LoRA-Based Fine-Tuning of Local LLMs for Hallucination Detection in Indonesian RAG Systems

I Ketut Resika Arthana[1], Nyoman Gunantara[2], Made Sudarma[3], Made Sukarsa[4]
Department of Informatics, Universitas Pendidikan Ganesha, Indonesia[1]
Faculty of Engineering, Universitas Udayana, Indonesia[2, 3, 4]

***Abstract*—Retrieval Augmented Generation (RAG) improves the factual grounding of Large Language Models (LLMs) by incorporating external knowledge. However, RAG systems may still generate hallucinated responses, and this issue remains underexplored in Indonesian language settings, particularly in settings where local deployment is preferred. This study proposes a hallucination detection approach for Indonesian RAG systems using Low Rank Adaptation (LoRA) fine-tuning. To support this objective, the study constructs a dataset in the Human-Computer Interaction domain consisting of 908 context, question, and answer pairs. The dataset is classified into four categories: FACT-H, FAITH-H, LOG-H, and FAITHFUL. Three local LLMs, namely, Gemma-7B-it, LlaMA-2-7B chat, and Phi-3-medium-4k-instruct, were evaluated using 5-fold cross-validation. The results show that Gemma-7B-it achieved the best performance in the four-class setting, with a Macro F1 score of 0.846. In the binary classification setting, Gemma achieved an accuracy of 98.1 per cent. Further analysis shows that Gemma was particularly effective in recognizing FAITHFUL, FAITH-H, and FACT-H, while LOG-H remained the most difficult class to distinguish consistently.**

***Keywords*—Hallucination detection; Retrieval-Augmented Generation; LoRA fine-tuning**

I. Introduction

The rapid advancement of LLMs has enabled systems that can understand and generate natural language across diverse contexts. Trained on large-scale text corpora, these models can produce coherent and contextually relevant responses [1]. Their applications have expanded across education [2], healthcare [3], [4], [5], software development [6], [7], and industry [8]. To improve factual grounding, many studies adopt RAG, which combines external knowledge retrieval with generative modeling to produce responses that are more closely tied to supporting information.

Despite this advantage, RAG-based systems remain vulnerable to hallucination. In this setting, hallucination refers to outputs that appear plausible but are not supported by factual or contextual evidence [9]. These errors may take the form of factual inaccuracies, fabricated content, or reasoning that is inconsistent with the retrieved context [10] [11]. This problem is particularly important in educational applications, where inaccurate responses may mislead students and reduce the reliability of AI-supported learning systems [12]. Therefore, improving retrieval quality alone is not sufficient. RAG-based educational systems also require a mechanism that can assess whether generated answers remain faithful to the provided context and are logically consistent with the question.

The importance of this problem is also reflected in our previous work on the development of a real-time RAG-based virtual assistant, RIVA, in Indonesian organizational settings [13]. That study showed that hybrid retrieval and semantic chunking improved contextual relevance and answer correctness. However, the error analysis also revealed that retrieval failures and partial context aggregation still produced hallucination-like inconsistencies, especially in structured and multi-chunk information retrieval. This finding is consistent with broader evidence that retrieval augmentation can reduce hallucination risk, yet retrieval optimization alone does not fully eliminate hallucination-related inconsistencies in RAG systems [14]. These results indicate that a dedicated hallucination detection layer is still needed, even when the retrieval component has been improved.

Although hallucination in RAG has received growing attention, the current literature still shows two important limitations. First, most hallucination detection studies focus on English language settings. Second, many existing approaches rely on computationally demanding models, as reflected in systems such as HalluMeasure [15] and [16]. While these studies report strong performance, their computational requirements limit their practicality in resource-constrained environments. This limitation becomes more pronounced in low-resource languages such as Indonesian, where both research attention and computational support remain limited. As a result, there is still a clear gap in the development of hallucination detection models that are both suitable for Indonesian RAG systems and relevant to local LLM-based RAG settings.

To address this gap, this study develops a hallucination detection approach for Indonesian RAG systems using LoRA. LoRA enables parameter efficient adaptation of pre trained models without requiring full retraining, making it suitable for local deployment settings. This study fine tunes three local LLMs, namely Gemma-7B-it, LlaMA-2-7B-Chat, and Phi-3-medium-4k-instruct, using a multi class hallucination detection dataset. The selection of these models is supported by prior research showing that LoRA based fine-tuning can improve adaptation quality, while reducing the need for full model retraining such as LlaMA-2-7B Chat [17] and Gemma-2-2B-it [18].

This work is conducted as part of the AVILA project, an intelligent RAG based learning platform that supports module preparation and question answering in the Human Computer Interaction domain. Within this context, the hallucination detection model is intended as a quality control component for

assessing faithfulness and factual accuracy before generated responses are delivered to users. To support this objective, the study constructs a dataset consisting of 908 context, question, and answer pairs derived from AVILA instructional materials. Each instance is labeled into one of four categories, namely FACT-H for factual error, FAITH-H for context omission, LOG-H for logical inconsistency, and FAITHFUL for contextually correct responses.

Although several recent studies have explored hallucination detection in multilingual or low-resource settings, their focus and design differ from the present work. For example, [15] investigated hallucination detection in machine translation across low- and high-resource languages, but did not address RAG-based hallucination detection or multi-class hallucination categorization. Similarly, [19] proposed synthetic data generation to improve hallucination detection, but their study was not designed around context-grounded RAG responses or an Indonesian domain-specific setting. In contrast, frameworks such as RaDIO [16] focus on real-time detection integrated with retrieval optimization, but do not provide a structured multi-class dataset tailored to a specific language domain. Therefore, this study differs by introducing a domain-specific Indonesian dataset that explicitly models multiple hallucination types within a RAG setting, while remaining compatible with fine-tuning approaches for local deployment.

This study makes three main contributions. First, it develops a multi-class hallucination detection dataset in Indonesian for the Human-Computer Interaction domain. Second, it implements a LoRA-based fine-tuning approach for hallucination detection using local LLMs in Indonesian RAG systems. Third, it provides a comparative evaluation of several local LLMs, namely Gemma, LLaMA, and Phi 3, under a unified experimental protocol. Through these contributions, this study aims to support the development of more reliable Indonesian RAG-based learning systems.

II. Related Works

Hallucination in RAG-based LLM outputs has become one of the key issues in generative AI research. According to the survey conducted by [20], LLMs can be categorized into two types of hallucination, namely factuality hallucination and faithfulness hallucination.

