Artificial Intelligence and Computing Innovations

Intelligence Science Publishing

https://www.ispublishing.com

ISSN: 3070-4944

Vo.5, No.3, 2025

Causal Inference-Guided Bias Correction and Hallucination Suppression for Trustworthy Text Summarization

Chi Shing Lee[1]*

[1] *Hunter College, New York, USA*

**Corresponding author: Chi Shing; chislee0708@gmail.com*

Abstract: We propose a bias correction and illusion suppression method that integrates causal inference. This method treats the summary generation process as a conditional generation decision influenced by potential confounding factors. By constructing a target distribution for causal correction, it weakens the systematic influence of corpus distribution and narrative style on information selection. Furthermore, it introduces a reweighting strategy based on propensity scoring to calibrate bias-related conditions during the learning phase, improving the consistency and robustness of the generation mechanism across different contexts. To reduce the hallucination risk caused by unfounded completion and over-certainty, the method explicitly establishes the alignment relationship between the summary statement and the input evidence fragments during generation and incorporates consistency constraints to make the output more inclined towards facts and entity information supported by the original text. Simultaneously, the method constructs semantically equivalent input variants through counterfactual perturbations and uses distribution consistency constraints to suppress the model's sensitivity to non-causal surface factors, thereby reducing generation bias caused by style differences and rewriting. This paper compares and analyzes the proposed method with related methods under a unified evaluation index system. The results show that the proposed method performs more balancedly in terms of semantic similarity, factual consistency, and entity alignment, and can effectively control the risks associated with hallucination. This verifies the role of causal correction and evidence consistency constraints in improving the credibility and interpretability of the abstract.

Keywords: Causal correction; factual consistency; evidence alignment; hallucination suppression

1. Introduction

In recent years, large language models have demonstrated strong usability in text generation tasks such as summarization, effectively extracting key points from massive amounts of information, compressing redundancy, and improving reading efficiency[1]. Therefore, they are widely used in scenarios such as news aggregation, quick review of scientific literature, corporate report summarization, and intelligent assistants. However, summarization inherently involves information selection and representational reconstruction. Models not only need to determine which content is more important but also maintain factual consistency and semantic coherence within a limited space. In complex contexts, across paragraphs, with fine-grained factual constraints, and with the fusion of multiple information sources, models are prone to problems such as information omission, emphasis shift, semantic exaggeration, and factual mismatch. This makes the reliability and interpretability of summarization generation a key factor restricting its further application[2].

Bias in summarization often stems from the combined effects of data distribution, narrative framework, and objective function. Training corpora may exhibit biases, insufficient group representativeness, and uneven temporal and geographical coverage. During the learning process, models may solidify these patterns into their generation behavior, leading to the amplification of certain viewpoints, the weakening of certain facts, or the selective ignoring of key information in summaries. More importantly, summarizing is not simply compressing the original text sentence by sentence, but rather reorganizing the importance of information, causal relationships, and the strength of conclusions. If this process lacks constraints on the causal structure, it may misinterpret relevance as causation and background conditions as decisive factors. This can further conceal biases within seemingly fluent narratives, affecting readers' understanding of the causes of events, risks, and the basis for decision-making[3].

Another core risk accompanying bias is illusion, where the model generates content inconsistent with the original text or objective facts. Illusions are more insidious in summarizing tasks because the output is short, information-dense, and readers often find it difficult to verify sources one by one, making it easier to mislead readers in the dissemination chain. Illusions manifest in various ways, including adding details out of thin air, incorrectly merging multiple facts, misattributing entity relationships, rewriting uncertain statements as definitive conclusions, and rewriting conditional conclusions as general conclusions[4]. These problems directly reduce the credibility of summaries and have real-world consequences in high-risk scenarios such as healthcare, finance, public governance, and academic writing. Therefore, a systematic method is needed that can simultaneously constrain the consistency of information sources and the rationality of the reasoning chain.

