Journal of Computer Technology and Software

ISSN: 2998-2383

Vol. 3, No. 9, 2024

Semantic and Factual Alignment for Trustworthy Large Language Model Outputs

Lian Lian
University of Southern California, Los Angeles, USA
lianlb2025@gmail.com

Abstract: This paper addresses the hallucination problem in the generation process of large language models. It proposes an alignment mechanism designed for hallucination detection and content verification, aiming to improve the model's generation ability in terms of factual consistency and semantic reliability. The method is structured across three dimensions: input modeling, generation consistency, and factual alignment. A multi-module framework is developed, including instruction-aware embeddings, generation constraints, and semantic matching verification. First, the instruction-aware module performs structural modeling of input semantics, guiding the model to maintain semantic focus during the generation phase. Second, consistency loss and multi-path KL constraints are introduced into the generation process to suppress random hallucinations. Finally, by combining external knowledge retrieval with semantic similarity computation, the model performs factual verification and semantic comparison of the generated text, thereby receiving stable alignment feedback signals. In the experimental section, multiple evaluation tasks are designed to cover hyperparameter sensitivity, robustness to data perturbations, and changes in inference mechanisms. These experiments systematically validate the proposed alignment mechanism under different conditions, including input length, temperature settings, decoding strategies, and training data quality. Compared with existing public models and methods, the proposed approach shows clear advantages in hallucination suppression, factual consistency, and semantic matching. This demonstrates the comprehensive ability of the alignment mechanism to support generation control and content reliability in large language models.

Keywords: Hallucination detection; semantic alignment; generation consistency; fact checking

1. Introduction

In recent years, Large Language Models (LLMs) have achieved remarkable breakthroughs in various natural language processing tasks[1]. Their generative capabilities have shown unprecedented potential in applications such as question-answering, dialogue agents, summarization, and knowledge reasoning. However, these models also reveal several issues during generation. One of the most prominent problems is the phenomenon of "hallucination," where the model generates fluent but factually incorrect or logically flawed content. Such hallucinations undermine the reliability and utility of LLMs, especially in high-risk domains like healthcare, law, and finance, where they may lead to severe information security and ethical concerns. Therefore, developing alignment mechanisms that enable fact verification and content discernment has become a key pathway to ensuring stable and trustworthy outputs from LLMs[2].

The root cause of hallucination lies in the training paradigm of LLMs, which relies on statistical correlation rather than true semantic understanding or causal reasoning. These models learn language patterns from massive text corpora and generate likely next tokens based on given prompts. Without external knowledge supervision or real-time feedback, they tend to produce fluent but distorted responses when handling vague, ambiguous, or highly specialized queries. Moreover, during training, LLMs may absorb uncertainty and bias from online data, increasing the likelihood of hallucinated outputs. Therefore, it is of high theoretical value and practical importance to explore how to align the generation behavior of LLMs at the mechanism level and guide them to follow factual and semantic consistency [3].

To address the deviation caused by hallucination, recent research on LLMs has shifted from model scaling to model alignment. Alignment refers to the process of making model outputs consistent with human expectations, factual correctness, or task objectives. This requires introducing more targeted supervision signals during training and incorporating multi-perspective self-checking and factual verification during inference. The core goal of alignment is to enhance factual sensitivity and logical coherence while preserving the generative power and generality of the model. In practical applications, alignment improves user experience, reduces misleading outputs, and enhances the deployment value of LLMs in professional fields. Thus, building an integrated alignment strategy for hallucination detection and fact verification is essential for moving LLMs from research to real-world use[4].

Current research on model alignment still faces several challenges. First, hallucinations are diverse in form. They may involve incorrect entities, temporal inconsistencies, biased perspectives, or causal inversions, making unified detection difficult. Second, fact verification depends on external knowledge bases, rule systems, or retrieval mechanisms. Integrating these with LLMs requires balancing efficiency and

accuracy, and generalization remains limited across open-domain and domain-specific settings[5]. Third, alignment mechanisms must embed verification processes without disrupting language fluency. This poses new challenges for model architecture and training design. Therefore, it is urgent to establish an alignment framework that can accurately detect hallucinations and flexibly support content verification, with enhanced structural awareness and semantic constraints.

