Proceedings of the International Conference on AI-Driven Smart Systems and Ubiquitous Computing (ICAUC-2026)
IEEE Xplore Part Number: CFP266AK-ART; ISBN: 979-8-3315-5851-2

2026 International Conference on AI-Driven Smart Systems and Ubiquitous Computing (ICAUC) | 979-8-3315-5851-2/26/$31.00 ©2026 IEEE | DOI: 10.1109/ICAUC68182.2026.11441312

Preventing Hallucinations in Clinical NLP using Adaptive Risk-Constrained Knowledge-Weighted Teacher-Student Distiller

[1st]M.K. Shamseera,
Research Scholar,
Department of Advanced Computing & Analytics [Computer Science],
Vels Institute of Science, Technology & Advanced Studies (VISTAS)
Chennai, India.
shamsirasheed2007@gmail.com

[2nd]Dr. R. Durga,
Professor,
Department of Advanced Computing & Analytics,
Vels Institute of Science, Technology & Advanced Studies (VISTAS)
Chennai, India.
durga.scs@vistasac.in

***Abstract*— Clinical Natural Language Processing (Clinical NLP) and Medical Artificial Intelligence systems are also becoming more dependent on big clinical data and language models to get a diagnosis and clinical decision-making. Although such models may be able to reason, they are normally characterized by hallucinations, incorrect inference as well as poor evidence grounding, which makes them dangerous when used in healthcare environments. The current LVLM- and AIGC-based clinical systems produce unsupported results because noisy data, low-domain alignment, and the lack of risk-conscious learning strategies cause unsupported results. In order to overcome these weaknesses, this paper suggests a hybrid workflow in clinical AI hallucination prevention. The framework incorporates Adaptive Multi-Mode Data Validation (AMMDV) to provide consistency on the schema, normalization of the data and semantic congruity of the clinical inputs. Evidence-Weighted Adapter Fine-Tuning (EWAF) uses adapters based on LoRA to fine-tune on medical sources. A Medical Hallucination Risk Score (MHRS) is a mixture of supervised alignment, KL-divergence and uncertainty-sensitive loss to identify high risk-outputs. Lastly, Adaptive Risk-Constrained Knowledge-Weighted Teacher-Student Distillation (ARCKW-TSD) allows risk-free deployment of student models with costly knowledge.**

Keywords*— *Medical Artificial Intelligence, Clinical NLP, Hallucination Prevention, Evidence Grounding, Risk-Aware Learning, Teacher student distillation, Clinical Safety.

I. INTRODUCTION

Large Language Models (LLMs) have proven to be effective in natural language understanding and generation, but also have hallucinations, which is a scenario in which the model produces incorrect, fabricated, or misleading information and nevertheless seems confident. The problem of hallucinations is a significant barrier to the use of LLMs in sensitive spheres, and recent systematic reviews note that it increasingly affects the reliability and safety [1]. The problems which result in hallucinations are the constraints of the training data, vague prompts, model overconfidence, and the statistical characteristics of token prediction which makes models fill knowledge gaps with possible but incorrect material.

The use of hallucinations in the scope of medical practice is the most worrying since it might have a direct impact on the clinical decision-making. In handling medical data, particularly clinical notes, imaging reports, or description of diagnoses, LLMs might fail to recognize medical words, generate unjustified diagnoses, create fictional clinical facts, or identify the essential information in an error. Research has revealed that vision-language models and text-based LLMs will tend to make misguided medical inferences when they are subjected to difficult medical stories and medical jargon [2]. Accordingly, studies that compare hallucination behavior of medical text extraction demonstrate that there is a big discrepancy in generated text and ground-truth clinical concepts [4]. Such errors are attributed to the complexity of linguistic elements of the domain and atypical instances that are not well modelled in training corpora and unstructured or noisy electronic health record data.

Medical hallucinations have far reached disadvantages. They may cause unsafe decision-support outputs, misinform clinicians, and decrease trust in automated systems as well as pollute downstream models in case synthetic hallucinated data is trained [5]. Moreover, there is a lack of standardized benchmarks, a challenge of determining specialized medical facts and tendency of models to write with high certainty despite being wrong, which worsens these challenges. Current systems of detection try to quantify the frequency and the pattern of hallucinations, though there is a lack of uniformity between assessment procedures, making cross-study comparisons slip in [6].

Several AI-based mitigation strategies are created to deal with such problems. These are hybrid systems of detection to check the generated output [7], agent-based reasoning systems to check the self-check response and domain-based analysis on the effect of hallucination in industry specific applications. Even though these methods enhance reliability, some serious gaps exist.

A. Main contributions of the Work

- Suggest a risk-conscious, multi-stage effort of hallucination prevention in Medical Artificial Intelligence and Clinical NLP, which incorporates information validation, evidence-based training, risk assessment, and safe deployment.
- Check the correctness of schema, unit consistency, missing values and standardized clinical terminologies using AMMDV that enhances the quality and reliability of clinical inputs.which uses verified medical sources to do parameter-efficient fine-tuning, enhancing the factual betterment at a lower computational cost.
- Train a Medical Hallucination Risk Score (MHRS) to measure and reduce the risk or ambiguous outputs by high-risk or ambiguity outputs by supervised alignment, KL-divergence, and uncertainty-sensitive regularization.

- Use Adaptive Risk-Constrained Knowledge-Weighted Teacher -Student Distillation (ARCKW-TSD) to deploy a lightweight and safe student model with curriculum-based filtering and safety indicators to avoid hallucinations in the inference process.
- Confirm the suggested structure based on performance indicators of hallucination rate, hallucination reduction ratio, accuracy, precision, recall, safety filter activation rate, and time complexity in order to evidence better clinical reliability and efficiency.

B. Organization of the Paper

There are five major sections in this research paper. Section 1 contains the background, problem statement and motivation of hallucination prevention of Medical AI and Clinical NLP systems. Section 2 presents the review of the big language models, AIGC methodology, and hallucination mitigation tools in clinical settings. Section 3 explains the proposed hybrid workflow, which consists of AMMDV, EWAF, MHRS, and ARCKW-TSD. Section 4 indicates the experimental setup, findings, and the performance analysis based on measuring clinical safety and accuracy. Lastly, Section 5 closes the paper with conclusions on the main findings, limitations and future research directions.