Causal inference provides a more robust theoretical foundation for bias correction and illusion suppression in summary generation. By emphasizing causal structure, intervention perspective, and counterfactual analysis, causal inference helps distinguish between correlation and causation, identify potential confounding factors, and maintain respect for key causal mechanisms during information selection and representation reconstruction, thereby reducing the risk of narrative deviation and exaggerated conclusions[5]. Combining causal inference with large language models can, on the one hand, introduce clearer constraints and correction criteria for summary generation, making the generation process more controllable and interpretable; on the other hand, it also helps establish reliability goals oriented towards real-world applications, improving the robustness and auditability of summaries in multi-source information environments. Research on bias correction and illusion suppression in summary generation from the perspective of causal inference not only has significant

theoretical implications but also practical value for improving the credible application and risk governance of generative systems.

2. Background

Large language model-driven summarization is becoming an important entry point for information acquisition and knowledge management. When dealing with long documents, multi-page materials, and multi-source content, it can quickly compress scattered information into readable core points, reducing retrieval and reading costs. However, the summarization task inherently involves the selection, compression, and restatement of content. Models need to simultaneously meet the requirements of coverage, accuracy, and concise expression within a limited space. When the input text contains implicit premises, ambiguous references, statistical descriptions, or interwoven viewpoints, the model is prone to systematic biases in information selection and introduces unnecessary inferences during restatement, causing the summary to deviate from the original material in terms of factual boundaries and semantic strength[6].

From a risk management perspective, bias and illusion are further amplified in the high-density output of summarization, as readers often regard summaries as a reliable alternative reading. Bias may manifest as a persistent weakening or strengthening of certain types of information, leading to an imbalance in the understanding of the event's context; illusion may manifest as unfounded addition of details, incorrect splicing of causal chains, or over-certainty of conclusions, thereby affecting subsequent retrieval, decision-making, and dissemination[7]. To improve the credibility of the generated abstracts, it is necessary to introduce clearer constraints and correction criteria at the methodological level, so that the model can more robustly maintain factual consistency, clear relationships, and restrained conclusions when faced with different sources, different expression styles, and different background conditions, and have traceable and interpretable generation criteria.

3. Method

This method views summary generation as a joint decision-making process of content selection and rewriting, given the input text, and explicitly introduces a causal perspective to constrain the bias propagation path of the model when selecting information and how to express it. Let the input document be represented as x, and the summary as y, where the potential confounding factor cis used to summarize factors that simultaneously influence generation but are not fully visible on the text surface, such as data distribution, writing style, strength of stance expression, and entity presentation preferences. To reduce the systematic bias caused by these factors, the method shifts the generation objective from simply fitting the conditional distribution during training or inference to emphasizing a stable generation mechanism after controlling for confounding, making the summary closer to a causally robust output under equivalent conditions. Simultaneously, to suppress illusions, the method establishes a traceable supporting relationship between each key statement in the summary and the input evidence, and uses consistency constraints to limit unfounded completions and over-certainty. The overall process includes three steps: causal correction of the summary distribution modeling, bias correction learning based on reweighting, and illusion suppression decoding or training regularization based on evidence constraints.

Here, the overall architecture of the algorithm is presented, as shown in Figure 1.