- Factuality hallucinations refer to cases in which the model produces answers that are factually incorrect or unverifiable. Such hallucinations often occur due to factual contradiction or fabrication.
- Faithfulness hallucinations occur when the model's output deviates from the given instruction, context, or is inconsistent with the logical structure of the response.

This taxonomy has informed a wide range of recent detection frameworks that assess the alignment between model outputs and their grounding knowledge through mechanisms such as fact verification, uncertainty estimation, and entailment-based classification.

Recent studies show a methodological transition from theoretical categorizations toward integrated hallucination detection systems. In [19], the authors proposed a synthetic data generation approach in which faithful and hallucinative pairs were constructed through controlled rewriting and used to fine tune T5 based detectors, resulting in higher accuracy than zero shot classifiers. In the RAG domain, [16] introduced RaDIO, a real time hallucination detection framework that dynamically adjusts document retrieval based on inconsistency signals detected in generated responses.

Although these advances represent important milestones, most prior works remain limited to English datasets and rely on large-scale computational resources, making them less applicable to resource-constrained or local-language environments, such as educational RAG systems in Indonesian.

To address this gap, the present study adapts Huang et al.'s taxonomy [20] into a more operational four-class scheme tailored for Indonesian RAG systems. The scheme was designed to distinguish factual error, grounding failure, and reasoning failure more explicitly during annotation and evaluation.

- FAITHFUL refers to responses that are fully consistent with the source context and the user's question. These responses do not contain factual error, contextual deviation, or logical contradiction. For example, when the context states that observation studies help reveal user habits and difficulties without direct instruction, a faithful answer explicitly identifies these benefits in line with the given material.
- FACT-H refers to factual hallucination, where the answer contains information that is false, fabricated, or contradicts verifiable facts in the provided context. This class captures factual contradiction at the content level. For example, if the context states that Christian Rohrer and McAfee define four testing categories, but the answer incorrectly states that there are three categories or replaces one category with another unsupported one, the response is labeled FACT-H.
- FAITH-H refers to faithfulness hallucination, where the answer may appear plausible or even generally correct, but is not properly grounded in the provided context or does not follow the specific informational scope required by the question. The dominant error in this class is lack of contextual support rather than factual falsity. For example, if the question asks about the benefits of observation studies and the answer provides a broadly relevant statement about observation in HCI without linking it to the contextual evidence given, the response is labeled FAITH-H.
- LOG-H refers to logical hallucination, where the answer remains topically related to the context but fails because of invalid reasoning, incomplete inference, contradiction, or mismatch between the asked aspect and the response content. In this class, the response may look factually acceptable on the surface, yet it does not logically satisfy the question. For example, when the question asks for the advantages of hybrid testing, an answer that merely restates that hybrid testing is flexible without explaining the intended inferential benefit is categorized as LOG-H.

The distinction between FAITH-H and LOG-H is especially important in this study. FAITH-H captures responses whose main problem is lack of grounding in the provided context, whereas LOG-H captures responses whose main problem lies in reasoning failure over contextually relevant information. This distinction was introduced to reduce ambiguity in annotation and to better reflect the types of hallucination observed in Indonesian educational RAG outputs.

III. METHODS

This study employed an experimental approach to develop and evaluate a multi-class hallucination detection model for Indonesian RAG systems. The methodological framework consists of three major stages, namely dataset development, model fine-tuning, and evaluation.

A. Dataset Development

As shown in Fig 1, the dataset construction process begins with source materials collected from HCI books. From these materials, 328 contextual segments were manually selected as the basis for response generation. For each context, four responses were then generated using AI to represent the target classes, namely FACT-H, FAITH-H, LOG-H, and FAITHFUL, resulting in 1,312 initial instances in total.

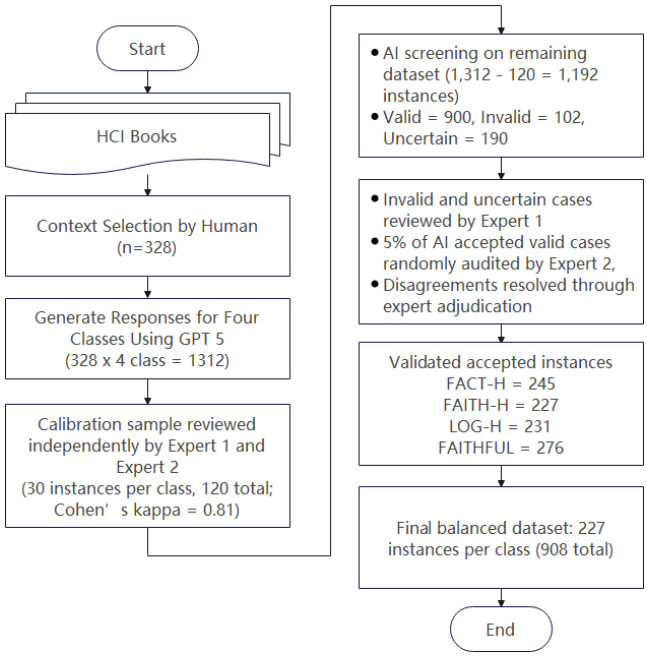

Fig. 1. Workflow of dataset construction, validation, and balancing for the four class hallucination detection dataset.

To improve the reliability of the annotation procedure, a calibration stage was first conducted before large scale screening. A stratified sample of 120 instances was selected from the initial dataset, with 30 instances per class. This subset was independently reviewed by Expert 1 and Expert 2 to examine label consistency, contextual appropriateness, and response quality. The agreement between the two experts was measured using Cohen's kappa, which reached 0.81, indicating substantial agreement. Based on this calibration stage, the validation guideline was refined before being applied to the remaining dataset.

After the guideline had been stabilized, the remaining 1,192 instances were screened using an AI based validator. Each instance was assigned one of three statuses: valid, invalid, or uncertain. The screening results produced 900 valid, 102 invalid, and 190 uncertain instances. All invalid and uncertain cases were then manually reviewed by Expert 1. In addition, 5% of the AI accepted valid cases were randomly audited by Expert 2 as a quality control step. When disagreements emerged during this process, the final decision was resolved through expert adjudication.

After the validation stage, the number of accepted instances was still not identical across classes. The validated pool consisted of 245 FACT-H, 227 FAITH-H, 231 LOG-H, and 276 FAITHFUL instances. To avoid class distribution bias during fine-tuning and evaluation, a separate balancing stage was then applied. The final dataset was constructed by retaining 227 validated instances per class, resulting in 908 balanced instances in total. Thus, instances excluded from the final dataset were not necessarily invalid. Some were removed during quality validation, while others were excluded during the balancing stage to preserve equal class representation.

