Hallucination Detection in Long-Form Text Generated by LLMs: A Benchmark and a Hyper-Relational Knowledge Graph Approach

Zituo Li, Guangzhou Chen, Jianbin Sun*, Qi Song and Kewei Yang
College of Systems Engineering, National University of Defense Technology
{lizituonudt, chen_gz, sunjianbin, songqi, kayyang27}@nudt.edu.cn

Abstract

Hallucination detection has received increasing attention, particularly in long-form text generation, where language models are more prone to producing factually inaccurate content. Previous studies reveal two limitations: (1) current benchmarks focus on short-form content, lacking the structural complexity required in long-form scenarios; (2) existing methods are constrained by coarse-grained consistency checks and fail to capture long-range and hyper-relational dependencies. To address these challenges, we provide LHD, a benchmark for long-form hallucination detection that contains diverse entity types and intricate factual dependencies spanning extended contexts. We further propose HRKG-HD, a black-box framework that models responses as fact-centric hyper-relational knowledge graphs and detects hallucinations through relation-aware multi-hop reasoning over these graphs. By linking distant facts through shared entities and qualifiers, this design enables a global and dependency-aware verification of factual consistency. Extensive experiments demonstrate that HRKG-HD not only outperforms existing baselines at both the passage-level and sentence-level, but also exhibits robust and consistent performance across various LLMs. The code and dataset will be released upon acceptance to facilitate reproducibility.

1 Introduction

Large language models (LLMs) have demonstrated remarkable generative capabilities across a variety of tasks, including summarization, dialogue and question answering (Gao et al., 2025; Staniszewski et al., 2025; Zhan et al., 2026, 2025). Despite their success, LLMs frequently generate hallucinations-outputs that are fluent and syntactically correct but factually incorrect or unverifiable (Huang et al., 2025; Zhang et al., 2025; Lyu et al., 2025). The risk

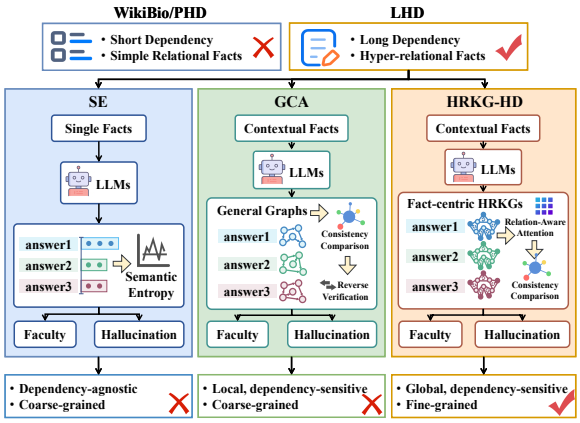

Figure 1: Comparison of hallucination detection benchmarks and methods. LHD, compared with WikiBio (Manakul et al., 2023) and PHD (Yang et al., 2023), incorporates long dependencies and hyper-relational structures. HRKG-HD extends beyond Semantic Entropy (SE) (Farquhar et al., 2024) and GCA (Fang et al., 2025) by constructing hyper-relational knowledge graphs and performing relatiin-aware message passing, enabling fine-grained, global, and dependency-sensitive detection.

is amplified in long-form generation, where multi-paragraph outputs introduce distant factual dependencies (Huang et al., 2025; Yang et al., 2025). Given the impact of hallucinations on trustworthiness in high-stakes applications that require rigorous factual accuracy (Chakraborty et al., 2025; Zhang et al., 2026; Lyu et al., 2026a; Kim et al., 2025), hallucination detection in long-form generation is necessary for safe and reliable deployment of LLMs in the real world. Recent advances in hallucination detection span both benchmark construction and methodological development. **In terms of benchmarks,** Su et al. (2024) introduce datasets for evaluating hallucinations in LLM-generated text. These datasets emphasize short factual claims, often within question answering and summarization settings. Other benchmarks extend coverage to longer and semantically diverse textual inputs

*Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 33477–33494

(Dong et al., 2024; Manakul et al., 2023). **In terms of methodology**, approaches such as (Min et al., 2023b; Chern et al., 2023; Bayat et al., 2023) detect hallucinations by aligning generated content with structured knowledge bases, but incomplete and stale coverage in these bases limits fine-grained verification. In contrast, zero-knowledge approaches offer greater adaptability by avoiding reliance on external retrieval. For example, Farquhar et al. (2024); Elaraby et al. (2023); Fang et al. (2025) assess internal consistency across multiple responses using entropy, entailment, or triple-level alignment. **However,** current research on long-form hallucination detection encounters two major limitations. **First,** existing long-form text benchmarks emphasize surface length rather than capturing long-range factual dependencies, leaving long-form text reasoning underexplored. **Second,** existing hallucination detection methods rely on locally scoped alignment rather than holistic document modeling, limiting the ability to model long-range dependencies and perform accurate factual reasoning. To the best of our knowledge, no existing benchmark or method explicitly addresses these challenges in long-form text generation by enabling fine-grained, global, and dependency-aware detection.

To address these challenges, we reformulate hallucination detection as structured reasoning over hyper-relational graphs rather than as isolated text spans. **First**, we introduce **L**ong-form **H**allucination **D**ataset **(LHD)**, a novel benchmark specifically designed for hallucination detection in long-form outputs. It provides fine-grained annotations and captures complex long-range factual dependencies across diverse scenarios. **Next**, motivated by the limited effectiveness of adapted baselines on LHD, we propose a framework named **H**yper-**R**elational **K**nowledge **G**raph-based **H**allucination **D**etection **(HRKG-HD)**. Our approach models both original and sampled responses as fact-centric hyper-relational knowledge graphs (HRKGs). Fact nodes act as semantic anchors, enabling a relation-aware graph attention network to propagate information across distant facts and identify subtle inconsistencies. HRKG-HD operates without external knowledge bases or access to model internals, making it suitable for black-box and resource-constrained settings. **In conclusion**, experiments demonstrate that LHD presents richer dependency structures, and that HRKG-HD consistently outperforms existing baselines, achieving state-of-the-art performance.

As shown in Fig. 1, our contributions are as follows:

- We construct LHD, a hallucination detection benchmark tailored for long-form outputs, incorporating long-range contextual dependencies.
- We propose HRKG-HD, a hallucination detection framework based on fact-centric HRKGs. It captures long-range factual dependencies and detects inconsistencies without external retrieval systems or model internal information.
- Experimental results show that HRKG-HD consistently outperforms all baselines across multiple benchmarks at both the passage-level and sentence-level.

2 Related Work

Hallucination Detection Benchmark. Existing benchmarks for hallucination detection primarily focus on sentence-level factual consistency, as exemplified by TruthfulQA (Lin et al., 2022a) and FactScore (Min et al., 2023b). As the field has progressed, a shift has emerged toward finer-grained, token-level evaluations, such as UHGEval (Liang et al., 2024) and HalOmi (Dale et al., 2023). However, they often struggle with broader aspects of text understanding. In particular, they fail to capture semantic coherence and logical consistency in complex, multi-layered narratives that are characteristic of long-form generation. To overcome these limitations, recent benchmarks like DiaHalu (Chen et al., 2024a) and HalluDial (Luo et al., 2024) adopt a more holistic perspective. They evaluate hallucinations at the dialogue level, accounting for multi-turn dynamics and dialogue-level coherence. Nonetheless, one critical dimension of long-form evaluation remains underexplored: the modeling of rich, long-range semantic dependencies. Without incorporating this dimension, current benchmarks struggle to reflect the real-world challenges of hallucination detection in extended texts (Chen et al., 2024b; Yang et al., 2025; Lyu et al., 2026b). *To bridge this gap, we introduce a new benchmark tailored for hallucinations in long-form text. It captures the essential characteristics of long-form , providing a more realistic foundation for evaluating detection methods.*

Hallucination Detection for Long-Form Text. Non-zero-knowledge approaches (Hu et al., 2024a; Li et al., 2024), which depend on external knowledge bases, are constrained by limited knowledge scope and coverage, resulting in lower accuracy in

long-form, fact-dense scenarios. In contrast, recent studies have begun to leverage the intrinsic reasoning abilities of LLMs. For example, Farquhar et al. (2024) measure semantic consistency by comparing entropy across multiple sampled outputs. HaloCheck employs sentence-level entailment to assess consistency between samples (Elaraby et al., 2023). Yang et al. (2023) reverse the generation process, reconstructing prompts from model-generated answers to evaluate alignment. Other works adopt similar generate-and-verify strategies, relying on natural language inference or answer overlap metrics (Honovich et al., 2021; Fabbri et al., 2022). Although effective for short-form generation, these techniques frequently falter in extended contexts. They struggle to connect facts spread across distant segments and frequently miss long-range factual dependencies. To address this, Fang et al. (2025) attempt to extract knowledge triples and structure them within graphs. While this improves contextual linkage, it oversimplifies complex relations into flat triples thereby sacrificing semantic precision and global information. *To overcome these limitations, we propose a novel approach based HRKG. This method is designed to better capture the rich and intricate factual structures in long-form text, enabling more robust and reliable hallucination detection across extended contexts.*

3 Long-Form Hallucination Detection Benchmark

We introduce the LHD, a large-scale benchmark specifically designed to characterize and detect hallucinations in entity-centric long-form generation. Unlike existing datasets that focus on short-span fact-checking, LHD centers on entity-based biographies and technical descriptions, where models must maintain factual consistency across extended contexts and complex relational dependencies. The overview is shown in Figure 2.

