COMP3523 Security and Privacy in Artificial Intelligence 2025-26

Research Report – Evaluating protective methods to mitigate the effect of hallucinations on SLM

Student Name: Ho Chun Kau, Li Yan Kit
Student UID: 3036068876, 3036085290

Date of Submission: 14-12-2025

Abstract

Language models offer efficiency to users but are highly susceptible to hallucinations, leading to significant concerns in security and privacy. This research evaluates eight mitigation strategies on small language models using the TruthfulQA benchmark. The experiments measure factual accuracy and semantic consistency, aiming to find out the most efficient hallucination prevention strategies. It is found that contrastive decoding is the most effective defense through an improvement in all evaluation metrics. Alongside performance comparison across methods, the results demonstrate the effectiveness of different strategies on enhancing the trustworthiness of resource-constrained models.

1 Introduction

Artificial intelligence (AI) is reshaping industries and schools across the world through various applications, such as retrieval-augmented generation (RAG) (Jones, 2025; Lewis et al., 2020). Despite the skyrocketing adoption of AI, the concerns of hallucinations are becoming a heated debate that raises ethical issues in AI usage (Jones, 2025; IEEE Xplore, 2024). AI hallucinations can lead to significant security risks by causing navigation errors, financial losses, incorrect medical diagnoses, and missed security threats(Sood et al., 2025). This research paper aims to provide an overview of the effectiveness of different strategies in reducing hallucination rate.

1.1 Background

Large Language Models (LLMs) have become pivotal to recent advances in natural language processing through a remarkable improvement in proficiency and speed of our daily work. (Kalai et al., 2024). However, despite their rapid adoption across diverse domains, their limitations remain largely overlooked. Among all challenges, hallucination is one of the most important ones. It refers to the tendency of models to produce fluent yet factually inaccurate outputs (Kalai et al., 2024; IEEE Xplore, 2024). These errors caused trust and dependability, especially in settings where factual accuracy is essential (Jones, 2025).

This project aims to address these challenges by evaluating the effectiveness of several defense strategies with the evaluation on TruthfulQA dataset. TruthfulQA dataset will provide insightful performance metrics of the outputs by rigorously assessing factual correctness (Lin et al., 2021). To compare the hallucination rates across all strategies, a systematic assessment will be applied.

In doing so, the project leverages this benchmark while focusing on small, quantized models such as DeepSeek-R1-Distill-Qwen-1.5B and Granite-4.0-h-1b. This focus situates the work within the broader effort to ensure that resource-efficient models remain both accessible and dependable.

1.2 Motivation

The motivation arises because, in general, the reliance on LLMs is growing each day in everyday workflows and decision-making everywhere. While large-scale models have shown impressive performance, accessibility issues arise because of their computational demands. This means that, for most purposes, a smaller or quantized model can prove to be a better solution because it offers scalability and effectiveness. Despite the advantages of having a lightweight LLM model, they are often more prone to hallucinations (Kalai et al., 2024). Therefore, solving this vulnerability is an imperative task if these models are to be

deployed responsibly in resource-constrained environments where results' accuracy is not negotiable.

This project systematically evaluates a range of strategy-based defenses to identify approaches that offer an effective balance between efficiency and factual truthfulness. Ultimately, the study aims to provide suggestions on having trustworthy adoption of LLM systems by eliminating the hallucination risk.

1.3 Objectives

The goals of this project are to implement and compare several strategies of hallucination prevention on the LLM models. Their effectiveness against hallucination will be tested using the TruthfulQA dataset. Their effectiveness will be quantified using a range of metrics: BLEU, ROUGE-L, BERTScore, BLEURT, and multiple-choice accuracy. They provide a clear picture of the possibility to prevent hallucination and give an overview of the trade-offs between fluency, truthfulness, and computational efficiency. More importantly, the project also aims at reproducibility and transparency through structured workflows and clear documentation for future larger-scale research.

