# A Trustworthy Federated Framework for Automated Medical Reporting via Two-Stage OCR and NLI-Based Risk Management

Yu-Chih Wei[1], Wei-Shan Lin[1,*] and Chia-Yun Hsu[1]

[1]National Taipei University of Technology, Taipei, Taiwan

**Abstract**

We propose a federated framework for automated ultrasound reporting that integrates two-stage Optical Character Recognition (OCR), Large Language Models (LLMs), and a Natural Language Inference (NLI) verification mechanism. Using Federated Proximal federated learning, the system enables cross-institutional learning while preserving data privacy and mitigating overfitting. In heterogeneous simulated settings, the federated approach reduced the Small Clinic's overfitted local baseline from 98% to 84% Word Accuracy at the best-performing federated round, while the held-out Medium Hospital node reached 89% Word Accuracy. Furthermore, the NLI module verifies the logical consistency between OCR-extracted evidence and generated text. A sentence-level evaluation of 959 claims derived from 300 clinical images showed that the NLI module detected 47 of 48 expert-annotated hallucinated claims, corresponding to a recall of 97.92%. However, it also produced 230 false positives, indicating a substantial physician review burden. This study presents an auditable prototype framework for trustworthy OCR-assisted ultrasound report drafting under heterogeneous clinical data settings.

**Keywords**

Federated Learning, Trustworthy AI, Two-Stage OCR, Natural Language Inference (NLI), Medical Imaging, Overfitting

## 1. Introduction

Since many legacy ultrasound systems lack direct DICOM export, clinical findings are manually transcribed. This resource-intensive process is prone to fatigue-induced errors, potentially compromising patient safety.

While Generative Artificial Intelligence (GenAI) and Large Language Models (LLMs) offer opportunities for automating such reports, their deployment in clinical workflows faces two key challenges. First, sensitive medical data is often restricted by privacy regulations, such as the Personal Data Protection Act (PDPA) of Taiwan [1] and the General Data Protection Regulation (GDPR) [2], leading to data silos. This limits training data availability in small-scale clinics and results in severe overfitting and poor generalization across diverse ultrasound device layouts [3, 4]. Second, LLMs may generate hallucinations, i.e., factually incorrect yet linguistically plausible statements [5], which reduce clinical reliability and increase the burden of expert review.

To address these challenges, we propose a federated framework for automated ultrasound reporting that integrates Optical Character Recognition (OCR)-based data extraction with Natural Language Inference (NLI)-based verification.

- **Federated Two-Stage OCR Pipeline:** We deploy a two-stage OCR pipeline across heterogeneous clinical nodes, including Large and Small hospital settings, while introducing a Medium Hospital as an unseen validation node for generalization evaluation. The pipeline consists of Character Region Awareness for Text Detection (CRAFT) [6] for text localization, YOLOv8 [7] for clinical symbol detection, and a Convolutional Recurrent Neural Network with Connectionist Temporal Classification (CRNN-CTC) [8, 9] for sequence recognition. Using the Federated

*TRUST-AI: The Second European Workshop on Trustworthy AI. Organized as part of the International Joint Conference on Artificial Intelligence - IJCAI/ECAI 2026. August 2026, Bremen, Germany.*

*Corresponding author.

vickrey@mail.ntut.edu.tw (Y. Wei); kellylin1050@gmail.com (W. Lin); annehsu0408@gmail.com (C. Hsu)

Proximal (FedProx) Federated Learning (FL) algorithm, the system aggregates cross-institutional knowledge without centralizing raw data, reducing local overfitting in data-scarce settings.

- **NLI-Based Risk Management Guardrail:** To improve the reliability of generated reports, we implement an NLI verification loop [10]. The OCR-extracted data are treated as the clinical premise, while generated report sentences are treated as hypotheses. The system evaluates whether each statement is supported by the extracted evidence and flags unsupported claims for human review.
- **End-to-End Trustworthy Assessment:** We evaluate the complete workflow. Results demonstrate that the federated approach successfully regularizes severely overfitted local models, achieving 84% stabilized Word Accuracy in sample-scarce settings and 89% generalization on an unseen node. Evaluating 959 claims from 300 images further demonstrates the effectiveness of the NLI guardrail in flagging unsupported statements. Here, trustworthiness is operationalized through privacy-preserving federated learning, NLI-based evidence verification, and human-in-the-loop (HITL) auditability.

