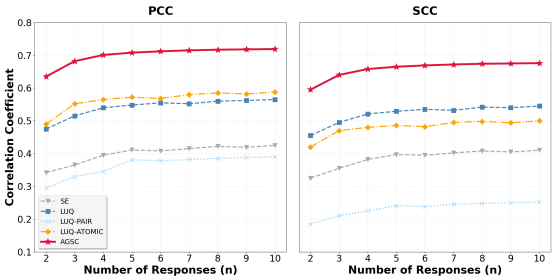

Figure 5: Impact of the number of sampled responses (n) on uncertainty estimation performance. The left and right plots show Pearson (PCC) and Spearman (SCC) correlations, respectively.

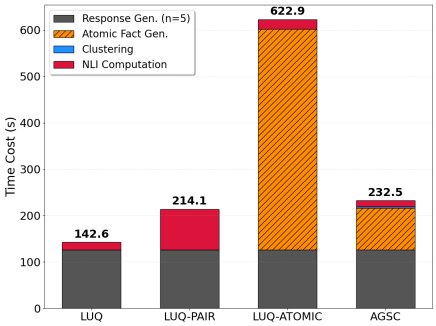

Figure 6: Time consumption breakdown of different uncertainty estimation methods (n=5).

5.3.3 Efficiency and Mechanism Analysis

Time Efficiency Analysis. Efficiency is a critical bottleneck for fine-grained semantic UQ. Figure 6 presents a fine-grained breakdown of computational latency. For **LUQ**, the total time is dominated by response generation T_{gen}. In **LUQ-Pair**, the quadratic complexity causes T_{nli} to become a new bottleneck. **LUQ-Atomic** introduces an additional atomic fact generation step T_{atom}. Since it indiscriminately decomposes every sentence, T_{atom} becomes the overwhelming dominant cost. In contrast, **AGSC** introduces a strategic optimization. Although it adds a clustering step T_{cluster}, this overhead is negligible. Crucially, by filtering out irrelevant neutral sentences, AGSC drastically reduces the volume of text requiring decomposition, thereby significantly cutting down T_{atom}.

Figure 7 illustrates the Time-Performance Trade-off, where AGSC occupies the optimal position on the Pareto frontier. Crucially, AGSC achieves superior performance compared to LUQ-Atomic while reducing the total inference time by approximately 60%, primarily by bypassing the decomposition of irrelevant neutral sentences.

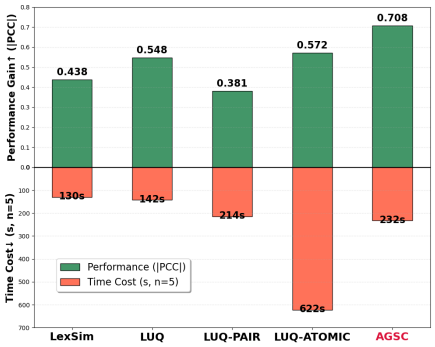

Figure 7: Cost-benefit analysis comparing performance gain versus time cost.

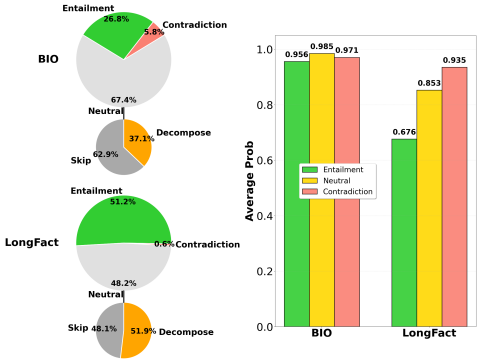

Figure 8: Analysis of NLI Distribution.

NLI Distribution and Adaptive Effectiveness. To understand the necessity of our adaptive mechanism, we analyze the NLI statistics in Figure 8. **Figure 8 (Left)** shows that the *Neutral* category constitutes a massive portion (40-70%) of the generated sentences. Within these neutral samples, $\sim$30-50% exhibit an *Uncertainty Tendency* that necessitates decomposition. Ignoring such a large segment leads to significant information loss.

Figure 8 (Right) presents the average probability of the corresponding category. We observe a distinct **probability saturation** on the BIO dataset: the NLI model assigns probabilities > 0.95 even for Neutral and Contradiction labels. This implies that the absolute value of P_N is insufficient to distinguish between Irrelevance and Uncertainty tendencies. AGSC addresses this by utilizing the *Entailment-Contradiction Gap* (Δ). By distinguishing tendencies based on polarity difference, AGSC triggers atomic decomposition for only $\sim$25% of the potential candidates compared to LUQ-Atomic. This selective granularity allows AGSC to achieve state-of-the-art performance with significantly reduced computational overhead.