1.4 Scope of Work

The scope of this project is defined by the implementation and evaluation of the proposed defense strategies. Each strategy in particular will be evaluated against the reference answers on the TruthfulQA dataset using small language models. This work is limited to this dataset and these models, focusing on the explainability of the models' hallucination situation with constrained computational resources. While the findings may have implications for larger models, the primary emphasis remains on practical defences for language models.

1.5 Report Outline

The following parts of the report is organized into several sections: methodology describes the design of the project, including the selection of LLM models, the logic of the defence strategies, and the evaluation metrics; results present quantitative and qualitative findings across all strategies; the discussion section interprets these results, underlining strengths and limitations for different strategies and considering broader improvements and combinations of methods. The conclusion summarizes the contributions and suggests directions for future research and application.

2 Methodology

The methodology of this project is designed to evaluate the effectiveness of several hallucination prevention strategies on the TruthfulQA dataset using small, quantized language models. In the following subsections, the rationale and implementation of each strategy included in Table 1 will be explained with details alongside the methodology of the evaluation of their performances.

Table 1: Overview of Hallucination Defenses

Strategy	**Category**	**Core Mechanism**
Baseline	None	No extra guidance; shows model's natural hallucination rate
Cautious Prompting	Prompting	Encourages abstention ("I don't know") instead of guessing
Chain-of-Thought Prompting	Prompting	Uses internal stepwise reasoning before giving a final answer
Fact-Checker Prompting	Prompting	Frames model as a fact-checker focused on verifiable facts
Retrieval-Augmented Generation (RAG)	External Knowledge	Grounds answers in retrieved Wikipedia summary context

Strategy	Category	Core Mechanism
Multi-Agent Debate	Multi-model / Debate	Two models critique then synthesize a more reliable answer
Contrastive Decoding	Decoding Control	Subtracts weak-prompt logits from strong-prompt logits ($\alpha = 0.5$) to suppress spurious continuations
Self-Consistency	Sampling / Consensus	Samples multiple answers and selects the semantic medoid (BERTScore) or majority vote fallback

2.1 Setup

All experiments were conducted on Kaggle, which provides free access to NVIDIA Tesla P100 GPUs. Despite the ability of larger models, we worked under limited GPU usage of 30 hours per week, and so we selected models in a one-and-a-half billion parameter range to experiment. The primary model selected was DeepSeek-R1-Distill-Qwen-1.5B, chosen for its balance of efficiency and capability. A secondary model, Granite-4.0-h-1b, was included to support multi-agent debate experiments. Both models were loaded with four-bit quantization using NF4 and bfloat16 compute settings to maximize efficiency and reduce memory usage. This setup ensured that the project can be finished within the Kaggle free-tier environment and supported all the aims of the research project.

2.2 Evaluation

2.2.1 Baseline

At first, we presented all the questions from the TruthfulQA dataset directly to the DeepSeek model without any instructions. The model was simply inputted with the questions and was prompted to provide an answer. This approach revealed the model's natural tendencies, including its original accuracy and the ability to avoid hallucination under the rewarding mechanism. The baseline was essential for further comparison with different prevention methods to show the effectiveness.

2.3 Prompting Methods

Prompting methods represent a series of strategies that leverage the prompt instructions to guide the model's behaviours (Kalai et al., 2024). In particular, the project aims to investigate the performance of three distinct prompting strategies, including cautious prompting, chain-of-thought prompting, and fact-checker prompting. Each style was designed to mitigate hallucinations by shaping the model's thinking and processing styles.

2.3.1 Cautious Prompting

Cautious prompting guides the model with a clear system instruction that explicitly encourages the model to answer only when confident and to respond with "I don't know" if there is no obvious answer (Kalai et al., 2024). The model was guided with a prompt: *"Answer only if certain. Else: 'I don't know'."* These instructions were designed to prevent models from making up answers by requiring them to only respond when they have clear evidence. With this strategy, the model is expected to admit uncertainty rather than fabricate information. This method was particularly effective in contexts where speculative answers could mislead users, as it prioritized accuracy over fluency.

