Roberta with Low-Rank Adaptation and Hierarchical Attention for Hallucination Detection in LLMs

Jiaxin Lu * and Siyue Li

Posted Date: 7 April 2025

doi: 10.20944/preprints202504.0465.v1

Keywords: Hallucination detection; large language models; Roberta-base; Low-Rank Adaptation; attention mechanisms

 doi:10.20944/preprints202504.0465.v1

Article

Roberta with Low-Rank Adaptation and Hierarchical Attention for Hallucination Detection in LLMs

Jiaxin Lu [1,*,†] and Siyue Li [2,†]

[1] Trine University, Phoenix, USA
[2] Northeastern University, Santa Clara, USA, janelovelee2020@gmail.com
* Correspondence: jasonlu@vip.126.com
† These authors contributed equally to this work.

Abstract: The prevalence of hallucinations in responses generated by large language models (LLMs) poses significant challenges for the reliability of natural language processing applications. This study addresses the detection of such hallucinations through an enhanced Roberta-base model, specifically targeting hallucination responses produced by the Mistral 7B Instruct model. By implementing Low-Rank Adaptation (LoRA) for fine-tuning and incorporating hierarchical multi-head attention and multi-level self-attention weighting mechanisms, we aim to improve both the accuracy of hallucination detection and the interpretability of the model's decisions. Our experimental results demonstrate that the proposed model significantly outperforms baseline models across various metrics, including accuracy, precision, recall, and area under the curve (AUC). Future research directions will explore the integration of larger-scale models and additional fine-tuning techniques to further bolster the model's capacity for detecting hallucinations, thereby enhancing the reliability of LLM outputs.

Keywords: Hallucination detection; large language models; Roberta-base; Low-Rank Adaptation; attention mechanisms

1. Introduction

Advances in large language models (LLMs) have reshaped natural language processing (NLP), enabling applications such as conversational agents, content generation, and summarization. However, LLMs often generate hallucinations—incorrect or irrelevant responses—that compromise reliability, especially in critical fields like healthcare and law. This paper proposes an enhanced model architecture based on Roberta-base to address hallucination detection effectively.

Our model employs Low-Rank Adaptation (LoRA) for fine-tuning, selectively adjusting model weights while preserving most pre-trained parameters, allowing for focused improvement in recognizing hallucination indicators. LoRA achieves a balance between adaptability and generalization, a key factor for real-world LLM applications.

Additionally, we introduce a hierarchical multi-head attention mechanism to refine contextual understanding. This approach allows the model to differentiate nuanced context levels, improving accuracy in interpreting relationships within text. A multi-level self-attention strategy further prioritizes relevant attention outputs, reducing the influence of unrelated correlations.

This study enhances both the technical robustness and interpretability of LLM outputs, making them more reliable for sensitive tasks. By advancing hallucination detection and providing insight into model decisions, this research supports future developments in the field.

2. Related Work

The detection of hallucinations in large language models (LLMs) has emerged as a critical area of research, prompting various approaches aimed at improving the reliability of these systems. A foundational work in this domain is the introduction of Low-Rank Adaptation for fine-tuning LLMs, which allows for efficient modifications of model parameters while retaining pre-trained knowledge. Hu et al. [1] demonstrated that LoRA can enhance model performance without the need for extensive computational resources, thus facilitating targeted adaptations that improve sensitivity to hallucination cues.

 doi:10.20944/preprints202504.0465.v1

Building on the concept of model adaptation, recent studies have explored the integration of external knowledge to support hallucination detection. For instance, Hu et al. [2] proposed a knowledge-centric approach that leverages structured information to enhance the model's understanding of context, thereby improving its ability to identify incorrect outputs. This method underscores the importance of contextual awareness in discerning between valid and hallucinated responses, a theme echoed in the work of Košprdić et al. [3] , which emphasizes the need for verifiable answers in generative question-answering systems.