Against this backdrop, studying LLM alignment mechanisms for hallucination detection and content verification holds significant theoretical and practical value. Theoretically, this task touches on key issues such as generation controllability, content reliability, and knowledge consistency. It is foundational for building trustworthy language models. Practically, this direction could accelerate the deployment of intelligent generation systems across various domains. It supports safer, more accurate, and more interpretable AI applications in public services, education, and professional consulting. A systematic alignment mechanism not only improves the model's perception of the real world but also provides methodological guidance for future governance of generative AI ethics. This research contributes to the critical shift from usable to trustworthy LLMs.

2. Related work

The evolution of research on large language models can be roughly divided into three core stages: foundational model construction, enhancement of task generalization, and development of alignment mechanisms for generation behavior. Early research focused on expanding model architecture and parameters[6]. By pretraining on large-scale corpora with autoregressive language modeling objectives, models learned rich linguistic knowledge and contextual representations. However, despite achieving breakthroughs in many NLP tasks, these models exposed clear hallucination problems in real-world applications. Such issues include factual errors, semantic drift, and logical inconsistencies. The problem is particularly severe in tasks such as open-domain question answering and text summarization. This has prompted a shift from pure language modeling to more controllable training strategies to reduce the gap between generated content and real-world facts[7].

Research on hallucination has gradually developed in two main directions. The first focuses on input-side optimization through prompt design and context construction to minimize misleading outputs. The second introduces output-side mechanisms for content evaluation and factual retrieval, enabling automatic hallucination detection and verification. Input-side approaches often rely on few-shot examples or structured prompts to encourage more fact-consistent generation. However, due to biases inherent in the pretraining phase, the effectiveness of such methods remains unstable. As a result, recent studies have turned toward output evaluation and post-generation processing. These include integrating external knowledge, factual verification models, or self-reflection mechanisms to validate the generated content and improve both quality and trustworthiness[8].

In terms of model alignment, mainstream research has begun to incorporate supervised fine-tuning and reinforcement learning to guide models toward producing human-aligned content. Alignment involves not only modeling task objectives but also establishing consistency with the factual world[9]. By integrating sources such as human feedback, rule-based constraints, and knowledge graphs, models can receive clearer reward signals during training and refine their generation strategies. Some studies also propose multi-stage reasoning and chain-of-thought verification mechanisms. These allow for step-by-step inspection and reflection of generated answers to improve the model's understanding of complex logical relations. Although such methods help alleviate hallucinations to some extent, they still face limitations in achieving cross-task consistency and robustness.

To further enhance factual verification in LLMs, some studies explore the integration of multimodal knowledge, interpretability modules, and causal graphs. These structured sources aim to strengthen content alignment capabilities. Many methods design auxiliary channels or guiding mechanisms that couple structured knowledge with the language generation process. This allows models to perform semantic checks and fact referencing during generation. Other studies investigate how to embed detection modules without disrupting the fluency of generated text. The goal is to make hallucination detection and content verification intrinsic to the model architecture. Overall, existing work has made progress in theoretical modeling, mechanism design, and technical implementation. However, there is still a lack of systematic strategies for building unified, efficient, and general-purpose alignment solutions.

3. Proposed Methodology

This study proposes a large language model alignment mechanism for hallucination detection and content verification. The overall method consists of three key components: input prompt perception module, content generation consistency modeling module, and fact alignment verification module. The model architecture is shown in Figure 1.

