Transactions on Computational and Scientific Methods | Vo. 5, No. 6, 2025
ISSN: 2998-8780
https://pspress.org/index.php/tcsm
Pinnacle Science Press

Context-Aligned and Evidence-Based Detection of Hallucinations in Large Language Model Outputs

Yuting Peng
New York University, New York, USA
yp2212@nyu.edu

Abstract: This paper addresses the issues of semantic drift and hallucinated information commonly found in the outputs of large language models. It proposes a detection framework based on fine-grained analysis. The framework consists of two main components: a Context-Aligned Representation module and a Layered Verification of Evidence module. The first module builds semantic alignment between the input and the generated text. It effectively identifies contextual shifts and logical inconsistencies. The second module segments the generated content into multiple semantic units and performs layer-by-layer verification using external knowledge. This enables the localization and modeling of potential hallucinated content. The model adopts a shared representation learning structure. It maintains strong semantic consistency modeling while improving the detection of implicit hallucinations in complex reasoning tasks. Systematic experiments on the TruthfulQA dataset show that the proposed method significantly outperforms existing mainstream detection models in terms of precision, recall, F1-score, and fact consistency. It demonstrates strong fine-grained awareness and cross-task stability. In addition, transfer evaluations in both open-domain and closed-domain scenarios, along with extended experiments under different domains and knowledge source conditions, further confirm the adaptability and practicality of the method in real-world generation settings. This approach provides solid technical support for improving the trustworthiness and safety of language model outputs. It also offers a structured solution for intelligent review of generated content.

Keywords: Semantic consistency detection, language model hallucination, external knowledge verification, fine-grained semantic modeling

1. Introduction

With the rapid advancement of artificial intelligence, large language models have achieved remarkable progress in natural language processing. They are widely used in tasks such as text generation, question answering, and dialogue systems. These models possess strong capabilities in language understanding and generation[1]. They can simulate human language expression across various contexts. However, as their application deepens, issues such as semantic drift and hallucinated content in model outputs have become increasingly prominent. These problems not only affect the practicality and reliability of the models but also pose challenges in information security, content regulation, and human-computer interaction. Therefore, fine-grained detection of deviations from true semantics or factual accuracy in model outputs has become a key focus in the study of trustworthy AI.

Semantic drift refers to inconsistencies between the generated text and the original input or real-world context at the semantic level. It often appears as topic deviation, referential confusion, or logical errors. Hallucinated

content refers to information that does not align with objective facts. This includes fabricated facts, fictitious events, or false inferences. These problems are especially serious in high-sensitivity scenarios such as news generation, medical question answering, and legal consultation. If not identified in time, they may cause misinformation or even trigger public opinion risks. As the size and generative power of language models increase, the generated texts become more natural and fluent. This makes hallucinated content more deceptive and harder to detect, imposing greater demands on detection techniques[2].

On the technical level, existing methods often rely on coarse-grained semantic consistency or fact-checking mechanisms to assess text credibility. These methods are often insufficient when dealing with complex contexts, multi-step reasoning, or implicit semantics. This is especially true when model outputs appear detailed and logically coherent. Coarse-grained methods struggle to capture subtle semantic drift or hallucination[3]. Fine-grained detection emphasizes detailed identification and analysis of semantic units across different levels of the text. This includes words, sentences, paragraphs, and inter-sentence relations. By building more sophisticated detection models, it is possible to identify deviations early and localize them precisely. This can improve overall detection accuracy and response speed.

Studying semantic drift and hallucinated content in language model outputs has both theoretical and practical value. On the one hand, in-depth research in this area can promote improvements in natural language understanding and generation mechanisms[4,5]. In turn, this supports the optimization of the models themselves and enhances the trustworthiness and semantic consistency of their outputs. On the other hand, from a societal perspective, fine-grained detection techniques can be applied to scenarios such as information review, content security, and academic writing assistance. These applications help limit the spread of false information and protect public understanding. In the context of increasingly digital regulation and governance, establishing automated and intelligent content detection mechanisms is essential for ensuring a healthy information ecosystem[6].

Therefore, building a fine-grained detection framework to address semantic drift and hallucination in language model outputs is not only a technical challenge but also a strategic imperative. It is essential for supporting the healthy development of AI and enhancing public trust in intelligent systems[7,8]. Such research will promote higher-quality and more reliable human-AI collaboration. It lays a solid foundation for the safe deployment of large language models. As AI becomes more deeply integrated with society, ensuring the truthfulness and consistency of generated content will be a critical task for the sustainable development of intelligent technologies.

2. Related work

2.1 Large Language Model

In recent years, large language models have demonstrated strong modeling capabilities and broad adaptability in the field of natural language processing. Trained on large-scale corpora, these models can learn rich linguistic structures, semantic knowledge, and reasoning abilities[9]. As a result, they show human-like performance in tasks such as text generation, contextual understanding, translation, and question answering. These models often contain billions or even tens of billions of parameters[10]. They rely on large-scale unsupervised pre-training combined with limited supervised fine-tuning. This enables them to transfer across tasks under a unified model framework, greatly advancing the generalization and efficiency of language technologies.

Despite these strengths, large language models still present non-negligible problems in content generation. Semantic drift and hallucinated information arise when the model, lacking explicit factual grounding, completes content based on statistical associations in the training data. This can result in logically fluent but factually inaccurate outputs. Such issues are especially evident in open-domain generation tasks. When facing vague questions or incomplete contexts, models tend to produce content that appears plausible but is

actually misleading[11]. These problems reveal the current limitations in semantic control and factual grounding. They also highlight the inadequacies in the model's internal mechanisms for capturing real-world meaning.

As large language models become increasingly embedded in real-world applications, their output raises new concerns regarding safety, interpretability, and controllability[12]. Improving the verifiability of model outputs and enforcing semantic consistency during generation are key research goals. Detecting and filtering potential false content has also become a critical task[13]. These challenges have prompted research efforts from multiple perspectives, including model architecture, training paradigms, and detection techniques. The aim is to develop more precise and effective methods of control and evaluation, in order to enhance the reliability and accountability of large language models[14].