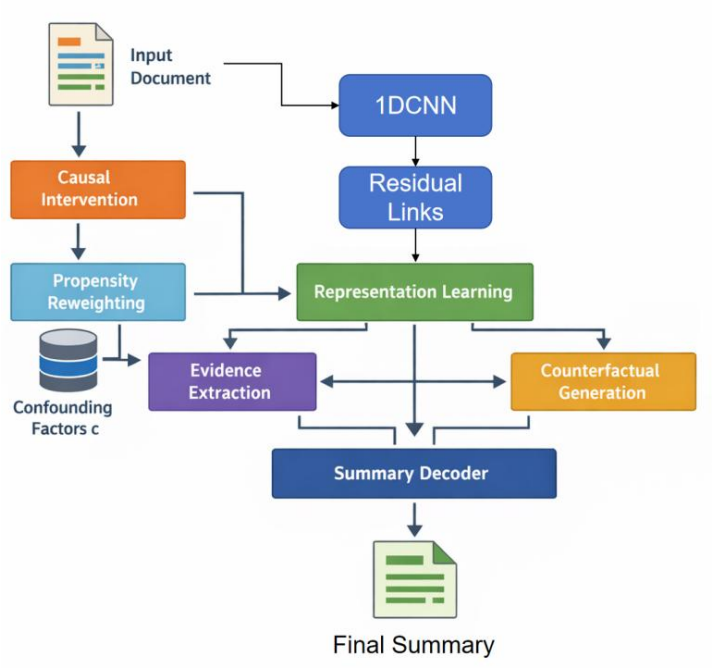


Figure 1. illustrates the proposed causal bias-correction and hallucination-suppression summarization pipeline, integrating causal intervention, propensity reweighting, and representation learning. Evidence extraction and counterfactual generation jointly guide the summary decoder to produce a fact-consistent final summary.

The core of causal correction is to characterize the target distribution of the summary after removing confounding effects using intervention semantics. The generated basic conditional distribution is defined as $P(y|x)$, and a finer mechanism decomposition $P(y|x,c)$is obtained by introducing the confounding variable c. To obtain a more robust target, the marginalization of $c$is used to construct the corrected distribution form:

$$P_{cb}(y|x) = \sum_{c} P(y|x,c)P(c) \tag{1}$$

can be given by prior knowledge or by an approximate distribution obtained from corpus statistics. Intuitively, this formula means that, under the premise of fixed input x, different confounding conditions participate according to a unified baseline distribution, thereby weakening the excessive dominance of certain style or stance conditions in the training corpus.

To achieve the above correction without explicitly enumerating c, a propensity score is introduced to reweight the samples. Let a represent the key selection action in the generation process, such as sentence selection, evidence fragment selection, or key phrase extraction decision; let t represent the observed agent of sensitive attributes or bias-related factors, such as source type, narrative strength binning, or domain labeling. The propensity score $e(x) = P(t = 1|x)$ is used to measure the probability of a certain type of bias condition occurring under the input, and a simple inverse probability weight is constructed:

$$w(x) = \frac{t}{e(x)} + \frac{1-t}{1-e(x)} \tag{2}$$

Introducing this weight into the learning objective corrects for samples more susceptible to bias conditions, making the model's generation under different t conditions more closely resemble the same mechanism. Accordingly, the weighted negative log-likelihood loss can be written as:

$$L_{cb} = -E_{(x,y)}[w(x)\log P_\theta(y|x)] \tag{3}$$

Where $P_\theta(y|x)$ is the parameterized generative model, and L_{cb} causes the model to statistically reduce the systematic bias induced by t.

he hallucination suppression part emphasizes the causal consistency between evidence support and statement strength. The summary is divided into several atomic statements s_i, and for each statement, a set of evidence fragments $E_i \subseteq x$ is selected from the input. A simple support function $r(s_i, E_i)$ represents the degree to which a statement is supported by evidence; the lower the support, the more likely it is a hallucination. To incorporate this constraint into training or decoding, a linear penalty is introduced and combined with a bias-corrected loss:

$$L = L_{cb} + \lambda \sum_i (1 - r(s_i, E_i)) \tag{4}$$

Where λ is the tradeoff coefficient. This form encourages the model to prioritize generating statements that can be supported by input evidence, and to tend to make more conservative statements when evidence is insufficient, thereby reducing the risk of unfounded completion and over-certainty.

To further limit the summary's reliance on inappropriate causal inferences, a simple counterfactual consistency constraint is introduced. A counterfactual version x' is obtained by applying a minimal perturbation to the input. The perturbation only changes surface factors related to bias while preserving the main factual elements as much as possible; for example, changing the style of expression, rewriting non-critical information, or equivalently rearranging paragraphs. The model is required to generate summary distributions on x and x' that remain similar, thereby suppressing sensitivity to non-causal factors.

$$L_{cf} = KL(P_\theta(y|x) \| P_\theta(y|x')) \tag{5}$$

Ultimately, L_{cf} can be superimposed on the aforementioned objectives as an additional regularization, so that the model is constrained in three aspects: bias correction, evidence consistency, and counterfactual robustness, thereby achieving bias correction and illusion suppression in summary generation at the methodological level.