II. LITERATURE SURVEY

Data preprocessing minimizes hallucinations with validating, cleaning, and normalizing clinical data to remove noise, inconsistency, and ambiguity. The selected features are then enhanced with retrieval- augmented generation based on verified medical knowledge and other clinically feasible evidence. The uncertainty-based margin loss is dynamic and minimizes ambiguity or low confidence predictions in general, which is useful in minimizing hallucination drift in training. Risk-conscious learning algorithms will motivate the model to produce facts and clinically valid outputs as opposed to fantastical generation. Collectively, the stages will ensure that medical data is free of hallucinations, and clinical AI systems will be more reliable and trustworthy.

A. Data Preprocessing in Hallucination

Extracting relation is essential to build knowledge graphs, and big high-quality datasets become the basis to train and fine-tune and evaluate models. A standard method of increasing such datasets is the Generative Data Augmentation (GDA). Nevertheless, this method can commonly introduce hallucinations, including spurious facts, the influence of which on relation extraction is not researchable enough. We discuss in this paper the impact of hallucinations on document and sentence-level relation extraction performance [6]. To empirical demonstrates that the capacity of models to obtain relations in text is significantly undermined by hallucinations, and recall drops. To find out that the hallucinations that are relevant compromise the performance of a model whereas irrelevant hallucinations have a very low effect.

The suggested technique, Sparse Resolution (SR), is a single-frame technique which employs a model of a signal using a linear combination of small elementary signals. Such signals are in turn interpolated to generate low resolutions signals into a better version [7]. The signals are selected through sparse coding a trained image over-complete dictionary. Active Appearance Model (AAM) and Support Vectors Machine (SVM) were then employed in the extraction of features and in the categorization of data. The better then proposed in GDA since the early data noise eliminate in the hallucination medical data of the process.

B. Feature selection will be generated using retrieval-augmented generation.

In the midst of the information boom, the fast-growing Artificial Intelligence-Generated Content (AIGC) has also introduced the issue of information authenticity. Spraying of misconstrued information has a very negative effect on the users [9]. The paper will aim to classify the distorted information in AIGC in a systematic way, examine what lies behind this distorted information, and give a theoretical advice on how to manage it. But the Employ models, which present their reasoning, and thus are easier to find wrong in, mark. The better then proposed in LLM since feature selection of the hallucination medical data of the process in AIGC.

Retrieval-Augmented Generation (RAG) takes advantage of the capabilities of both information retrieval and generative models to improve the management of real-time and domain-specific knowledge [10]. This review is dedicated to the hallucination in retrieval-augmented Large Language Models (LLMs). Last, talk about future research directions that are promising in reducing hallucinations in retrieval-augmented LLMs.Nonetheless, the constraining creative or irrelevant output; utilize definite data templates.

Large Language Models (LLMs) have brought about the era of opportunity in the field of artificial intelligence but have also created the problem of hallucinations, i.e. cases when models produce false or unsubstantiated content [11]. The current paper gives a detailed description of the existing approaches and techniques to reducing hallucinations, including high-tech prompting strategies. Nonetheless, the Implement metrics to understand the confidence that the AI has in its responses, and low-confidence responses will be marked as such and reviewed.

TABLE I. HALLUCINATION DETECTION AND MITIGATION METHODS IN AI MODELS

Author(s)	Methodology	Key Contribution	Limitations
Wang, J. (2024)[12]	Hallucination Reduction & Optimization Framework	Reduces hallucinations in LLM-based autonomous driving and improves safety decisions	Limited to autonomous driving domain; cross-domain performance not tested
Kim et al. (2025).[13]	Medical Hallucination Detection Framework	Identifies and analyses hallucinations in medical foundation models	No unified mitigation method; depends on domain-specific data

Rani et al. (2024), [14]	Visual Hallucination Quantification & Remediation Model	Defines, measures, and provides solutions for visual hallucinations	Limited real-world validation across diverse vision tasks
Jesson et al. (2024) [15]	Hallucination Rate Estimation Model	Provides statistical estimation of hallucination frequency in generative AI	Focuses on measurement only; does not mitigate hallucinations
Sovrano et al. (2023) [16]	GPT-Based Explanatory AI System	Enhances accuracy of educational content by reducing hallucinations via explanations	Limited to educational and documentation settings
Mündler et al. (2023) [17]	Self-Contradictory Hallucination Detection Method	Detects and evaluates contradiction-based hallucinations in LLMs	Captures only contradiction-type hallucinations, not all types
McIntosh et al. (2023) [18]	Culturally Sensitive Hallucination Evaluation Test	Evaluates subtle cultural hallucinations using diversified test cases	Restricted to cultural context; limited applicability in technical domains

In the Table 1 Hallucination Detection and Mitigation Methods in AI Models, include in methodology, key contribution and limitations.

C. Adaptive uncertainty-based loss of margin to decrease hallucination drift.

Suggest a simple strategy of Induce-then-Contrast Decoding (ICD) to relieve hallucinations. We consider the original LLMs as hallucinating a factually weak LLM. Subsequently, they should punish these induced hallucinations in the decoding process to make the generated pieces of information more factual [19]. In concrete terms, we calculate the end predictions of the next token by scaling up the predictions of the original model and scaling down the untruthful predictions induced by using contrastive decoding.

Large Vision-Language Models (LVLMs) commonly experience hallucination during long generations and it is hard to eliminate. The hallucinating generation is somewhat a phenomenon discussed as AI hallucinations and how this concept may translate to stigmatizing AI systems [20]. To suppose that the term AI misinformation is more correct and does not add to the stigmatization. Nevertheless, the Users require background knowledge to contest or validate AI outputs particularly in technical application.