## 2. Related Work

### 2.1. Federated Learning in Medical Imaging

FL addresses the "data silo" problem while ensuring privacy compliance (e.g., GDPR/PDPA). A primary challenge in cross-institutional settings is the "domain shift" caused by heterogeneous devices and UI variability [3]. Since non-IID data distributions often lead to model instability and severe overfitting in small-scale clinics [4], our study employs the FedProx algorithm to ensure robust weight aggregation across diverse nodes.

### 2.2. Two-Stage OCR for Clinical Scenes

Modern Scene Text Recognition (STR) decouples text localization from recognition. Algorithms like CRAFT identify character regions using heatmaps [6]. Integrating YOLOv8 is essential to distinguish medical markers from alphanumeric text, which traditional engines often miss [7]. Our framework combines CRAFT and YOLOv8 for localization, followed by a CRNN-CTC network for sequential recognition [8, 9].

### 2.3. Hallucination Mitigation via NLI

Generative hallucinations remain a critical barrier to clinical LLM adoption [5]. While Retrieval-Augmented Generation (RAG) grounds models in factual evidence, it lacks a formal verification step to ensure accurate data reflection [11]. NLI evaluates logical consistency by classifying the relationship between source data (premise) and generated claims (hypothesis) [10]. This study integrates NLI with the TAIDE model [12] to provide a verifiable and auditable clinical reporting workflow.

## 3. System Methodology

To address the dual challenges of data privacy and generative hallucinations, we propose a dual-guardrail architecture. The framework consists of a federated two-stage OCR pipeline and an NLI-based verification loop to support clinical accountability. The overall system architecture is illustrated in Figure 1.

### 3.1. Federated Two-Stage OCR Pipeline

Medical ultrasound images contain critical clinical measurements. To comply with data protection laws and mitigate local overfitting in data-scarce clinics, we implement a FL approach utilizing the FedProx