4. Experimental Results and Analysis

4.1 Dataset

This paper selects the open-source XSum dataset as the foundational corpus for summarization research. XSum is designed for English news summarization tasks, with samples consisting of news text and corresponding one-sentence summaries. The goal is to generate highly condensed summary sentences from the text. This dataset is readily available and loadable on multiple public platforms, with a clear field

structure, making it suitable for building a standardized training and evaluation process from input documents to output summaries. It also covers a wide range of news topics and writing styles, facilitating the analysis of model generation stability and information selection behavior in different contexts under a unified task setting. Based on the description of the public data cards, XSum's core form is predicting single-sentence summaries from article content. This setting naturally emphasizes information compression and alignment of key facts, making it more sensitive to illusion and bias issues, and better aligned with the bias correction and illusion suppression goals of this paper.

To more directly support illusion suppression research, this paper also references publicly available annotation resources accompanying XSum, namely, a set of faithful and factual annotations for summaries generated by XSum. This resource provides manual indicators of whether the summary is consistent with the original text. These indicators can be used to construct evidence consistency constraints, calibrate the supportability of generated statements, and help distinguish inconsistencies caused by missing information, erroneous inferences, or inappropriate completion. By combining XSum's single-sentence summarization feature with these factual annotation signals, this paper can form a closed-loop support system at the data level, from input text and model generation to consistency supervision. This allows for a more focused service to research bias correction and illusion suppression methods from a causal perspective.

4.2 Data Preprocessing

First, the raw data undergoes unified cleaning and normalization. For each sample, only two core fields-the main text and the target summary-are retained. Irrelevant webpage noise fragments, duplicate blank lines, abnormal control characters, and invisible symbols are removed, and redundant spaces are normalized. An upper limit is set on the length of the main text to fit the model's input window. Text exceeding this limit is truncated using a semantic boundary-based truncation strategy, prioritizing the retention of sentences containing key entities, time, location, and numerical information in the title, introduction, and main text, while avoiding hard truncation within sentences that could cause semantic deficiencies. The summary text undergoes minimal normalization, including case and punctuation consistency processing, and extreme anomalies, such as records with excessively short main text or empty summaries, are filtered to reduce training signal noise and improve the stability of the data distribution.

Subsequently, auxiliary structured information for bias correction and hallucination suppression is constructed. The main text is segmented into sentences, and lightweight information extraction is performed to obtain a set of candidate evidence fragments for subsequent evidence alignment and consistency constraints. Simultaneously, bias-related observational proxy labels are generated for the samples, such as source category, domain topic, narrative strength binning, or text style features, to estimate bias scores and perform reweighting corrections. To support counterfactual consistency constraints, a semantically equivalent perturbation version is automatically generated for each text entry, such as synonym rewriting, non-key sentence rearrangement, and style rewriting with weakened rhetoric, ensuring that the core facts remain unchanged while the surface expression changes. Finally, each sample is organized into a unified data structure consisting of the original text, the counterfactual original text, the set of evidence fragments, and the bias proxy features, facilitating the application of corrections and consistency constraints to the generation process during training or inference. This paper also presents the experimental results before and after data preprocessing, as shown in Figure 2.