2.2 The "hallucination" problem of large language models

The hallucination problem in large language models refers to the generation of plausible but factually incorrect information in the absence of real-world grounding. This issue arises when the model produces content that appears coherent and contextually appropriate but does not align with verifiable facts or objective reality[15]. It is not limited to standard text generation tasks but is also frequently observed in applications such as question answering, summarization, and code generation. These tasks often require precise factual references and logical consistency, which large language models may fail to uphold due to their reliance on statistical correlations rather than true comprehension. Unlike humans, who can cross-reference external sources and apply common sense or domain-specific reasoning, these models generate outputs based on patterns learned from large text corpora[16]. When the input prompt lacks sufficient detail, or when the topic falls outside the model's training data distribution, the risk of hallucination increases significantly. As a result, the model may introduce invented details, fabricate entities or events, or present misleading interpretations as factual content. This makes hallucination one of the most critical and persistent challenges in the development and deployment of language models, especially in high-stakes domains such as healthcare, law, finance, and scientific research, where factual accuracy is non-negotiable and misinformation can lead to harmful outcomes. Addressing hallucination is essential not only for improving the reliability of model outputs but also for ensuring their safe and responsible use in real-world applications.

Hallucinated content often shows high linguistic coherence and logical consistency. Ordinary users may struggle to judge its truthfulness, which increases the risk of misinformation during dissemination. Compared to simple grammatical errors or obvious model failures, hallucinations are more hidden and deceptive. They reduce model credibility and make comprehension and judgment more difficult for users. In multi-turn dialogue and summarization tasks, models may introduce incorrect details during local reasoning. This can lead to semantic drift that users may not easily notice. Therefore, identifying such fluent but fabricated content is key to improving the reliability of language models.

Researchers have proposed various strategies to mitigate hallucinations. These include knowledge-enhanced models, external fact-checking, and improved training objectives. However, practical use still faces trade-offs between accuracy and computational cost. In complex contexts or open domains, hallucination detection remains a challenging task. Identifying gradually accumulated misleading content at a fine-grained level requires a combination of semantic consistency checks, context-aware analysis, and fact-based comparison. In-depth research on this issue will directly affect the safe deployment and widespread use of large language models. It is also critical to advancing intelligent systems toward greater trustworthiness.

3. Method

This study proposes a fine-grained detection framework for identifying semantic drift and hallucinated content in outputs generated by large language models. The overall approach introduces innovations from

two dimensions: semantic modeling and multi-layer verification. First, in terms of semantic modeling, a Context-Aligned Representation mechanism (CAR) is proposed. It constructs alignment vectors between the input and the generated text. This captures potential semantic shifts and enables consistency measurement both within and across sentences. Second, for the verification mechanism, a Layered Verification of Evidence module (LVE) is introduced. This module decomposes the generated text into hierarchical semantic units. It then verifies each layer using external knowledge sources. This enables precise localization and identification of hallucinated content. These two modules work together to significantly enhance the system's ability to detect subtle anomalies under complex contexts. They offer technical support for building trustworthy applications based on large language models. The model architecture is shown in Figure 1.

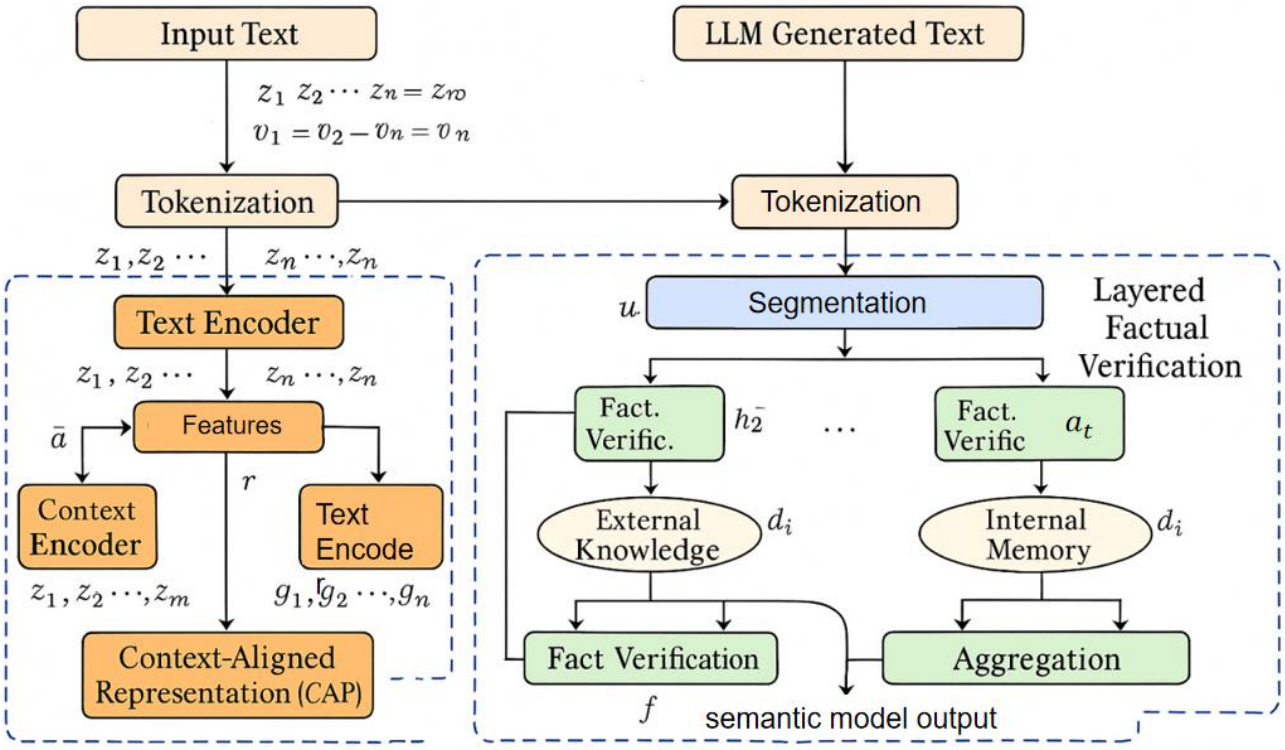

Figure 1. Overall model architecture diagram

3.1 Context-Aligned Representation

The contextual alignment representation mechanism proposed in this section aims to accurately model the correspondence between the input text and the generated text at the semantic level. By explicitly aligning the two, the local semantic shift and overall topic drift in the generated content can be effectively identified. This mechanism not only focuses on the surface matching at the word level, but also emphasizes the establishment of stable contextual associations in the deep semantic space, thereby improving the perception of fine-grained anomalies. Its module architecture is shown in Figure 2.