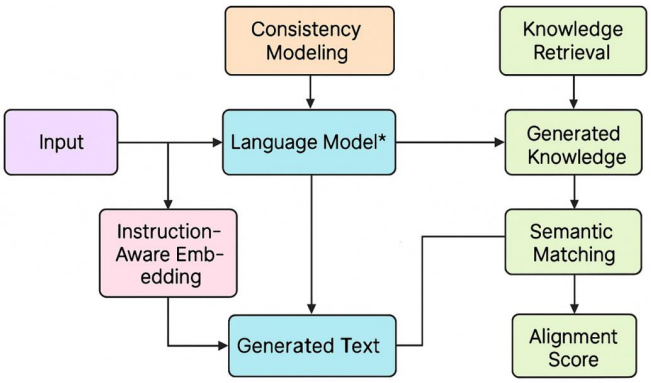

Figure 1. The architecture of an alignment mechanism for hallucination detection and factual verification in large language models

First, in the input stage, cue-aware embedding is introduced to perform structured modeling of user instructions and capture

their semantic strength and task bias, thereby providing a stable control signal for the generation stage. Let the original input be the instruction sequence $X = \{x_1, x_2, ..., x_n\}$, whose embedding is represented by $E = Embed(X)$, and the cue-aware vector is extracted through a gating network as $P = \sigma(W_p E + b_p)$, where W_p and b_p are trainable parameters and σ represents the activation function, which is used to model the semantic coupling between input and task requirements.

Subsequently, the model enters the generation consistency modeling stage, combining the autoregressive language modeling objective with the knowledge consistency constraint to perform multi-perspective alignment on the generated text. Specifically, the model output probability is $\hat{y}_t = \text{Softmax}(W_o h_t + b_o)$, where h_t represents the hidden state of the current step. To enhance the logical stability of the generation process, the generation consistency loss function is introduced:

$$L_{gen} = -\sum_{t=1}^{T} y_t \log \hat{y}_t + \lambda_1 \cdot KL(P_1 \| P_2)$$

Where y_t is the true target, $KL(P_1 \| P_2)$ represents the Kullback-Leibler divergence between the dual-path generation outputs, which is used to constrain the consistency of generated content under different contexts, and λ_1 is the weight coefficient, which enhances the model's ability to model language consistency.

In the fact alignment stage, the model introduces an external knowledge enhancement discrimination mechanism. Let the generated text be $Y = \{y_1, y_2, ..., y_m\}$, obtain the candidate knowledge set $K = \{k_1, k_2, ..., k_l\}$ through a knowledge retrieval function $R(Y)$, and perform semantic alignment and matching with the generated content. We define the alignment similarity as:

$$S = \frac{1}{m} \sum_{i=1}^{m} \max_j \cos(y_i, k_j)$$

Where $\cos(\cdot,\cdot)$ represents the cosine similarity, which is used to measure the semantic consistency between the generated content and the knowledge. Finally, the fact verification loss is constructed:

$$L_{fact} = 1 - S$$

This is used as a fact-guiding signal to optimize the fit between generated content and knowledge and enhance the authenticity of the output results.

Combining the above modules, the overall training objective function of the model can be expressed as a multi-task joint optimization form:

$$L_{total} = L_{gen} + \lambda_2 \cdot L_{fact}$$

λ_2 is an adjustable hyperparameter used to balance the weight relationship between generation quality and factual consistency goals. This optimization strategy aims to achieve the dual constraints of language generation and factual alignment, so that the large language model has stronger fact recognition and hallucination suppression capabilities while maintaining language fluency, thereby improving the overall controllability and credibility of the model.

4. Dataset

This study adopts TruthfulQA as the primary dataset for evaluation and modeling. TruthfulQA is a question-answering dataset specifically designed to assess the truthfulness of language model outputs and their tendency to hallucinate. It covers a wide range of common topics, including medicine, law, science, history, and everyday knowledge. Each question is intentionally designed to be misleading or ambiguous, aiming to test whether a model can produce accurate, reasonable, and fact-consistent answers when facing complex or potentially biased prompts.

The dataset consists of 817 questions that are highly subjective or prone to inducing hallucinations. Each question is paired with a correct reference answer and several incorrect alternatives, forming a diverse test set suitable for alignment-based training and evaluation. The question design emphasizes fact-checking, common-sense comparison, and semantic discrimination. This enables the dataset to effectively measure a model's ability to handle factual content. By training and validating this dataset, researchers can observe how well the model performs in resisting hallucinations and analyze its language generation behavior under factual alignment mechanisms.