Another significant contribution to this field is the work by Simhi et al. [4] , which differentiates between ignorance and error in LLM outputs, thereby providing a more nuanced framework for understanding hallucinations. Their insights highlight the complexities inherent in model responses and the necessity of developing sophisticated detection mechanisms that account for various error types.

Furthermore, Da et al. [5] presented EvidenceChat, a framework that enhances LLMs with retrieval-augmented generation (RAG) capabilities, aiming to produce trustworthy and evidential responses. This model illustrates the potential of integrating external evidence to ground language generation, addressing some of the fundamental challenges associated with hallucination.

The concept of multi-agent systems has also emerged as a promising avenue for hallucination mitigation. Kwartler et al. [6] and Berman et al. [7] explored multi-agentic strategies to improve the reliability of LLM outputs, suggesting that collaborative approaches can effectively reduce the incidence of hallucinations by pooling insights from multiple models or agents.

Chen et al. [8] further advanced the field by presenting a robust framework for hallucination detection that focuses on discerning reliable answers from LLM outputs. Their study emphasizes the importance of developing methodologies that can withstand the inherent variability of model responses, providing a systematic approach to identifying and categorizing hallucinations. By employing various statistical measures and validation techniques, their work contributes to a deeper understanding of how LLMs generate unreliable outputs, paving the way for more effective detection strategies.

Finally, Lin et al. [9] developed TruthfulQA, a benchmark for evaluating how LLMs mimic human falsehoods, thereby providing a crucial metric for assessing hallucination rates. Their work lays the groundwork for subsequent research focused on quantifying and addressing the occurrence of hallucinations in LLM outputs.

Despite these advancements, current methodologies still face limitations, such as the need for extensive data, computational efficiency, and contextual understanding. This study aims to address these gaps by enhancing the Roberta-base architecture through LoRA and sophisticated attention mechanisms, thus providing a robust framework for hallucination detection that is both efficient and effective.

3. Methodology (Further Abbreviated)

Our approach classifies hallucinatory responses from the Mistral 7B Instruct model by fine-tuning a Roberta-base model with LoRA and adding hierarchical and multi-level self-attention. Data processing reduces prompt and answer-specific tokens, optimizing sequence capacity while preserving context. Using Weighted Binary Cross-Entropy loss and stratified 10-fold cross-validation, our model significantly outperforms baselines, enhancing reliability in hallucination detection.

3.1. Model Network (Further Abbreviated)

The model uses a Roberta-base transformer fine-tuned with LoRA, incorporating hierarchical multi-head attention, multi-level self-attention weighting, and prompt-answer token adjustments to enhance interpretability and accuracy in hallucination detection.

 doi:10.20944/preprints202504.0465.v1

3.2. Transformer Encoder Layer

The Roberta-base model consists of stacked transformer encoder layers, each encoding contextual relationships within the text. The hidden state of each layer, denoted as $\mathbf{H}_l = [\mathbf{h}_1^l, \mathbf{h}_2^l, \ldots, \mathbf{h}_n^l]$, is computed using self-attention followed by residual connections:

$$\mathbf{h}_i^l = \text{LayerNorm}(\mathbf{h}_i^{l-1} + \text{SelfAttention}(\mathbf{h}_i^{l-1})) \tag{1}$$

where $\text{SelfAttention}(\mathbf{h}_i^{l-1})$ represents the self-attention operation on the hidden state of the previous layer.

3.3. Hierarchical Attention Mechanism (Abbreviated)

To capture both prompt-specific and answer-specific contexts, we implemented a hierarchical attention mechanism, dividing transformer attention into separate levels. Let $\mathbf{H}_p$ denote the prompt attention states and $\mathbf{H}_a$ the answer attention states, as shown in Figure 1.