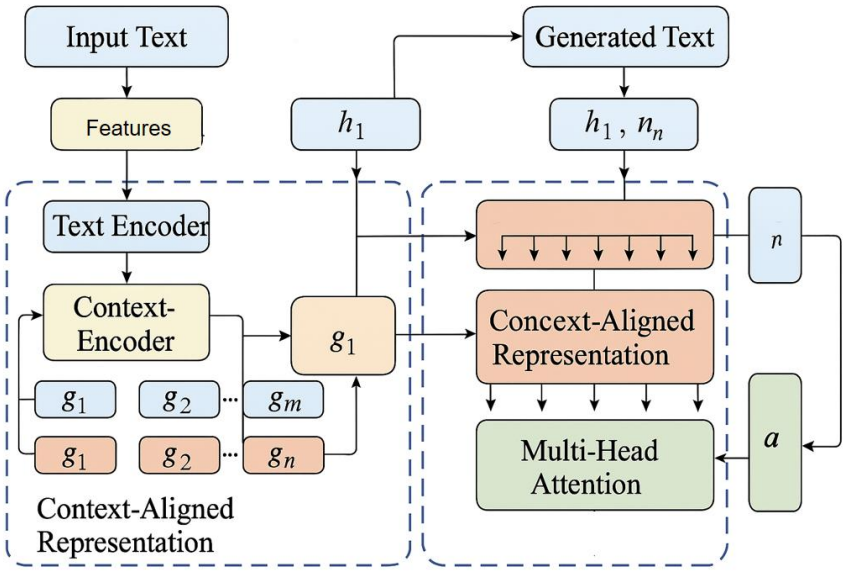

Figure 2. CAR module architecture

First, the input text $X = \{x_1, x_2, ..., x_m\}$ and the generated text $Y = \{y_1, y_2, ..., y_n\}$ are mapped to corresponding representation sequences $H_X = \{h_1^X, h_2^X, ..., h_m^X\}, H_Y = \{h_1^Y, h_2^Y, ..., h_n^Y\}$ through a shared encoder. In order to further model the semantic association between the two, an attention mechanism is introduced to construct a context alignment matrix $A \in R^{m \times n}$, where each element is defined as $A_{ij} = sim(h_i^X, h_j^Y), sim(\cdot)$, which is a standard cosine similarity or a trainable similarity function.

Based on the alignment matrix $\tilde{h}_j^Y = \sum_{i=1}^{m} A_{ij} h_i^X$, we calculate the weighted semantic mapping between each position of the generated text and the input to form a contextual alignment vector $d_j = h_j^Y - \tilde{h}_j^Y$. This alignment vector is used to capture whether the generated content is fully aligned with the original input, thereby assisting in detecting semantic drift. On this basis, a semantic difference representation is constructed to measure the local semantic deviation between the current generated content and the input. The overall alignment error can be obtained by reducing the difference vectors of all positions:

$$D = \frac{1}{n} \sum_{j=1}^{n} \| d_j \|_2$$

This scalar can be used as an indicator of the degree of semantic consistency and provide a quantitative basis for subsequent offset judgment.

In order to further model the deep semantic association between contexts, a gating mechanism is introduced to dynamically adjust the importance of input and generated information. Specifically, a gating vector $g_j = \sigma(W_g[h_j^Y; \tilde{h}_j^Y] + b_g)$ is defined, where $[\cdot;\cdot]$ represents vector concatenation, W_g, b_g is a trainable parameter, and σ is a Sigmoid activation function. The final context fusion representation is defined as:

$$r_j = g_j \otimes h_j^Y + (1 - g_j) \otimes \tilde{h}_j^Y$$

Where $\otimes$ represents element-by-element multiplication. This representation can simultaneously capture the semantic features of the generated text itself and its dependence on the input context, enhancing the context sensitivity of the representation.

To provide a stable semantic representation for overall alignment modeling, sequence-level compression is also required to aggregate the fusion representation $\{r_1, r_2, ..., r_n\}$ into a unified vector r. This paper uses a multi-head self-attention aggregation mechanism to calculate the global representation:

$$r = MultiHeadAttn(R, R, R)$$

Where $R \in R^{n \times d}$ represents the matrix composed of all fused vectors. This global semantic vector is used as the final input of the semantic consistency detection module, providing a basis for subsequent tasks such as offset classification and anomaly recognition. This mechanism can capture semantic differences at different granularities and is particularly suitable for stable alignment analysis of diversified generated texts.

3.2 Layered Verification of Evidence module

In order to achieve fine-grained identification of fictional content in the output of large language models, this section introduces a hierarchical evidence verification module designed to conduct detailed factual consistency analysis at multiple levels of granularity. The primary objective of this module is to identify and isolate segments of generated text that may deviate from established facts or introduce hallucinated information, which can often appear coherent yet lack real-world grounding. By decomposing the generated content into smaller semantic units—such as phrases, sentences, or logical segments—the module enables more precise analysis of the factual validity of each part. Furthermore, it incorporates evidence from multiple external sources, including structured databases, knowledge graphs, and verified textual references, to support or challenge the factual claims made in the output. This multi-source approach ensures that the verification process is not overly dependent on a single knowledge base, which could be incomplete or biased. Instead, it leverages a diverse range of knowledge resources to assess consistency, cross-check facts, and detect inaccuracies that may not be evident through semantic analysis alone. The hierarchical design of the module allows it to operate progressively, starting from basic fact-checking at the sentence level and advancing to more complex relational and contextual consistency evaluations across larger text spans. The overall module architecture, as illustrated in Figure 3, reflects this layered verification logic, offering a robust framework for detecting hallucinated content with high granularity and contextual awareness.