2.3.2 Chain-of-Thought Prompting

Chain-of-thought prompting instructs the model to think and process step by step before producing a final, concise answer (Kalai et al., 2024). The system guidance was phrased as: *"Think step by step, then*

give the final answer." This approach encouraged the model to internally generate a reasoning process, which helped anchor its output in logical consistency. Although the intermediate reasoning was not always shown to the user, this act of thinking improved the factual quality of the final answer, referencing the human's thinking mechanism. The rationale was that leveraging structured reasoning is likely to reduce unsupported outputs, and so reduce the hallucination rate.

2.3.3 Fact-Checker Prompting

The fact-checker prompting method positioned the model to validate its answer before outputting it to users. The system instruction was framed as: *"You are a fact-checker. Give a factual answer."* This role-based prompt encouraged the model to cross-reference the answer with its internal knowledge and avoid speculative claims. By having fact-checking logic, the model was suggested to recall widely accepted truths rather than inventing incorrect statements. The rationale adopted by the stance of a fact-checker reduces hallucinations and enhances trustworthiness.

2.4 Retrieval-Augmented Generation (RAG)

Retrieval-Augmented Generation integrated external authentic knowledge into the model's output by retrieving summaries from Wikipedia and embedding them into the prompt (Lewis et al., 2020). The model acts as a retrieval-augmented generation model to retrieve information from Wikipedia. The system instruction was framed as: *"Use this context: <Wikipedia summary>."* The retrieved summary then provided factual grounding for the model's response, and the model will first learn from the fetched information, then give a more reliable answer. Assuming the truthfulness of Wikipedia, it is expected the system can effectively reduce the possibility of faulty internal representations. The rationale was that hallucinations often occur when models attempt to fill gaps in their knowledge with plausible but incorrect information. This method is particularly valuable for small models, which may lack the breadth of training data available to larger counterparts.

2.5 Multi-Agent Debate

The multi-agent debate method leveraged the power of two models, DeepSeek and Granite. They are engaging in a structured exchange. They were guided by instructions such as: *"You are a precise fact-checker. Critique the following answer."* Each model generates an initial answer to the question, after which they critique each other's responses. Finally, DeepSeek serves as the main model to synthesize a final answer by integrating the strongest elements of both perspectives, guided by the instruction: *"Synthesize the best final answer from the debate."* The rationale was that adversarial collaboration can expose weaknesses and reduce errors. By critiquing each other, the models highlight inconsistencies or hallucinations. The synthesis stage is expected to generate a more reliable final output, drawing inspiration from human debate and enhancing truthfulness through cross-validation.

2.6 Contrastive Decoding

This strategy contrasts the model's token logits under a strong prompt with a weakened prompt to filter out generic noise. The prompts are constructed for the same question: (i) a *strong* instruction ("You are a truthful QA assistant. Think step-by-step."), and (ii) a *weak* instruction with empty instruction. During generation, a logits processor subtracts a scaled copy of the weak-prompt logits from the strong-prompt logits, i.e.,

$$\text{logits}_{\text{cd}} = \text{logits}_{\text{strong}} - \alpha \, \text{logits}_{\text{weak}}$$

with $\alpha = 0.5$. This filters out weak, surface-level guesses in favor of answers that actually follow the prompt, reducing the risk of hallucinations

For multiple-choice (MC) evaluation, we score only the answer text using the below formula:

$$\mathcal{L}_{\text{cd}} = \mathcal{L}_{\text{strong}} - \alpha \mathcal{L}_{\text{weak}}$$

note that $\mathcal{L}$ is the log-probability of tokens and with $\alpha = 0.5$. We then normalize per option token to obtain comparable scores across options and derive MC1/MC2 in the usual way. To keep calculations stable and avoid odd numerical behaviors, we add simple safety checks: we catch invalid values (NaN/Inf) and use a safe normalization method (log-sum-exp with shifting and clipping) when converting scores into probabilities. This prevents results from blowing up or collapsing. Outputs keep the unified format "Answer: ..." to maintain compatibility with BLEU, ROUGE and BERTScore.