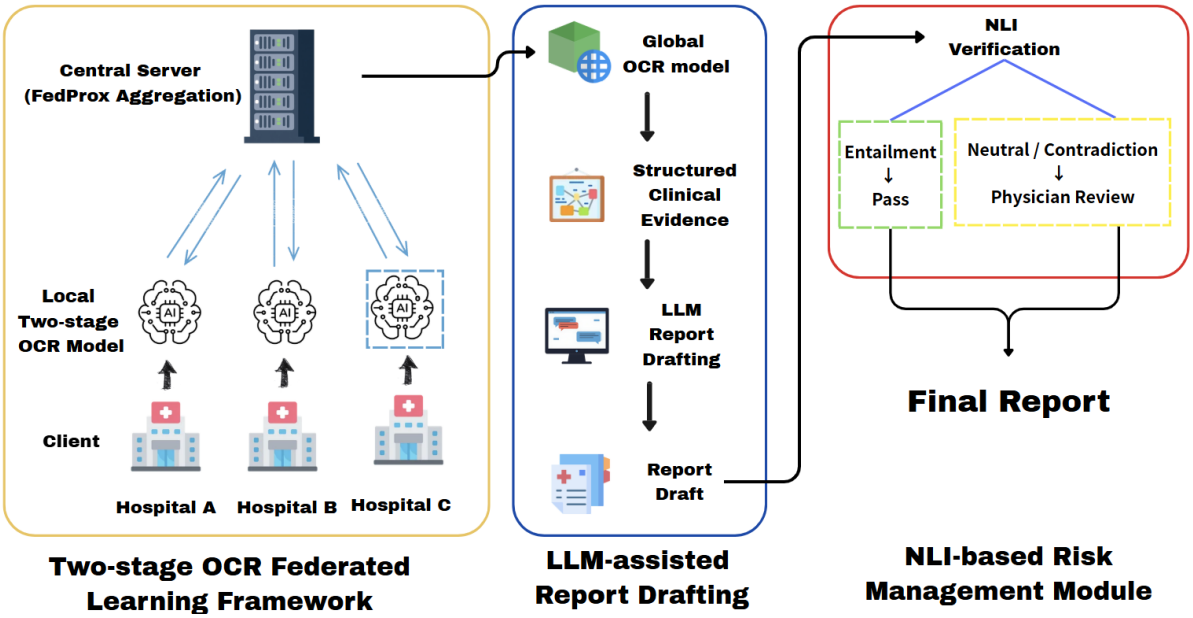


**Figure 1:** Overview of the proposed trustworthy federated framework for automated ultrasound report drafting. Clinical nodes train local two-stage OCR models, which are aggregated through FedProx to reduce overfitting in data-scarce settings. The extracted OCR evidence supports LLM-based report drafting, while the NLI module flags unsupported statements for physician review before the final report is produced.

algorithm [13] to aggregate knowledge across heterogeneous clinical nodes without centralizing raw data.

Our setup simulates three network nodes representing diverse environments: a Large Hospital (primary data source), a Small Clinic (sample scarcity), and a Medium Hospital (unseen distribution shifts).

At each local node, we deploy a two-stage trainable OCR pipeline optimized for heterogeneous ultrasound interfaces:

1. **Region Localization:** We utilize the CRAFT algorithm [6] to dynamically locate character regions amidst complex ultrasound background noise. Concurrently, a YOLOv8 model [7], fine-tuned on manually annotated ultrasound images, detects clinic-specific symbols (e.g., measurement cursors) that traditional OCR engines miss.
2. **Sequence Recognition:** The localized text crops are normalized and fed into a CRNN-CTC [8, 9] network. This network extracts sequential features and translates them into clinical abbreviations and numerical values.

During the federated communication rounds, each participating training node (Large Hospital and Small Clinic) sends its local weight updates ($\Delta w_i$) to a central server for FedProx-based aggregation. Through this process, the sample-scarce Small Clinic can benefit from the feature representations learned by the Large Hospital node, helping improve the generalization of the global OCR model.

### 3.2. NLI-Based Risk Management Guardrail

While the federated OCR pipeline provides privacy-preserving data extraction, the subsequent LLM-based report generation (e.g., TAIDE [12]) introduces the risk of hallucinated statements. Such unsupported inferences reduce clinical reliability and increase the burden of expert validation. To mitigate this risk, we implement a post-hoc NLI risk management loop [10, 14].

The verification process is formalized as follows:

- **Premise ($P$):** The OCR-extracted medical data.

- **Hypothesis ($H$):** The LLM-generated diagnostic sentences.

The NLI module evaluates the logical relationship between $P$ and $H$, classifying each statement into three categories: *Entailment* (the hypothesis is supported by the premise), *Neutral* (the hypothesis introduces unverified information), or *Contradiction* (the hypothesis conflicts with the premise). Outputs classified as *Neutral* or *Contradiction* are flagged for mandatory human review to ensure unverified content is evaluated by an expert. We adopt a general-domain NLI model (XLM-RoBERTa fine-tuned on XNLI) as a deployable baseline without resource-intensive domain retraining. For practical deployment, the localized interface (Figure 2) explicitly separates OCR-extracted conclusions from AI-generated recommendations, triggering a HITL review mechanism for NLI-flagged content to ensure clinical accountability.


**Figure 2:** Overview of the automated ultrasound diagnostic report interface. Key translated UI components include: (1) Diagnostic Conclusion: Indicates the OCR-extracted finding, such as "Mild Fatty Liver (FLM)". (2) AI-Generated Recommendation: Displays the LLM-drafted patient education and follow-up advice grounded in the clinical findings.

## 4. Experimental Results and Analysis

We evaluate the proposed framework across two primary dimensions: (1) the federated OCR pipeline's ability to mitigate local overfitting and improve cross-device generalization, and (2) the NLI guardrail's ability to flag potentially unsupported generated statements for human review.

### 4.1. Experimental Setup and Datasets

To simulate a heterogeneous clinical network, we deployed the federated framework across three virtualized nodes with varying data distributions. The overall dataset comprises **300** ultrasound images. Across all nodes, the data was partitioned into training (80%), validation (10%), and testing (10%) sets.