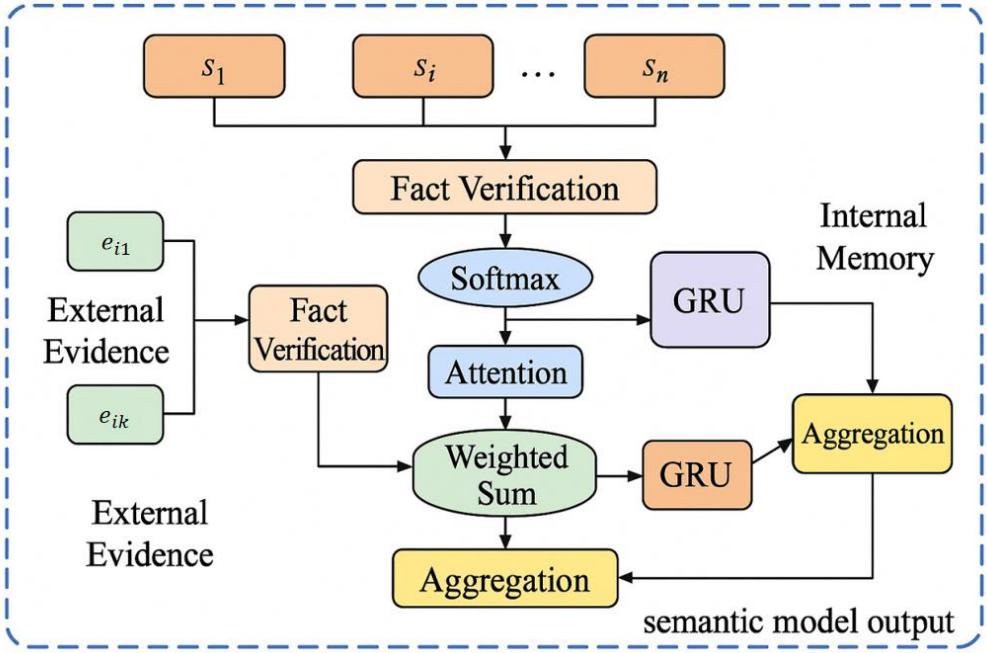

Figure 3. LVE module architecture

First, the generated text is represented as a fragment sequence $Y = \{s_1, s_2, ..., s_n\}$, where each fragment s_i is encoded as a representation vector h_i and input into the verification module for hierarchical modeling. At the first level, for each fragment, the most relevant external evidence set $E = \{e_{i1}, e_{i2}, ..., e_{ik}\}$ is obtained through the retrieval module, and each evidence fragment e_{ij} is represented as a vector v_{ij} to form a candidate set for alignment.

In order to evaluate the consistency between the fragment and the evidence, a semantic matching function is introduced to score the relevance between the two, in the form of:

$$a_{ij} = \cos(h_i, v_{ij})$$

Where $\cos(\cdot)$ represents the cosine similarity. The scores of all evidences are normalized to obtain the normalized attention weights:

$$w_{ij} = \frac{\exp(a_{ij})}{\sum_{j'=1}^{k} \exp(a_{ij'})}$$

Based on this, the weighted evidence vector representation corresponding to the fragment is constructed:

$$v_i = \sum_{j=1}^{k} w_{ij} \cdot v_{ij}$$

This representation serves as the factual comparison semantic basis for the segment.

Furthermore, a segment-level consistency vector $d_i = h_i - \tilde{v}_i$ is constructed to measure the semantic difference between the current generated content and its evidence. By taking the norm of the difference vector, the consistency offset value of the segment is defined as:

$$\delta_i = \| d_i \|_2$$

The offset values of all fragments are aggregated to generate global deviation features for the final judgment output. Inside the model, a memory mechanism module M is introduced to model the cumulative information of semantic consistency across fragments. Its update method is:

$$M^{(i)} = GRU(M^{(i-1)}, [h_i, \tilde{v}_i])$$

To capture potential cross-segment semantic relationships and contradictions between contexts.

Finally, the consistency offset value δ_i and memory state $M^{(i)}$ of all fragments will be sent to the aggregation function for overall evaluation, and the overall credibility score or abnormal prompt signal of the generated text will be output. This hierarchical verification module is modeled at both the local and global levels, effectively improving the system's recognition depth and semantic tracking capabilities for fictional content.

4. Experimental Results

4.1 Dataset

This study uses TruthfulQA as the primary dataset to support the detection of semantic drift and hallucinated content in outputs generated by large language models. TruthfulQA is a dataset specifically designed to evaluate the factual accuracy and truthfulness of language models. It includes questions across various categories such as health, law, history, and science. The content is complex and diverse, making it both challenging and suitable for real-world applications.

A key feature of this dataset is its use of highly leading questions. These are likely to trigger hallucinated or factually incorrect answers from language models. This provides ample abnormal samples for detection algorithms. Each sample typically contains a question, several model-generated answers, and corresponding factuality labels. This structure supports the training and evaluation of fine-grained detection methods.

In addition, TruthfulQA uses a fact-based consistency scoring system. This offers a reliable reference for assessing the correctness of model outputs. By using this dataset, the study focuses on semantic alignment and fact verification under complex conditions. It provides strong support for the development and testing of the proposed algorithm.

4.2 Experimental setup

To validate the effectiveness of the proposed fine-grained detection method, this study builds an experimental setup using the TruthfulQA dataset. A mainstream pre-trained language model is used as the source for text generation. After receiving a question, the model generates an answer, which is then processed by the proposed detection framework. The text passes through the Context-Aligned Representation module (CAR) and the Layered Verification of Evidence module (LVE). The final output includes predictions of semantic consistency and factual deviation.

During training and evaluation, all texts are preprocessed in a standardized manner. Embedding representations are obtained through a shared encoder. External knowledge retrieval is supported by an open-source knowledge base to simulate real-world application scenarios.

For evaluation, four core metrics are used to reflect detection performance. These include Precision, Recall, F1-score, and Fact-Consistency Rate. The first three metrics measure the accuracy and coverage of abnormal content detection. Fact-Consistency Rate assesses the model’s overall ability to align with true semantics. It is suitable for evaluating the real-world effectiveness of multi-source evidence verification. The table below presents the experimental settings used in this study. The main training configuration is shown in Table 1.