3.1 Long-Form Text Generation

Entity Selection. To enable analysis of diverse hallucination patterns across knowledge familiarity, we follow Yang et al. (2023) to select 300 entities spanning varying frequencies and domains. Specifically, entities are drawn from three frequency bands: low (<100K), medium (100K-1M), and high (>1M), based on their occurrence in the Wikipedia corpus.

Data Generation and Sampling Strategy. Each sample in LHD is a Wikipedia-style article generated by an LLM prompted with a target entity. We design a prompt constraint strategy that generates long-form text exhibiting the following characteristics: (1) sufficient length for long-form text ($\geq$ 512 tokens) (Tay et al., 2020; Lin et al., 2022b), (2) high factual density, and (3) long-range factual dependencies. The prompts are designed to elicit role-rich events, qualifier-rich relations, and narrative coherence. For each entity prompt, we first generate one response as the original response. We then produce 5 additional sampled responses using the same prompt with stochastic decoding (temperature = 1.0), in order to capture diverse realizations of the model's output distribution. These sampled responses serve as diverse reference outputs, enabling consistency-based comparison with the original response for subsequent hallucination detection. See Appendix A.1 for full prompt templates.

3.2 Human Annotation

Annotator Selection. To ensure high-quality annotations, we first distributed detailed task guidelines and administered a qualification test before the annotation phase. The submissions were manually evaluated, and 3 annotators were selected from an initial pool of 10 based on annotation accuracy.

Annotation Process. The sentence-level annotation was carried out in two stages. First, qualified annotators label hallucinations against preselected references and mark unverifiable sentences. Second, these sentences were reexamined using broader external sources retrieved from credible online platforms. Each sample was independently annotated by multiple annotators, with disagreements resolved via discussion, yielding fine-grained supervision for evaluation. At the passage-level, a sample is labeled as hallucinated if any of its constituent sentences is annotated as hallucinated; otherwise, it is considered factually consistent.

Annotation Consistency. To evaluate inter-annotator consistency, we calculate Fleiss's Kappa (Fleiss, 1971), a statistical measure of agreement among multiple raters. The resulting Kappa score of 0.75 indicates a substantial level of agreement among the annotators.

3.3 Quality Evaluation and Statistics

Quality Evaluation.

To ensure a fair comparison and disentangle the

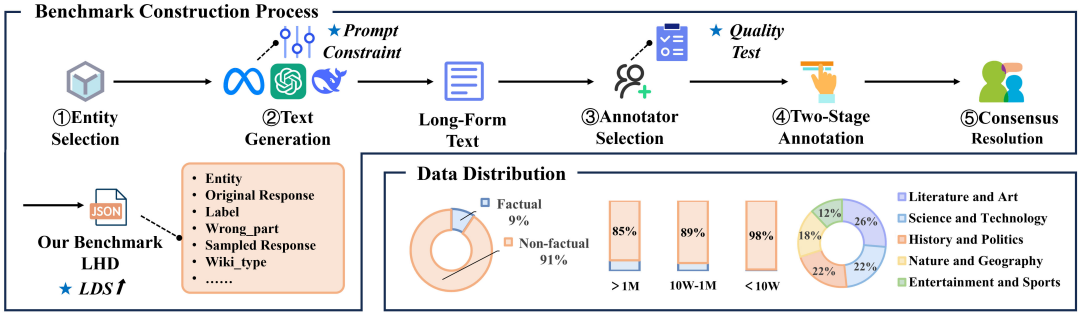

Figure 2: Benchmark construction process and data distribution.

impact of text length from structural complexity, we construct two controlled variants of PHD in Table 1: (1) Long-PHD, which only enforces a length constraint (≥ 512 tokens), and (2) Structure-PHD, which only imposes structural constraints to encourage factual dependencies.

We evaluate the quality of the generated texts with a particular focus on long-range dependency characteristics. We adopt the Long Dependency Score (LDS) (Chen et al., 2024b), which quantifies the richness and coherence of long-distance dependencies.

Table 1: Comparison of long-range dependency across different benchmarks. **LDS** (Long Dependency Score) measures the richness and coherence of factual linkages; higher values (⇑) indicate more complex dependencies. Bold denotes the best performance.

Benchmark	Avg. token count	LDS ⇑
PHD	111	0.96
WikiBio	229	2.89
Long-PHD	587	7.51
Structure-PHD	226	10.42
LHD	**644**	**29.39**

As shown in Table 1, LHD achieves a substantially higher density of long-range dependencies with LDS = 29.39, surpassing Structure-PHD and PHD by approximately 3× and 30×, respectively. Notably, increasing sequence length alone is insufficient: despite a comparable length (587 tokens), Long-PHD attains a much lower LDS (7.51), indicating that unconstrained long-form generation leads to sparse factual connectivity. While Structure-PHD improves LDS through structured prompting, it is limited by shorter context. These results highlight that our dual-constraint design promotes genuinely interconnected factual structures rather than redundant verbosity. A more granular analysis of these dependencies in the context of hallucinations is provided in Section 5.3.

Statistics. Figure 2 summarizes the dataset statistics and distributions of the LHD benchmark; further details are provided in Appendix A.4.

4 Method

We propose HRKG-HD, a framework for long-form hallucination detection based on consistency across sampled responses. Given a query, we obtain an original response R_o and n sampled responses $\{R_j\}_{j=1}^{n}$ under identical prompts. We extract hyper-relational facts from all responses, construct their corresponding HRKGs, and perform relation-aware reasoning to compute fact-level consistency. Hallucinations are identified based on aggregated consistency scores. An overview of the pipeline is shown in Figure 3.

4.1 Hyper-Relational Fact Extraction

Following Hu et al. (2024b), we use an instruction-tuned LLM to extract hyper-relational facts from both original and sampled responses via prompt-based parsing, leveraging its ability to model context-dependent structures in long-form text (Min et al., 2023a). Each fact $f \in \mathcal{F}^H$ is represented as $f = (s, r, o, (a_i, v_i)_{i=1}^{n})$, where (s, r, o) denotes the primary triple with $s, o \in \mathcal{E}$ and $r \in \mathcal{R}$, and (a_i, v_i) encodes qualifier pairs capturing contextual attributes.

To ensure quality, we apply a two-stage procedure consisting of extraction followed by LLM-based validation, which filters ill-formed or inconsistent tuples. We adopt a unified few-shot prompting scheme; full templates are provided in Appendix A.5.

4.2 Fact-Centric HRKG Construction

To explicitly model dependencies among extracted facts, we construct a fact-centric HRKG. Unlike

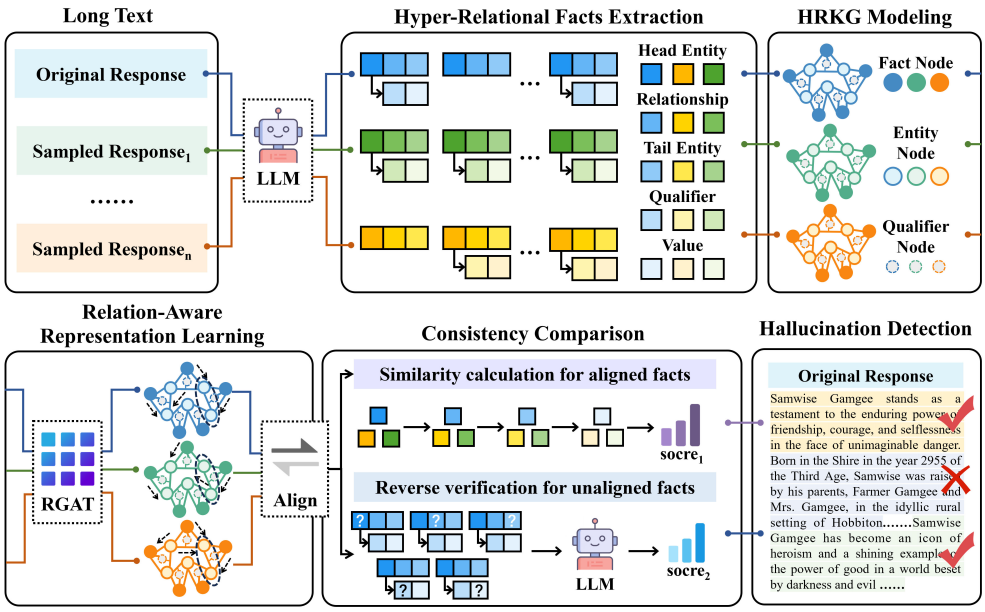

Figure 3: HRKG-HD framework. (1) Hyper-Relational Fact Extraction: Hyper-relational facts are extracted from both the original and sampled responses and transformed into structured representations. (2) Fact-centric Hyper-Relational Knowledge Graph Modeling: The extracted facts are organized into an HRKG that captures long-range contextual and conditional dependencies by representing each fact as a semantic anchor node and linking them via shared entities. (3) Relation-Aware Representation Learning on HRKG: A Relation-aware Graph Attention Network (RGAT) is applied to the HRKG to learn expressive node embeddings that encode contextual and semantic dependencies. (4) Consistency Comparison: Hallucinations are detected by measuring the semantic consistency of aligned facts across multiple HRKGs, based on their RGAT-derived embeddings.