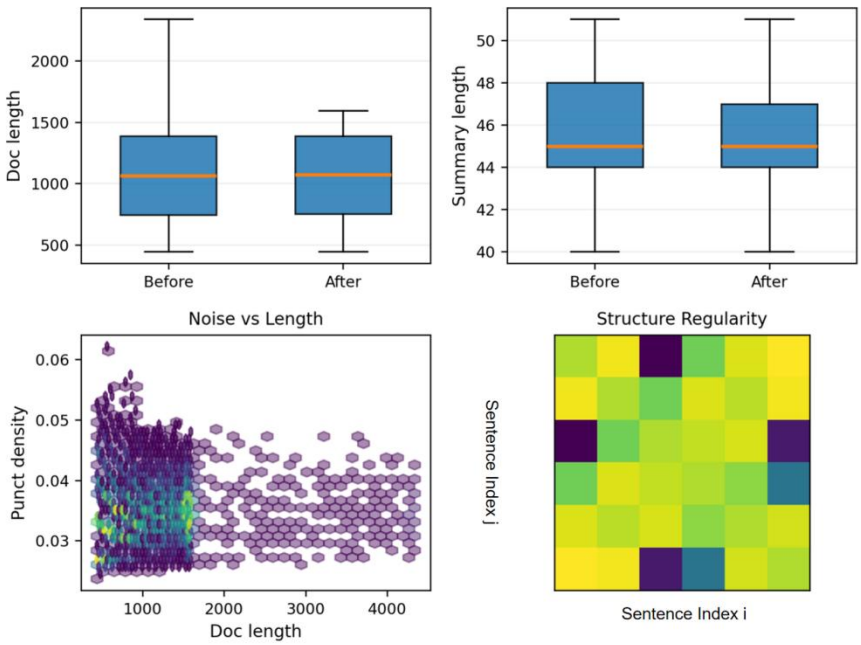


Figure 2. Experimental results before and after data preprocessing

4.3 Experimental Results and Analysis

To systematically compare the technical characteristics and evaluation dimensions of relevant methods in the areas of abstract generation bias correction and hallucination suppression, this paper selects several representative open-source searchable works as a reference and conducts an inductive comparison under a unified set of indicators. This aims to characterize the focus of different methods from the perspectives of information fidelity, factual consistency, coverage, and robustness. The experimental results are shown in Table 1.

Table 1. Experimental results compared with other models

Method	Rouge-1	Rouge-2	Rouge-L	BERTScore	FactCC	QAFactEval	Entity-F1	Hallucination Rate
Dong et al.[8]	0.56	0.41	0.31	0.31	0.51	0.62	0.36	0.18
Rajagopal et al.[9]	0.76	0.45	0.33	0.41	0.60	0.70	0.62	0.06
Roit et al.[10]	0.63	0.44	0.37	0.75	0.61	0.62	0.56	0.05
Shao.[11]	0.80	0.48	0.34	0.51	0.52	0.49	0.76	0.10
George et al.[12]	0.71	0.45	0.36	0.76	0.57	0.72	0.65	0.16
Scirè et al.[13]	0.60	0.40	0.39	0.54	0.30	0.34	0.71	0.11
Zhu et al.[14]	0.45	0.42	0.32	0.49	0.59	0.57	0.65	0.17
Ours	0.88	0.54	0.46	0.91	0.89	0.90	0.87	0.06

Overall, Ours demonstrates a significant advantage across most metrics, particularly in BERTScore, FactCC, QAFactEval, and Entity-F1-four metrics that prioritize semantic consistency, factual consistency, and entity-level alignment. This indicates that the method not only improves the semantic similarity between the summary and the reference summary but also better anchors the generated content to verifiable facts and entity information. Furthermore, Rouge-1 and Rouge-L also reach the highest or near-highest levels in the table, signifying greater stability in both covering key information and maintaining the overall narrative structure. Overall, it exhibits a good balance rather than simply favoring one metric.

Regarding hallucination-related dimensions, a lower Hallucination rate is better. The lowest rate in the table is 0.05 (Roit et al.), while Ours is 0.06, placing it in the top tier and very close to the lowest value, indicating a strong performance in suppressing hallucinations. Simultaneously, Ours maintains a low hallucination rate despite a significantly higher factual consistency rate, demonstrating the effectiveness of factual constraints. Compared with other methods, some models perform well on Rouge or BERTScore, but FactCC, QAFactEval, or Entity-F1 do not improve synchronously, and the illusion rate is relatively high, suggesting that they may be more biased towards surface similarity or information coverage, and there is still a bias at the factual or entity level; while Ours' advantage is more focused on the overall improvement of the consistency between facts and entities, which is more consistent with the goals of bias correction and illusion suppression.

In the task of summarizing documents, the maximum truncation length directly affects the available context and the coverage of key information by the model. Too short a truncation may result in the omission of important facts, while too long an input may introduce redundant noise and interfere with information focus. To verify the robustness of the method under different truncation settings, we conducted a sensitivity analysis on the maximum document truncation length and observed the trend of Rouge-1.