- **Large Hospital (Medical Center):** Utilizes a high-quality dataset (80% of total data) with diverse machine layouts, acting as the primary feature extractor.
- **Small Clinic (Local Hospital):** Simulates a resource-constrained environment using a publicly available liver ultrasound dataset from Kaggle [15]. Lacking text annotations, we constructed proxy textual labels using clinical terminology to simulate OCR pretraining conditions for representation learning, rather than as clinically validated ground truth. This node is highly susceptible to overfitting.

- **Medium Hospital (Validation Node):** To evaluate the system's adaptability to unseen data distributions, we construct a third node by combining samples from both the large and small hospital datasets, complemented by data augmentation techniques such as noise injection and geometric transformations. This node simulates intermediate data characteristics and is used solely for generalization evaluation. Crucially, this node does not participate in the federated training process and is utilized purely as an independent, held-out evaluation node to assess cross-distribution generalization.

We define Round 0 as the pre-federation local evaluation stage. Starting from Round 1, local updates are aggregated via FedProx. Despite an initial performance drop due to non-IID data distributions, the global model gradually adapts and improves its generalization across subsequent communication rounds.

The federated OCR training employed the FedProx algorithm for a total of 13 communication rounds, with local epochs set to 10. We evaluate the OCR performance comprehensively using Word Accuracy, Character Error Rate (CER), Precision, Recall, and F1-Score. To assess the framework at its best observed state, we report the performance at the best-performing round (Round 11), where the validation performance reached its peak among the 13 communication rounds. A slight performance decrease was observed in the subsequent rounds. Specifically, CER is calculated based on the Levenshtein distance:

$$\text{CER} = \frac{S + D + I}{N}, \tag{1}$$

where $S$, $D$, and $I$ represent the number of substitutions, deletions, and insertions required to match the ground truth, and $N$ is the number of characters in the ground truth.

## 4.2. Mitigating Overfitting via Federated Aggregation

In traditional localized training, the sample-scarce Small Clinic suffered from severe overfitting. Constrained by its limited training set, the local model simply memorized the local data, achieving an inflated initial baseline Word Accuracy of 98% and a near-zero Character Error Rate (CER) at Round 0. However, this perfectly overfitted local model lacked the ability to generalize across varying font styles and interface noises typical of different ultrasound manufacturers. We explicitly compare against this local-only training to illustrate how federated aggregation acts as a regularizer.

By integrating the FedProx FL mechanism, the global model bridges the knowledge gap across heterogeneous training nodes and penalizes local memorization. As illustrated in Figure 3, the global model demonstrates a progressive generalization trend across 13 communication rounds. After an initial sharp drop at Round 1—reflecting the severe non-IID data mismatch when aggregation begins—the global model gradually recovers and adapts to shared representations. The model achieved its best observed validation performance at Round 11, after which a slight performance degradation suggests that early stopping may be necessary to prevent over-adapting to local noise.

To further quantify this regularization effect, Figure 4 compares the initial local baseline (Round 0) against the optimally aggregated global model (Best Round 11). After federated aggregation, the Small Clinic's Word Accuracy reached 84% at the best-performing round, accompanied by an F1-Score of 85%. Although this represents a numerical decrease from its initial overfitted peak of 98%, the result is consistent with reduced local memorization under the simulated federated setting. This suggests that FedProx aggregation helped reduce reliance on localized noise patterns and encouraged more transferable feature representations.

The best-performing federated model reached 89% Word Accuracy on the held-out Medium Hospital node. Since the Medium Hospital node did not participate in federated training, this result suggests potential cross-distribution generalization in the simulated heterogeneous setting. However, further validation with real multi-institutional data and external benchmarks is required.