conventional entity-centric KGs that treat relations as edges between entities, our design represents facts as dedicated nodes, enabling explicit modeling of inter-fact dependencies and preserving qualifier-level structure.

Each hyper-relational fact is instantiated as a fact node v_f. We connect v_f to its head entity s and tail entity o via typed edges ("has-head", "has-tail"). For each qualifier (a_i, v_i), we add an edge (v_f, v_i) with type a_i. This induces connectivity between fact nodes through shared entities and qualifiers, forming a heterogeneous graph that supports multi-hop dependency propagation.

Formally, the graph is defined as $\tilde{\mathcal{G}}^H = (\mathcal{V}, \mathcal{T})$, where $\mathcal{V}$ contains both entity and fact nodes, and $\mathcal{T}$ denotes typed edges derived from $\mathcal{F}^H$. For each query, we construct one graph $\tilde{\mathcal{G}}o^H$ for the original response R_o and a set of graphs $\{\tilde{\mathcal{G}}_j^H\}_{j=1}^n$ for the sampled responses $\{R_j\}_{j=1}^n$.

4.3 Relation-Aware Reasoning on HRKG

We apply a relation-aware graph attention network (RGAT) over the HRKG to learn context-aware node representations. The model performs multi-hop, relation-aware message passing over semantically related nodes, capturing long-range dependencies. The overall procedure consists of three stages:

Node and Relation Initialization. We initialize node and relation embeddings to capture their respective semantic taxonomies. For each entity node $v \in \mathcal{V}_e$, the initial embedding $\mathbf{h}_v^{(0)}$ is derived from the Sentence-BERT (Wang et al., 2020) encoding of its surface name. For a fact node $v_f \in \mathcal{V}_f$, we construct a holistic representation by encoding the concatenated textual sequence of its core triplet (s, r, o) and auxiliary qualifiers $\{(a_i, v_i)\}_{i=1}^k$. This ensures that $\mathbf{h}_{v_f}^{(0)}$ encapsulates the complete atomic semantics of the assertion. Relations $r \in \mathcal{R}$ are projected into a learnable latent space $\mathbb{R}^d$.

Relation-Aware Attention. We employ relation-aware attention to model heterogeneous edge types. For each node v, attention over its neighbors $u \in \mathcal{N}(v)$ is computed conditioned on node features and relation type r_{vu}. The attention coefficient is defined as:

$$\alpha_{vu} = \frac{\exp\left(\text{LeakyReLU}\left(\mathbf{a}_{r_{vu}}^\top \left[\mathbf{W}_{\phi(v)}\mathbf{h}_v \| \mathbf{W}_{\phi(u)}\mathbf{h}_u\right]\right)\right)}{\sum_{k \in \mathcal{N}(v)} \exp\left(\text{LeakyReLU}\left(\mathbf{a}_{r_{vk}}^\top \left[\mathbf{W}_{\phi(v)}\mathbf{h}_v \| \mathbf{W}_{\phi(k)}\mathbf{h}_k\right]\right)\right)}. \quad (1)$$

Here, $\mathbf{W}_{\phi(\cdot)}$ denotes a type-specific projection

that maps different node types into a shared space, and $\mathbf{a}_r$ is a relation-specific attention vector.

Node Representation Update. Given attention weights α_{vu}, node representations are updated in a layer-wise manner:

$$\mathbf{h}_v^{(l+1)} = \sigma\left(\sum_{u \in \mathcal{N}(v)} \alpha_{vu} \mathbf{W}_{r_{vu}} \mathbf{h}_u^{(l)}\right), \quad (2)$$

where σ is a non-linear activation function. After L layers, node representations capture multi-hop semantic and relational dependencies for downstream hallucination detection.

4.4 Graph-based Consistency Comparison

We detect hallucinations by measuring consistency between hyper-relational facts in the original and sampled HRKGs.

Fact Alignment. A fact node v_i^o is aligned with v_{ij}^s if: (1) they share the same head entity, $s_i = s_{ij}$; (2) their relation embeddings $\mathbf{r}_i$ and $\mathbf{r}_{ij}$ exhibit high semantic similarity, i.e., $S(\mathbf{r}_i, \mathbf{r}_{ij}) > \theta_r$, where $S(\cdot, \cdot)$ denotes cosine similarity; and (3) they share at least one common qualifier attribute, i.e., $A_i \cap A_{ij} \neq \emptyset$, where $A_i = \{a_k \mid (a_k, v_k) \in q_i\}$ and $A_{ij} = \{a_k \mid (a_k, v_k) \in q_{ij}\}$.

Consistency Scoring. For aligned facts, we compute a consistency score based on tail entities and shared qualifiers:

$$C_i = \alpha \cdot S(\mathbf{o}_i, \mathbf{o}_{ij}) + \beta \cdot \frac{1}{|\mathcal{V}_s|} \sum_{(v_k, v_k') \in \mathcal{V}_s} S(\mathbf{v}_k, \mathbf{v}_k'), \quad (3)$$

where $\mathcal{V}_s$ denotes shared qualifier values. We set $\alpha = \beta = 1$.

For unaligned facts, we apply LLM-based reverse verification (Fang et al., 2025) to obtain a reliability score M_i.

Final Decision. Each fact is assigned a final score:

$$C_i^f = \begin{cases} C_i, & \text{if aligned} \\ M_i, & \text{otherwise} \end{cases}. \quad (4)$$

A sentence is classified as hallucinatory if its fact-level scores fall below a threshold θ_{resp}. A passage is flagged if any sentence is classified as hallucinatory.

5 Experiments

5.1 Experimental Setup

Datasets. We evaluate HRKG-HD on two benchmark groups: (1) Wikipedia-style generation benchmarks, including PHD, WikiBio, and Structure-PHD; (2) extended-response benchmarks, including HaluQuestQA(Sachdeva et al., 2025), Long-PHD, and LHD. These datasets cover diverse factuality settings with varying response lengths, dependency structures, and domains. More Details are provided in Appendix A.6.

Baselines. To validate the effectiveness of our approach, we include baselines from two categories: (1) Long-form specific methods: GCA, Error-Informed Refinement (EIR) (Sachdeva et al., 2024), and the Decomposition & Aggregation Model (D&A Model) (Liu et al., 2025). All of which are explicitly designed for hallucination detection in long-form texts. (2) General-purpose methods: Semantic Entropy (SE) , Kernel Language Entropy (KLE) (Nikitin et al., 2024), SelfCheckGPT-BS (SelfCk-BS) (Manakul et al., 2023), and RVQG (Yang et al., 2023), which are widely applicable to hallucination detection across tasks. The threshold selection procedure for HRKG-HD is detailed in Appendix A.7.

Implementation Details. We use DeepSeek-V3 for fact extraction, question reconstruction, and all baseline methods, with temperature set to 0.0 for reproducibility. RGAT is parameterized but untrained, and that the entire module operates solely in forward inference mode.

Evaluation Metrics. All evaluations are conducted at both the sentence and passage levels. A sentence or passage is labeled as hallucinated if its consistency score falls below the threshold. Performance is measured using F1 and Accuracy (ACC) at both levels.

5.2 Main Results

First, Table 2 presents the performance of all methods across benchmarks, evaluated at both passage-level and sentence-level. The results show that HRKG-HD consistently outperforms all baselines across datasets. We further analyzed the density of long-dependency hallucinations in LHD and conducted experiments in Sec. 5.3. These improvements demonstrate the effectiveness of fact-centric hyper-relational graph modeling with relation-aware reasoning. Second, A notable exception is observed on HaluQuestQA. The EIR marginally exceeds HRKG-HD due to its dataset-specific optimization. This advantage, however, is task-dependent and limited to QA-style settings. On other benchmarks, EIR shows consistently lower performance. In contrast, HRKG-HD

Table 2: Comparison of hallucination detection performance (HRKG-HD vs. baselines).