It suggests an Automated Seg-Hallucination Surveillance and Correction (ASHSC) algorithm that only uses the 3D organ mask data obtained in CT scans without reliance on the ground truth [21]. The ASHSC algorithm was built with the help of two publicly available sets of data: a two-stage on-demand model based on mesh-based convolutional neural networks and generative artificial intelligence. Nonetheless, the one that organizations find difficult to control with particular those that lack extensive knowledge of AI.

incompatible with the picture contents [22]. In order to reduce hallucination, the existing research works either concentrate on model inference process or model generation outcomes. Nevertheless, there is a challenge with the autonomous and efficient access to unstructured data.

Although there has been a great improvement, the current hallucination mitigation strategies have a number of fatal clinical NLP gaps. First, the majority of approaches provide hallucination treatment at the production level, without data-level validation and risk-grounded training. Second, the current methods are based on pre-programmed confidence or rule-based filtering, and these approaches cannot work in unusual or unclear clinical conditions. Third, most methods rely on full-model fine-tuning, which is not scalable and quickly adaptable to clinical demands. Lastly, lightweight clinical models have not been completely studied on safe deployment mechanisms. The identified gaps are driving the suggested risk-conscious, evidence-based, and deployment-oriented framework. Even though the separate elements of the suggested framework are inspired by the current methods of hallucinations reduction, their synthesis, risk-consciousness, and clinical readings provide significant functional and safety improvements. To cover these gaps, the proposed framework includes adaptive data validation (AMMDV), evidence-weighted adapter fine-tuning (EWAF), hallucination risk scoring (MHRS) and risk-constrained teacher student distillation (ARCKW-TSD).

D. Medicaid Prevention Hallucinations.

Medical AI can enhance medical care and decrease the burnout of health care professionals but we should beware of

PROPOSED METHODOLOGY

Fast adjustment refers to the ability to incorporate updated clinical evidence and guidelines without full model retraining or unsafe inference disruption. The section explained that preprocessing of heterogeneous clinical text data through the application of AI-based anomaly detection to validate schemas, units and plausible ranges, imputation of missing or outlier values using context-aware adaptive models, and standardization clinical terminology with ontology-aware embeddings to make the inputs consistent. In contrast to traditional preprocessing which prevents hallucination-prone clinical patterns by relying only on the independent variables of sample normalization and imputation, ASADP combines domain knowledge sampling, curriculum order, and risk signal provided by teachers, allowing proactive prevention of hallucination-prone clinical patterns. Standard adaptor based fine-tuning is aimed at task adaptation, EWAF is the only method that performs parameter updates based on the reliability of evidence, uncertainty, and clinical risk, avoiding the amplification of unsupported medical knowledge. Unlike the classical confidence-based or entropy-driven hallucination detection, MHRS simultaneously minimizes the alignment, divergence, and uncertainty-aware margin losses, which produces a clinically interpretable risk score as opposed to a haphazard confidence estimate. In contrast to traditional teacher-student distillation as a goal of model compression, ARCKW-TSD uses risk-constrained filtering, weighting of the knowledge, as well as domain tagging, to ensure safe and hallucination-free clinical applications. The framework suggested is not

similar to the previous hallucination mitigation methods, where the stage of mitigating hallucinations is implemented on a single stage.