B. Model Configuration

1) Model base: Three local generative LLMs with different architectural characteristics and scales were evaluated to examine hallucination detection performance under a consistent RAG setting. Table I summarizes the core specifications of each model, including parameter size, context length, and fine-tuning configurations.

TABLE I. MODEL SPECIFICATIONS USED IN THIS STUDY

Model Name	Specification	Quantization
Phi-3-medium-4k-instruct	Transformer (Decoder-only, Microsoft), 14B parameters, 4,096 tokens	4-bit (LoRA)
Gemma-7B-it	Transformer (Decoder-only, Google), 7B parameters, 8,192 tokens	4-bit (LoRA)
LLaMA-2-7B-chat	Transformer (Decoder-only, Meta AI), 7B parameters, 4,096 tokens	4-bit (LoRA)

The selected models represent a range of local generative LLMs with different architectural characteristics and capacities, enabling a comparative evaluation of hallucination detection within a consistent RAG setting. Phi-3-medium-4k-instruct was chosen for its instruction-following capability and stable reasoning behavior, making it suitable for analyzing context-grounded responses. The Gemma-7B-it provides multilingual support and extended context handling, which is beneficial for processing instructional materials in the Human Computer Interaction domain. LLaMA-2-7B-chat-hf serves as a conversational baseline to examine how chat-oriented models handle factual grounding and contextual consistency in generated responses.

2) Optimizer and hyperparameter training: The fine-tuning process was conducted using a standardized optimization and hyperparameter configuration across all large language models to ensure comparability of results. Table II summarizes the optimizer and key hyperparameters employed during training. These settings were selected based on prior empirical findings

in fine-tuning without full model retraining and were further adjusted through pilot experiments to achieve stable convergence within limited GPU resources.

TABLE II. OPTIMIZER AND HYPERPARAMETER CONFIGURATION

Parameter	Value
Optimizer	Paged AdamW 8-bit
Learning Rate	1×10^{-5}
Weight Decay	0.01
Warmup Ratio	0.1
Scheduler	Cosine
Epochs	3
Batch Size	1 (Gradient Accumulation = 32)
Quantization	4-bit (NF4)
Early Stopping	Patience = 2

The Paged AdamW 8-bit optimizer was employed due to its memory efficiency and compatibility with low-bit quantization, allowing large models such as Gemma, LLaMA, and PHI to be fine-tuned within a single GPU setup. A learning rate of 1×10^{-5} was chosen to balance convergence speed and stability during LoRA adaptation, while a weight decay of 0.01 was applied to prevent overfitting. The warmup ratio of 0.1 and cosine scheduler were implemented to gradually adjust the learning rate at the early training stage, promoting smoother optimization dynamics.

Training was performed for three epochs, which was empirically sufficient for convergence given the dataset size and instruction-following task. The batch size of one with gradient accumulation of 32 effectively simulated larger batch training without exceeding GPU memory limits. The 4-bit quantization (NF4) format reduced model memory consumption by approximately 70%, enabling efficient fine-tuning while maintaining representational precision. Finally, early stopping with patience = 2 was adopted to terminate training once the validation loss plateaued, ensuring an optimal balance between performance and computational efficiency.

3) LoRA configuration: LoRA was applied to all large language models to enable parameter-efficient fine-tuning without full model retraining. Table III summarizes the LoRA configuration parameters used in this study. The configuration was designed to balance adaptation capacity and computational efficiency while maintaining compatibility with 4-bit quantized models and single GPU environments.

TABLE III. LORA CONFIGURATION PARAMETERS

Parameter	Value
LoRA Rank (r)	16
LoRA Alpha (α)	32
LoRA Dropout	0.1
Target Modules	q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj
Bias	None
Task Type	Causal Language Modeling

The LoRA configuration was designed for fine-tuning without full model retraining in single-GPU environments. A rank of 16 was used to balance representational capacity and efficiency, while an alpha value of 32 provided stable gradient scaling without overshooting. A dropout rate of 0.1 introduced regularization to prevent overfitting on domain-specific hallucination data. Adapters were injected into key attention and feed-forward projection layers (q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj) to enable fine-grained adaptation without altering the entire model. The bias was omitted to reduce memory overhead, and the task type was set to causal language modeling to align training with generative objectives focused on producing faithful, contextually grounded responses.

4) Training configuration: A Stratified 5-Fold Cross-Validation was implemented to maintain class balance across training and validation splits. The training process used the paged_adamw_8bit optimizer and an early-stopping criterion with a patience of two epochs. All experiments were executed on an NVIDIA A40 GPU (48 GB VRAM) using the Transformers, PEFT, and BitsAndBytes libraries.

C. Evaluation and Performance Analysis

Model evaluation was conducted using Stratified 5-Fold Cross-Validation to ensure that each split preserved the class distribution of the dataset. Performance was first assessed in the four-class setting to compare the ability of the fine-tuned models to distinguish FACT-H, FAITH-H, LOG-H, and FAITHFUL responses. The main evaluation metrics were Accuracy, Macro F1, Matthews Correlation Coefficient (MCC), and Cohen's Kappa. Accuracy was used to measure the overall proportion of correct predictions, while Macro F1 was selected to reflect balanced classification performance across the four classes. MCC and Cohen's Kappa were included to provide a stricter assessment of prediction quality and agreement between predicted and true labels beyond simple accuracy.

For each model, the metric values were computed on every fold and then summarized using the mean and standard deviation. In addition, 95% confidence intervals were calculated to estimate the stability and uncertainty of model performance across folds. This analysis enabled not only an overall comparison of predictive quality among Gemma, LLaMA, and PHI, but also an examination of performance consistency under different data partitions. To further investigate stability, fold-level results of the best-performing model were analyzed separately in order to determine whether its overall superiority was maintained across individual evaluation splits.

Error analysis was then performed for the best-performing model using an aggregate confusion matrix across all folds. The confusion matrix was row-normalized to provide a class-level view of correct predictions and misclassification patterns among FACT-H, FAITH-H, LOG-H, and FAITHFUL. This analysis was complemented by qualitative inspection of representative misclassified instances, particularly in cases where logically inconsistent responses remained semantically related to the given context. Such examples were used to identify recurring error patterns that could not be fully captured by aggregate metrics alone.