	Method	PHD		WikiBio		Structure-PHD		LongPHD		HaluQuestQA		LHD	
		F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC
Passage-level	GCA	62.4	64.7	73.1	75.3	41.5	43.7	62.1	64.3	88.9	90.3	60.8	63.0
	KLE	61.1	63.0	69.5	71.0	58.2	60.0	59.8	61.7	87.2	88.7	54.4	56.3
	SE	57.2	59.3	53.4	55.7	51.9	54.0	55.1	57.0	78.4	80.0	49.2	51.3
	RVQG	58.5	60.7	68.4	70.3	43.2	45.3	51.2	53.3	94.8	96.0	34.2	36.7
	SelfCK-BS	51.4	53.0	49.2	51.0	52.8	54.7	51.8	53.7	91.5	93.3	53.5	55.3
	EIR	58.1	60.3	64.9	66.7	53.6	55.7	56.2	58.0	**96.2**	**97.7**	51.1	53.0
	D&A Model	52.3	54.3	50.4	52.0	52.6	54.7	52.9	54.7	80.9	82.3	53.8	55.7
	HRKG-HD	**69.5**	**71.3**	**78.9**	**80.7**	**67.4**	**69.3**	**68.2**	**70.0**	95.8	97.3	**66.1**	**68.0**
Sentence-level	GCA	68.6	70.7	74.4	76.3	68.1	70.0	67.5	69.7	93.1	94.7	66.2	68.3
	KLE	66.8	68.7	74.8	76.7	63.8	65.7	64.2	66.3	90.4	92.0	59.4	61.0
	SE	57.0	59.0	60.1	61.7	57.2	59.0	60.5	62.3	83.7	85.3	56.1	58.0
	RVQG	64.2	66.3	73.2	75.0	48.1	50.0	56.2	58.0	96.5	97.7	50.5	52.3
	SelfCK-BS	56.5	58.3	55.2	57.0	57.4	59.3	56.4	58.0	94.4	96.0	58.2	60.0
	EIR	63.2	65.0	69.3	71.0	58.2	60.0	61.1	62.7	**96.9**	**98.0**	55.4	57.3
	D&A Model	57.8	59.7	54.8	56.7	56.1	58.0	57.0	58.7	84.9	86.7	57.3	59.0
	HRKG-HD	**75.1**	**76.7**	**84.3**	**86.0**	**73.2**	**75.0**	**73.9**	**75.7**	**96.9**	**98.0**	**72.4**	**74.3**

Table 3: Dependency-length distribution of hallucinations across benchmarks.

Benchmark	Long-dependency density	Short-dependency density	Zero-dependency density
PHD	0.17	0.64	0.19
WikiBio	0.19	0.55	0.26
Long-PHD	0.24	0.57	0.19
Structure-PHD	0.32	0.32	0.36
LHD	**0.60**	**0.28**	**0.12**

achieves stable and superior results across both QA and encyclopedic domains, confirming its stronger generalization capability.

5.3 Effectiveness on Long-Dependency Hallucinations

Dependency Distribution Across Benchmarks. To quantify the extent to which hallucinations rely on contextual information, we measure the dependency length of each hallucinated fact as the average shortest fact-path distance (FD) between the hallucinated fact and its relevant supporting facts in the HRKG graph.

Based on FD, hallucinations are categorized into three buckets: (1) *independent* (FD=0), (2) *short-dependency* (FD=1–2), and (3) *long-dependency* (FD$\geq$3), which requires multi-hop semantic reasoning.

Bucket-wise Performance Comparison. As shown in Table 3, hallucinations in LHD exhibit a substantially higher concentration of long-range dependency. Specifically, 60% of hallucinations in LHD fall into the long-dependency bucket, significantly exceeding all other benchmarks. In contrast, prior datasets such as PHD and WikiBio are dominated by short-dependency cases, indicating a limited requirement for long-range factual reasoning. These statistics demonstrate that LHD poses a more challenging evaluation setting that emphasizes long-context factual consistency. To further examine whether HRKG-HD is particularly effective for long-dependency hallucinations, we evaluated HRKG-HD and all baselines within each dependency bucket in Table 4.

Results show that while all methods perform comparably in the independent and short-dependency buckets, HRKG-HD achieves a clear and consistent advantage in the long-dependency bucket. This confirms that HRKG-HD is especially well-suited for detecting hallucinations that require cross-sentence and multi-hop factual reasoning, rather than isolated fact verification.

Representation and Structural Analysis. To evaluate the HRKG-HD capacity to capture long-range factual dependencies, we compare two input strategies: (1) using the entire long-form text, and (2) dividing the passage into multiple shorter segments. The embeddings of fact nodes extracted

Table 4: Performance comparison under different dependency-length buckets.

	GCA		LKE		SE		RVQG		SelfCK-BS		EIR		D&A Model		HRKG-HD	
Bucket	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC
Long-dependency	67.2	69.4	55.8	58.3	53.1	55.0	28.5	32.2	54.9	57.2	51.2	53.3	53.4	55.5	77.5	79.4
Short-dependency	65.8	67.9	65.8	67.9	59.8	61.9	54.1	56.0	62.4	64.3	61.2	63.1	62.4	64.3	67.1	69.0
Zero-dependency	62.0	63.9	62.0	63.9	59.2	61.1	59.2	61.1	62.0	63.9	59.2	61.1	62.0	63.9	62.0	63.9

from original response and five sampled responses are visualized by t-SNE. Detailed procedures for the t-SNE analysis are provided in Appendix A.8. As shown in Figure 4 (a) and (b), the fact nodes from the entire passage exhibit a more cohesive distribution compared to segmented inputs. The results confirm the ability of HRKG-HD to capture long-range dependencies. We analyze how

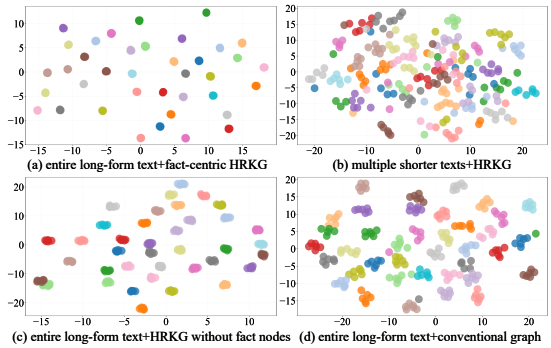

Figure 4: Visualization of long-range dependency modeling by HRKG-HD. Each color represents the embedding of an entity.

different graph construction strategies affect node representations by comparing three variants: (1) fact-centric HRKG, (2) HRKG without fact nodes, and (3) conventional graph. The node distribution is shown in Figure 4 (a), (c), and (d). A similar pattern can be observed in a randomly selected subset of 20 responses in the Appendix A.8. These observations suggest that fact-centric HRKG effectively integrates the features of neighboring nodes into each node's representation, blending information for every node.

5.4 Ablation Study

To evaluate the contribution of each core module in HRKG-HD, we conduct ablation studies. As shown in Table 5, **-Fact node** removes virtual fact nodes, degrading the graph to entity-relation form. The consistent performance drop confirms their importance for long-range dependency modeling. **-HRKG** discards qualifiers, reducing hyper-relational facts to flat triples. This results in a notable decline, highlighting the semantic value of qualifiers. **-RGAT** replaces relation-aware attention with vanilla GAT, indicating that relation-specific reasoning is critical. **-RV** disables the reverse verification module. This shows that RV compensates for the limitations of alignment-based matching.

5.5 Model-Agnostic Robustness Analysis

To evaluate the generalizability of HRKG-HD across different model architectures, we further generate samples using three distinct LLMs. Specifically, we utilize DeepSeek-V3, LLaMA2-Chat-7B, and Qwen3-32B, with identical prompt templates and entity selection criteria to ensure experimental consistency. As shown in Figure 5, HRKG-HD achieves stable results. This demonstrates its robustness across different model architectures. In contrast, methods such as RVQG demonstrate high sensitivity to LLMs. This instability likely arises from their dependence on model-specific generation heuristics.

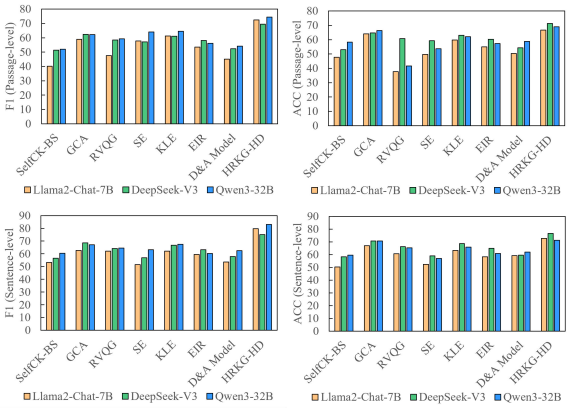

Figure 5: Performance of HRKG-HD on various LLMs.