Table 1: Training Configuration

Component	**Configuration**
Dataset	TruthfulQA
Language Model	chatglm3-6b
Knowledge Source	Open-source external corpus
Embedding Encoder	Transformer-based shared encoder
Evaluation Metrics	Precision, Recall, F1-score, Fact-Consistency Rate

4.3 Experimental Results

1) Comparative experimental results

First, this paper gives the comparative experimental results with other models. The experimental results are shown in Table 2.

Table 2: Comparative experimental results

Method	**Precision**	**Recall**	**F1-Score**	**Fact-Consistency Rate。**
DetectGPT[17]	72.4	68.3	70.3	65.8
Fast-detectgpt[18]	75.1	70.7	72.8	69.4
HaluEval[19]	78.6	74.9	76.7	71.5
FACTScore[20]	79.3	76.1	77.7	73.0
Ours	84.7	81.2	82.9	78.6

From the overall comparison results, the method proposed in this study outperforms existing public approaches across all four core metrics. It demonstrates stronger overall performance in detecting semantic consistency and factual correctness. In particular, it achieves an F1-score of 82.9 and a Fact-Consistency Rate of 78.6, which are significantly higher than those of other methods. This indicates that the proposed method performs better in balancing detection accuracy and coverage. It also shows greater stability in identifying semantic drift and hallucinated content in outputs generated by large language models.

Compared with methods such as DetectGPT and Fast-detectgpt, the proposed approach improves the modeling of semantic correspondence between input and output through the Context-Aligned Representation (CAR) module. This significantly enhances the detection of topic shifts and incoherent context. Since those baseline models mostly rely on global features or coarse-grained statistical patterns to identify anomalies, they often miss subtle logic inconsistencies or local drift. The fine-grained modeling in this method effectively addresses this limitation.

In comparison with methods like HaluEval and FACTScore, which introduce external knowledge for verification, the proposed Layered Verification of Evidence (LVE) module achieves a higher fact-consistency rate. This result suggests that multi-level, multi-source verification strategies can better analyze the factual accuracy of generated content. The method increases sensitivity to hallucinated details and improves detection in complex question answering tasks that involve multi-step reasoning.

Moreover, this method improves precision while maintaining a high recall rate. This reflects its ability to detect more abnormal segments while reducing false positives. Such performance is important for real-world scenarios like content review and safety monitoring. It also shows that fine-grained, structured detection strategies are more robust and interpretable when handling hallucinations in large language models. These findings provide a solid technical and experimental foundation for future research.

2) Analysis of the contribution of temporal consistency to model checking effect

This paper first gives an analysis of the contribution of temporal consistency to model detection effect, and the experimental results are shown in Figure 4.

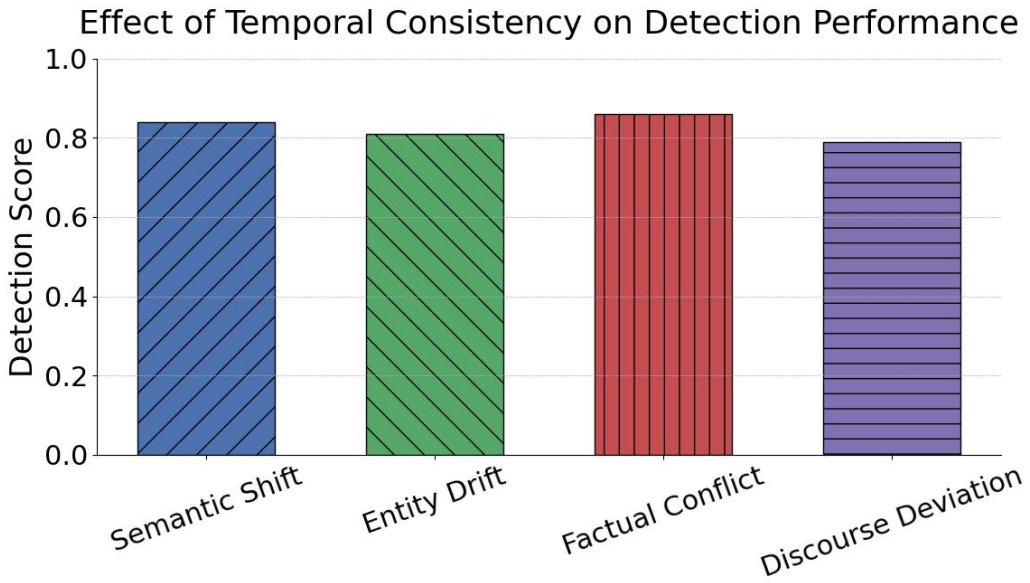

Figure 4. Analysis of the contribution of temporal consistency to model checking effect

As shown in the experimental results in Figure 4, after introducing the temporal consistency modeling mechanism, the proposed method performs stably across various types of semantic anomaly detection tasks. The overall detection scores remain at a high level. This indicates that the mechanism significantly contributes to improving the model's overall judgment ability. It achieves the highest detection score in the Factual Conflict scenario, showing that temporal modeling is especially effective in capturing fact conflicts caused by distorted temporal reasoning.

In the Semantic Shift and Entity Drift scenarios, the model also demonstrates strong detection performance. This suggests that temporal consistency helps not only in identifying breaks in logical sequence but also in tracking the evolution of meaning and entities within the text. By modeling the state changes of generated content over time, the model can more accurately detect subtle shifts caused by changes in context.

In comparison, the detection score for the Discourse Deviation task is slightly lower, though still at a relatively high level. This indicates that while the temporal mechanism aids global semantic coherence, it may be less effective in handling more complex discourse-level jumps or logical gaps. Addressing such issues may require stronger cross-paragraph modeling or deeper pragmatic understanding.

In summary, the introduction of temporal consistency enhances the model's ability to capture structural deviations in generated text. It provides clear benefits in detecting local factual inconsistencies, entity drift, and time-dependent logical errors. These findings confirm the importance of incorporating temporal logic into content detection models and offer technical support for building robust and fine-grained content monitoring systems.