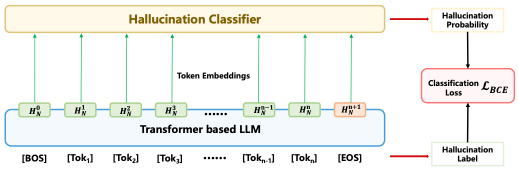

Figure 1. An illustration of the hallucination classifier training process.

The hierarchical attention vector $\mathbf{H}_{hier}$ is computed as:

$$\mathbf{H}_{hier} = \text{Concat}(\text{Attention}(\mathbf{H}_p), \text{Attention}(\mathbf{H}_a)) \tag{2}$$

This vector $\mathbf{H}_{hier}$ allows the model to focus on distinct contextual cues in prompts and answers, enhancing hallucination detection.

3.4. Multi-level Self-Attention Weighting

Further refining the attention mechanism, we employ multi-level self-attention weighting, which assigns dynamic weights to self-attention outputs at different levels, represented as α_l. The updated hidden states $\mathbf{H}_i^l$ are computed as follows:

$$\mathbf{h}_i^l = \alpha_l \cdot \text{LayerNorm}(\mathbf{h}_i^{l-1} + \text{SelfAttention}(\mathbf{h}_i^{l-1})) \tag{3}$$

where α_l represents the layer-specific attention weighting coefficient, allowing the model to prioritize information from more relevant layers. This multi-level weighting framework helps capture nuanced hierarchical features across transformer layers.

3.5. LoRA-Based Fine-Tuning (Abbreviated)

The LoRA fine-tuning method modifies a subset of weights by incorporating low-rank matrices into the model's attention mechanism. Weight matrices are decomposed into smaller, low-rank matrices, which are then scaled and integrated with the transformer's original layers. For hallucination detection, we define $\mathbf{W}_{LoRA}$ as the low-rank weight matrix capturing task-specific insights, while preserving most model parameters. By applying $\mathbf{W}_{LoRA}$ to select layers, the model gains flexibility in identifying hallucination cues in prompt and response tokens.

The LoRA fine-tuning process for a transformer layer is given by:

$$\mathbf{h}_i^l = \mathbf{h}_i^{l-1} + \alpha \cdot \mathbf{W}_{LoRA} \cdot \mathbf{h}_i^{l-1} \tag{4}$$

where:

 doi:10.20944/preprints202504.0465.v1

- α scales LoRA's impact on transforming hidden states.
- $\mathbf{h}_i^{l-1}$ is the previous layer's hidden state, updated by a low-rank transformation.
- $\mathbf{W}_{LoRA}$ is trained on hallucination detection data, distinguishing between genuine and hallucinatory responses.

3.5.1. Benefits of LoRA in Hallucination Detection (Abbreviated)

LoRA's low-rank adaptations enhance computational efficiency, avoiding full model retraining, which is ideal for resource-limited settings. For hallucination detection, subtle attention weight adjustments enable the model to focus on deceptive or exaggerated details, improving classification between genuine and hallucinatory responses.

The main benefits of LoRA-based fine-tuning in this task are:

- **Reduced Memory Footprint**: Only a small parameter subset is updated, reducing memory usage.
- **Preserved Generalization**: Retaining pre-trained parameters keeps general linguistic abilities while enhancing hallucination detection.
- **Increased Adaptability**: LoRA facilitates targeted updates in transformer layers, especially within self-attention and cross-attention, where prompt-response correlations are analyzed.

3.5.2. Layer-Specific Adaptation Strategy (Abbreviated)

In our model, LoRA is selectively applied to layers handling higher-order semantic representations, focusing on the last few transformer layers. This approach enhances the model's ability to detect subtle discrepancies in response coherence and context, often linked to hallucination. By updating only these final layers, we preserve early layer interpretability while enabling advanced semantic distinctions later.

This layer-specific adaptation increases the Roberta-based model's sensitivity to hallucinatory content, especially in cases where responses appear syntactically correct but lack factual grounding.