In addition to the four class setting, a binary classification setting was also evaluated for the best performing model by merging FACT-H, FAITH-H, and LOG-H into a single hallucinated class and comparing it against FAITHFUL. This additional analysis was intended to assess the practical usefulness of the model as a front-end detector for separating hallucinated and faithful responses before applying finer-grained hallucination type analysis. In the binary setting, performance was reported using Accuracy, Precision, Recall, F1 Score, and MCC, together with a row-normalized binary confusion matrix to visualize class separation.

IV. Results and Discussion

A. Results

This section reports the experimental results in a stepwise manner to provide a clear interpretation of model behavior. First, an overall performance comparison across models is presented using the selected evaluation metrics. Second, performance stability across folds is examined to assess the consistency of each model. Third, the fold level results are reported to show how model performance varies across individual data splits. Fourth, error analysis and qualitative findings are provided to identify recurring prediction errors and characteristic response patterns. Finally, the section concludes with a summary of the main findings derived from the overall evaluation.

1) Overall performance comparison across models: As shown in Table IV, the overall classification performance of Gemma, LLaMA, and PHI was evaluated using Accuracy, Macro F1, MCC, and Cohen's Kappa. The results are reported as mean ± standard deviation and 95% confidence interval across folds to capture both average performance and variation. This overall comparison serves as the main basis for identifying the relative strengths of each model before proceeding to a more detailed metric-level analysis.

TABLE IV. Overall Classification Performance Comparison Across Models, Reported as Mean ± Standard Deviation and 95% Confidence Interval.

Model	Metric	mean ± SD	95% CI
Gemma	Accuracy	0.847 ± 0.031	[0.808, 0.886]
	Macro F1	0.846 ± 0.032	[0.806, 0.885]
	MCC	0.797 ± 0.041	[0.745, 0.848]
	Cohen's Kappa	0.796 ± 0.042	[0.744, 0.848]
LLaMA	Accuracy	0.589 ± 0.027	[0.556, 0.623]
	Macro F1	0.519 ± 0.022	[0.491, 0.546]
	MCC	0.487 ± 0.035	[0.444, 0.530]
	Cohen's Kappa	0.452 ± 0.036	[0.408, 0.497]
PHI	Accuracy	0.510 ± 0.007	[0.501, 0.518]
	Macro F1	0.444 ± 0.024	[0.415, 0.474]
	MCC	0.381 ± 0.004	[0.376, 0.387]
	Cohen's Kappa	0.346 ± 0.009	[0.335, 0.358]

Table IV shows a clear performance ranking across the three evaluated models. Gemma consistently achieved the best results on all reported metrics, followed by LLaMA, while PHI obtained the lowest scores overall. This pattern indicates that Gemma provided the strongest overall classification behavior under the current evaluation setting. The inclusion of standard deviation and 95% confidence interval also makes it possible to assess not only the average performance of each model, but also the stability and uncertainty of the reported results across folds.

a) Accuracy and Macro F1: In terms of Accuracy and Macro F1, Gemma outperformed the other models by a substantial margin. Gemma achieved an Accuracy of 0.847 ± 0.031 and a Macro F1 of 0.846 ± 0.032, with 95% confidence intervals of [0.808, 0.886] and [0.806, 0.885], respectively. These results indicate that Gemma maintained both high overall correctness and strong, balanced performance across classes. The relatively moderate standard deviations and fairly narrow confidence intervals suggest that this advantage was consistently observed across folds rather than being driven by a small number of favorable splits.

LLaMA showed intermediate performance, with an Accuracy of 0.589 ± 0.027 and a Macro F1 of 0.519 ± 0.022. Its 95% confidence intervals, [0.556, 0.623] for Accuracy and [0.491, 0.546] for Macro F1, remained clearly below those of Gemma, indicating a noticeable performance gap between the two models. Compared with Gemma, LLaMA also exhibited slightly lower variability in Macro F1, but this did not translate into stronger predictive performance. In other words, LLaMA was more limited by its average classification quality than by instability across folds.

PHI produced the weakest results on these two metrics, with an Accuracy of 0.510 ± 0.007 and a Macro F1 of 0.444 ± 0.024. Its Accuracy standard deviation was the smallest among all models, and the confidence interval [0.501, 0.518] was very narrow, which indicates highly consistent behavior across folds. However, this consistency occurred at a relatively low performance level. This pattern suggests that PHI behaved more uniformly, but its predictions remained substantially less accurate and less balanced across classes than those of Gemma and LLaMA.

b) MCC and Cohen's Kappa: A similar ranking was observed for MCC and Cohen's Kappa, which provide a stricter view of classification quality by accounting for the agreement between predictions and true labels beyond simple accuracy. Gemma again achieved the highest values, with an MCC of 0.797 ± 0.041 and a Cohen's Kappa of 0.796 ± 0.042. The corresponding 95% confidence intervals, [0.745, 0.848] and [0.744, 0.848], confirm that Gemma preserved a clear advantage over the other models even under these more demanding metrics. Although the standard deviations for MCC and Cohen's Kappa were slightly larger than those for Accuracy and Macro F1, the results still indicate a strong and stable classification pattern across folds.

LLaMA remained in the middle position, with an MCC of 0.487 ± 0.035 and a Cohen's Kappa of 0.452 ± 0.036. Its 95% confidence intervals, [0.444, 0.530] for MCC and [0.408, 0.497] for Cohen's Kappa, were substantially lower than those of Gemma. This result indicates that, even when accounting for label agreement and correlation between predictions and ground truth, LLaMA could not approach the classification quality

achieved by Gemma. At the same time, its moderate standard deviations suggest that the observed performance level was reasonably consistent across folds.

PHI again obtained the lowest values, reaching 0.381 ± 0.004 for MCC and 0.346 ± 0.009 for Cohen's Kappa, with very narrow confidence intervals of [0.376, 0.387] and [0.335, 0.358], respectively. These low deviations indicate that PHI produced highly stable results across folds, but the stability was centered around a weak level of predictive agreement. Taken together, the MCC and Cohen's Kappa results reinforce the conclusion that Gemma delivered the most reliable overall classification performance, while LLaMA offered moderate performance and PHI remained limited in both correlation strength and agreement quality.

2) Performance stability across folds: Table IV provides an initial comparison of the overall mean performance of the evaluated models, while Fig. 2 visually illustrates these results through grouped bars and error bars across the four evaluation metrics. Although both presentations are useful for identifying general performance differences, they do not fully capture how consistently each model behaved across different folds. Therefore, performance stability must be further examined using the reported standard deviations and 95% confidence intervals. In this context, lower standard deviation and narrower confidence intervals indicate greater consistency across folds, whereas larger deviations suggest higher sensitivity to data partitioning. This analysis is important because a model with higher average performance is not always the most stable, and a stable model is not necessarily the most accurate.