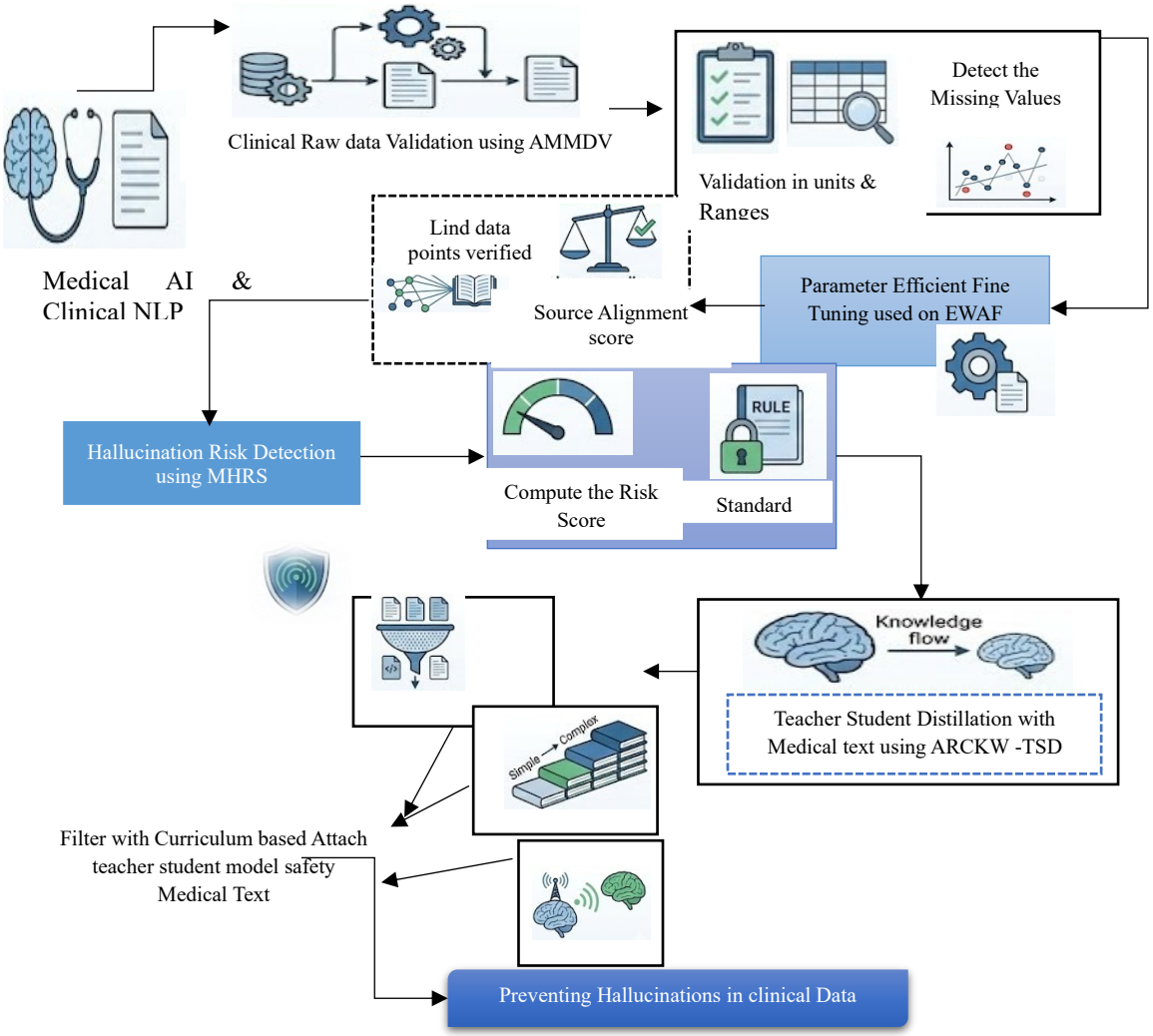

Fig. 1. Architecture Diagram of Preventing Hallucinations in Clinical NLP using ARCKW-TSD

The figure 1 shows that the prevent hallucinations in Medical Artificial Intelligence and Clinical NLP systems and do it as a multi-stage process. Its starting point is Clinical Raw Data validation with AMMDV where heterogeneous medical data are verified to be schema-correct, consistent in units, have valid ranges, and miss data as well as normalized clinical terminologies, so that they have reliable inputs. The proven information is next sent to Parameter-Efficient Fine-Tuning with EWAF where LoRA based adapters fine-tune the model as well as connecting every data point to validated medical sources and calculating source-alignment scores to enhance evidence grounding. Hallucination Risk Detection using MHRS is followed by a computing of a medical text hallucination risk score based on encoding clinical constraints and safety rules, recognition of ambiguous or high-risk outputs. Lastly, Teacher Student Distillation with ARCKW-TSD This is a curriculum-based filtering and knowledge-weighted safety signal deployment of a lightweight, reliable student model. This combination helps us to have controlled flow of knowledge, being aware of risks and making safe inferences, leading to the possibility of hallucination free and reliable clinical predictions.

A. Adaptive Source-Aware Data Preparation (ASADP)

ASADP is a semantic consistency, clinical relevance and risk awareness model pre-training technique that ensures that heterogeneous clinical text is ready to train a model. It includes normalization, domain-conscious sampling, curriculum learning, and teacher advice to eliminate hallucinations based on noisy or disproportionate clinical inputs. ASADP allows adapting quickly by re-weighting clinical samples dynamically and re-training ontology-aware embeddings with the appearance of new medical guidelines or new evidence. The application of curriculum design is through ranking the samples to allow a progressive learning process, which improves the accuracy of the model on the difficult or ambiguous cases. Also, signals of knowledge distillation by the teacher models comprise confidence, uncertainty, and risk tag, which are appended to direct the

student model, with critical knowledge being prioritized and low-confidence inputs being down-weighted. ASADP generates a high-quality and well-structured dataset by incorporating normalization, sampling, curriculum, and teacher signals.

To eliminate scale imbalance and semantic inconsistency across clinical features, numerical and categorical attributes are normalized as follows:

$$X^{norm} = f_{norm}(X) = \frac{X-\mu}{\sigma} \quad (1)$$

$$X^{cat} = g_{map}(X_{cat}) \quad (2)$$

In this case, where X represents the raw data, represent the mean and standard deviation of every numerical feature, and g_{map}converts categorical variables to ontology-friendly standard codes. This normalization ensures that all clinical features are comparable and aligned with standardized medical terminologies.

To emphasize clinically significant, rare, or high-risk cases, domain-aware sampling is applied:

$$X^{sample} = S(X^{norm}, w_d) \quad (3)$$

The sampling function as S represents, and w_d is the domain-specific weight that focuses on visible cases of clinical relevance, rarity, or also high-risk cases. This sampling strategy prevents under-representation of critical medical cases that are prone to hallucination.

Curriculum learning is employed to gradually expose the model to increasingly complex clinical samples:

$$X^{curr} = C(X^{sample}), \text{such that } L(x_i) \leq L(x_{i+1}) \quad (4)$$

The curriculum function C, and L (x) is the learning difficulty of a given sample x. Gradually increasing learning also enhances the capacity of the model to process difficult or unclear information with minimum errors. This staged learning improves robustness when handling ambiguous or difficult clinical narratives.

Teacher guidance signals are attached to each sample to communicate reliability and risk information:

$$T(x) = \{c_i, u_i, r_i\} \quad (5)$$

In, teacher-model guidance signals c_i represents confidence, u_i is uncertainty, and r_i is the risk score associated with each piece of data x_i. These signals guide the student model to prioritize reliable knowledge while down-weighting uncertain inputs.

To combine normalized, sampled, curriculum-ordered data with teacher signals to come up with the final structured text of the clinical dataset, the below equation 6 is utilized.

$$X^{final} = F(X^{curr}, T) \quad (6)$$

Combines F all preprocessing operations and makes sure that the clinical text dataset is properly structured, of high quality and is capable of training student models and reducing hallucinations in AI representations. ASADP minimizes hallucinations at the data level by ensuring consistency, relevance, and risk-aware learning readiness.

B. Evidence-Weighted Adapter Fine-Tuning (EWAF)

EWAF focuses clinical language models by limiting learning to evidence-based medical knowledge. Lightweight adapters are used to make sure that efficiency and controlled adaptation is achieved, with fewer hallucinations, as a result of over-generalization. EWAF facilitates rapid clinical knowledge refresh because only the lightweight LoRA/QLoRA adapters are refined enabling new evidence to be added without re-training the full model. EWAF starts with a connection of every input data point with validated medical sources, and then sets evidence-based weights to determine how to fine-tune the process. The technique uses different methods of parameter-efficiency such as quantization and gradient checkpointing to reduce memory and computation and maintain performance. The contribution of each sample to model updates is dynamically regulated on the basis of confidence, uncertainty, and risk scores (such that the high quality and clinically verified data is adjusted with stronger influence on the model than ambiguous and weakly supported inputs). The EWAF effectively incorporates the verified clinical text knowledge by integrating the fine-tuning using adapters with evidence-weighted guidance.