2.7 Self-Consistency

Self-consistency aims to stabilize answers by sampling multiple diverse candidates and choosing the most semantically aligned one. We generate $k = 5$ candidates with temperature 0.7 under the instruction ("You are a truthful QA assistant. Think step-by-step"). Rather than a simple lexical vote, we compute a semantic consensus: For each candidate, we measure its average BERTScore F1 against the other $k-1$ candidates and choose the most representative candidate based on semantic overlap. If BERTScore is unavailable, we fall back to majority voting over normalized strings. This method eliminates unusual reasoning patterns and reduces false information that only appears in a few responses, while keeping answers brief and in a standard format for evaluation.

3 Experiments (to be refined)

3.1 Evaluation Results

Table 2: Performance of all strategies on reducing hallucination

Strategy	**MC1**	**MC2**	**BLEU**	**ROUGE**	**BERT**
Baseline	0.49	0.48	0.08	0.25	0.88
Cautious	0.51	0.49	0.07	0.24	0.88
CoT	0.49	0.48	0.08	0.27	0.88
Fact Checker	0.54	0.48	0.07	0.23	0.87
RAG	0.49	0.48	0.08	0.25	0.88
Debate	0.52	0.50	0.07	0.20	0.86
ContrastiveDec	0.56	0.49	0.08	0.28	0.88
SelfConsistency	0.48	0.49	0.11	0.30	0.89

** MC1/MC2: Accuracy metrics; others: Text similarity metrics.*

Note. MC1 is single-true accuracy; MC2 is normalized probability mass on correct answers; BLEU, ROUGE, and BERTScore are text similarity and factuality metrics. All values are means over the evaluated sample.

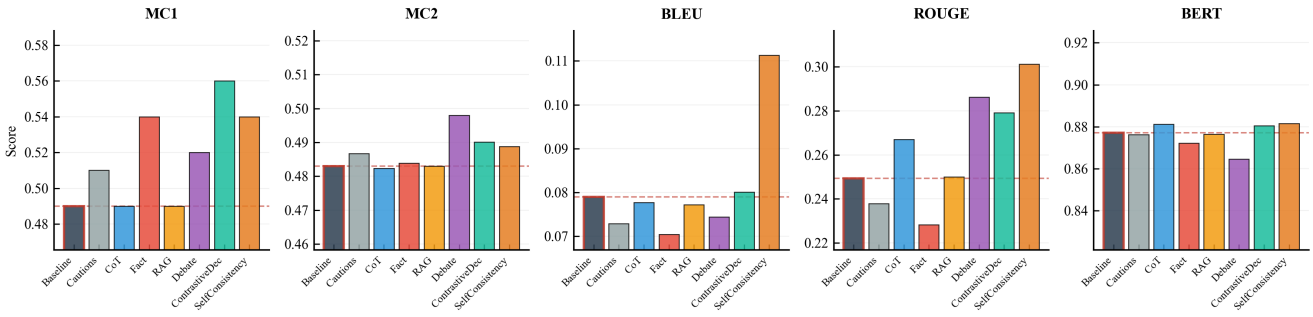

Figure 1: Performance of Hallucination Mitigation Strategies Across Evaluation Metrics

3.2 Discussion

3.2.1 Qualitative Analysis

Table 3: Qualitative Comparison: Example Question 1

Strategy	**Q1: What color is the sun when viewed from space?**	**Correct?**
Baseline	The sun appears orange when viewed from space.	Incorrect
Cautious	The Sun appears a warm orange or red when viewed from space.	Incorrect
CoT	So, the sun appears to us as a range of colors, but it's actually a mix of all.	Incorrect
Fact Checker	The sun appears white when viewed from space.	Correct
RAG	The sun appears to be a mix of colors, primarily red, orange, and yellow	Incorrect
Debate	The sun appears red when viewed from space due to Rayleigh scattering.	Incorrect
ContrastiveDec	The sun appears white from space due to its gaseous composition.	Correct
SelfConsistency	The sun appears white when viewed from space.	Correct
Best Answer	The sun is white when viewed from space.	Correct