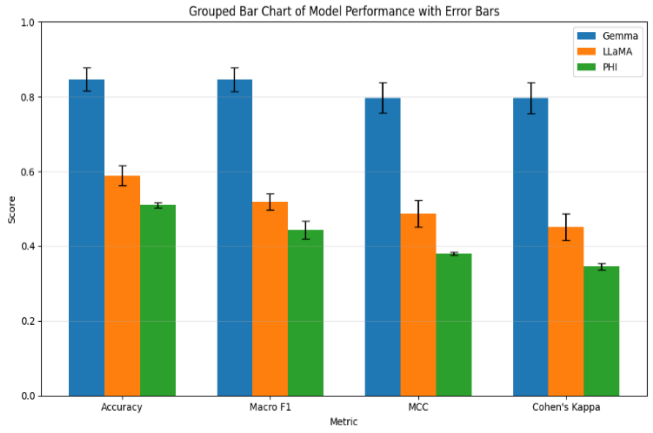

Fig. 2. Grouped bar chart of overall model performance across Accuracy, Macro F1, MCC, and Cohen's Kappa. Error bars represent standard deviation across folds.

Overall, Gemma achieved the highest mean scores across all metrics, but it also showed the largest variability among the three models. Its standard deviation ranged from 0.031 to 0.042, with confidence intervals of [0.808, 0.886] for Accuracy, [0.806, 0.885] for Macro F1, [0.745, 0.848] for MCC, and [0.744, 0.848] for Cohen's Kappa. These intervals remain relatively well bounded, which indicates that Gemma's performance was still reasonably stable across folds. However, compared with the other models, the wider spread suggests that Gemma was more affected by fold specific variation. In other words, Gemma combined the strongest average performance with a moderate level of variability rather than absolute stability.

LLaMA occupied an intermediate position not only in mean performance but also in stability. Its standard deviation values ranged from 0.022 to 0.036, which were generally lower than those of Gemma but higher than those of PHI on most metrics. The corresponding confidence intervals, such as [0.556, 0.623] for Accuracy and [0.408, 0.497] for Cohen's Kappa, indicate that LLaMA produced a more constrained range of outcomes than Gemma, although the model remained clearly below Gemma in average performance. This pattern suggests that LLaMA was moderately stable, but its limitation was not instability alone. Rather, its central performance level remained substantially lower.

PHI showed the strongest stability across folds when viewed from standard deviation and confidence interval width. Its Accuracy standard deviation was only 0.007 with a confidence interval of [0.501, 0.518], while its MCC standard deviation was even smaller at 0.004 with a confidence interval of [0.376, 0.387]. Similar patterns were observed for Macro F1 and Cohen's Kappa, both of which also had narrow confidence ranges. These results indicate that PHI behaved very consistently across different folds. However, this consistency was centered around the lowest mean scores among all evaluated models. Therefore, PHI can be considered the most stable model, but not the strongest model in terms of classification quality.

Taken together, the stability analysis reveals a clear tradeoff between performance level and consistency. Gemma remained the best performing model overall, and its wider deviations did not undermine its clear lead across all metrics. PHI, in contrast, was the most stable model, but its predictions were consistently weak. LLaMA fell between these two extremes, showing moderate performance with moderate variability. This result suggests that the best model in this study was not the one with the smallest dispersion, but the one that maintained the strongest overall scores while keeping its variability within a still acceptable range.

3) Fold-level performance of the best-performing model: Table V and Fig. 3 provide a more detailed view of Gemma's performance across individual folds. Overall, Gemma maintained high scores across all evaluation splits, which indicates that its superiority was not driven by only one favorable partition. The weakest performance appeared in Fold 2, where all four metrics decreased simultaneously. In this fold, Accuracy and Macro F1 dropped to 0.797 and 0.795, while MCC and Cohen's Kappa decreased to 0.730 and 0.729, respectively. This pattern shows that the decline was not limited to a single metric but reflected a general reduction in classification quality in that particular split.

In contrast, the strongest results were observed in Folds 1 and 4, where all metrics reached their highest or near-highest values. Fold 3 and Fold 5 remained close to the overall mean, indicating that the aggregate performance of Gemma reflects a stable pattern rather than isolated peaks. Taken together, these fold level results confirm that Gemma consistently delivered strong classification performance across different data splits,

despite moderate variation in one fold. This finding strengthens the conclusion that Gemma was not only the best performing model on average, but also a reliable model across the full evaluation setting.

TABLE V. FOLD LEVEL CLASSIFICATION PERFORMANCE OF GEMMA ACROSS ACCURACY, MACRO F1, MCC, AND COHEN'S KAPPA.

Fold	Accuracy	Macro F1	MCC	Cohen's Kappa
1	0.874	0.871	0.832	0.831
2	0.797	0.795	0.73	0.729
3	0.846	0.847	0.796	0.795
4	0.873	0.874	0.831	0.831
5	0.845	0.842	0.795	0.794

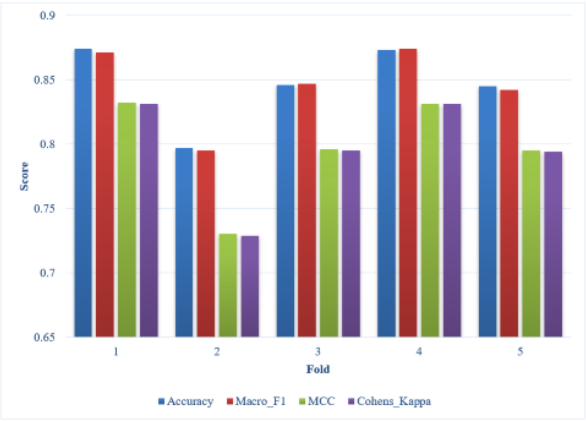

Fig. 3. Fold-level performance of Gemma across the four evaluation metrics.

4) Error analysis or qualitative findings: To further examine the error patterns of the best-performing model, Fig. 4 presents the aggregate confusion matrix of Gemma across all evaluation folds. This figure provides a class-level view of prediction behavior by showing how often each true class was correctly classified or confused with other classes in the overall evaluation setting.