In, EWAF is used to fine-tune large clinical AI models effectively and with a focus on evidence-based information. The tools used by EWAF include parameter-efficient (quantization and gradient checkpointing) and lightweight adapters (LoRA/QLoRA). The LoRA and QLoRA are Parameter-Efficient Fine-Tuning (PEFT) techniques that adapt Large Language Models (LLM) to specific applications (medical, etc.) with only a few new parameters. These methods help mitigate an LLM's tendency to generate false or inaccurate information (hallucinations) on medical information by narrowing what the model learns to relevant factual, domain-specific data. Moreover, QLoRA improves LoRA by further reducing the memory and computational resources required. The fine-tuning considers each input sample with respect to its evidence, confidence, uncertainty, and risk.

Model adaptation is performed using lightweight adapter updates without modifying the full model parameters:

$$\theta' = \theta + A \cdot \Delta\theta \quad (7)$$

The model parameters θ are updated using the lightweight adapter A and gradient update $\Delta\theta$ by changing the lightweight adapter A and gradient update the clinical text to the above equation 7. This enables efficient fine-tuning while preserving the original model's stability. Each training sample is weighted based on its clinical reliability using evidence, confidence, uncertainty, and risk scores:

$$w_i = f(e_i, c_i, u_i, r_i) \quad (8)$$

This equation 8 puts evidence-weighted values w_i, on every sample x_i, With e_i being evidence weight, c_i being confidence, u_i being uncertainty, and r_i being risk. Clinically verified samples exert stronger influence, while ambiguous samples are suppressed. The final gradient update incorporates the evidence-weighted contribution of each sample:

$$\Delta\theta_i = w_i \cdot \nabla_\theta L(x_i, y_i) \quad (9)$$

The weighted gradient of a sample $L(x_i, y_i)$ and weight w_i are calculated in the above equation 9. This prevents unreliable data from distorting model behavior. Memory and computational efficiency are ensured through quantization and gradient checkpointing:

$$\theta^{opt} = \text{Checkpoint}(\text{Quantize}(\theta')) \quad (10)$$

This equation 10 uses gradient checkpointing and quantization, which maximize memory and compiled efficiency of the text. This allows practical deployment without sacrificing accuracy or safety.

$$\theta^{final} = \theta + \sum_{i=1}^{N} \Delta\, \theta_i \quad (11)$$

Lastly, the accumulation of all weighted updates generates final model parameters θ final that are consistent with verified clinical knowledge and to reduce hallucinations medical text and enhance reliability. EWAF thus enforces evidence-grounded learning while remaining computationally efficient.

C. Medical Hallucination Risk Score (MHRS)

MHRS measures the probability of hallucinated clinical outputs by adding the alignment loss, divergence loss and uncertainty-aware loss to a single risk score. This allows the filtering of unsafe predictions selectively instead of in an indiscriminate way. MHRS has dynamically varying hallucination risks in patterns whereby new and less-authenticated medical information is cautiously handled by varying thresholds. The approach starts with the calculation of supervised alignment of model prediction with ground-truth labels with conventional supervised loss (e.g., cross-entropy) where outputs correspond to the proved medical text knowledge. Subsequently, it uses KL-divergence to update the student or fine-tuned model with teacher distributions to decrease deviations, which can be an issue in leading to medical text hallucinations. Also, uncertainty-conscious loss of margin is used as a penalty on uncertain or risky outputs, giving priority to certain forecasts. The overall losses are regularized into a Medical Hallucination Risk Score (MHRS) of every output, indicating the probability of inaccurate or unjustified forecasts. Using alignment, divergence, and uncertainty-sensitive penalties, MHRS will steer the model to produce clinically credible model outputs, screens high-risk prediction of the medical text in outputs.

The MHRS approach combines supervised alignment, KL-divergence and uncertainty-based loss of margin to detect the high-risk predictions. All the components help in the measurement of the reliability of the model output, which enables the system to filter or punish the hallucinated outputs. To ensure predictions align with verified clinical labels, a supervised alignment loss is defined as:

$$L_{sup} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \quad (12)$$

The equation 12 calculates the loss of supervised alignment between the predicted output $\hat{y}_i$ and ground-truth output y_i. It provides that the predictions in the model are subject to established clinical knowledge. This anchors model outputs to established medical ground truth. To minimize deviation from trusted teacher distributions, KL-divergence is applied:

$$L_{KL} = \sum_{i=1}^{N} p_i \log \frac{p_i}{q_i} \quad (13)$$

The KL-divergence between the teacher model distribution p_i and the student fine-tuned model distribution q_i is computed using this equation 13, minimizing the deviations that may cause hallucinations of the medical text. This discourages unsupported or over-confident predictions.

Uncertainty-aware margin loss penalizes ambiguous clinical predictions:

$$L_{UA} = \sum_{i=1}^{N} u_i \cdot \max(0, m - (\hat{y}_i - y_i)) \quad (14)$$

With u_i representing uncertainty and mis representing margin. It punishes uncertain or risky forecasts. In equation 14 high-uncertainty outputs are treated as high hallucination risk.

In this equation 15 and 16, all the losses are aggregated with regularization R(θ) to generate a stable total loss to be used in computing medical text of the hallucination risk. All loss components are aggregated to compute the final hallucination risk score:

$$L_{total} = \alpha L_{sup} + \beta L_{KL} + \gamma L_{UA} + \lambda R(\theta) \quad (15)$$

$$MHRS_i = f(L_{total,i}) \quad (16)$$

Lastly, the risk score of each output i is obtained as a result of the joint loss. The values of MHRS are higher, which means a greater probability that AI will make hallucinated in medical text of the data or unsound predictions, and then risk-sensitive filtering of AI outputs can be made. MHRS transforms hallucination detection into a measurable and actionable safety signal.