5.6 Extraction-Noise Robustness Analysis

We evaluate the robustness of HRKG-HD against fact extraction errors by introducing controlled noise into the extracted graphs. The results of Appendix A.9 show that randomly dropping facts or corrupting qualifiers leads to a gradual decline in

Table 5: Ablation study of HRKG-HD components on hallucination detection performance.

	Variant	PHD		WikiBio		Structure-PHD		LongPHD		HaluQuestQA		LHD	
		F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC
Passage-level	w/o Fact node	60.5	62.7	62.4	64.7	60.8	62.7	61.2	63.0	76.8	78.5	59.5	61.3
	w/o HRKG	59.2	61.3	68.3	70.2	58.9	60.7	59.2	61.0	73.0	74.6	56.8	58.7
	w/o RGAT	51.4	53.3	70.5	72.7	50.1	52.0	50.8	52.7	64.2	65.8	48.2	50.0
	w/o RV	67.4	69.3	74.2	76.1	64.1	66.0	64.8	66.7	81.5	83.2	60.9	62.7
	HRKG-HD	**69.5**	**71.3**	**78.9**	**80.7**	**67.4**	**69.3**	**68.2**	**70.0**	**95.8**	**97.3**	**66.1**	**68.0**
Sentence-level	w/o Fact node	63.8	65.7	66.4	68.5	63.1	65.0	63.4	65.3	79.2	80.8	62.9	64.7
	w/o HRKG	61.9	63.7	70.2	72.3	60.1	62.0	61.2	63.0	72.5	74.1	58.2	60.0
	w/o RGAT	53.8	55.7	69.6	71.8	53.4	55.3	54.1	56.0	68.9	70.6	51.5	53.3
	w/o RV	69.8	71.7	74.6	76.5	66.4	68.3	67.4	69.3	80.1	81.9	63.8	65.7
	HRKG-HD	**75.1**	**76.7**	**84.3**	**86.0**	**73.2**	**75.0**	**73.9**	**75.7**	**96.9**	**98.0**	**72.4**	**74.3**

model performance. Nevertheless, HRKG-HD consistently outperforms all baselines, a result that can be attributed to its RV mechanism that compensates for fact extraction errors. We also compare different LLMs as extractors and find that the more powerful extractors produced higher ACC.

6 Conclusion

In this work, we detected hallucination in long-form text generation by introducing LHD, a benchmark capturing long-range factual dependencies, and HRKG-HD, a zero-knowledge, black-box framework that models outputs as fact-centric HRKGs. Through multi-hop reasoning, HRKG-HD enables global, dependency-aware consistency verification. Experiments show that it outperforms state-of-the-art baselines and generalizes well across different LLMs.

7 Limitations

Our framework has two main limitations. First, as input length and fact count grow, hyper-relational graph construction and multi-hop reasoning incur substantial memory and computational overhead. Second, the multi-stage workflow, though model-agnostic, increases complexity and computational cost, potentially limiting applicability in large-scale or latency-sensitive production settings. Future work may explore lightweight graph representations, efficient verification strategies, and more scalable retrieval mechanisms.

8 Acknowledgments

This work is sponsored in part by the National Natural Science Foundation of China under Grant No. 72231011, 72401287 and 72471238.

References

Farima Fatahi Bayat, Kun Qian, Benjamin Han, Yisi Sang, Anton Belyy, Samira Khorshidi, Fei Wu, Ihab Ilyas, and Yunyao Li. 2023. Fleek: Factual error detection and correction with evidence retrieved from external knowledge. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 124–130.

Neeloy Chakraborty, Melkior Ornik, and Katherine Driggs-Campbell. 2025. Hallucination detection in foundation models for decision-making: A flexible definition and review of the state of the art. *ACM Computing Surveys*.

Kedi Chen, Qin Chen, Jie Zhou, He Yishen, and Liang He. 2024a. Diahalu: A dialogue-level hallucination evaluation benchmark for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9057–9079.

Longze Chen, Ziqiang Liu, Wanwei He, Yinhe Zheng, Hao Sun, Yunshui Li, Run Luo, and Min Yang. 2024b. Long context is not long at all: A prospector of long-dependency data for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8222–8234.

I Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, Pengfei Liu, and 1 others. 2023. Factool: Factuality detection in generative ai–a tool augmented framework for multi-task and multi-domain scenarios. *arXiv preprint arXiv:2307.13528*.

David Dale, Elena Voita, Janice Lam, Prangthip Hansanti, Christophe Ropers, Elahe Kalbassi, Cynthia Gao, Loïc Barrault, and Marta Costa-jussà. 2023. Halomi: A manually annotated benchmark for multilingual hallucination and omission detection in machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 638–653.

Zican Dong, Tianyi Tang, Junyi Li, Wayne Xin Zhao, and Ji-Rong Wen. 2024. Bamboo: A comprehensive benchmark for evaluating long text modeling capacities of large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2086–2099.

Mohamed Elaraby, Mengyin Lu, Jacob Dunn, Xueying Zhang, Yu Wang, Shizhu Liu, Pingchuan Tian, Yuping Wang, and Yuxuan Wang. 2023. Halo: Estimation and reduction of hallucinations in open-source weak large language models. *arXiv preprint arXiv:2308.11764*.

Alexander Richard Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. Qafacteval: Improved qa-based factual consistency evaluation for summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601.

Xinyue Fang, Zhen Huang, Zhiliang Tian, Minghui Fang, Ziyi Pan, Quntian Fang, Zhihua Wen, Hengyue Pan, and Dongsheng Li. 2025. Zero-resource hallucination detection for text generation via graph-based contextual knowledge triples modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23868–23877.

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

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Mingqi Gao, Xinyu Hu, Xunjian Yin, Jie Ruan, Xiao Pu, and Xiaojun Wan. 2025. Llm-based nlg evaluation: Current status and challenges. *Computational Linguistics*, pages 1–27.

Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. 2021. Q2: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7856–7870.

Xiangkun Hu, Dongyu Ru, Lin Qiu, Qipeng Guo, Tianhang Zhang, Yang Xu, Yun Luo, Pengfei Liu, Yue Zhang, and Zheng Zhang. 2024a. Knowledge-centric hallucination detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6953–6975.

Xiangkun Hu, Dongyu Ru, Lin Qiu, Qipeng Guo, Tianhang Zhang, Yang Xu, Yun Luo, Pengfei Liu, Yue Zhang, and Zheng Zhang. 2024b. Refchecker: Reference-based fine-grained hallucination checker and benchmark for large language models. *arXiv preprint arXiv:2405.14486*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Yubin Kim, Hyewon Jeong, Shan Chen, Shuyue Stella Li, Mingyu Lu, Kumail Alhamoud, Jimin Mun, Cristina Grau, Minseok Jung, Rodrigo Gameiro, and 1 others. 2025. Medical hallucinations in foundation models and their impact on healthcare. *arXiv preprint arXiv:2503.05777*.

Ningke Li, Yuekang Li, Yi Liu, Ling Shi, Kailong Wang, and Haoyu Wang. 2024. Drowzee: Metamorphic testing for fact-conflicting hallucination detection in large language models. *Proceedings of the ACM on Programming Languages*, 8(OOPSLA2):1843–1872.

Xun Liang, Shichao Song, Simin Niu, Zhiyu Li, Feiyu Xiong, Bo Tang, Yezhaohui Wang, Dawei He, Cheng Peng, Zhonghao Wang, and 1 others. 2024. Uhgeval: Benchmarking the hallucination of chinese large language models via unconstrained generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5266–5293.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022a. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252.

Tianyang Lin, Yuxin Wang, Xiangyang Liu, and Xipeng Qiu. 2022b. A survey of transformers. *AI open*, 3:111–132.

Siyi Liu, Kishaloy Halder, Zheng Qi, Wei Xiao, Nikolaos Pappas, Phu Mon Htut, Neha Anna John, Yassine Benajiba, and Dan Roth. 2025. Towards long context hallucination detection. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7827–7835.

Wen Luo, Tianshu Shen, Wei Li, Guangyue Peng, Richeng Xuan, Houfeng Wang, and Xi Yang. 2024. Halludial: A large-scale benchmark for automatic dialogue-level hallucination evaluation. *arXiv preprint arXiv:2406.07070*.

Guangtao Lyu, Xinyi Cheng, Qi Liu, Chenghao Xu, Jiexi Yan, Muli Yang, Fen Fang, and Cheng Deng. 2026a. Towards interpretable hallucination analysis and mitigation in lvlms via contrastive neuron steering. *arXiv preprint arXiv:2602.00621*.