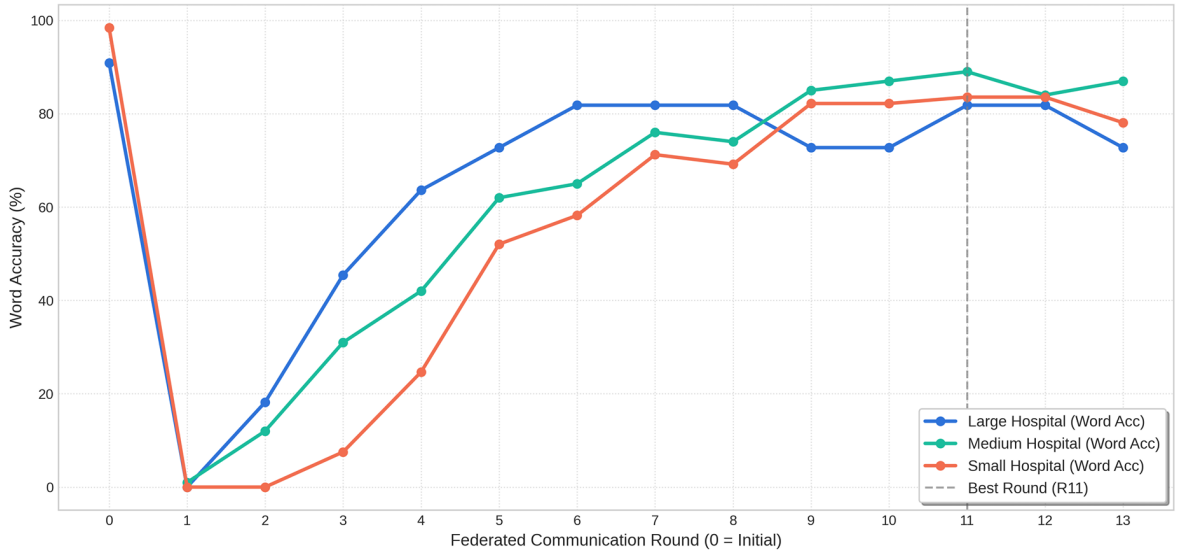


**Figure 3:** Global Model Generalization: Word Accuracy Trend across 13 federated communication rounds. Round 0 represents local pre-federation performance (severe overfitting in the Small Clinic). The sharp drop at Round 1 reflects the initial non-IID mismatch after aggregation, followed by gradual recovery in validation performance.

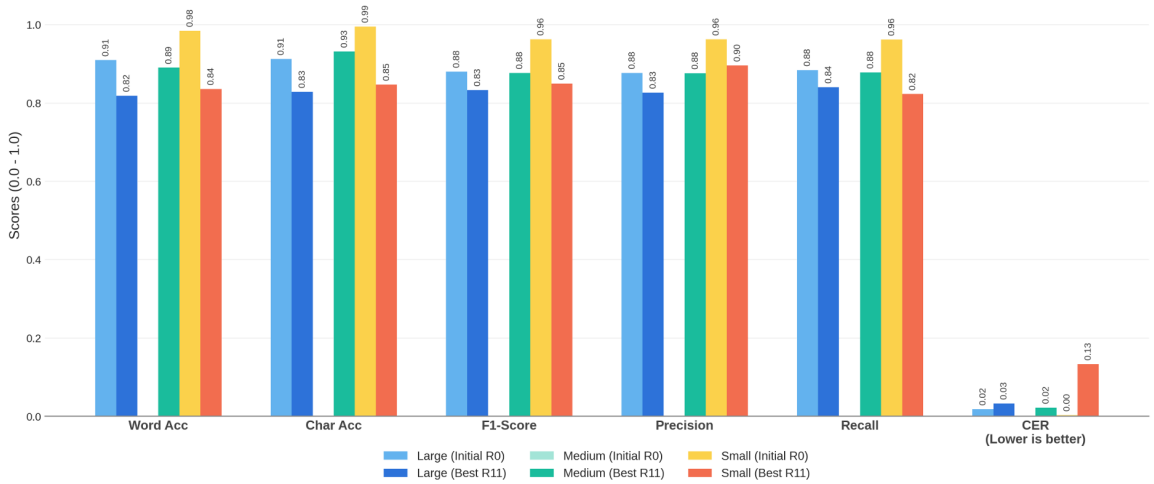


**Figure 4:** Performance Comparison (Initial vs. Best Round 11). The federated approach reduces the Small Clinic's overfitted local baseline and leads to more stable validation performance at the best-performing federated round.

### 4.3. Quantitative and Qualitative Assessment via NLI Guardrail

To evaluate the NLI guardrail's effectiveness, we conducted a sentence-level quantitative analysis. We extracted 959 LLM-generated claims derived from 300 clinical ultrasound images. The 959 claims were validated by a senior radiologist to establish the ground truth for hallucination detection. Annotations involved sentence-by-sentence comparisons between the federated OCR evidence and the LLM-generated reports to ensure clinical reliability.