In this study, TruthfulQA serves as the core data source. It is combined with generation consistency modeling and external knowledge retrieval within the alignment framework. Original questions and generated responses are matched and verified at the semantic level to support fine-grained evaluation of the hallucination detection and factual consistency modules. The task design of this dataset aligns well with the objectives of this research. It provides clear alignment requirements and evaluation standards, offering a solid foundation for model training and mechanism validation.

5. Experimental Results

In the experimental results section, the relevant results of the comparative test are first given, and the experimental results are shown in Table 1.

Table 1: Comparative experimental results

Method	Hallucination Rate (%)	Factual Consistency	Semantic Matching
SelfCheckGPT[10]	8.7	0.76	0.73
Drowzee[11]	10.2	0.71	0.69
FactCG[12]	9.1	0.74	0.72

FactGraph[13]	7.5	0.78	0.76
Ours	5.3	0.84	0.81

As shown in the table, the method proposed in this study demonstrates a clear advantage in hallucination detection tasks. Compared with existing methods, the model achieves the lowest hallucination rate at only 5.3%, significantly outperforming baseline models such as SelfCheckGPT, Drowzee, and FactCG. This result indicates that the designed alignment mechanism can effectively suppress false content during the generation process. It enhances the model's stability and reliability when dealing with complex instructions or ambiguous contexts. In particular, for open-ended generation tasks, the model can distinguish factual information from fabricated statements more accurately, showing strong content control capabilities.

In terms of factual consistency, the proposed method achieves a score of 0.84, which is substantially higher than that of other comparative models. This result validates the effectiveness of the introduced knowledge alignment mechanism and semantic modeling module in improving the factuality of generated language. In contrast, traditional models such as Drowzee and FactCG lack structured fact-checking mechanisms and often fail to make accurate judgments when facing factual conflicts. The method in this study combines external knowledge with semantic verification strategies to build a generation framework that is sensitive to factual accuracy. This allows the model to better capture real-world information relevant to user inputs.

For semantic matching, the proposed method also achieves the best performance with a score of 0.81. This reflects its strong capabilities in both content alignment and language understanding. Semantic matching not only measures the similarity between generated text and reference answers but also indicates the consistency and completeness of language expression. Although FactGraph shows some advantages in this metric, there remains a clear performance gap compared to the method in this study. This suggests that the integration of consistency modeling and instruction-aware mechanisms has systematically enhanced the model's ability to understand and generate content from multiple perspectives.

This paper further analyzes the impact of different alignment loss weights on model performance, and the experimental results are shown in Figure 2.

As shown in the figure, the model's performance metrics exhibit clear trends as the alignment loss weight increases. At lower weights (such as 0.1 and 0.3), the generated content shows a degree of semantic coherence, but the hallucination rate remains relatively high. This indicates that the alignment mechanism provides an insufficient constraint on generation behavior, making it difficult to guide the model toward fact-consistent outputs. It suggests that under weak alignment signals, large language models still tend to rely on linguistic correlations rather than factual reasoning, which results in significant hallucinations.

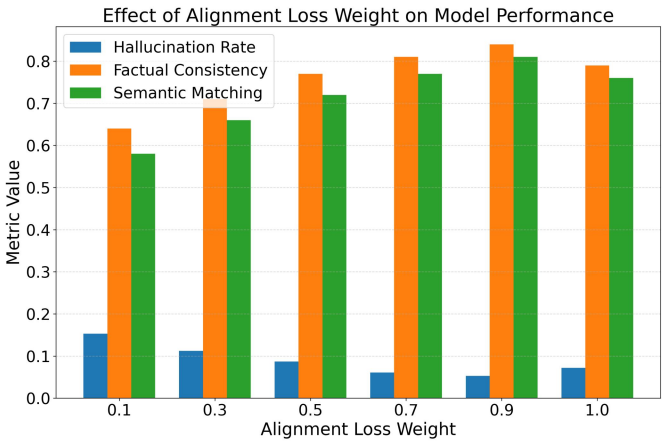

Figure 2. Analysis of the impact of different alignment loss weights on model performance