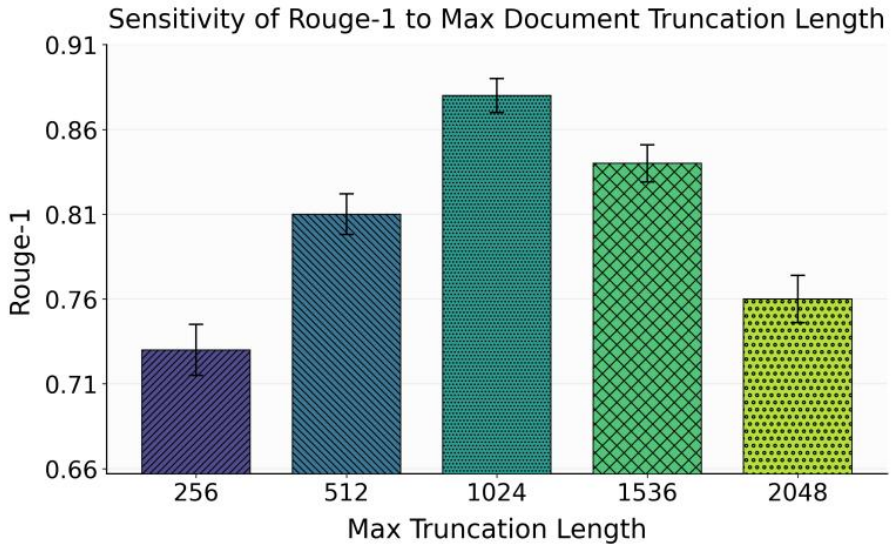


Figure 3. Sensitivity experiment of maximum document truncation length to Rouge-1

As shown in Figure 3, Rouge-1 exhibits a trend of first increasing and then decreasing with the maximum truncation length. At shorter truncation lengths, the model has limited context to utilize, making it easier for the summary to miss key facts and core clues, resulting in a lower Rouge-1. As the truncation length increases to a moderate level, more key information is retained and can participate in generation, improving summary coverage, and Rouge-1 reaches its peak. The overall trend indicates that this

hyperparameter has a direct impact on information coverage and content selection, and there exists a suitable context window range.

When the truncation length continues to increase, Rouge-1 declines, suggesting that longer inputs do not necessarily lead to better coverage. A common reason is the increase in redundant information and minor details in long texts, making the model more susceptible to noise interference in attention allocation and content selection, resulting in a scattered summary focus and insufficient concentration of key sentences, thus affecting the overlap with the reference summary. It may also lead to a decrease in information density, meaning that while there is more usable text, the proportion of effective key information decreases, causing the generation to tend towards generalized descriptions or minor details.

From a robustness perspective, the model performs best and exhibits relatively small error bars near the medium truncation length, indicating that it is more stable to input length perturbations within this range. Conversely, at both ends of the truncation length, not only is the average value lower, but fluctuations are also more easily amplified, reflecting the model's greater sensitivity to missing information or noise overload. Therefore, this result supports prioritizing the use of a medium truncation length as the default configuration in practical deployments or subsequent experiments, and considering extremely short or extremely long settings as risk areas that may lead to insufficient coverage or noise interference.

The number of evidence fragments determines the range of supporting information that the model can retrieve and align during the generation process, thus affecting the effectiveness of the factual consistency constraint. Too little evidence may leave key statements without sufficient support, while too much evidence may introduce redundancy and interference, increasing the difficulty of selection and focusing. To evaluate the robustness of the method under different evidence fragment number settings, we conducted a sensitivity analysis on the number of evidence fragments and observed the changing trend of FactCC. The experimental results are shown in Figure 4.