3) Comparison of fictional content detection capabilities based on domain division

This paper also compares the fictitious content detection capabilities based on domain division, and the experimental results are shown in Figure 5.

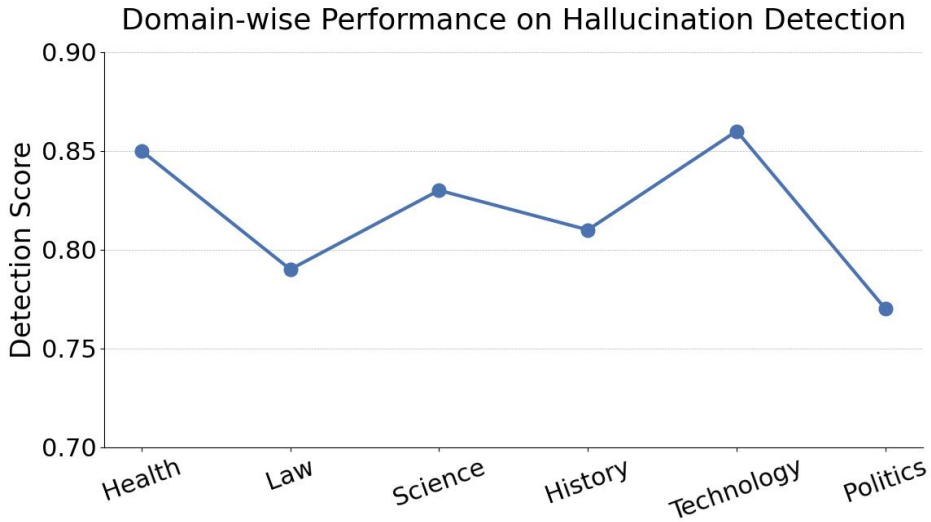

Figure 5. Comparison of fictional content detection capabilities based on domain division

As shown in the results of Figure 5, the proposed method displays varying performance across hallucination detection tasks in different domains, while maintaining high overall detection quality. These variations reflect the differences in semantic structure, expression patterns, and context dependence of hallucinated information across domains. Such differences pose distinct challenges for detection models. The model achieves higher detection scores in the Health and Technology domains, indicating that it is more accurate in identifying semantic drift and factual errors in technical and highly structured texts.

In the Technology domain, generated content often contains clearer causal chains and more well-defined terminology. This allows the model to better use context and external knowledge for consistency checking. Similarly, in the Health domain, the determinism of terminology and the stability of background knowledge make hallucinated content easier to identify. This confirms the high adaptability and sensitivity of the proposed Context-Aligned Representation (CAR) and Layered Verification of Evidence (LVE) mechanisms when applied to well-structured data.

In contrast, performance is weaker in the Politics and Law domains, with the lowest detection score observed in political discourse. This may be due to the frequent use of metaphor, subjective inference, and vague factual boundaries in such texts, which makes hallucination detection more difficult. Moreover, content in these domains often depends heavily on context and temporal dynamics. A single instance of semantic drift may not capture the full scope of factual contradiction, so the model requires deeper reasoning to assess fact consistency.

These results further indicate that different domains place different demands on detection capabilities. They also show that the proposed method has strong generalization ability across domains. However, hallucination detection can be further improved by introducing domain adaptation techniques or enhancing domain-specific knowledge modeling. This provides a direction for improving model robustness and reliability in complex real-world scenarios.

4) The impact of external knowledge source quality on detection accuracy

This paper also examines the impact of the quality of external knowledge sources on the detection accuracy of hallucinated content in the output of large language models. The reliability and completeness of external knowledge play a critical role in supporting factual verification processes, especially when the model output includes information that requires grounding in real-world facts. In the context of the proposed detection

framework, external knowledge sources are used to validate specific claims or assertions made by the model. When these sources are inconsistent, sparse, outdated, or contain noise, they can compromise the ability of the system to correctly identify fictional or misleading content. On the other hand, high-quality knowledge sources that are well-structured, comprehensive, and up to date provide a more solid foundation for evaluating the factual integrity of the generated text. This relationship between knowledge source quality and detection performance is particularly important in domains where factual accuracy is essential, such as healthcare, law, and scientific communication. The analysis presented in this paper highlights this correlation, offering insights into how variations in the credibility and granularity of external information can influence the effectiveness of hallucination detection. The corresponding experimental setup and results that illustrate these findings are presented in Figure 6.

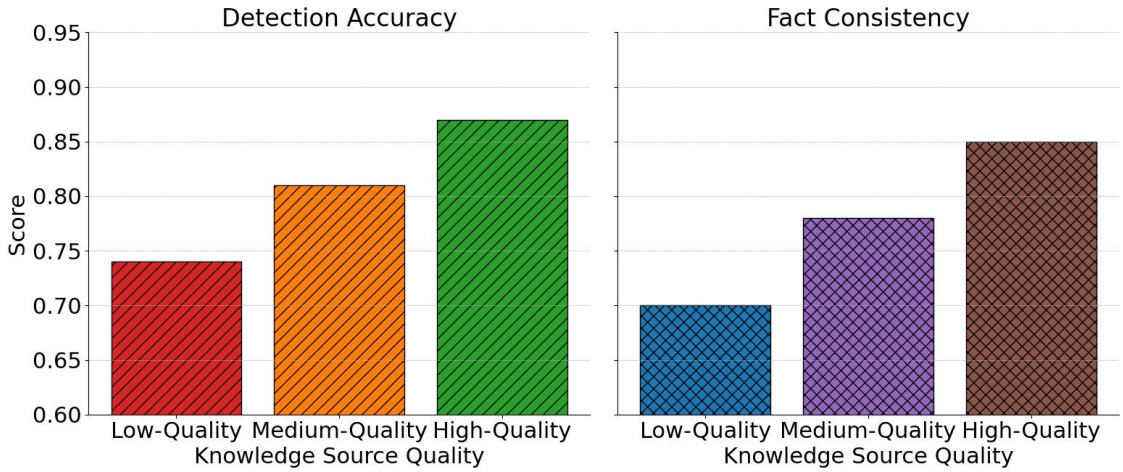

Figure 6. The impact of external knowledge source quality on detection accuracy

As shown in the results of Figure 6, the quality of external knowledge sources has a significant impact on the model's detection accuracy and fact consistency. As knowledge quality increases from low to high, the model shows steady improvement on both core metrics. This confirms the importance of high-quality knowledge in supporting content verification and reducing hallucination risks. The trend clearly indicates that the credibility of external evidence directly affects the model's ability to identify hallucinated content. It plays a critical role in the fact verification module.