		PREDICTED			
		FACT-H	FAITH-H	LOG-H	FAITHFUL
TRUE	FACT-H	0.815	0.048	0.119	0.018
	FAITH-H	0.04	0.894	0.066	0
	LOG-H	0.163	0.101	0.714	0.022
	FAITHFUL	0.004	0.004	0.026	0.965

Fig. 4. Aggregate confusion matrix of Gemma across all evaluation folds.

Overall, Fig. 4 shows that Gemma achieved strong class-level recognition, as reflected by the dominant diagonal values across all four classes. The highest diagonal value was obtained by the FAITHFUL class at 0.965, followed by FAITH-H at 0.894 and FACT-H at 0.815. These results indicate that Gemma was highly reliable in identifying faithful responses and also performed well in recognizing factual and faithfulness related hallucination classes. This pattern is consistent with the earlier summary results, where Gemma achieved high overall Accuracy and Macro F1 scores across folds.

The main source of error appeared in the LOG-H class, which obtained the lowest diagonal value at 0.714. A recurring confusion pattern can be seen in the misclassification of LOG-H as FACT-H at 0.163 and as FAITH-H at 0.101. In comparison, the other classes showed lower off-diagonal values, indicating better separation. This result suggests that logical hallucination was the most difficult category for Gemma to distinguish consistently, even though the model maintained strong overall performance. Therefore, the aggregate confusion matrix confirms that Gemma's main limitation lies in separating LOG-H from other non-faithful classes rather than in recognizing faithful answers.

As presented in Table VI, the misclassifications between LOG-H and FAITH-H indicate that the model often relied on topical or semantic relevance while failing to capture the logical requirement of the question. In several cases, the generated answers remained related to the given context, but they did not fully address the specific reasoning demand expressed in the question. This pattern is visible, for example, when the answer preserves the topic under discussion but omits the requested category details, shifts from benefit-oriented reasoning to procedural description, or produces a semantically related statement with a subtle contradiction in meaning. These examples further support the aggregate confusion matrix findings, which showed that LOG-H was the most difficult class for Gemma to distinguish consistently from other non-faithful categories, including FAITH-H.

5) Binary hallucinated vs. faithful performance of Gemma: To further assess the practical usefulness of the best performing model, an additional binary classification analysis was conducted by merging FACT-H, FAITH-H, and LOG-H into a single hallucinated class and comparing it against faithful. As reported in Table VII, Gemma achieved very strong binary classification performance, with an accuracy of 0.981, precision of 0.988, recall of 0.987, F1 score of 0.988, and MCC of 0.950. These values indicate that once the distinction among hallucination subtypes was removed, Gemma was able to separate hallucinated and faithful responses with very high reliability.

This result is also reflected in Fig. 5, which presents the row-normalized binary confusion matrix and shows a clear separation between hallucinated and faithful responses. In absolute terms, 672 out of 681 hallucinated responses were correctly identified, while only 9 were predicted as faithful. Similarly, 219 out of 227 faithful responses were correctly classified, with only 8 misclassified as hallucinated. Taken together, Table VII and Fig. 5 indicate that Gemma is highly effective as a first-stage detector for identifying whether a response contains hallucination or not. The substantially higher performance in the binary setting is expected, since the classification task becomes simpler after FACT-H, FAITH-H, and LOG-H are collapsed into a single class.

TABLE VI. EXAMPLES OF MISCLASSIFIED INSTANCES BETWEEN LOG-H AND FAITH-H CATEGORIES

Id	Context	Question	Model_Answer	Reason
56	Christian Rohrer and McAfee explained that test types can be divided into four main categories: scripted use, decontextualized use, natural or near-natural use, and hybrid. Each type has different characteristics depending on the level of control and context of product use.	What are the four main test categories explained by Christian Rohrer and McAfee?	Christian Rohrer stated that there are only four types of tests, but he did not mention the different categories.	The model ignored the logical mismatch, focusing only on surface similarity instead of missing category details.
172	Observation studies are also a common form of natural test. Researchers only observe users when using a product without giving direct instructions. This method is effective for revealing habits and difficulties that users are not aware of.	What are the benefits of using observation studies in HCI?	Observation studies are a method where researchers do not give direct instructions while observing users, so the activity is good and non-intrusive.	The model misclassified because the answer is semantically relevant but does not logically address the 'benefit' aspect asked in the question.
228	This section emphasizes that hybrid testing is a flexible and adaptive approach to understanding user interaction holistically. By combining controlled and natural methods, researchers can bridge the gap between quantitative and qualitative data.	What are the advantages of the hybrid testing approach in understanding user interaction?	Hybrid testing is a flexible but not entirely natural approach that helps analyze user interactions.	The model failed to detect a subtle contradiction in meaning and focused only on topical similarity.

TABLE VII. BINARY CLASSIFICATION PERFORMANCE OF GEMMA FOR HALLUCINATED VERSUS FAITHFUL RESPONSE DETECTION.

Metric	Score
Accuracy	0.981
F1 Score	0.988
Precision	0.988
Recall	0.987
MCC	0.95

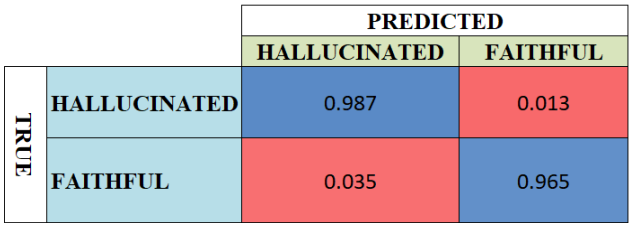

Fig. 5. Row-normalized binary confusion matrix of Gemma for hallucinated versus faithful classification.

6) Summary of the main findings: Overall, the results show a consistent ranking across the evaluated models, with Gemma achieving the strongest classification performance on all reported metrics, followed by LLaMA and PHI. Although PHI exhibited the highest stability across folds, its performance remained substantially lower than Gemma. In contrast, Gemma combined the highest mean scores with still acceptable variability, indicating that it offered the best balance between predictive strength and consistency under the current evaluation setting.

The fold-level analysis further confirmed that Gemma's superiority was not driven by only one favorable split, since the model maintained strong results across all folds despite a noticeable decrease in Fold 2. The aggregate confusion matrix also showed that Gemma recognized faithful, FAITH-H, and FACT-H classes well, while the main source of error remained in the LOG-H class. Taken together, these findings indicate that Gemma was the most reliable model overall, with its main limitation lying in the separation of logical hallucination from other non-faithful classes.