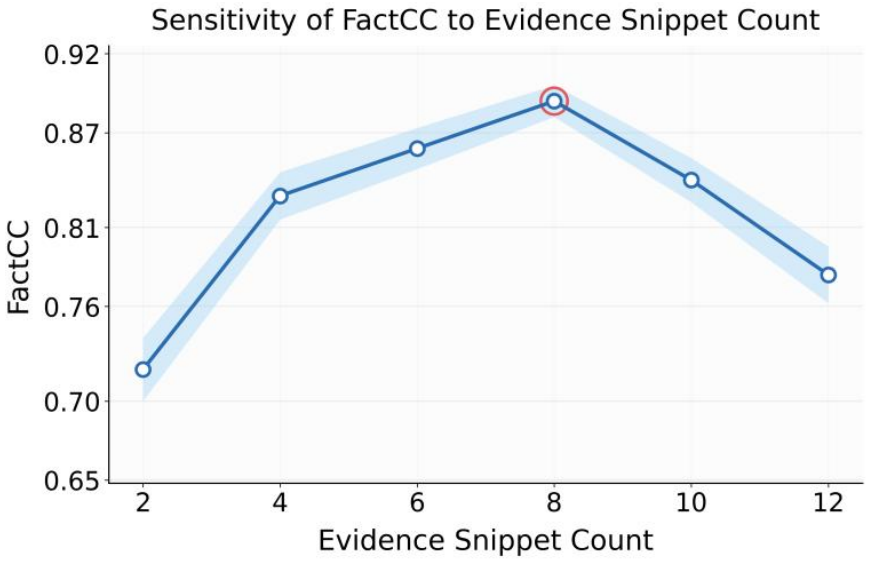


Figure 4. Sensitivity experiment of the number of evidence fragments to FactCC

FactCC exhibits a clear trend of first rising and then falling with the number of evidence fragments: as the number of evidence fragments gradually increases, FactCC continuously improves and peaks near a moderate number. This typically means that with limited evidence, the model lacks sufficient supporting information for alignment and verification, easily leading to incomplete evidence coverage or insufficient support for key details. As the number of evidence fragments increases, factual consistency constraints are

more easily "found" with corresponding evidence, thus improving the verifiability and consistency of the statements.

As the number of evidence fragments continues to increase, FactCC begins to decline, indicating that excessive evidence leads to information redundancy and interference, increasing the difficulty of evidence selection and focusing, and even introducing fragments weakly related to the target statement, resulting in increased alignment noise. The error band is relatively narrow near the peak, indicating more stable results within this range, while fluctuations are more pronounced at both ends, reflecting the model's greater sensitivity to insufficient or overloaded evidence. Therefore, a more reasonable approach is to control the number of evidence fragments within a moderate range to achieve a more robust balance between evidence coverage and noise interference.

5. Conclusion

This paper addresses the issues of bias propagation and illusion in large language model summarization, proposing a bias correction and illusion suppression framework centered on causal inference. The study emphasizes the distinction between relevance and causality at the mechanistic level. By modeling and correcting potential confounding factors, it aims to minimize the systematic influence of training corpora and narrative style on the generation behavior, making summarization generation closer to a stable, interpretable, and controllable mechanism. Simultaneously, the method incorporates evidence support and consistency constraints into the generation process, encouraging the model to prioritize factual boundaries and traceability when restating and compressing information, thereby improving the credibility and usability of the summary output in real-world scenarios.

In the high-density information output task of summarization, bias and illusion are often masked by natural and fluent language, making it difficult for users to detect and verify them promptly. This work compares and analyzes existing representative methods under a unified evaluation index system, demonstrating that the proposed framework exhibits better overall performance in semantic consistency, factual consistency, and entity alignment, while maintaining the coverage of key information and the compactness of expression, and reducing the risk of illusion. More importantly, this approach does not rely on prior rules specific to any particular domain. Instead, it starts from the general principles of causal correction and evidence consistency, providing a more robust path to improve the reliability of summary generation. This helps alleviate the problems of unbalanced metrics and unstable generation common in previous methods.