With low-quality knowledge sources, the model performs poorly. Both Detection Accuracy and Fact Consistency remain at low levels. This suggests that unstable, fragmented, or noisy knowledge interferes with the model's judgment and may even lead to false guidance. As a result, hallucinated segments may go undetected or be incorrectly identified as factual. This issue is especially common in open-domain or multi-source knowledge retrieval tasks. Therefore, improving knowledge quality is a fundamental requirement for building robust detection systems.

Under medium-quality knowledge sources, the model’s performance shows noticeable improvement, reflecting the importance of having access to reasonably accurate and well-structured external information when verifying the factual consistency of generated content. In this context, “medium-quality” refers to knowledge that, while not exhaustive or perfectly curated, maintains a general level of correctness and organization sufficient to support basic fact-checking processes. Such knowledge may include partially complete entries, moderate noise levels, or limited contextual depth, yet still offers enough substance for the model to conduct preliminary assessments of semantic and factual alignment. This setting highlights the practical reality in which most real-world applications operate, where high-quality, comprehensive knowledge bases may not always be available. The model’s ability to make use of these imperfect but usable

resources demonstrates the flexibility and resilience of the proposed detection framework. In particular, it confirms that the Layered Verification of Evidence (LVE) module is capable of functioning effectively under non-ideal conditions. The module shows a degree of fault tolerance and adaptability, allowing it to filter, match, and assess facts even when the supporting knowledge is incomplete or unevenly distributed. However, it is also evident that such performance remains influenced by the limitations in evidence coverage and precision, which can constrain the model's ability to detect more subtle or complex hallucinations. As such, while medium-quality knowledge can serve as a useful baseline for verification, achieving higher accuracy in detection still depends on the availability of more reliable and expansive external sources.

When high-quality knowledge is used, the model achieves the best results in both metrics. Detection Accuracy exceeds 0.87, showing the strong constraining effect of complete and credible knowledge systems on generated content. This result further confirms that the proposed method, when combined with knowledge integration mechanisms, can accurately identify hallucinations in complex contexts. It enhances overall detection stability and generalization. This finding provides both theoretical and experimental support for developing knowledge-enhanced detection frameworks.

5) Evaluation of transfer performance of detection algorithms in open and closed domains

Finally, this paper presents a comprehensive evaluation of the migration performance of the proposed detection algorithm across open-domain and closed-domain scenarios. This evaluation aims to assess the adaptability and generalization capability of the algorithm when applied to different types of language generation environments. Open-domain tasks typically involve a wide range of topics with less structured input and more variability in context, which can increase the likelihood of hallucinated content due to the model's broader generative freedom and reduced grounding. In contrast, closed-domain tasks are generally confined to specific subject areas with well-defined boundaries, stable terminology, and more predictable knowledge structures, offering a more controlled setting for detection. Evaluating the detection algorithm in both of these domains is essential for understanding its robustness and practical utility in real-world applications, where models are deployed in diverse and dynamic contexts. The migration performance analysis investigates how well the detection system maintains its effectiveness when transitioning between these domains, and whether its core components—such as semantic alignment and evidence verification—remain functional and reliable across varying semantic and structural conditions. The corresponding experimental setup and results that demonstrate this cross-domain evaluation are illustrated in Figure 7.

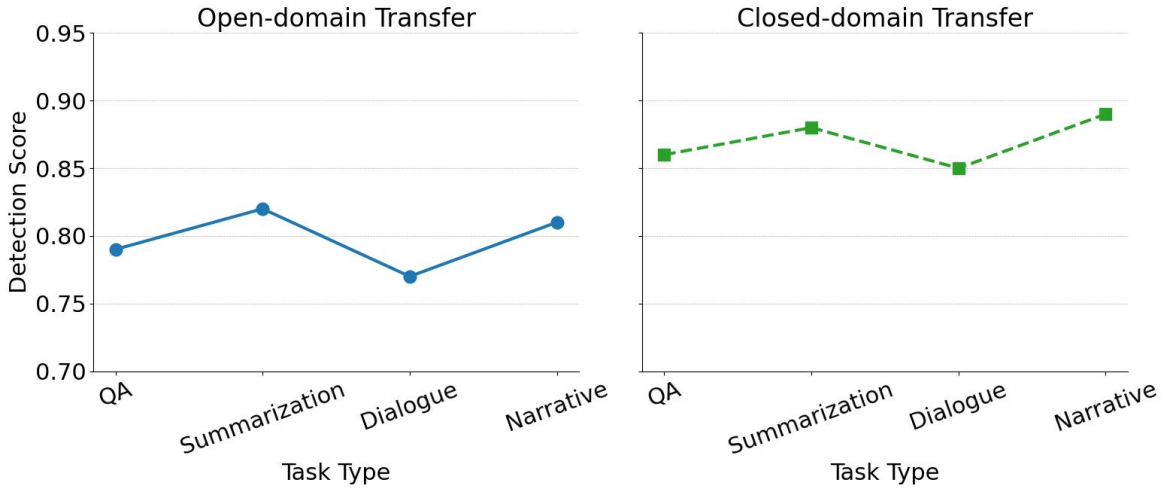

Figure 7. Evaluation of transfer performance of detection algorithms in open and closed domains

As shown in the experimental results in Figure 7, the proposed detection algorithm demonstrates strong transferability in both open-domain and closed-domain settings. However, the performance differences reveal distinct demands on model structure and semantic generalization across environments. In open-domain tasks, the model performs consistently, with high detection scores in Summarization and Narrative tasks. This suggests that the Context-Aligned Representation (CAR) and Layered Verification of Evidence (LVE) modules are highly adaptable. They can effectively detect semantic drift and hallucinated content in texts without explicit domain boundaries.