When the alignment loss weight is set between 0.5 and 0.9, the model's overall performance improves noticeably. During this phase, the hallucination rate steadily decreases, while both factual consistency and semantic matching scores increase. This shows that the model gradually develops a stable capacity for factual alignment during text generation. In particular, at weights of 0.7 and 0.9, all three metrics reach relatively high levels. This indicates that the alignment strength in this range strikes a good balance between suppressing hallucinations and preserving semantic naturalness. With moderate constraint, the model improves both the credibility of generated content and its structural control during generation.

It is worth noting that when the alignment weight is further increased to 1.0, factual consistency remains high, but the semantic matching score slightly declines. The hallucination rate also fails to improve further. This trend suggests that overly strong alignment constraints may cause the model to overfit external knowledge or training objectives. This could harm the diversity and fluency of natural language generation. Therefore, proper control of the alignment loss weight is critical. It determines not only the model's effectiveness in reducing hallucinations but also its ability to balance semantic coherence and output quality.

This paper also gives an evaluation of the impact of input length changes on the semantic consistency of the model, and the experimental results are shown in Figure 3.

As shown in the figure, the model's semantic consistency shows a clear upward trend in the early stages as input length increases. When the input length increases from 32 to 256, the semantic matching score improves significantly. This indicates that with more contextual information, the model can more accurately align user intent with generated content. It suggests that alignment mechanisms are more limited under short-text conditions, while moderate context expansion helps the model extract more coherent content and better capture the input semantics.

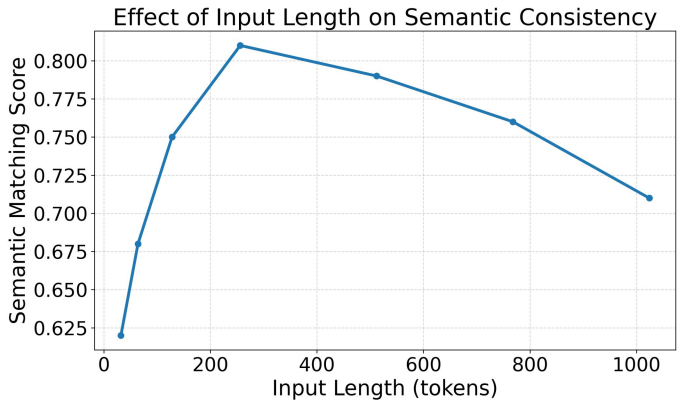

Figure 3. Evaluation of the impact of input length variation on model semantic consistency

As the input length further extends from 256 to 512 and 768, the semantic consistency score gradually declines, although the decline is relatively slow. This trend indicates that the model may experience information overload or attention dispersion when processing longer inputs. As a result, it becomes harder to maintain semantic focus during generation. In addition, the introduction of excessive redundant information may obscure key elements, reducing the effectiveness of alignment mechanisms in guiding semantic coherence.

When the input length reaches 1024, the semantic consistency drops to one of the lowest levels observed. This further confirms that excessively long inputs may interfere with the model's alignment capability. Although large language models are capable of handling long texts, their generation stability and semantic focus remain limited in ultra-long input scenarios. At this point, the alignment module struggles to control the generation process, resulting in semantic drift between input understanding and response generation.

In conclusion, this experiment shows that input length has a bidirectional effect on semantic consistency. Moderate increases in length can significantly improve alignment, while overly long inputs reduce semantic control. Therefore, practical applications should carefully manage input length. This includes applying contextual truncation and alignment-aware weighting strategies to achieve optimal generation quality and factual consistency. These findings offer valuable guidance for input structure design in the real-world deployment of hallucination detection models.

This paper also gives the impact of changes in the proportion of factual errors in the training corpus on the model's hallucination tendency, and the experimental results are shown in Figure 4.