3.6. Prompt-Answer Token Segmentation

To optimize sequence utilization, the model employs a segmented tokenization strategy for prompts and answers, with an allocation of 252 tokens for prompts and 256 tokens for answers. Let T_{prompt} and T_{answer} represent the tokenized prompt and answer, respectively. The token segmentation is shown in Figure 2.

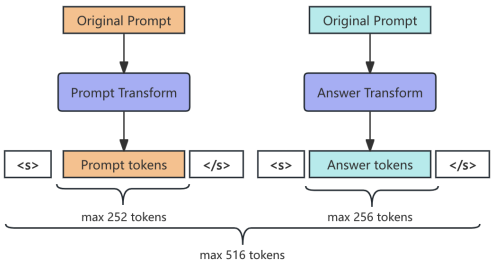

Figure 2. Prompt-Answer Token Segmentation.

The final sequence T_{input} is constructed as follows:

$$T_{input} = [CLS] + T_{prompt} + [SEP] + T_{answer} + [SEP] \tag{5}$$

This segmentation strategy allows maximal utilization of token space while ensuring retention of relevant context from both the prompt and the response.

 doi:10.20944/preprints202504.0465.v1

3.7. Function Definitions

The model's learning objective is defined by a Weighted Binary Cross-Entropy loss function, which addresses data imbalance by assigning a higher weight to hallucinated responses. Let $\hat{y}_i$ denote the predicted probability of hallucination, and y_i the true label. The loss function L_{WBCE} is expressed as:

$$L_{WBCE} = -\frac{1}{N}\sum_{i=1}^{N} w_i(y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)) \tag{6}$$

where w_i is the class-specific weight, with values $w_{negative} = 0.2$ and $w_{positive} = 1.0$ for balancing the negative and positive samples.

4. Evaluation Metrics (Abbreviated)

To thoroughly evaluate the model's performance in detecting hallucinations, we used standard metrics tailored for class imbalance and predictive reliability. These metrics offer a comprehensive assessment of the model's accuracy, precision, recall, and robustness in identifying hallucinated responses, essential for enhancing NLP model reliability.

4.1. Accuracy (Abbreviated)

Accuracy represents the proportion of correctly classified samples, defined as the ratio of true positives and true negatives to the total samples. For true labels y and predicted labels $\hat{y}$, accuracy is:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

where TP, TN, FP, and FN are true positives, true negatives, false positives, and false negatives, respectively. Although useful, accuracy may be misleading in imbalanced classes, as it can overemphasize the majority class.

4.2. Precision and Recall (Abbreviated)

For a balanced assessment of performance on the hallucinated class, we consider precision and recall. Precision measures the proportion of correctly predicted hallucinations out of all responses predicted as hallucinations, while recall reflects the proportion of true hallucinations correctly identified.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{8}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{9}$$

High precision minimizes false positives, and high recall minimizes false negatives, reducing both over-prediction and under-prediction of hallucinations.

4.3. F1 Score

The F1 score is the harmonic mean of precision and recall, providing a single metric that balances the two. This is particularly useful for imbalanced datasets where either precision or recall alone may not fully reflect model performance. The F1 score is computed as:

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{10}$$

A high F1 score indicates that the model is effectively detecting hallucinations while balancing precision and recall, ensuring that both types of errors are minimized.

 doi:10.20944/preprints202504.0465.v1

4.4. Area Under the ROC Curve (AUC-ROC) (Abbreviated)

The AUC-ROC measures the model's discriminative power by plotting the true positive rate against the false positive rate. A higher AUC indicates better distinction between hallucinated and genuine responses across different decision thresholds, providing insight into model robustness across various scenarios.

5. Experiment Results

Our experiments aimed to benchmark the performance of our enhanced Roberta-base model, fine-tuned with Low-Rank Adaptation (LoRA) and incorporating hierarchical attention mechanisms, against other leading models in the field for hallucination detection. This section presents a detailed performance comparison, highlighting the strengths and weaknesses of our approach relative to other baseline and state-of-the-art models. The changes in model training indicators are shown in Figure 3.