6 Conclusion

We propose **AGSC** to address efficiency and topic heterogeneity challenges in long-form UQ. By integrating an **Adaptive Granularity Strategy** for selective decomposition with **GMM-based Semantic Clustering** for theme-aware aggregation, AGSC ensures stable uncertainty estimates. Experiments on BIO and LongFact demonstrate that AGSC achieves SOTA performance while reducing inference latency by $\sim$60% compared to full atomic decomposition.

Limitations

Dependency on NLI Calibration. The core of our *Adaptive Granularity* mechanism relies on the NLI model's ability to accurately classify the "Neutral" category. If the NLI model is miscalibrated—specifically, if it misclassifies subtle contradictions as neutral (false negatives) or fails to detect the entailment relationship in paraphrased texts—the adaptive trigger will fail. For instance, in highly technical domains (e.g., medical or legal texts), general-purpose NLI models (like DeBERTa-v3-trained on MNLI) often lack the domain knowledge to distinguish between a *factual contradiction* and a *neutral supplement*, potentially causing AGSC to incorrectly skip hallucinated content.

Vulnerability to "Echo Chamber" Hallucinations. Like all consistency-based methods, AGSC assumes that truth converges while hallucinations diverge. However, for certain widespread misconceptions or when the LLM has been poisoned with specific incorrect data during pretraining, the model may consistently generate the same hallucination across all K samples. In such cases of *systematic error*, high semantic consistency will result in a low uncertainty score, misleadingly indicating high factuality. AGSC measures *consensus*, not absolute *truth*.

Structural Constraints. AGSC is optimized for fact-heavy, descriptive texts (e.g., biographies, clinical notes). It may struggle with content where the "unit of truth" is not sentence-based, such as **mathematical reasoning** or **code generation**. In these cases, GMM clustering on sentence embeddings may break the logical flow, and NLI models are ill-equipped to verify logic-based entailment.

Linguistic Constraints. Our experiments are conducted exclusively in English using English-

centric embedding (GTE) and NLI models. The effectiveness of GMM-based *semantic clustering* relies on the quality of the embedding space. In low-resource languages where embedding models may exhibit anisotropy or poor semantic separability, the learned clusters (and thus the theme-aware aggregation weights) can become unstable, which may degrade the robustness of the final uncertainty estimate.

Ethics Statement

Reliability vs. Truth. Users must be aware that a low uncertainty score computed by AGSC indicates that the model is *confident and consistent*, but not necessarily *correct*. Deploying this metric in high-stakes decision-making systems (e.g., automated medical diagnosis or legal advice) carries the risk of *automation bias*, where users might over-trust the system based on a high consistency score derived from shared hallucinations.

Bias Propagation from Backbone Models. AGSC leverages NLI models and embedding models that are known to inherit social biases (e.g., gender, race, or religious stereotypes) from their training data. If an NLI model exhibits bias (e.g., consistently classifying premises involving certain demographics as "contradictions" or "neutral" without logical basis), AGSC will propagate this bias into the final uncertainty score. This could lead to systematically higher uncertainty estimates for content related to marginalized groups.

Artifacts and Licensing. In this study, we utilize several open-source scientific artifacts, including the Llama-3 , Qwen-2.5 , and DeBERTa-v3 models, as well as the BIO and LongFact datasets. All artifacts are used in strict accordance with their respective original licenses. Our proposed AGSC framework and its implementation code are provided in the supplementary materials.

AI Usage Statement. AI assistants were primarily employed for language polishing and as an auxiliary tool for generating specific code modules within the AGSC framework. It is important to emphasize that all AI-generated content and code were rigorously reviewed and manually verified by the authors to ensure technical accuracy, integrity, and the absence of bias. The final manuscript and implementation represent the original intellectual contributions of the authors.

Acknowledgements

We appreciate the valuable discussion from the anonymous reviewers. This work was supported by the Solfeggio Ear-Training Intelligent Robot and Cloud Platform RD Project for Music Education (No. 2024CXY0102), the 3D Visualization Digital Twin Integrated Control System Project(No. 2023CXY0111), the Pre-research Project for Introduced Talents of Minjiang University (No. MJY25025), the Public Technology Service Platform Project of Xiamen City(No. 3502Z20231043), and the Fujian Provincial Science and Technology Major Project (No. 2024HZ022003).