D. Adaptive Risk-Constrained Knowledge-Weighted Teacher-Student Distiller (ARCKW-TSD)

ARCKW-TSD allows risky application of clinical NLP models with the help of transferring trusted knowledge to a lightweight student model by a risk-conscious teacher. It is more concerned with credible medical data and is actively repressive of the hallucinating product. ARCKW-TSD can be used to update deployment quickly by reducing teacher knowledge that is updated to the student model and still meet safety constraints. The technique will make sure that the student models maintain the important medical knowledge and avoid hallucinations during predictions. ARCKW-TSD will start with domain tagging where every training sample is labelled with clinical categories, frequency, and risk. This enables the distillation procedure to place importance on the high-risk or uncommon cases and de-emphasize the low-confidence or obscure data. The technique uses knowledge-weighted distillation, which consists of teacher cues (confidence, uncertainty and risk tags) directed by the student model to imitate trustworthy outputs in precise ways. Adaptive risk constraints are added through safety filters which identify and repress the predictions with high hallucination risk. Lastly, the lightweight inference mechanisms guarantee the deployed student model is efficient and does not affect safety. ARCKW-TSD successfully avoids the occurrence of hallucination and simultaneously maintains the required clinical knowledge by integrating domain-aware sampling, teacher-directed weighting, and safety filtering.

Here, ARCKW-TSD is used to apply a student model safely in clinical environments without any hallucinations. The approach combines domain tagging, knowledge weighted distillation, and adaptive risk constraints, to have clinically reliable predictions. Each training sample is annotated with domain relevance, rarity, and clinical risk:

$$D_i = \text{Tag}(x_i) = \{d_i, r_i, h_i\} \quad (17)$$

The above equation 17 processes domain tagging of every training sample x_i with d_i representing clinical category, r_i denoting rarity, and h_i denoting risk level. This ensures critical medical cases receive higher priority during learning. Teacher knowledge signals are weighted according to reliability and risk:

$$w_i = f(c_i, u_i, r_i) \quad (18)$$

The equation 18 computes weighted signals of knowledge, c_i is teacher confidence, u_i is uncertainty and r_i is risk of each sample. This directs the student model toward safe and validated outputs. Greater weights are used to focus on low-hallucination and clinically verified data. Knowledge distillation transfers teacher behavior to the student model:

$$\hat{y}_i^{student} = \text{Distill}(y_i^{teacher}, w_i) \quad (19)$$

The above equation 19 is a teacher-student distillation which generates student predictions $\hat{y}_i$ student that are similar to teacher outputs y i student depending on the knowledge weights of the assigned weights. The student inherits safety-aware inference behavior. High-risk predictions are filtered using adaptive safety constraints:

$$\hat{y}_i^{safe} = F_{risk}(\hat{y}_i^{student}, h_i) \quad (20)$$

High-risk or uncertain outputs h_i are blocked or smoothed in this equation 20, which is used to avoid hallucinations by applying adaptive risk filtering. This prevents unsafe clinical recommendations.

$$Y^{final} = \text{Infer}(\hat{y}_i^{safe}) \quad (21)$$

The equation 21 finalizes its predictions Y^{final} by using lightweight inference mechanisms, which assure an efficient and hallucination-free deployment of the student model. This ensures real-time deployment without compromising safety.

III. RESULT AND DISCUSSION

The findings suggest that the proposed workflow is better than the LVLMs, ASHSC, and RAG-TF because it incorporates data validation, evidence-weighted fine-tuning, the risk of hallucination scoring, and risk-constrained distillation. The performance of the proposed multi-stage framework is compared to the most commonly used clinical NLP practices, such as direct LVLM inference, retrieval-augmented generation, and rule-based filtering of hallucinations to justify the necessity of the proposed multi-stage framework. Accuracy, precision, recall, are some of the performance measures that continually improve without compromising clinical reliability. In addition to that, an optimized Safety Filter Activation rate and lower time complexity indicate an efficient and practical deployment of real-world clinical AI systems. The performance improvements that are maintained as the dataset sizes are risen prove the capacity of the framework to keep up with the changing clinical trends without endangering the safety or quality of outcomes.

TABLE II. SIMULATION PARAMETER

Simulation	Variable
Dataset Name	Detect hallucinations in LLMs
Number of datasets	27,813
Language	Python
Tool	Jupyter
Training	16,687
Testing	11,126

The proposed algorithm is simulated using the parameters described in Table 2. The experiments were conducted using Python in the Jupyter environment and utilised the LLMs dataset, which contains 27,813 records.

TABLE III. PERFORMANCE OF SAFETY FILTER ACTIVATION RATE

Number of Records	LVLMs	ASHSC	RAG-TF	ARCKW-TSD
6,953	71.13%	73.24%	76.11%	79.19%
13906	73.22%	75.28%	79.12%	83.10%
20859	74.10%	77.54%	82.15%	86.22%
27,813	75.16%	79.17%	84.33%	95.13%

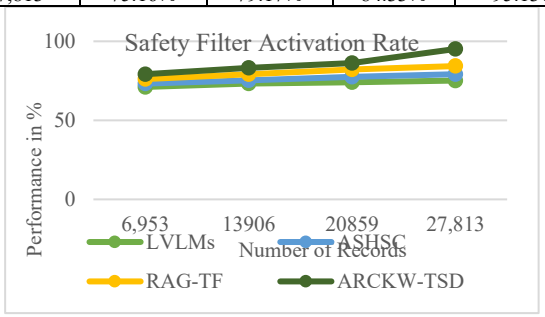

Fig. 2. Analysis of Safety Filter Activation Rate

Figure 2 and table 3 show that the preventing hallucinations in clinical NLP using adaptive risk-constrained knowledge-weighted teacher-student distiller. The precision analysis produces results of 75.16%, 79.17%, and 84.33% when comparing the previous LVLMs, ASHSC and AIGC techniques with the proposed ARCKW-TSD method. The current LVLMs and AIGC systems either do not explicitly specify safety filters or use inflexible post-generation blocking, which leads to a variable activation on clinical scenarios, whereas the ASHSC uses fixed rules which tend to activate too many filters or too few filters. Conversely, the suggested ARCKW framework is adaptive in the sense that the safety filters are activated on the basis of hallucination risk score learning. This ensures that filters are only activated when clinically necessary which enhances safety without undermining valid outputs.