We compared expert annotations with NLI predictions to assess hallucination interception. As shown in Table 1, among 959 generated sentences, the baseline LLM produced 48 hallucinated claims. The NLI module flagged 47 of these hallucinated claims as unsupported, corresponding to a recall of 97.92%. One hallucinated claim was not flagged by the guardrail and was therefore counted as a false negative. However, this configuration produced 230 false positives (16.97% precision), meaning many supported claims were still routed for physician review, highlighting a clear trade-off between minimizing missed hallucinations and increasing workload. Therefore, the current NLI module should be interpreted as a

high-recall screening tool rather than an automatic acceptance mechanism. Its main role is to reduce missed hallucinations, but this comes at the cost of lower precision and additional human review.

**Table 1**
Confusion Matrix of NLI Hallucination Interception over 959 Claims.

| | **Expert: Hallucinated** | **Expert: Supported** |
|---|---|---|
| **NLI: Flagged for Review** | True Positive: 47 | False Positive: 230 |
| **NLI: Not Flagged** | False Negative: 1 | True Negative: 681 |

Beyond quantitative metrics, we perform a qualitative analysis to understand the NLI module's behavior in clinical scenarios. As shown in Table 2, the system evaluates the logical consistency between the clinical Premise extracted by the OCR and the Hypotheses generated by the LLM.

**Table 2**
NLI Verification Case Studies. These logs illustrate how the NLI module flags unsupported or weakly grounded statements. Translated for publication.

| **LLM Generated Hypothesis** | **NLI Verdict** | **Conf.** | **Action** |
|---|---|---|---|
| Based on the test results, you have been diagnosed with mild fatty liver. | Entailment | 0.999 | **Passed** |
| You recently may feel that there is excessive bowel gas; follow-up is advised. | Neutral | 0.959 | **Flagged** |
| You recently underwent a cholecystectomy and are now in the recovery period. | Neutral | 0.955 | **Flagged** |

These cases illustrate the NLI module's ability to identify unsupported inferences, such as subjective symptoms (Case 2) or unrecorded surgical history (Case 3). Because these claims are not directly supported by the OCR-extracted evidence, they were classified as *Neutral* and routed for HITL review. With an average of approximately 3.2 diagnostic statements per image, the sentence-level evaluation suggests that the proposed guardrail can help screen unsupported claims before physician validation.

Furthermore, we acknowledge an inherent limitation of the pipelined architecture: OCR error propagation remains a challenge. Although our current NLI evaluation utilizes 100% accurate OCR extractions as premises to strictly isolate the hallucination detection performance, in real-world deployments, if the OCR stage extracts incorrect evidence, the downstream NLI module will evaluate against a flawed premise. This emphasizes the absolute necessity of human oversight. To reduce this practical risk, the HITL framework does not rely solely on NLI-triggered flags. Instead, the entire document, including the OCR-extracted clinical findings and AI-generated recommendations, remains fully editable by certified medical professionals, providing an additional layer of oversight in high-stakes clinical scenarios.

## 4.4. Human-in-the-Loop and System Auditability

When a high-risk claim is intercepted by the NLI guardrail, our system refrains from deploying another autonomous AI agent for self-correction, as recursive generation risks introducing secondary hallucinations. Instead, to better align with trustworthy AI principles (e.g., the EU AI Act's requirement for human oversight), the intercepted claims are routed to a strict HITL interface.

As illustrated in Figure 5, the Chinese-language HITL interface highlights NLI-flagged statements using warning icons. These flags correspond to claims classified as *Neutral* or *Contradiction* by the NLI module. Physicians can revise, remove, or confirm the flagged statements based on clinical judgment. Importantly, the interface separates OCR-extracted findings from LLM-generated recommendations, while keeping the full report editable to ensure that final diagnostic authority remains with certified medical professionals.