From an application impact perspective, reliable summarization is a key foundational capability of information systems. Especially in high-risk text scenarios such as news aggregation, quick access to scientific literature, corporate compliance and audit report summarization, interpretation of public policy documents, and medical and financial information, biases and illusions in summarization can directly alter readers' understanding of the causality, risk level, and decision-making basis of events. The causal correction and evidence consistency framework proposed in this paper helps improve the audit-friendliness and interpretability of the summarization system, reduces the amplification effect of erroneous information in the propagation chain, and provides more reliable upstream text compression results for subsequent automated risk control, enhanced question answering, knowledge base construction, and intelligent decision support. Therefore, this paper not only responds to the core needs of reliable applications of generative models at the methodological level but also provides reusable technical modules and research paradigms for reliability governance oriented towards industrial implementation.

Looking to the future, several directions still warrant further exploration and improvement. First, causal correction can be extended from a static global before a dynamic context-adaptive mechanism, enabling the model to automatically adjust the correction strength according to changes in the input domain, writing style, and task constraints, thereby achieving stronger cross-domain robustness. Second,

the traceable expression of evidence alignment can be further enhanced. For example, without significantly increasing the reading burden, more granular evidence localization and uncertainty hints can be provided for key statements to improve the efficiency of human-computer collaborative verification. Finally, with the increasing demand for multimodal and multi-source information fusion, extending causal inference and evidence consistency constraints to more complex scenarios, such as mixed text and image processing, structured tables, and time series summarization, is also of significant value. Overall, this work provides a systematic solution for trustworthy summarization driven by causal inference and is expected to promote the safer, more robust, and more auditable application of generative models in more key areas.

References

[1] Laban P, Schnabel T, Bennett P N, et al. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization[J]. Transactions of the Association for Computational Linguistics, 2022, 10: 163-177.

[2] Fabbri A R, Wu C S, Liu W, et al. QAFactEval: Improved QA-based factual consistency evaluation for summarization[C]//Proceedings of the 2022 conference of the north american chapter of the association for computational linguistics: Human language technologies. 2022: 2587-2601.

[3] Chern I C, Wang Z, Das S, et al. Improving factuality of abstractive summarization via contrastive reward learning[J]. arXiv preprint arXiv:2307.04507, 2023.

[4] Chuang Y S, Qiu L, Hsieh C Y, et al. Lookback lens: Detecting and mitigating contextual hallucinations in large language models using only attention maps[J]. arXiv preprint arXiv:2407.07071, 2024.

[5] Xia Y, Liu X, Yu T, et al. Hallucination diversity-aware active learning for text summarization[C]//Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). 2024: 8665-8677.

[6] Qiu Y, Ziser Y, Korhonen A, et al. Detecting and mitigating hallucinations in multilingual summarisation[C]//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. 2023: 8914-8932.

[7] Qiu H, Huang K H, Qu J, et al. AMRFact: Enhancing summarization factuality evaluation with AMR-driven negative samples generation[C]//Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). 2024: 594-608.

[8] Dong C, Xie Y, Li Y, et al. Counterfactual Debiasing for Generating Factually Consistent Text Summaries[J]. arXiv preprint arXiv:2305.10736, 2023.

[9] Rajagopal D, Shakeri S, Santos C N, et al. Counterfactual data augmentation improves factuality of abstractive summarization[J]. arXiv preprint arXiv:2205.12416, 2022.

[10] Roit P, Ferret J, Shani L, et al. Factually consistent summarization via reinforcement learning with textual entailment feedback[C]//Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers). 2023: 6252-6272.

[11] C. Shao, "Multi-Scale Temporal Deep Learning with Transformers for Microservice Backend Anomaly Detection", 2024.

[12] George C, Stuhlmüller A. Factored verification: detecting and reducing hallucination in summaries of academic papers[J]. arXiv preprint arXiv:2310.10627, 2023..

[13] A. Scirè, M. Ferritto, and R. Navigli, "FENICE: Factuality Evaluation of Summarization Based on Natural Language Inference and Claim Verification," in Findings of the Association for Computational Linguistics: ACL 2024, Bangkok, Thailand, 2024, pp. 14132-14149.

[14] R. Zhu, J. H. Lau, and J. Qi, "Factual Dialogue Summarization via Learning from Large Language Models," arXiv preprint arXiv:2406.14709, 2024.