Guangtao Lyu, Xinyi Cheng, Chenghao Xu, Qi Liu, Muli Yang, Fen Fang, Huilin Chen, Jiexi Yan, Xu Yang, and Cheng Deng. 2025. Revealing perception and generation dynamics in lvlms: Mitigating hallucinations via validated dominance correction. *arXiv preprint arXiv:2512.18813*.

Guangtao Lyu, Qi Liu, Chenghao Xu, Jiexi Yan, Muli Yang, Xueting Li, Fen Fang, and Cheng Deng. 2026b. Revealing and enhancing core visual regions: Harnessing internal attention dynamics for hallucination mitigation in lvlms. *arXiv preprint arXiv:2602.15556*.

Potsawee Manakul, Adian Liusie, and Mark 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.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2023a. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2):1–40.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023b. 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.

Alexander Nikitin, Jannik Kossen, Yarin Gal, and Pekka Marttinen. 2024. Kernel language entropy: Fine-grained uncertainty quantification for llms from semantic similarities. *Advances in Neural Information Processing Systems*, 37:8901–8929.

Rachneet Sachdeva, Yixiao Song, Mohit Iyyer, and Iryna Gurevych. 2024. Fine-grained hallucination detection and mitigation in long-form question answering. *arXiv e-prints*, pages arXiv–2407.

Rachneet Singh Sachdeva, Yixiao Song, Mohit Iyyer, and Iryna Gurevych. 2025. Localizing and mitigating errors in long-form question answering. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20437–20469.

Konrad Staniszewski, Szymon Tworkowski, Sebastian Jaszczur, Yu Zhao, Henryk Michalewski, Łukasz Kuciński, and Piotr Miłoś. 2025. Structured packing in llm training improves long context utilization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25201–25209.

Weihang Su, Changyue Wang, Qingyao Ai, Yiran Hu, Zhijing Wu, Yujia Zhou, and Yiqun Liu. 2024. Unsupervised real-time hallucination detection based on the internal states of large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 14379–14391.

Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler. 2020. Long range arena: A benchmark for efficient transformers. *arXiv preprint arXiv:2011.04006*.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in neural information processing systems*, 33:5776–5788.

Joonho Yang, Seunghyun Yoon, Hwan Chang, Byeongjeong Kim, and Hwanhee Lee. 2025. Hallucinate at the last in long response generation: A case study on long document summarization. *arXiv preprint arXiv:2505.15291*.

Shiping Yang, Renliang Sun, and Xiaojun Wan. 2023. A new benchmark and reverse validation method for passage-level hallucination detection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3898–3908.

Lixin Zhan, Yukun Du, Jie Jiang, Yingmei Wei, Tianjian Zhou, and Ziyuan Yang. 2025. Bgc-net: Bilateral graph convolutional network for weakly-supervised semantic segmentation of large-scale point clouds. *IEEE Transactions on Circuits and Systems for Video Technology*.

Lixin Zhan, Jiang Jie, Tianjian Zhou, Yukun Du, Yan Zheng, and Xuehu Duan. 2026. P-slcr: Unsupervised point cloud semantic segmentation via prototypes structure learning and consistent reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 12349–12357.

Caiqi Zhang, Chang Shu, Ehsan Shareghi, and Nigel Collier. 2025. All roads lead to rome: Graph-based confidence estimation for large language model reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 31802–31812.

Caiqi Zhang, Ruihan Yang, Xiaochen Zhu, Chengzu Li, Tiancheng Hu, Yijiang River Dong, Deqing Yang, and Nigel Collier. 2026. Confidence estimation for llms in multi-turn interactions. *arXiv preprint arXiv:2601.02179*.

Table 6: Details of prompt templates for long text generation.

Prompt Type	Prompt Template
Text generation	You are a factual knowledge writer tasked with composing a long, coherent, and information dense paragraph about a given entity. The paragraph must integrate a high volume of real-world facts into a logically structured and semantically rich narrative. Your writing should simulate a high quality encyclopedia entry or scholarly exposition, optimized for use in large language model training and evaluation. **Requirements**: 1. The paragraph must exceed 512 tokens. 2. The content must be factually accurate, verifiable, and reflect contemporary domain knowledge. Ground all statements in known or widely accepted truths. 3. Incorporate a large number of factual knowledge triples—(head entity, relation, tail entity)—but do not present them explicitly. Instead, weave them naturally into the prose using fluent academic English. 4. Go beyond simple relations: include hyper-relational knowledge, such as events involving multiple roles (agents, objects, locations, time), attributes on relations (e.g., causality, temporality, negation), and multi-hop dependencies. These must be deeply embedded in the narrative structure. And these contains multiple triples or hyper-relational reasoning chains, and the reasoning path spans multiple paragraphs; 5. The overall paragraph must be highly coherent: ensure local sentence transitions are smooth, and global discourse structure builds toward a cumulative understanding of the entity. 6. Do not use bullet points, lists, or section headers. All content should be integrated into a continuous, scholarly-style paragraph. 7. All sentences must be complete sentences, and the last sentence must end with a period. Here are two few-shot examples: **<Entity>**: Apollo Program **<Response>**: The Apollo Program, ...seismometers, and core sample drills, with Apollo 17 hosting the only geologist astronaut, Harrison Schmitt. **<Entity>**: Louvre Museum **<Response>**: The Louvre Museum,.....accessibility, and transhistorical juxtaposition were foregrounded over nationalist teleologies. **<Entity>**: target_entity **<Response>**:

A Appendix

A.1 Long-Form Text Generation

Table 6 presents the prompt template used for generating long-form factual texts based on a given entity. To ensure that the outputs exhibit high factual density, structural coherence, and complex inter entity reasoning, the prompt adopts an instruction-based design augmented with explicit writing constraints and quality requirements. Following the data generation and sampling strategy described in Section 3.1, we use DeepSeek-V3.2 to generate responses, treating the recorded response for each prompt as the original response and generating multiple sampled responses via stochastic decoding (temperature = 1.0).

A.2 Annotator Selection and Background

To ensure high quality annotations for hallucination detection, we implemented a rigorous qualification phase before the formal annotation process. All candidates were required to read a comprehensive annotation guideline, which included labeled examples and descriptions of common error types, as shown in Table 7. They were then asked to complete a qualification test (QT) consisting of 10 long-form texts generated by large language models (LLMs), each accompanied by relevant reference sources (e.g., Wikipedia articles). For each long text content, candidates were instructed to label it as factual, non-factual, or unverifiable, and to highlight specific hallucinated content where applicable. The qualification test was distributed via the "Wenjuanxing" platform, which provides similar functionality to Amazon Mechanical Turk. The

Table 7: Annotation guideline.

Components	Detailed instruction
Label Definition	**Factual**: A long text content in which all sentences are factually accurate and fully supported by reliable sources (e.g., Wikipedia, academic literature, official websites). There are no unverifiable or false statements. The long text content represents entirely trustworthy information without exception. **Non-factual (Hallucination)**: A long text content that contains at least one sentence that is verifiably false or fabricated. Any confirmed factual error—regardless of scope or importance—qualifies the entire long text content as non-factual. Such hallucinations may involve incorrect claims, distorted relationships, invented entities, or misrepresented facts. **Unverifiable**: A long text content that contains one or more sentences whose factuality cannot be confirmed or refuted using accessible reliable sources, while all other sentences are factually accurate. There must be no verifiably false statements in the long text content. The unverifiable content typically involves obscure, ambiguous, or niche information beyond verifiable coverage.
Annotation Labels	Each long text content should be labeled with one of the following: Factual/ Non-factual/ Unverifiable. In addition, for Non-factual and Unverifiable long text contents, annotators should highlight the specific span(s) that are incorrect or unverifiable.
Annotation Procedure	**Step 1: Reference Checking** For each sentence, consult preselected reference sources, such as Wikipedia and relevant domain databases. If evidence is not found, proceed to Step 2. **Step 2: Internet Verification** Use web search (e.g., Google) to look for reliable sources only, including: Official government/organizational websites; Peer-reviewed academic publications; Reputable media outlets (e.g., BBC, New York Times). Only content from the first page of search results is considered valid to ensure consistency and credibility.
Special Instructions	•Always consider the sentence within its paragraph context, especially when evaluating entity mentions or temporal references. •Be cautious with partial truths: a sentence may contain both factual and hallucinated components; label based on the most severe factual error. •Unverifiable $\neq$ hallucinated: if no clear evidence exists either way, mark it as unverifiable, not non-factual.
Examples	**Example 1:** **Entity**: Apollo Program **Text**: The Apollo Program was initiated in 1961 by NASA as the United States' third human spaceflight initiative. Its most famous mission, Apollo 11, successfully landed astronauts Neil Armstrong and Buzz Aldrin on the Moon in July 1969, with Michael Collins remaining in lunar orbit. The mission used the Saturn V launch vehicle and relied on guidance systems developed by MIT's Instrumentation Laboratory. .....The program concluded with Apollo 17 in 1972. **Label**: Factual **Reason**: All sentences are verifiable via NASA archives and widely cited historical records; no unverifiable or fabricated claims. **Example 2:** **Entity**: Louvre Museum **Text**: The Louvre Museum was founded in 1750 as a royal art collection for the exclusive use of the French monarchy. It was Napoleon Bonaparte who opened it to the public during the height of his reign in 1804. ..... The famous Mona Lisa painting was acquired through a donation from an Italian noble family in the early 19th century. **Label**: Non-factual **Reason**: The Louvre was not founded in 1750 but opened as a public museum in 1793. The Mona Lisa was brought by Leonardo himself, not donated by a noble family. Passage contains at least two verifiably false statements.
Time Management and Rotation	Annotators are encouraged to rotate between entities of different frequency levels (low/ mid / high) to minimize cognitive fatigue and maintain labeling quality.

labeling task in the QT matched the task performed during the formal annotation stage, as shown in Table 8.