To track AI-generated errors and subsequent physician corrections, the system provides version control and audit logging. As depicted in Figure 6, deleted AI-generated content is shown with red

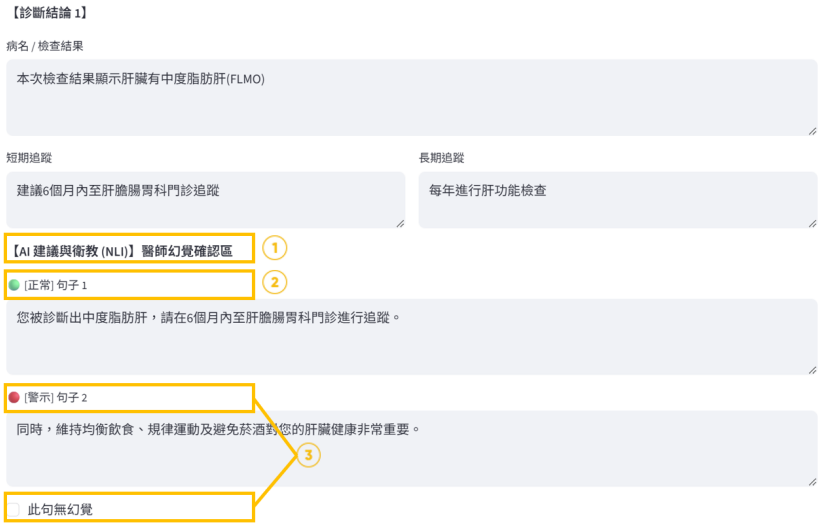


**Figure 5:** The HITL interface with NLI guardrail. (1) Hallucination confirmation zone. (2) Verified claim marker. (3) Warning flag for unsupported statements with manual override.

strikethroughs, while physician-authored edits are highlighted in green. The log also records timestamps and reviewer identity.

These visual indicators in Figure 6 enable reviewers to trace how each AI-generated statement is modified during the clinical validation process and distinguish between AI-generated drafts and human-authored corrections, supporting clinical accountability and auditability.


**Figure 6:** System audit log showing (1) Revision timestamp and physician identity, (2) Version diff legend (red for deleted, green for added text), and (3) Audit trail of precise physician corrections to the AI-generated draft.

## 4.5. System Efficiency and Latency Analysis

To assess practical feasibility, we evaluated end-to-end latency on a local node (dual Intel Xeon CPUs, 128GB RAM, four NVIDIA Tesla T4 GPUs). OCR extraction and NLI verification were executed locally to ensure privacy, while diagnostic reports were generated via a TAIDE-based API built upon a Llama-3.1-405B-Instruct-FP8 model without further fine-tuning.

Compared to approximately 180 seconds required for manual reporting, our pipeline completes the process in 6 seconds per image, with only a 0.4-second overhead introduced by the NLI verification module. This reduces the initial drafting time to under 7 seconds per patient, enabling physicians to shift their focus from manual data entry to clinical validation. These results demonstrate the practical feasibility of integrating the proposed framework into real-world clinical workflows.

## 5. Discussion: Regulatory and Ethical Alignment

The integration of FL and NLI-based verification establishes a multi-layered trust model that aligns with the "Ethics Guidelines for Trustworthy AI" and key requirements under the EU AI Act. Specifically, our framework focuses on:

- **Data Sovereignty:** FedProx ensures sensitive ultrasound pixel data remains within the hospital intranet, adhering to GDPR and local privacy laws.
- **Human Oversight:** GenAI is treated strictly as an augmented decision-support tool. The interactive PDF editor and mandatory NLI-flagging ensure final diagnostic authority remains with the physician.
- **Traceability:** The automated version control provides a detailed audit trail of all AI-generated draft modifications, fulfilling EU AI Act documentation requirements. Furthermore, our risk management logic flags even non-contradictory but ungrounded statements as high-risk, as diagnostic ambiguity could impact clinical accountability.
- **Failure Boundaries:** The system may fail in cases where OCR extraction is incomplete or when clinical statements require multi-hop reasoning beyond sentence-level verification. These limitations highlight the importance of maintaining human oversight in high-risk scenarios.