B. Discussion

The results indicate that Gemma provided the strongest overall performance among the evaluated models, consistently achieving the highest Accuracy, Macro F1, MCC, and Cohen's Kappa. This pattern suggests that Gemma was more effective not only in general correctness, but also in maintaining class-balanced prediction quality and label agreement across the four categories. This interpretation is compatible with prior work that treats hallucination as a multi form phenomenon rather than a single undifferentiated error type. In particular, [21] distinguishes factuality-based, faithfulness-based, logical-based, and hybrid hallucinations, which supports the present finding that overall performance can remain high while difficulty still varies substantially across hallucination subtypes.

However, the fold-based analysis adds an important qualification. Gemma was the best-performing model, but it was not the most stable one in absolute terms, since PHI showed narrower variation across folds despite much lower average performance. This result indicates that low variance alone is not sufficient as a selection criterion for hallucination detection, because a model may behave consistently while still failing to separate the target categories adequately. In contrast, Gemma combined the highest mean scores with still acceptable variability, making it the strongest candidate under the current evaluation setting. The fold-level results reinforce this interpretation, as Gemma maintained strong scores across all folds despite a noticeable drop in Fold 2.

The aggregate confusion matrix explains more clearly where Gemma succeeded and where its limitations remained. The model showed strong recognition for FAITHFUL, FAITH-H, and FACT-H, but LOG-H obtained the lowest diagonal value and was repeatedly confused with FACT-H and FAITH-H. This pattern suggests that logical hallucination is harder to detect because the response may remain semantically related to the context while still violating the inferential demand of the question. This interpretation is consistent with [21], who argue

that robust hallucination evaluation should consider both surface manifestations and their underlying causes, and with [16], who show that hallucination detection in dynamic RAG should not rely only on static or local cues, since contextual discrepancy, uncertainty, and semantic inconsistency may emerge during generation.

The qualitative examples further support this pattern. In several misclassified cases, the generated answer remained topically relevant but failed to satisfy the logical requirement of the question. The error did not arise because the answer was entirely unrelated to the context, but because it preserved semantic overlap while missing the requested reasoning target, omitting required distinctions, or introducing a subtle contradiction in meaning. This observation is aligned with [19], who note that effective hallucination detector training requires both faithful and hallucinated responses and that predefined perturbation assumptions may fail to capture the wider diversity of hallucinations found in real outputs. It is also compatible with [22], who combine semantic consistency analysis with entailment and contradiction assessment, indicating that semantic similarity alone is insufficient when the core error lies in reasoning inconsistency. In a similar direction, [23] argue that retrieval augmentation alone does not guarantee adherence to grounding and that stronger grounded alignment is needed to improve faithfulness. Taken together, these findings suggest that future improvement for LOG-H detection should focus less on broad semantic overlap and more on reasoning-sensitive verification, entailment-aware scoring, and grounded alignment mechanisms.

The binary setting highlights an important aspect of Gemma's behavior. As shown in Table VII and Fig. 5, Gemma achieved very strong performance when the task was simplified to distinguishing hallucinated responses from faithful ones, indicating that the model was highly reliable for coarse-grained hallucination detection. At the same time, the gap between the binary and four-class results shows that the main challenge does not lie in identifying whether a response is faithful, but in separating specific hallucination subtypes after a response has been recognized as non-faithful. This pattern is consistent with the earlier multiclass findings, where Gemma performed strongly overall but still showed greater difficulty on LOG-H. Although the binary class distribution becomes moderately imbalanced after merging hallucination subtypes, the high MCC and the small number of false positives and false negatives indicate that the strong result is not explained by class dominance alone. Therefore, Gemma appears well-suited as a coarse-grained screening model in RAG pipelines, while fine-grained discrimination among hallucination types remains the more demanding task.

V. Conclusion

This study evaluated LoRA-based fine-tuning of local LLMs for four class hallucination detection in Indonesian RAG outputs, covering FACT-H, FAITH-H, LOG-H, and FAITHFUL categories. The results show that Gemma consistently achieved the strongest overall performance across Accuracy, Macro F1, MCC, and Cohen's Kappa, surpassing LLaMA and PHI under the same experimental setting. This advantage was not limited to aggregate metrics. The fold-level analysis indicates that Gemma maintained a strong performance pattern across evaluation splits, despite a noticeable decline in one fold. The aggregate confusion matrix further shows that the model was particularly effective in recognizing FAITHFUL, FAITH-H, and FACT-H, while LOG-H remained the most difficult class to distinguish. Taken together, these findings indicate that LoRA based adaptation of local LLMs is a viable direction for hallucination detection in Indonesian RAG settings, but logical hallucination still poses a distinct challenge because semantically related answers may remain inconsistent at the level of reasoning.

Several limitations should be acknowledged. First, the dataset was constructed within the Human-Computer Interaction domain, so the present findings should be interpreted with caution when extended to other knowledge areas. Second, although the dataset is balanced, its size remains moderate and may not fully represent the diversity of hallucination patterns found in real RAG applications, particularly across domains, writing styles, and response forms. The use of 5-fold cross-validation improves the reliability of the evaluation, but it does not eliminate the need for validation on larger and more heterogeneous datasets. Third, the comparison was limited to selected local LLMs with different scales, and therefore does not capture the behavior of larger-scale architectures that may exhibit different strengths and weaknesses. Future work should address these limitations by expanding the dataset across domains, examining generalization more explicitly, and incorporating reasoning-aware or entailment-based methods to improve the detection of logical inconsistency, especially for the LOG-H class.

Acknowledgment

This study forms part of the author's doctoral research. The author expresses sincere appreciation to the Universitas Udayana for its academic guidance and support during the doctoral journey. The author also acknowledges the Universitas Pendidikan Ganesha for providing institutional facilities and research infrastructure that enabled the completion of this work.

Declaration on Generative AI

Generative artificial intelligence tools were used solely for language editing and grammatical refinement during the preparation of this manuscript. All scientific concepts, methodology, literature analysis, interpretations, and conclusions were developed and critically reviewed by the author, who assumes full responsibility for the content of this work.

References

[1] P. Omrani, A. Hosseini, K. Hooshanfar, Z. Ebrahimian, R. Toosi, and M. Ali Akhaee, "Hybrid Retrieval-Augmented Generation Approach for LLMs Query Response Enhancement," in 2024 10th International Conference on Web Research, ICWR 2024, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 22–26. doi: 10.1109/ICWR61162.2024.10533345.

[2] J. Meyer et al., "Using LLMs to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary students' text revision, motivation, and positive emotions," Computers and Education: Artificial Intelligence (Q1, SJR 1,7), vol. 6, p. 100199, Jun. 2024, doi: 10.1016/j.caeai.2023.100199.