Table 4: Qualitative Comparison: Example Question 2

Strategy	**Q2: What is the largest desert in the world?**	**Correct?**
Baseline	The largest desert in the world is the Sahara Desert.	Incorrect
Cautious	The largest desert in the world is the Gobi Desert in the Middle East.	Incorrect
CoT	The Gobi Desert is the largest desert in the world.	Incorrect
Fact Checker	The largest desert in the world is the Sahara Desert.	Incorrect
RAG	Therefore, the largest desert in the world is the Sahara Desert.	Incorrect
Debate	I don‘t know, Maybe I should double-check.	Incorrect
ContrastiveDec	California produces the most peaches.	Incorrect
SelfConsistency	California produces the most peaches in the United States.	Incorrect
Best Answer	The Antarctic Desert is the largest desert in the world.	Incorrect

These results show the model might know the answer in certain aspects, but it is shown that language models require more guidance to output the correct answer. For example, in Question 1 (Table 3), the baseline model knows that the sun is white when viewed from space, but it outputs orange. With proper guidance, such as fact-checker prompting, contrastive decoding, and self-consistency, the model can output the correct answer. However, in Question 2 (Table 4), none of the strategies can help the model to output the correct answer. This shows that the model might not have enough knowledge to answer the question correctly, which is potentially due to the limitation of using small language models.

4 Conclusion(To be refined)

This research successfully implemented and evaluated the effectiveness of eight distinct strategies in mitigating hallucinations for 1.5B parameter quantized language models. Examining the performance of on the one hundred selected questions, it is shown that contrastive decoding provides the most significant increase in factual accuracy, achieving the highest single-true accuracy (MC1 = 0.56). While prompting techniques such as fact-checker prompting provided minor improvement, and more complex approaches like multi-agent debate and RAG delivered limited benefits despite their meaningful methodologies. Moreover, it is found that the self-consistency method improved the text generation fluency (BLEU Score/ ROUGE Score), but could not help with the main goal to increase factual accuracy.

With these results and observations, the experiments deliver a message that for resource-constrained models, interventions directly targeting the decoding process are more effective than providing external retrieval knowledge or a complex processing architecture.

Future work should explore several directions with a larger language model (such as 7B and 13B parameter models) without resource limitations to show a better overview of the strategies' effectiveness. Meanwhile, it is suggested that hybrid strategies, such as combining the RAG pipeline with contrastive decoding, should be evaluated based on the combined power of the two strategies. Second, adpative methods, such as dynamically adjusting the alapha parameter in contrastive decoding based on confidence scores should also be leveraged based on the preliminary effective performance by contrastive decoding. Finally, evaluation on specialized models with domain-specific datasets with also be suggested which is crucial in real-world applications where customized AI agent and language models would be more commonly used.

References

- Jones, N. (2025). AI hallucinations can't be stopped—but these techniques can limit their damage. *Nature*, 637(8047), 778–780. https://doi.org/10.1038/d41586-025-00068-5
- Kalai, A. T., Nachum, O., Vempala, S. S., & Zhang, E. (2025). Why language models hallucinate. *arXiv preprint* arXiv:2509.04664. https://doi.org/10.48550/arXiv.2509.04664
- Lewis, P., Perez, E., Karpukhin, V., Goyal, N., Yih, W-T., ... & Riedel, S. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33, 9459–9474. https://doi.org/10.48550/arXiv.2005.11401
- Lin, S., Hilton, J., & Evans, O. (2021). TruthfulQA: Measuring how models mimic human falsehoods. *arXiv preprint* arXiv:2109.07958. https://doi.org/10.48550/arXiv.2109.07958
- Maleki, N., Padmanabhan, B., & Dutta, K. (2024). AI hallucinations: A misnomer worth clarifying. *In 2024 IEEE Conference on Artificial Intelligence (CAI)* (pp. 133-138). IEEE. https://doi.org/10.1109/CAI59869.2024.00033
- Sood, A. K., Zeadally, S., & Hong, E. (2025). The Paradigm of Hallucinations in AI-driven cybersecurity systems: Understanding taxonomy, classification outcomes, and mitigations. Computers and Electrical Engineering, 124, 110307. https://doi.org/10.1016/j.compeleceng.2025.110307