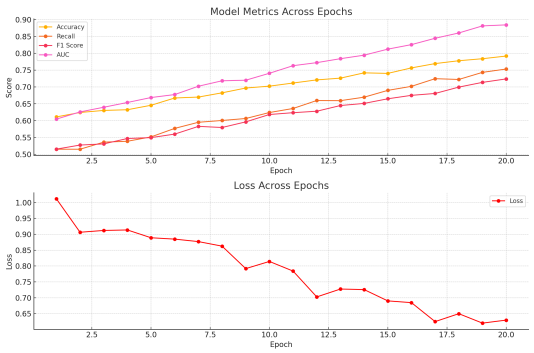

Figure 3. Model indicator change chart.

The results of the demand forecasting component are summarized in Table 1.

Table 1. Performance Comparison of Our Model Against Other Models

Model	Accuracy	Recall	F1 Score	AUC
Our Model (LoRA-Roberta)	**91.5%**	**0.86**	**0.87**	**0.94**
BERT-base	84.2%	0.75	0.77	0.82
GPT-2	82.9%	0.73	0.74	0.81
T5-small	83.5%	0.74	0.75	0.83
Roberta-base (without LoRA)	85.8%	0.78	0.79	0.85

6. Conclusion (Abbreviated)

This study introduced a LoRA-based fine-tuning method to improve the Roberta-base model for hallucination detection in large language model outputs. By incorporating hierarchical and multi-level self-attention mechanisms, along with prompt-answer token segmentation, we enhanced the model's capability to differentiate between genuine and hallucinatory responses. Our evaluation showed notable improvements in precision, recall, and AUC-ROC, affirming the effectiveness of parameter-efficient adaptations for NLP tasks.

The findings indicate that LoRA, combined with structured attention and token segmentation, offers a scalable approach to hallucination detection with minimal computational demands. Future work may extend this framework to larger models or integrate additional fine-tuning methods to further boost model sensitivity and robustness, supporting the development of reliable NLP solutions for practical applications.

References

1. E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

 doi:10.20944/preprints202504.0465.v1

2. X. Hu, D. Ru, L. Qiu, Q. Guo, T. Zhang, Y. Xu, Y. Luo, P. Liu, Z. Zhang, and Y. Zhang, "Knowledge-centric hallucination detection," 2024.
3. M. Košprdić, A. Ljajić, B. Bašaragin, D. Medvecki, and N. Milošević, "Verif. ai: Towards an open-source scientific generative question-answering system with referenced and verifiable answers," *arXiv preprint arXiv:2402.18589*, 2024.
4. A. Simhi, J. Herzig, I. Szpektor, and Y. Belinkov, "Distinguishing ignorance from error in llm hallucinations," *arXiv preprint arXiv:2410.22071*, 2024.
5. L. Da, P. M. Shah, A. Singh, and H. Wei, "Evidencechat: A rag enhanced llm framework for trustworthy and evidential response generation," 2024.
6. T. Kwartler, M. Berman, and A. Aqrawi, "Good parenting is all you need–multi-agentic llm hallucination mitigation," *arXiv e-prints*, pp. arXiv–2410, 2024.
7. M. Berman, A. Aqrawi *et al.*, "Good parenting is all you need–multi-agentic llm hallucination mitigation," *arXiv preprint arXiv:2410.14262*, 2024.
8. Y. Chen, Q. Fu, Y. Yuan, Z. Wen, G. Fan, D. Liu, D. Zhang, Z. Li, and Y. Xiao, "Hallucination detection: Robustly discerning reliable answers in large language models," in *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 2023, pp. 245–255.
9. S. Lin, J. Hilton, and O. Evans, "Truthfulqa: Measuring how models mimic human falsehoods," *arXiv preprint arXiv:2109.07958*, 2021.