References

Jinze Bai, Shuai Bai, Yun-Hsuan Sung, Tian-Yi Tang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Christopher M. Bishop. 2006. *Pattern Recognition and Machine Learning*. Springer, New York, NY, USA.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, et al. 2023. A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109*.

Armen Der Kiureghian and Ove Ditlevsen. 2009. Aleatory or epistemic? does it matter? *Structural Safety*, 31(2):105–112.

Jinhao Duan, Hao Cheng, Shiqi Wang, Chenan Wang, Alex Zavalny, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. 2023. Shifting attention to relevance: Towards the uncertainty estimation of large language models. *arXiv preprint arXiv:2307.01379*.

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)*, pages 5050–5063, Bangkok, Thailand. Association for Computational Linguistics.

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 (EACL)*.

Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, et al. 2023. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 56(Suppl 1):1513–1589.

Yonatan Geifman and Ran El-Yaniv. 2017. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, pages 4878–4887.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. 2023. A survey of language model confidence estimation and calibration. *arXiv preprint arXiv:2311.08298*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Dan Hendrycks and Kevin Gimpel. 2017. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations (ICLR)*.

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):217–223.

Yuheng Huang, Jiayang Song, Zhijie Wang, Huaming Chen, and Lei Ma. 2023. Look before you leap: An exploratory study of uncertainty measurement for large language models. *arXiv preprint arXiv:2307.10236*.

Eyke Hüllermeier and Willem Waegeman. 2021. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110:457–506.

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

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. *arXiv preprint arXiv:2205.14334*.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2024. Generating with confidence: Uncertainty quantification for black-box large language models. *Transactions on Machine Learning Research*.

Chen Ling, Xujiang Zhao, Jiaying Lu, Chengyuan Deng, Can Zheng, Junxiang Wang, Tan Tan, Xiaoqiang Huang, Wei Xu, 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 (NAACL)*.

Andrey Malinin and Mark J. F. Gales. 2021. Uncertainty estimation in autoregressive structured prediction. In *International Conference on Learning Representations (ICLR)*.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore. Association for Computational Linguistics.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grošberger. 2018. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861.

Meta AI. 2024. Introducing the Llama 3 family of models. https://ai.meta.com/blog/meta-llama-3/. Meta AI Blog.

Sabrina J. Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. 2022. Reducing conversational agents' overconfidence through linguistic calibration. *Transactions of the Association for Computational Linguistics*, 10:857–872.

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. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Vladislav Nikitin, Jannik Kossen, and Yarin Gal. 2024. Kernel language entropy: Fine-grained uncertainty quantification for LLMs from semantic similarities. In *Advances in Neural Information Processing Systems (NeurIPS)*.

OpenAI. 2025. Introducing GPT-4.1 in the API. https://openai.com/index/gpt-4-1/. OpenAI Blog, published 2025-04-14, accessed 2025-12-17.

Patrick Schober, Christa Boer, and Lothar A Schwarte. 2018. Correlation coefficients: Appropriate use and interpretation. *Anesthesia & Analgesia*, 126(5):1763–1768.

Katherine Tian, Eric Mitchell, Huaxiu Yao, Christopher D Manning, and Chelsea Finn. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442.

Lifu Tu, Rui Meng, Shafiq Joty, Yingbo Zhou, and Semih Yavuz. 2025. Investigating factuality in long-form text generation: The roles of self-known and self-unknownan. In *Proceedings of the 2nd Workshop on Uncertainty-Aware NLP (UncertaiNLP 2025)*, pages 322–336, Suzhou, China. Association for Computational Linguistics.

Artem Vazhentsev, Ekaterina Fadeeva, Rui Xing, Gleb Kuzmin, Ivan Lazichny, Alexander Panchenko, Preslav Nakov, Timothy Baldwin, Maxim Panov, and Artem Shelmanov. 2025. Unconditional truthfulness: Learning unconditional uncertainty of large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 35661–35682, Suzhou, China. Association for Computational Linguistics.

Chengbing Wang, Yang Zhang, Wenjie Wang, Xiaoyan Zhao, Fuli Feng, Xiangnan He, and Tat-Seng Chua. 2025. Think-while-generating: On-the-fly reasoning for personalized long-form generation. *arXiv preprint arXiv:2512.06690*.

Chengbing Wang, Wuqiang Zheng, Yang Zhang, Fengbin Zhu, Junyi Cheng, Yi Xie, Wenjie Wang, and Fuli Feng. 2026. Perm: Psychology-grounded empathetic reward modeling for large language models. *arXiv preprint arXiv:2601.10532*.

Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Cheng Ji, et al. 2023. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. *arXiv preprint arXiv:2309.05660*.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and Quoc V. Le. 2024. Long-form factuality in large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 80756–80827. Curran Associates, Inc.

Yuxin Xiao, Paul Pu Liang, Umang Bhatt, Willie Neiswanger, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2022. Uncertainty quantification with pre-trained language models: A large-scale empirical analysis. *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7273–7284.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2023. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint arXiv:2310.04937*.

Yingqing Yuan, Linwei Tao, Haohui Lu, et al. 2025. KG-UQ: Knowledge graph-based uncertainty quantification for long text in large language models. In *Companion Proceedings of the ACM Web Conference (WWW)*.

Caiqi Zhang, Fangyu Liu, Marco Basaldella, and Nigel Collier. 2024a. LUQ: Long-text uncertainty quantification for LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Qianchi Zhang, Hainan Zhang, Liang Pang, Yongxin Tong, Hongwei Zheng, and Zhiming Zheng. 2025. Less is more: Compact clue selection for efficient retrieval-augmented generation reasoning. *arXiv preprint arXiv:2502.11811*.

Qianchi Zhang, Hainan Zhang, Liang Pang, Hongwei Zheng, and Zhiming Zheng. 2024b. Adacomp: Extractive context compression with adaptive predictor for retrieval-augmented large language models. *arXiv preprint arXiv:2409.01579*.

Qianchi Zhang, Hainan Zhang, Liang Pang, Hongwei Zheng, and Zhiming Zheng. 2026. Stable-rag: Mitigating retrieval-permutation-induced hallucinations in retrieval-augmented generation. *arXiv preprint arXiv:2601.02993*.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, et al. 2023. Siren's song in the ai ocean: A survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

A Adaptive Soft Clustering Algorithm

To address the varying number of sentences and semantic densities in long-form generation, we designed a GMM-based adaptive clustering algorithm. It integrates UMAP dimensionality reduction, a heuristic estimation for the maximum number of clusters ($K_{\max}$), and BIC-based model selection. The detailed procedure is outlined in Algorithm 1.

B Qualitative Case Study

In this section, we provide a qualitative analysis of real-world examples from the **LongFact** dataset (Topic: "Black Swan" movie) to demonstrate the effectiveness of AGSC's components and discuss their limitations.

B.1 Adaptive Granularity in Action

Standard sentence-level methods treat all "Neutral" NLI predictions equally. However, our Adaptive Granularity strategy distinguishes between "Irrelevant" content (to be skipped) and "Uncertain/Complex" content (to be decomposed) based on the entailment-contradiction gap (Δ). Figure 9 illustrates how these decisions are made.

B.2 Impact of GMM Semantic Clustering

Long-form responses often mix multiple semantic themes. We illustrate how GMM-based soft clustering provides theme-aware weights that stabilize aggregation under topic heterogeneity (Figure 10).

B.3 Failure Mode Analysis: Clustering Instability

While semantic clustering improves robustness, it is not infallible. Figure 11 describes a typical failure mode caused by ambiguous theme boundaries in the embedding space.

Algorithm 1 Adaptive Soft Clustering with UMAP and BIC

Require: Generated responses $\mathcal{R} = \{r_1, \ldots, r_n\}$; Embedding model $\mathcal{M}_{\text{emb}}$; UMAP target dimension $D' = 32$; BIC improvement threshold $\epsilon = 0.01$; Global max clusters $K_{\text{limit}} = 15$.

Ensure: Soft membership matrix $P \in \mathbb{R}^{N \times K}$, Cluster centers μ.

Step 1: Preprocessing & Embedding
Split all responses into a global sentence set $\mathcal{S} = \{s_1, \ldots, s_N\}$, where $N = \sum |r_i|$.
Compute embeddings: $X \leftarrow \mathcal{M}_{\text{emb}}(\mathcal{S})$, where $X \in \mathbb{R}^{N \times D}$.
Step 2: Dimensionality Reduction
if $D > D'$ **then**
    Apply UMAP: $X_{\text{reduced}} \leftarrow \text{UMAP}(X, \text{n\_components} = D')$.
else
    $X_{\text{reduced}} \leftarrow X$.
end if
Step 3: Heuristic $K_{\max}$ Selection
Calculate density constraint: $K_{\text{density}} \leftarrow \max(2, \lfloor N/3 \rfloor)$.
Calculate logarithmic constraint: $K_{\log} \leftarrow \lfloor \log_2(N) \rfloor + 1$.
Determine search limit:

$$K_{\max} \leftarrow \min(K_{\text{limit}}, K_{\text{density}}, K_{\log})$$

Step 4: BIC-based GMM Selection
Initialize $BIC_{\text{last}} \leftarrow \infty$, BestModel $\leftarrow$ None.
for $k = 2$ **to** $K_{\max}$ **do**
    Fit GMM with k components on X_{reduced}: $\mathcal{G}_k$.
    Calculate $BIC_{\text{current}} \leftarrow \text{BIC}(\mathcal{G}_k, X_{\text{reduced}})$.
    // Check relative improvement
    if BestModel is None **or** $BIC_{\text{current}} < BIC_{\text{last}} - \epsilon \cdot |BIC_{\text{last}}|$ **then**
        BestModel $\leftarrow \mathcal{G}_k$.
        $BIC_{\text{last}} \leftarrow BIC_{\text{current}}$.
    else
        break *// Stop if improvement is negligible*
    end if
end for
Step 5: Final Projection
Compute membership probabilities P using BestModel on X_{reduced}.
return P, BestModel.μ.

Adaptive Granularity Decisions

Case 1: The "Skip" Decision (Irrelevance Tendency)
Sentence: "In summary, 'Black Swan' is a layered, dark exploration of the pursuit of perfection..."
Context: This sentence appears at the end of a response as a concluding remark.
NLI Analysis: When compared to other factual responses, this sentence yields a **Neutral** prediction because it expresses a subjective summary not explicitly present in other samples.
Adaptive Trigger: The model detects a low Entailment-Contradiction gap ($\Delta \leq 0.1$), correctly identifying this as phatic/opinionated content.
Action: **SKIP**. This prevents the model from hallucinating atomic facts for subjective opinions, reducing noise.

Case 2: The "Decompose" Decision (Uncertainty Tendency)
Sentence: "Directed by Darren Aronofsky and co-written by him and Mark Heyman."
Comparison Target: Another response lists writers as "Mark Heyman, Andres Heinz, John McLaughlin", without mentioning Aronofsky as a co-writer.
NLI Analysis: The sentence contains mixed veracity. "Directed by Aronofsky" is Entailment, but "co-written by him" is not supported by the target response (**Neutral**).
Adaptive Trigger: The conflict creates a high gap ($\Delta > 0.1$).
Action: **Decompose** into atomic facts:

1. ***Directed by Darren Aronofsky*** $\rightarrow$ Supported
2. ***Co-written by Darren Aronofsky*** $\rightarrow$ Contradiction/Neutral
3. ***Co-written by Mark Heyman*** $\rightarrow$ Supported

Figure 9: Examples of the Skip vs. Decompose logic in AGSC's adaptive granularity module.

Semantic Clustering for Theme-aware Weighting

Target Sentence (s): 'The film stars Natalie Portman as Nina Sayers..." (from Response 1)
Without Clusteringt (Standard LUQ):
s is compared against *all* sentences in Response 2:

- vs. 'Black Swan is a horror film..." $\rightarrow$ Neutral (Noise)
- vs. 'The film stars Natalie Portman..." $\rightarrow$ **Entailment**
- vs. 'Plot Summary: The film follows..." $\rightarrow$ Neutral (Noise)

Result: The strong entailment signal is diluted by multiple Neutral scores from irrelevant sections (Plot, Intro).

With GMM Clustering (AGSC):
s is mapped to the 'Cast & Characters" cluster.

- **High Weight:** 'The film stars Natalie Portman..." (Cluster Match)
- **Low Weight:** 'Black Swan is a horror film..." (Different Cluster)

Result: The uncertainty score is derived primarily from the semantically relevant counterpart, resulting in a high-confidence **Entailment** score.

Figure 10: Comparison of NLI targets before and after GMM semantic clustering.

Clustering Failure Analysis

Example: "The film explores themes of ambition and duality."
Scenario: If sentence embeddings do not separate "themes" from "plot summary" cleanly, the soft memberships may spread across multiple clusters or assign low mass to the intended theme.
Consequence: Theme-aware aggregation may underweight this sentence (treating it as minor context), leading to a higher uncertainty estimate even when the statement is factual.
Result: This is a clustering/representation limitation: the method can mis-estimate uncertainty due to unstable theme weights, even though the NLI comparison target (whole reference responses) remains unchanged.

Figure 11: Analysis of a GMM clustering failure due to cross-cluster thematic ambiguity.