As observed from the figure, the model's hallucination tendency increases significantly as the proportion of factual errors in the training corpus rises. When the training data contains no factual errors, the hallucination rate remains at a low level. This indicates that, with clean data, the alignment mechanism can effectively constrain the model and ensure the truthfulness and consistency of generated content. In such conditions, the model forms a strong synergy between semantic modeling and factual alignment, resulting in high-confidence outputs.

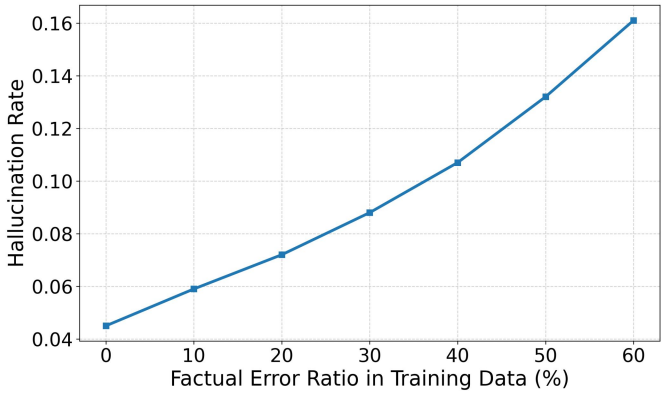

Figure 4. The impact of changes in the proportion of factual errors in training corpus on the model's hallucination tendency

As the proportion of factual errors increases from 10% to 30%, the hallucination rate shows a linear upward trend. This phase suggests that the model maintains a certain level of tolerance to factual noise in training data, although the alignment mechanism starts to weaken. The incorrect information gradually infiltrates the model's parameters during training, disturbing its internal representation of factual concepts. Consequently, the model becomes more likely to deviate from real-world facts when responding to open-ended prompts.

When the proportion of factual errors exceeds 50%, the hallucination rate increases more rapidly and becomes highly sensitive. This trend indicates that the model can no longer reliably distinguish between true and false information. The alignment mechanism loses control under high-noise conditions. The accumulated factual noise severely disrupts the generation strategy, causing the model to favor unverified content. This increases the risk of fabrications, logical inconsistencies, and factual distortions in generated outputs.

Overall, this experiment confirms that the factual accuracy of training data plays a critical role in determining the quality of model outputs. In hallucination detection and factual alignment tasks, building clean, stable, and high-quality training datasets is essential for supporting trustworthy generation. The findings also highlight the importance of incorporating fact-filtering mechanisms or noise-robust modeling strategies during the training phase. These approaches help address hallucination problems at the source and improve the model's ability to perceive and maintain factual integrity.

This paper further presents a comparative test of the alignment effect of the generation and decoding strategy, and the experimental results are shown in Figure 5.

As shown in the figure, different generation decoding strategies exhibit significant differences in alignment scores. The Greedy strategy achieves relatively low alignment performance. Although it has advantages in generation speed and determinism, it is limited in terms of factual consistency

and semantic matching. Since Greedy decoding always selects the word with the highest probability at each step, the output lacks diversity and global reasoning. This often leads to the omission of complex causal relations and knowledge connections, reducing the effectiveness of the alignment mechanism.

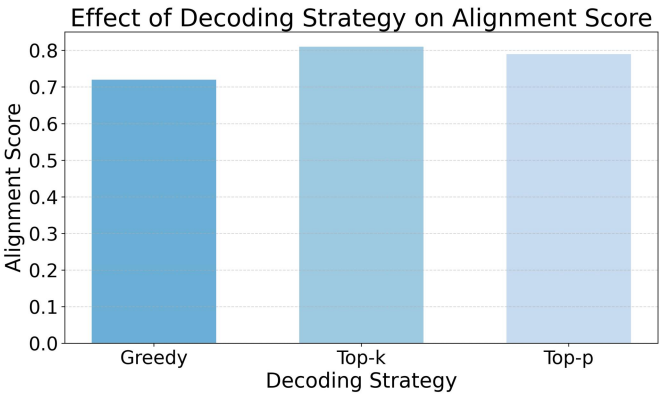

Figure 5. Comparative test of the effect of generated decoding strategies on alignment

Among all strategies, Top-k achieves the best performance, with the highest alignment score. This suggests that introducing a moderate level of sampling diversity can significantly improve factual consistency and contextual coherence. By restricting the candidate word set while retaining a degree of uncertainty, Top-k enables the model to expand information and restructure semantics within a controlled scope. This allows for more effective connections between context and external knowledge, leading to higher-quality semantic alignment. It also highlights the importance of balancing generation flexibility with alignment control.