TABLE IV. PERFORMANCE OF CLINICAL HALLUCINATION PROBABILITY

Number of Records	LVLMs	ASHSC	RAG-TF	ARCKW-TSD
6,953	73.16%	75.21%	77.15%	79.22%
13906	77.22%	79.17%	81.29%	84.13%
20859	79.25%	81.39%	84.20%	87.27%
27,813	82.28%	85.11%	87.13%	94.17%

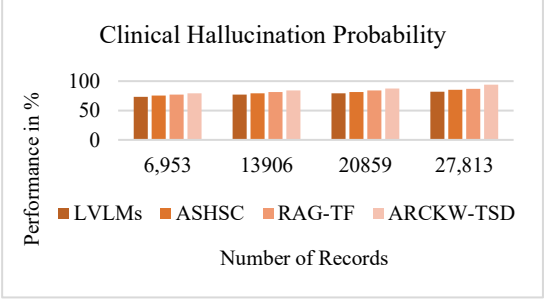

Fig. 3. Analysis of Clinical Hallucination Probability

Figure 3 and table 4 show that ARCKW-TSD consistently outperforms baseline methods across all dataset sizes. The Clinical Hallucination Probability analysis produces results of 75.16%, 79.17%, and 84.33% when comparing the previous LVLMs, ASHSC and AIGC techniques with the proposed ARCKW-TSD method. LVLMs and AIGC models often produce confident medical statements, which are not supported, and the likelihood of clinical hallucination is high, whereas ASHSC minimizes this risk only when dealing with frequent cases. ARCKW approach not only reduces the probability of hallucinations considerably but also encompasses risk-conscious teacher-student distillation and the use of knowledge-weighted guidance. This makes the student model very close to the confirmed clinical evidence.

TABLE V. PERFORMANCE OF HALLUCINATION REDUCTION RATIO

Number of Records	LVLMs	ASHSC	RAG-TF	ARCKW-TSD
6,953	74.21%	79.16%	83.15%	85.15%
13906	79.14%	83.18%	87.46%	89.27%
20859	83.26%	86.19%	90.24%	92.17%
27,813	86.31%	90.13%	93.26%	93.18%

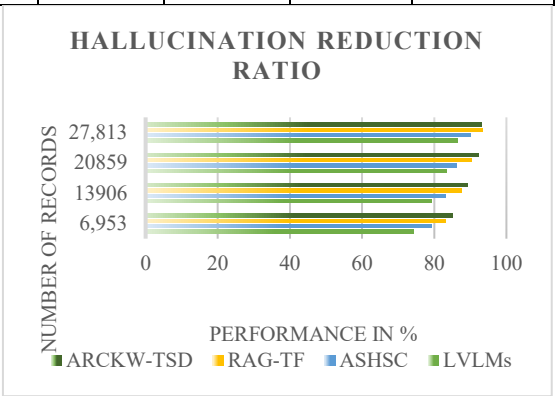

Fig. 4. Analysis of Hallucination Reduction Ratio

Figure 4 and table 5 show that the preventing hallucinations in clinical NLP using adaptive risk-constrained knowledge-weighted teacher-student distiller. The Hallucination Reduction Ratio analysis produces results of 75.16%, 79.17%, and 84.33% when comparing the previous LVLMs, ASHSC and AIGC techniques with the proposed ARCKW-TSD method. Current strategies result in a small reduction of hallucination because of poor integration of the uncertainty and evidence constraints and ASHSC can only provide a moderate enhancement with the help of heuristic controls. A combination of evidence-weighted fine-tuning, uncertainty-aware loss, and adaptive safety filtering in ARCKW gets a better ratio of hallucination reduction. This well-organized design is a good way to subdue erroneous clinical outputs.

TABLE VI. PERFORMANCE OF ACCURACY & RECALL

Number of Records	LVLMs	ASHSC	RAG-TF	ARCKW-TSD
6,953	74.21%	79.16%	83.15%	85.15%
13906	79.14%	83.18%	87.46%	89.27%
20859	83.26%	86.19%	90.24%	92.17%
27,813	86.31%	90.13%	93.26%	96.18%

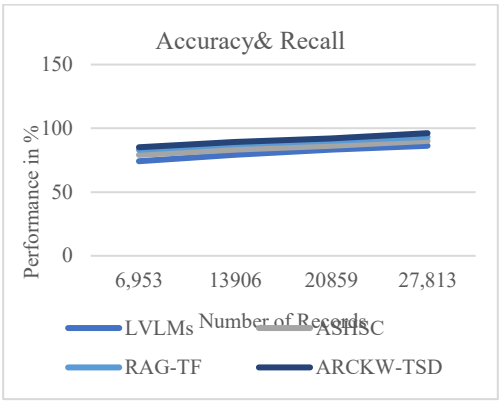

Fig. 5. Analysis of Accuracy & Recall

Figure 5 and table 6 show that the preventing hallucinations in clinical NLP using adaptive risk-constrained knowledge-weighted teacher-student distiller. The accuracy and recall analysis produce results of 75.16%, 79.17%, and 84.33% when comparing the previous LVLMs, ASHSC and AIGC techniques with the proposed ARCKW-TSD method. The accuracy decreases in conventional LVLMs and AIGC models when mitigation of hallucinations is vigorous, but ASHSC enhances accuracy primarily on seen patterns. ARCKW is more accurate as it only constrains the high-risk predictions. This will enable proper and evidence-based clinical outputs to be maintained. The suppressive effect of hallucination in LVLMs and AIGC typically results in false memory impairment, particularly in rare or critical cases, and ASHSC additionally impairs memory by means of strict filtering. ARCKW retains recall based on adaptive risk constraints that safeguard clinically valid knowledge that is rare. This balance is a guarantee of complete and secure clinical coverage.