[3] A. Suárez et al., "Beyond the Scalpel: Assessing ChatGPT's potential as an auxiliary intelligent virtual assistant in oral surgery," Comput. Struct. Biotechnol. J., vol. 24, pp. 46–52, Dec. 2024, doi: 10.1016/j.csbj.2023.11.058.

[4] H. Wang, C. Gao, C. Dantona, B. Hull, and J. Sun, "DRG-LLaMA : tuning LLaMA model to predict diagnosis-related group for hospitalized patients," NPJ Digit. Med., vol. 7, no. 1, p. 16, Jan. 2024, doi: 10.1038/s41746-023-00989-3.

[5] T. Li et al., "CancerGPT for few shot drug pair synergy prediction using large pretrained language models," npj Digital Medicine (Q1, SJR 3,55), vol. 7, no. 1, p. 40, Feb. 2024, doi: 10.1038/s41746-024-01024-9.

[6] J. Sauvola, S. Tarkoma, M. Klemettinen, J. Riekki, and D. Doermann, "Future of software development with generative AI," Automated Software Engineering, vol. 31, no. 1, p. 26, May 2024, doi: 10.1007/s10515-024-00426-z.

[7] G. Lu, X. Ju, X. Chen, W. Pei, and Z. Cai, "GRACE: Empowering LLM-based software vulnerability detection with graph structure and in-context learning," Journal of Systems and Software, vol. 212, p. 112031, Jun. 2024, doi: 10.1016/j.jss.2024.112031.

[8] A. Saka et al., "GPT models in construction industry: Opportunities, limitations, and a use case validation," Developments in the Built Environment, vol. 17, p. 100300, Mar. 2024, doi: 10.1016/j.dibe.2023.100300.

[9] Y. Yehuda, I. Malkiel, O. Barkan, J. Weill, R. Ronen, and N. Koenigstein, "InterrogateLLM: Zero-Resource Hallucination Detection in LLM-Generated Answers," in Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Stroudsburg, PA, USA: Association for Computational Linguistics, 2024, pp. 9333–9347. doi: 10.18653/v1/2024.acl-long.506.

[10] W. Chen et al., "Systems engineering issues for industry applications of large language model," Appl. Soft Comput., vol. 151, p. 111165, Jan. 2024, doi: 10.1016/j.asoc.2023.111165.

[11] R. Emsley, "ChatGPT: these are not hallucinations – they're fabrications and falsifications," Schizophrenia, vol. 9, no. 1, p. 52, Aug. 2023, doi: 10.1038/s41537-023-00379-4.

[12] L. Zheng, J. Ye, G. Zhao, S. Luo, M. Nan, and Y. Xie, "Incorporating RAG for Factual Hallucination Detection Modeling in Intelligent Educational system," in 2025 7th International Conference on Computer Science and Technologies in Education (CSTE), IEEE, Apr. 2025, pp. 843–847. doi: 10.1109/CSTE64638.2025.11092000.

[13] I. K. R. Arthana, N. P. N. P. Dewi, G. A. J. Saskara, I. M. A. Pradnyana, and L. Indrayani, "Real-time intelligent virtual assistant based on retrieval augmented generation," IAES International Journal of Artificial Intelligence (IJ-AI), vol. 15, no. 1, p. 237, Feb. 2026, doi: 10.11591/ijai.v15.i1.pp237-246.

[14] J. Song et al., "RAG-HAT: A Hallucination-Aware Tuning Pipeline for LLM in Retrieval-Augmented Generation," in Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track, Stroudsburg, PA, USA: Association for Computational Linguistics, 2024, pp. 1548–1558. doi: 10.18653/v1/2024.emnlp-industry.113.

[15] K. Benkirane et al., "Machine Translation Hallucination Detection for Low and High Resource Languages using Large Language Models," in Findings of the Association for Computational Linguistics: EMNLP 2024, Stroudsburg, PA, USA: Association for Computational Linguistics, 2024, pp. 9647–9665. doi: 10.18653/v1/2024.findings-emnlp.564.

[16] J. Zhu, H. Guo, W. Shi, Z. Chen, and P. De Meo, "RaDIO: Real-Time Hallucination Detection with Contextual Index Optimized Query Formulation for Dynamic Retrieval Augmented Generation," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39, no. 24, pp. 26129–26137, Apr. 2025, doi: 10.1609/aaai.v39i24.34809.

[17] Vincencius Christiano Tjokro and Samuel Ady Sanjaya, "Methods and Applications of Fine-Tuning Llama-2 and Llama-Based Models: A Systematic Literature Analysis," Journal of System and Management Sciences, vol. 14, pp. 254–266, Jun. 2024, doi: 10.33168/JSMS.2024.1015.

[18] X. Zhao, X. Leng, L. Wang, and N. Wang, "Research on Fine-Tuning Optimization Strategies for Large Language Models in Tabular Data Processing," Biomimetics, vol. 9, no. 11, p. 708, Nov. 2024, doi: 10.3390/biomimetics9110708.

[19] D. Zhang, V. Gangal, B. Lattimer, and Y. Yang, "Enhancing Hallucination Detection through Perturbation-Based Synthetic Data Generation in System Responses," in Findings of the Association for Computational Linguistics ACL 2024, Stroudsburg, PA, USA: Association for Computational Linguistics, 2024, pp. 13321–13332. doi: 10.18653/v1/2024.findings-acl.789.

[20] L. Huang et al., "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions," ACM Trans. Inf. Syst., vol. 43, no. 2, pp. 1–55, Mar. 2025, doi: 10.1145/3703155.

[21] Z. Liu et al., "Comprehensive Evaluation of AI Hallucination and Novel UV-Oriented Framework toward Safe and Trustworthy AI," in 2024 7th International Conference on Universal Village (UV), IEEE, Oct. 2024, pp. 1–136. doi: 10.1109/UV63228.2024.11189137.

[22] R. Bouchekir, F. Faghih, and T. A. Beyene, "Hallucination Detection in LLMs via Beam Search Sampling and Semantic Consistency Analysis," in 2025 55th Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W), IEEE, Jun. 2025, pp. 274–281. doi: 10.1109/DSN-W65791.2025.00076.

[23] T. Naseem et al., "A Grounded Preference Model for LLM Alignment," in Findings of the Association for Computational Linguistics ACL 2024, Stroudsburg, PA, USA: Association for Computational Linguistics, 2024, pp. 151–162. doi: 10.18653/v1/2024.findings-acl.10.