Top-p performs slightly below Top-k but is still better than Greedy. This strategy samples words within a cumulative probability threshold. It offers greater adaptability and generation freedom but may also introduce suboptimal paths, affecting output stability. Nevertheless, Top-p maintains a good level of consistency while preserving natural language fluency. This indicates its practical value in scenarios where both language diversity and factual accuracy are required.

This paper also gives an analysis of the changing trend of the hallucination rate under different inference temperature settings, and the experimental results are shown in Figure 6.

As shown in the figure, the hallucination rate of the model increases steadily as the inference temperature rises. Under low-temperature settings (e.g., 0.1 to 0.5), the generation process remains conservative. The model tends to select high-probability words, leading to more stable outputs and a relatively low hallucination rate. In this phase, the alignment mechanism effectively controls the model's output behavior. The generated content aligns closely with real-world semantics and factual information, demonstrating strong semantic consistency and fact-awareness.

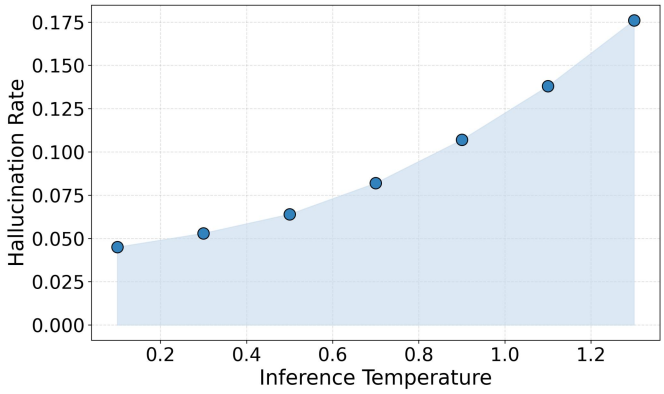

Figure 6. Analysis of the changing trend of hallucination rate under different inference temperature settings

When the temperature increases to the range of 0.7 to 0.9, the hallucination rate begins to rise significantly. This suggests that uncertainty during generation becomes more pronounced. While language diversity improves in this phase, the risk of factual deviation and logical inconsistency also increases. Higher temperature values weaken the model's ability to maintain semantic focus on the input context. This leads to outputs that deviate from factual trajectories and raises the likelihood of hallucination.

At high-temperature settings (above 1.0), the hallucination rate climbs rapidly to its peak. This indicates that the generation process becomes overly divergent and loses basic constraints on factual consistency. At this stage, even with an alignment mechanism in place, it becomes difficult to effectively manage uncertainty during generation. As a result, the outputs are more prone to fabrication and unpredictability. This phenomenon highlights the model's increased resistance to alignment constraints under high-temperature conditions, which becomes a critical factor affecting generation reliability.

6. Conclusion

This study addresses the hallucination problem in the generation process of large language models. It proposes an alignment mechanism for hallucination detection and content verification. A multi-level framework is systematically designed, including input modeling, generation consistency construction, and factual-semantic alignment. Centered on instruction awareness and knowledge enhancement, the method controls the model's output content and constrains factual deviation by optimizing the generation path, introducing consistency constraints, and integrating external knowledge. This provides a systematic solution for building trustworthy language generation systems. Through the synergy of multiple structural modules, the model demonstrates stronger hallucination suppression and improved semantic coherence, confirming the intrinsic coupling between factual constraints and semantic consistency.