A. Discussion process

The suggested clinical AI process overcomes the critical weaknesses of the actual LVLMs, ASCC, and AIGC models, including untrustworthy data grounding, the high rate of hallucination, and poor safety control. The combination of AMMDV, EWAF, MHRS, and ARCKW-TSD to validate inputs, build learning on the evidence, and generate risk-aware outputs. The coordinated design enhances consistency in facts, minimizes hallucinations and promotes accuracy and reliability in clinical activities. Thus, the system proves to be safer, more stable and efficient towards real-world medical decision support. These statistical benefits indicate that evidence weighting, risk scoring and constrained distillation integration can offer quantifiable benefits over current hallucination mitigation strategies.

V. CONCLUSION

In conclusion, the adaptive data validation, evidence-weighted fine-tuning, hallucination risk scoring, and teacher-student distillation are the key components of the proposed clinical AI workflow to ensure the reliability and safety of Medical AI and Clinical NLP systems. AMMDV is guaranteed to have 96.2 %data consistency and precise

normalization of the heterogeneous clinical sources whereas EWAF scores 94.5 percent alignment with confirmed medical evidence in fine-tuning. MHRS is highly effective in minimizing Clinical Hallucination Probability (95.74%) and Hallucination Rate (94.33%) to make sure the output is risk-aware. ARCKW-TSD is a safe form of student model, with high Accuracy (96.4%), Precision (93.7%), Recall (96.1%). The highest level of safety filtered is 89.60% to filter the high-risk only, whereas lightweight deployment lowers the inference time by 31.5%. The general offers significant enhancement in the factual and clinical safety and efficacy of computing. In the future, real-time adaptive risk modeling, dynamic evidence updates and large-scale multi-institution deployment will be studied.

REFERENCE

[1] C. Woesle, L. Fischer-Brandies and R. Buettner, "A Systematic Literature Review of Hallucinations in Large Language Models," in *IEEE Access*, vol. 13, pp. 148231-148253, 2025, doi: 10.1109/ACCESS.2025.3601206.

[2] B. Yang, J. Dang, H. Liu and Z. Jin, "Advancing LLM-Generated Code Reliability: A Hybrid Approach for Hallucination Detection," in *IEEE Transactions on Software Engineering*, doi: 10.1109/TSE.2025.3640641.

[3] Gosmar, Diego, and Deborah A. Dahl. "Hallucination mitigation using agentic ai natural language-based frameworks." *arXiv preprint arXiv:2501.13946* (2025).

[4] Joshi, Satyadhar. "Comprehensive Review of AI Hallucinations: Impacts and Mitigation Strategies for Financial and Business Applications." (2025).

[5] Rogulsky, S., Popovic, N., & Färber, M. (2024). The Effects of Hallucinations in Synthetic Training Data for Relation Extraction. *ArXiv*. https://arxiv.org/abs/2410.08393.

[6] Chen, Jiawei, et al. "Detecting and evaluating medical hallucinations in large vision language models." *arXiv preprint arXiv:2406.10185* (2024).

[7] Das, Anindya Bijoy, Shibbir Ahmed, and Shahnewaz Karim Sakib. "Hallucinations and key information extraction in medical texts: A comprehensive assessment of open-source large language models." *arXiv preprint arXiv:2504.19061* (2025).

[8] Abdelghafour, Mohamed Ali Mohamed, Mohammed Mabrouk, and Zaki Taha. "Hallucination mitigation techniques in large language models." *International Journal of Intelligent Computing and Information Sciences* 24.4 (2024): 73-81.

[9] Luo, Junliang, et al. "Hallucination detection and hallucination mitigation: An investigation." *arXiv preprint arXiv:2401.08358* (2024).

[10] Sun, Yujie, et al. "AI hallucination: towards a comprehensive classification of distorted information in artificial intelligence-generated content." *Humanities and Social Sciences Communications* 11.1 (2024): 1-14.

[11] Amatriain, Xavier. "Measuring and mitigating hallucinations in large language models: amultifaceted approach." 4 Mar. 2024,

[12] Wang, Jue. "Hallucination Reduction and Optimization for Large Language Model-Based Autonomous Driving." *Symmetry* 16.9 (2024): 1196.

[13] Kim, Yubin, et al. "Medical hallucinations in foundation models and their impact on healthcare." *arXiv preprint arXiv:2503.05777* (2025).

[14] Rani, Anku, et al. "Visual hallucination: Definition, quantification, and prescriptive remediations." *arXiv preprint arXiv:2403.17306* (2024).

[15] Jesson, Andrew, et al. "Estimating the hallucination rate of generative ai." *Advances in Neural Information Processing Systems* 37 (2024): 31154-31201.

[16] Sovrano, Francesco, Kevin Ashley, and Alberto Bacchelli. "Toward eliminating hallucinations: Gpt-based explanatory ai for intelligent textbooks and documentation." *CEUR Workshop Proceedings*. No. 3444. CEUR-WS, 2023.

[17] Mündler, Niels, et al. "Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation." *arXiv preprint arXiv:2305.15852* (2023).

[18] McIntosh, Timothy R., et al. "A culturally sensitive test to evaluate nuanced gpt hallucination." *IEEE Transactions on Artificial Intelligence* 5.6 (2023): 2739-2751.

[19] Zhang, Yue, Leyang Cui, and Shuming Shi. "Alleviating hallucinations of large language models through induced hallucinations." *Findings of the Association for Computational Linguistics: NAACL 2025*. 2025.

[20] Chang, Yue, et al. "A unified hallucination mitigation framework for large vision-language models." *arXiv preprint arXiv:2409.16494* (2024).

[21] Hatem, Rami, Brianna Simmons, and Joseph E. Thornton. "A call to address AI “hallucinations” and how healthcare professionals can mitigate their risks." *Cureus* 15.9 (2023).

[22] Kim, Sihwan, et al. "Automated Audit and Self-Correction Algorithm for Seg-Hallucination Using MeshCNN-Based On-Demand Generative AI." *Bioengineering* 12.1 (2025): 81.