Submissions were manually reviewed and compared against expert annotations. Only annotators who achieved over 80% agreement with expert labels were selected. Out of 10 applicants, 3 were ultimately qualified.

The selected annotators each hold at least a master's degree and have academic backgrounds in

Table 8: Input and output format of the annotation task.

Category	Description
Input	1. Long-form text generated by LLM 2. Reference materials (Wikipedia articles) 3. Web browser access (used only in the second stage of annotation)
Output	1. Label (Factual / Non-factual / Unverifiable) 2. Highlighted evidence (text identified as incorrect or unverifiable)

linguistics, computer science, or related disciplines. They are experienced researchers in the fields of linguistics and natural language processing and are proficient in using online tools such as Google and Bing for fact verification. Their diverse yet professional expertise ensured both the quality and reliability of the annotations. Furthermore, both annotators demonstrated strong communication and collaboration skills, which contributed to annotation consistency across the process.

A.3 Annotation process

To ensure annotation quality and consistency, we adopted a two-stage annotation procedure. In the first stage, annotators labeled hallucinations by consulting preselected reference materials, including Wikipedia and domain-specific databases. To account for long-range dependencies, annotators were allowed to review adjacent paragraphs when evaluating sentence-level claims. If a claim could not be conclusively verified or refuted using the provided references, it was marked as unverifiable.

In the second stage, annotators reevaluated these unverifiable cases using broader information sources. Evidence was collected via Internet searches, limited to credible and authoritative outlets such as official websites, academic publications, and reputable news media. To ensure source reliability, only content appearing on the first page of Google search results was considered admissible. After review, each sentence was labeled as either factual or non-factual, and annotators were instructed to highlight any specific spans deemed inaccurate or unverifiable.

Each long-form text was annotated independently by multiple annotators. Disagreements were resolved through group discussion and cross review. This protocol enabled the collection of fine-grained, sentence-level annotations that support high resolution evaluation of hallucination detection methods.

A.4 Dataset Statistics

Hallucinations are prevalent in long-form text generated by LLMs. Among all annotated sentences, only 9% are factual, whereas 91% are non-factual. Furthermore, hallucination rates remain high across all entity frequencies: 85% (high), 89% (medium), and 98% (low). These results indicate that LLMs exhibit widespread hallucination when generating long-form content, especially in low-resource knowledge settings and rare entity scenarios. In addition, LHD covers a wide range of domains, including Literature and Art (26%), Science and Technology (22%), History and Politics (22%), Nature and Geography (18%), and Entertainment and Sports (12%). This domain diversity makes the benchmark more generalizable and suitable for evaluating hallucination detection in different contexts.

A.5 Prompt Templates for Fact Extraction and Verification

Table 9 shows the prompt templates used for hyper-relational fact extraction and verification. For extraction, the prompt adopts a detailed instructional format accompanied by an incontext example to illustrate the desired structure and semantic granularity of the output. Each fact is extracted as a base triple (head, relation, tail) along with one or more qualifiers that capture contextual attributes such as time, location, method, and role. This design encourages fine-grained and faithful representa-tion of complex events and multi entity interactions. In contrast, the prompt for fact verification assumes that a candidate hyper-relational knowledge graph has already been extracted from the original text. It focuses on detecting and correcting semantic inconsistencies, vague entity mentions (e.g., pronouns), or factual deviations from the source passage. Unlike the extraction prompt, the verification prompt does not include any few shot examples, and requires the model to directly output the corrected knowledge triples without additional commentary.

A.6 Dataset Details

We provide detailed descriptions of all datasets used in our experiments: (1) PHD. A dataset of 300 Wikipedia-style passages generated by GPT-3.5, each prompted with a target entity. Human annotators provide passage-level hallucination labels. (2) WikiBio. A dataset of 238 passages gen-

Table 9: Details of prompt templates for hyper-relational fact extraction and verification.

Prompt Type	Prompt Templates
Hyper-relational fact extraction	In a hyper-relational knowledge graph, each fact is not only represented by a ("head entity", "relation", "tail entity") triple, but also supplemented with one or more contextual attributes that describe time, location, participants, purposes, methods, etc. These attributes enrich the factual semantics and provide a more complete representation of real world events or relations. Given a piece of text, please extract all such hyper-relational facts. Requirements: 1. The triple and its attributes are as fine-grained and faithful to the original text as possible. 2. No pronouns or vague references should be used in the head or tail entities. When there is ambiguous reference in the text, the closest subject is preferred. 3. If multiple facts are mentioned, they must be extracted separately. 4. If the same entity appears in different contexts, it needs to be split into different facts. 5. Attributes must be directly inferred from the text (e.g., time, place, role, purpose, etc.). 6.The time format is unified into numbers (e.g. "1905"→"1905") 7. Relations should be verb-centric (e.g., "published in", "awarded for") Here is an in-context example: <Text>: John Russell Reynolds ....the poet John Keats (1848). <Response>: ["base_triple": ["Microsoft", "headquarteredin", "Redmond, Washington"], "qualifiers": ["key": "type", "value": "globaltechnologygiant"], "..."qualifiers": ["key": "method", "value": "acquisitionofGreatPlainsSoftware"]] <Text>init_text
Hyper-relational fact verification	The following is the hyper-relational knowledge graph you extracted from the text, but there are still some errors in it. For example, the semantics of the triples and their attributes are different from the corresponding parts in the original text, or there are pronouns in the triples. Please check and correct them. <Initial prompt>p <Triples>t\ Please output all corrected triples directly ,including changed and unmodified ones. Don't output any other words. Here is an in-context example: <Initial prompt>: ["base_triple": ["Microsoft", "headquarteredin", "Redmond, China"], "qualifiers": ["key": "type", "value": "globaltechnologygiant"], ... "base_triple": ["Windows", "revolutionized", "personalcomputing"], "qualifiers": ["key": "result", "value": "acquisitionofGreatPlainsSoftware"]]

erated by GPT-3. The original annotations are at the sentence level; following Yang et al. (2023), we aggregate them into passage-level labels for consistency with other datasets. (3) LHD. A dataset of 300 Wikipedia-style passages with human provided passage-level annotations. Compared to PHD and WikiBio, LHD emphasizes richer contextual dependencies to evaluate long-range hallucination detection. (4) Long-PHD. A length controlled variant of PHD. We modify the prompts to instruct GPT-3.5 to generate passages of at least 512 tokens, using the same entities as PHD but without structural constraints, resulting in longer, free-form passages. (5) Structure-PHD. A structure controlled variant of PHD. Prompts are modified to encourage explicit sectioning and hierarchical structure, inducing long-range factual dependencies and increasing dataset difficulty. (6) HaluQuestQA. A long-form question answering dataset with span-level hallucination annotations. It contains pairs of human-written and model-generated answers, enabling evaluation of hallucination detection beyond Wikipedia-style passages.

A.7 Baselines and Setting Thresholds

We compare our method against six representative baselines covering diverse paradigms of hallucination detection. (1) RVQG, a black-box

consistency-based method that reconstructs questions from text and verifies answers via LLM responses; (2) Semantic Entropy (SE), a gray-box uncertainty-based method that measures entropy over LLM-generated answers to assess factual validity; (3) Self-CheckGPT-BS (SelfCk-BS), a black-box consistency-based variant using BERTScore to evaluate consistency between original and sampled responses; (4) GCA, a long-form graph-based method that detects hallucinations via triple-level claim alignment; (5) Error-Informed Refinement (EIR), a fine-grained long-form method that performs span-level hallucination detection and provides error-aware feedback for refining model outputs; and (6) Decomposition & Aggregation Model (D&A Model), a long-context method that extends encoder-only models by chunking long inputs and aggregating chunk-level representations via attention. These baselines span gray-box vs. black-box settings, as well as uncertainty-based, consistency-based, graph-based, fine-grained, and long-context paradigms, ensuring that our comparisons are broadly representative.