However, in the Dialogue task, the detection score in the open domain shows a noticeable decline. This may result from the higher uncertainty and contextual jumps typical of open-domain conversations. The lack of clear factual grounding increases the difficulty of modeling local logic and entity reference. This reduces the stability of hallucination detection. These findings suggest that improving context retention and coreference resolution is critical for better detection performance in multi-turn dialogue and open-ended QA tasks.

In contrast, all four tasks in the closed domain maintain high detection scores. This shows that when texts have clear boundaries, structured formats, and stable knowledge support, the proposed method can fully leverage its fine-grained analysis and multi-source verification strengths. In particular, detection performance in Narrative and Summarization tasks is significantly higher in the closed domain. This confirms that semantic consistency modeling is more reliable when the context is stable.

Overall, the experiment confirms that the model has good transfer and generalization ability in real-world applications. It performs consistently in closed scenarios and shows room for improvement in open tasks. Future work may explore dynamic knowledge retrieval and cross-paragraph semantic chain modeling to enhance robustness and detection depth in semantically complex open-domain environments.

5. Conclusion

This paper proposes a fine-grained detection framework to address semantic drift and hallucination in the outputs of large language models. The framework integrates a Context-Aligned Representation mechanism and a Layered Verification of Evidence module. It conducts analysis from two perspectives: semantic consistency and factual consistency. By modeling deep semantic relationships between the input and the generated text, the method enhances the system ' s ability to detect complex linguistic structures and ambiguous expressions. This provides a technical foundation for identifying potential hallucinations and semantic deviations. In terms of design strategy, the approach emphasizes multi-granular and multi-level understanding and verification of content. It fully leverages contextual structure and external knowledge to detect inconsistent and inaccurate information in a segmented and layered manner. This mechanism is especially effective in open-ended generation scenarios that involve multi-step reasoning and cross-sentence logic. It helps reduce false positives and false negatives commonly found in coarse-grained detection methods. Experimental results show that the proposed method has strong transferability across various dimensions and task types. It adapts well to diverse and complex language generation environments, demonstrating high practical value.

From an application perspective, the proposed detection framework plays a significant role in enhancing the safety and controllability of language generation systems. As large language models are widely deployed in sensitive domains such as news generation, medical question answering, educational services, and legal consultation, ensuring the truthfulness and consistency of generated content has become a core concern. This study provides a scalable and integrable solution for building trustworthy language systems. It has the potential for broad application in content review, text verification, and human-computer interaction, supporting the development of more reliable and responsible AI systems. Future research will extend the framework's adaptability to multilingual, multimodal, and cross-domain settings. It will explore efficient integration with external resources such as knowledge graphs and fact retrieval systems to improve its handling of hallucinations under dynamic knowledge conditions. In addition, the introduction of a warning

and feedback mechanism during the generation process will be considered. This will enable a closed-loop system that connects generation, detection, and optimization. The goal is to improve the semantic quality and factual consistency of generated content at the source and to promote the development of more trustworthy and explainable language generation technologies.

References

[1] Zhang, Yue, et al. "Siren's song in the AI ocean: a survey on hallucination in large language models." arXiv preprint arXiv:2309.01219 (2023).

[2] Tonmoy, S. M., et al. "A comprehensive survey of hallucination mitigation techniques in large language models." arXiv preprint arXiv:2401.01313 6 (2024).

[3] Dahl, Matthew, et al. "Large legal fictions: Profiling legal hallucinations in large language models." Journal of Legal Analysis 16.1 (2024): 64-93.

[4] Liu, Hanchao, et al. "A survey on hallucination in large vision-language models." arXiv preprint arXiv:2402.00253 (2024).

[5] Dhuliawala, Shehzaad, et al. "Chain-of-verification reduces hallucination in large language models." arXiv preprint arXiv:2309.11495 (2023).

[6] Huang, Lei, et al. "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions." ACM Transactions on Information Systems 43.2 (2025): 1-55.

[7] Chen, Yuyan, et al. "Hallucination detection: Robustly discerning reliable answers in large language models." Proceedings of the 32nd ACM International Conference on Information and Knowledge Management. 2023.

[8] Azamfirei, Razvan, Sapna R. Kudchadkar, and James Fackler. "Large language models and the perils of their hallucinations." Critical Care 27.1 (2023): 120.

[9] Hadi, Muhammad Usman, et al. "A survey on large language models: Applications, challenges, limitations, and practical usage." Authorea Preprints 3 (2023).

[10] Shen, Tianhao, et al. "Large language model alignment: A survey." arXiv preprint arXiv:2309.15025 (2023).

[11] Wang, Lei, et al. "A survey on large language model based autonomous agents." Frontiers of Computer Science 18.6 (2024): 186345.

[12] Guo, Taicheng, et al. "Large language model based multi-agents: A survey of progress and challenges." arXiv preprint arXiv:2402.01680 (2024).

[13] Wang, Yufei, et al. "Aligning large language models with human: A survey." arXiv preprint arXiv:2307.12966 (2023).

[14] Jiang, Chaoya, et al. "Hallucination augmented contrastive learning for multimodal large language model." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

[15] McKenna, Nick, et al. "Sources of hallucination by large language models on inference tasks." arXiv preprint arXiv:2305.14552 (2023).

[16] Manakul, Potsawee, Adian Liusie, and Mark JF Gales. "Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models." arXiv preprint arXiv:2303.08896 (2023).

[17] Mitchell, Eric, et al. "Detectgpt: Zero-shot machine-generated text detection using probability curvature." International Conference on Machine Learning. PMLR, 2023.

[18] Chaka, Chaka. "Detecting AI content in responses generated by ChatGPT, YouChat, and Chatsonic: The case of five AI content detection tools." Journal of Applied Learning and Teaching 6.2 (2023): 94-104.

[19] Li, Junyi, et al. "Halueval: A large-scale hallucination evaluation benchmark for large language models." arXiv preprint arXiv:2305.11747 (2023).

[20] Min, Sewon, et al. "Factscore: Fine-grained atomic evaluation of factual precision in long form text generation." arXiv preprint arXiv:2305.14251 (2023).