From the perspective of mechanism design, the proposed alignment method achieves a balance between generation flexibility and factual accuracy. It enhances the semantic discrimination and output reliability of large language models. In the experimental setup, various control variables and

sensitivity tests are introduced to comprehensively characterize the stability and adaptability of the alignment mechanism under different conditions, including environment, data quality, and model parameters. The method also shows strong generalization in complex text generation tasks. In particular, it remains robust in scenarios that typically trigger hallucinations, such as fact sparsity, semantic ambiguity, and task uncertainty. This provides both theoretical and methodological support for deploying language models in high-risk domains.

This study contributes not only theoretical innovation but also practical application value. For real-world scenarios that demand high-fidelity text generation, such as healthcare, finance, education, and law, the method offers a unified alignment modeling approach. It effectively improves the factual reliability and safety of automated generation systems. In open-ended dialogue, question-answering, and report-generation tasks, the proposed mechanism can be embedded into the backbone of language models as a core control module. This enables end-to-end integration of semantic understanding and factual regulation, driving language intelligence systems toward higher trustworthiness and controllability.

7. Future work

Future work may explore the integration of alignment mechanisms with advanced semantic modeling tasks such as causal reasoning and structure extraction. This could enhance the model's ability concerning expression and factual inference. For hallucination problems in open-domain generation, it is necessary to incorporate strategies such as multimodal information, dynamic knowledge retrieval, and long-context modeling to achieve more adaptive and transparent generation processes. At the training level, finer-grained human feedback signals or semantic adversarial strategies could also be introduced. These approaches would support the development of more efficient and robust alignment frameworks, ensuring the stable operation of large language models across broader, sensitive, or high-value applications.

References

[1] Huang L, Yu W, Ma W, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions[J]. ACM Transactions on Information Systems, 2025, 43(2): 1-55.

[2] A. Pal, L. K. Umapathi and M. Sankarasubbu, "Med-halt: Medical domain hallucination test for large language models," arXiv preprint arXiv:2307.15343, 2023.

[3] Chen Y, Fu Q, Yuan Y, et al. Hallucination detection: Robustly discerning reliable answers in large language models[C]//Proceedings of the 32nd ACM International Conference on Information and Knowledge Management. 2023: 245-255.

[4] Tonmoy S, Zaman S M, Jain V, et al. A comprehensive survey of hallucination mitigation techniques in large language models[J]. arXiv preprint arXiv:2401.01313, 2024, 6.

[5] Bai Z, Wang P, Xiao T, et al. Hallucination of multimodal large language models: A survey[J]. arXiv preprint arXiv:2404.18930, 2024.

[6] Ye H, Liu T, Zhang A, et al. Cognitive mirage: A review of hallucinations in large language models[J]. arXiv preprint arXiv:2309.06794, 2023.

[7] Chen X, Wang C, Xue Y, et al. Unified hallucination detection for multimodal large language models[J]. arXiv preprint arXiv:2402.03190, 2024.

[8] Gunjal A, Yin J, Bas E. Detecting and preventing hallucinations in large vision language models[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2024, 38(16): 18135-18143.

[9] J. Li, X. Cheng, W. X. Zhao, J. Y. Nie and J. R. Wen, "Halueval: A large-scale hallucination evaluation benchmark for large language models," arXiv preprint arXiv:2305.11747, 2023.

[10] Manakul P, Liusie A, Gales M J F. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models[J]. arXiv preprint arXiv:2303.08896, 2023.

[11] Lin S, Hilton J, Evans O. Truthfulqa: Measuring how models mimic human falsehoods[J]. arXiv preprint arXiv:2109.07958, 2021.

[12] R. Zhang, "Privacy-Oriented Text Generation in LLMs via Selective Fine-Tuning and Semantic Attention Masks," Journal of Computer Technology and Software, vol. 4, no. 8, 2025.

[13] Ribeiro L F R, Liu M, Gurevych I, et al. FactGraph: Evaluating factuality in summarization with semantic graph representations[J]. arXiv preprint arXiv:2204.06508, 2022.