The methods SE and SelfCk-BS generate likelihood scores indicating the probability that a sample is a hallucination, rather than assigning explicit labels denoting factuality. To align these methods with our task, we adopt a thresholding strategy analogous to our own. The main difference is that our method estimates the likelihood of a sample being factual. For each method and dataset, we calculate the mean (μ) and variance (σ^2) of the hallucination likelihood scores, then identify the optimal threshold within the interval $[\mu, \mu + 3\sigma]$ that maximizes evaluation performance. No method receives special or dataset-specific tuning.

A.8 Representation Visualization Analysis

To further validate HRKG-HD's capability in capturing long-range dependencies and constructing effective graph structures, we conducted two complementary experiments.

First, we randomly sampled 20 long-form responses from LHD to evaluate two input strategies. In the full-length strategy, the entire response was provided to the LLM for fact extraction. In the segmented strategy, the text was divided into shorter segments for independent extraction. The primary difference lies in the long-range factual dependencies: splitting text into independent chunks removes structural links and restricts HRKG-HD's propagation to local neighborhoods. Conversely, a full-text graph preserves cross-paragraph connections, enabling multi-hop aggregation over long-range dependencies.

The fact node vectors from the original and five sampled responses were normalized and visualized via t-SNE. Points in Figures 4 and 6 represent fact-node embeddings after relation-aware multi-hop propagation. Each color denotes an aligned fact cluster (the same fact across responses), while different colors represent distinct facts. Effective long-range modeling should make embeddings of the same fact more consistent after propagation. As shown in Figure 6, the full-length strategy produces a more cohesive embedding distribution, demonstrating its superiority.

Second, we examined three graph construction variants. The fact-centric HRKG explicitly introduces fact nodes, linking entities to their corresponding facts and qualifiers. The fact-node-removed HRKG eliminates fact nodes, retaining only direct entity relationships. The traditional relational graph uses a standard triple-based structure without higher-order features. Comparative results in Figure 6 show that the fact-centric HRKG yields more compact distributions with higher cross-response similarity. These findings indicate that the fact-centric HRKG effectively integrates global contextual features, leading to more stable, semantically consistent representations.

A.9 Analysis of Hyper-Relational Fact Extraction

To assess the robustness of HRKG-HD against fact extraction errors, we introduced controlled noise into hyper-relational graphs and analyzed the impact. Two noise-injection strategies are employed: fact-dropping, which simulates information loss by randomly removing 0–50% of core triples while preserving qualifiers, and qualifier-corruption, which simulates LLM-based inaccuracies by replacing qualifier values or types. For both, corruption intensity varied from 0% to 50%, meaning up to half of the extracted content was altered.

All evaluations were conducted on the same hardware to ensure comparability. We used 50 samples randomly drawn from LHD, with accuracy (ACC) as the metric. Each experiment was repeated five times to ensure reliability and statistical robustness.

As shown in Figure 7, HRKG-HD exhibits significant performance advantages in noisy envi-

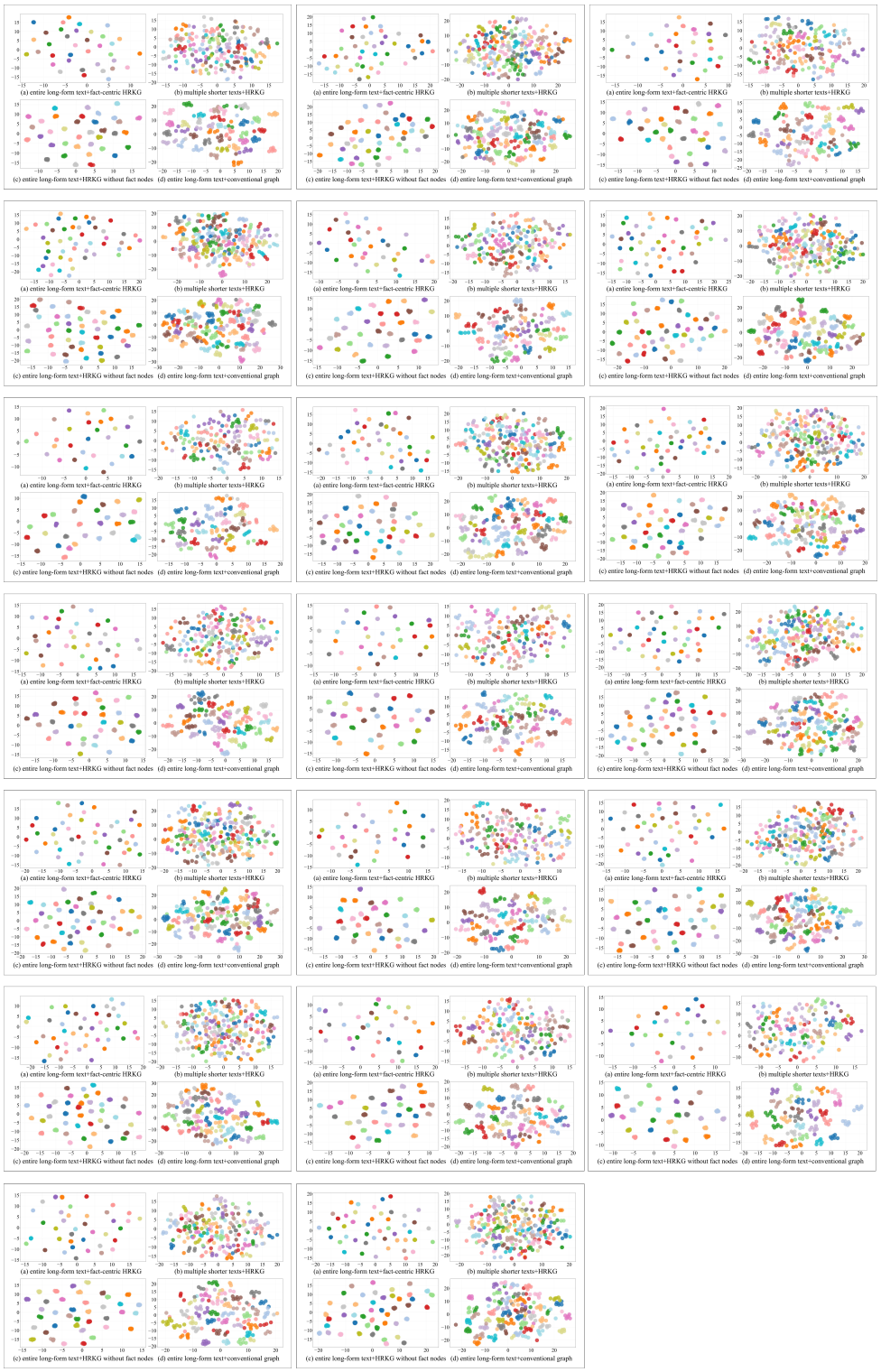

Figure 6: Representation visualization on a randomly selected subset.

ronments. Even with 40% fact discarding, the model maintains 63.0% accuracy. This advantage likely stems from the RV compensation mechanism; when facts are missed, the RV process verifies undetected information. However, excessive factual errors hinder HRKG-HD from constructing a complete hyper-relational knowledge graph, causing a substantial accuracy decline. Thus, maximizing extraction accuracy is essential for HRKG-HD effectiveness.

We also compared the effectiveness of different LLMs as extractors. Specifically, GLM-8B, GLM-32B, and GLM-235B were used to extract hyper-relational facts, while all other aspects of

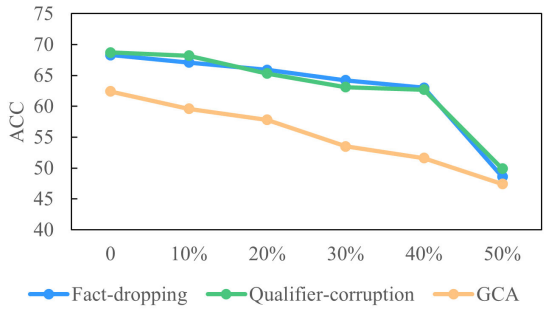

Figure 7: Performance of HRKG-HD under different noise conditions.

the experimental setup remained identical to those described above.

As shown in Table 10, more powerful LLMs consistently achieve higher ACC. These results confirm that the quality of the fact extractor plays a critical role in downstream performance.

Table 10: Details of prompt templates for qualifier in reverse validation.

Model	Qwen-8B	Qwen-30B	Qwen-235B
ACC	63.00	66.33	70.33