## 6. Conclusion and Future Work

In this study, we developed a federated reporting framework featuring a dual-guardrail architecture for automated ultrasound report drafting. By implementing the FedProx algorithm, our system addresses the "data silo" problem and mitigates local overfitting in resource-constrained clinics. In the simulated evaluation, the federated approach reduced the overfitted local baseline, with the Small Clinic reaching 84% Word Accuracy at the best-performing round and the held-out Medium Hospital node reaching 89% Word Accuracy. Furthermore, the NLI module flagged 47 of 48 expert-annotated hallucinated claims, corresponding to a recall of 97.92%. However, it also produced 230 false positives, indicating that the current high-recall configuration may increase physician review workload.

Several limitations remain in the current prototype, particularly regarding deployment realism and calibration of NLI confidence scores: the setup relies on simulated nodes rather than real multi-institutional deployment, the sentence-level NLI module does not model inter-sentence dependencies, and the high-risk threshold is implicitly based on classification labels rather than a calibrated probabilistic boundary.

Future work will extend the framework to other imaging modalities (e.g., MRI and X-ray), integrate Hardware Security Modules (HSM) with Fully Homomorphic Encryption (FHE) [16] to provide mathematically provable security against gradient leakage attacks, and implement adaptive early stopping mechanisms [17] to automatically determine sufficient communication rounds, eliminating the need for manual aggregation testing. Overall, this architecture provides a preliminary prototype for auditable OCR-assisted clinical report drafting, while further benchmark comparison, threshold calibration, and real-world validation remain necessary.

## Acknowledgments

This research was partially supported by the National Science and Technology Council (NSTC) of Taiwan under Grant No. NSTC 114-2637-8-027-001. We thank the collaborating medical professionals for their domain expertise and feedback during the system evaluation phase.

## Declaration on Generative AI

*(By using the activity taxonomy in ceur-ws.org/genai-tax.html):*
During the preparation of this work, the author(s) used LLMs to translate the English manuscript, perform grammar checks, and format the LaTeX source code. After using these tools, the author(s) reviewed, edited, and validated the content, taking full responsibility for the publication's content.

## References


[1] Ministry of Justice, Personal data protection act, 2015. URL: https://law.moj.gov.tw/ENG/LawClass/LawAll.aspx?pcode=I0050021, republic of China (Taiwan).

[2] European Union, General data protection regulation (gdpr), 2016.

[3] H. Guan, P.-T. Yap, A. Bozoki, M. Liu, Federated learning for medical image analysis: A survey, Pattern Recognition 151 (2024) 110424.

[4] J. O. du Terrail, et al., Flamby: Datasets and benchmarks for cross-silo federated learning in realistic healthcare settings, in: Advances in Neural Information Processing Systems (NeurIPS), volume 35, 2022, pp. 5315–5334.

[5] Z. Ji, et al., Survey of hallucination in natural language generation, ACM Computing Surveys (2023).

[6] Y. Baek, et al., Character region awareness for text detection, in: Proceedings of the CVPR, 2019, pp. 9365–9374.

[7] G. Jocher, et al., Yolov8 by ultralytics, 2023. URL: https://github.com/ultralytics/ultralytics.

[8] B. Shi, X. Bai, C. Yao, An end-to-end trainable neural network for image-based sequence recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence (2016).

[9] A. Graves, et al., Connectionist temporal classification: labelling unsegmented sequence data, in: ICML, 2006.

[10] A. Williams, et al., A broad-coverage challenge corpus for sentence understanding through inference, in: NAACL-HLT, 2018.

[11] P. Lewis, et al., Retrieval-augmented generation for knowledge-intensive nlp tasks, in: NeurIPS, 2020.

[12] National Science and Technology Council (Taiwan), Taide: Trustworthy ai dialogue engine, 2023. URL: https://taide.tw/, accessed: 2026-04.

[13] T. Li, et al., Federated optimization in heterogeneous networks, in: MLSys, 2020.

[14] Y. Liu, et al., Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019). URL: https://arxiv.org/abs/1907.11692.

[15] Orvile, Annotated ultrasound liver images dataset, https://www.kaggle.com/datasets/orvile/annotated-ultrasound-liver-images-dataset, 2023.

[16] J. Zhang, et al., Secure federated learning via privileged trusted execution environments and homomorphic encryption, IEEE TIFS (2021).

[17] A. Al-Qerem, et al., A novel federated learning framework for medical imaging: Resource-efficient approach combining pca with early stopping, Computers in Biology and Medicine (2025).