PrivNurse AI: Revolutionizing Clinical Documentation with On-Device Intelligence

Wei-Lin Wen, Yu-Yao Tsai

***Abstract*— The healthcare system is drowning in paperwork. Healthcare professionals spend an alarming 57.8% of their scheduled patient time in electronic health records (EHR), creating a documentation burden that contributes to widespread burnout and compromises patient care quality. PrivNurse AI introduces a groundbreaking solution that harnesses Gemma-3n's revolutionary on-device capabilities to transform clinical documentation through three core functionalities: intelligent consultation nursing note summarization, comprehensive discharge note generation, and seamless speech-to-text transcription for nursing records. By operating entirely offline with privacy-first architecture, our system addresses the critical gap between AI innovation and healthcare implementation while maintaining the explainability and trust essential for clinical adoption.**

I. Introduction

The healthcare industry faces an unprecedented documentation crisis that threatens the very foundation of patient care. Electronic health records (EHRs) are increasingly linked with documentation burden resulting in clinician burnout [1], creating a vicious cycle where tools designed to improve healthcare efficiency have become barriers to quality care. Recent studies reveal that physicians across all specialties spent 3.4 hours per eight hours of scheduled patient time in the EHR during patient scheduled hours [1], effectively transforming healthcare professionals into data entry clerks rather than caregivers.

This documentation burden extends beyond mere inconvenience—it represents a systemic threat to healthcare sustainability. Physician burnout rates have reached critical levels, with documentation workload being identified as a primary contributing factor. The ripple effects cascade through the entire healthcare ecosystem: reduced patient interaction time, increased medical errors due to fatigue, and a mass exodus of healthcare professionals from the profession.

A. The Promise and Peril of AI in Healthcare

While generative AI offers unprecedented potential to alleviate documentation burdens, existing solutions face critical limitations in healthcare environments. Cloud-based AI systems, despite their capabilities, introduce privacy concerns that are incompatible with HIPAA requirements and patient confidentiality standards. The medical field demands not just accurate AI assistance, but solutions that operate within the strict privacy and security frameworks that protect patient data.

Furthermore, interpretability and explainability are crucial for establishing user trust, ensuring legal and ethical compliance, and fostering broader social acceptance in medical environments [2]. Physicians are more inclined to trust and implement AI in clinical practice if they perceive the results as being more explainable or comprehensible [2], highlighting the critical need for transparent AI systems in healthcare applications.

B. Our Innovation: Breaking the Privacy-Performance Paradox

PrivNurse AI represents a paradigm shift in healthcare AI implementation by leveraging Gemma-3n's groundbreaking architecture to deliver enterprise-grade AI capabilities entirely on-device. This approach eliminates the traditional trade-off between AI sophistication and privacy protection, enabling healthcare institutions to harness cutting-edge AI while maintaining complete data sovereignty.

Our system addresses three critical clinical documentation challenges that consume the majority of healthcare professionals' time:

1) **Consultation nursing note summarization** - transforming complex consultation records into concise, actionable summaries
2) **Discharge note generation** - synthesizing comprehensive patient records into structured discharge summaries
3) **Voice-to-text transcription** - enabling hands-free nursing documentation through advanced speech recognition

II. Related Work

A. Healthcare Language Models and Clinical Applications

The integration of large language models in healthcare has gained significant momentum, with specialized models demonstrating superior performance on medical benchmarks compared to general-purpose alternatives. Notable developments include domain-specific architectures like GatorTron and Med-PaLM, which have shown proficiency in clinical language understanding and medical reasoning tasks [3], [4]. These systems excel in diverse healthcare applications including clinical entity recognition, medical relation extraction, and diagnostic reasoning [5]. Despite these advances, most healthcare LLM implementations depend on cloud-based infrastructure, creating substantial barriers for clinical adoption due to privacy regulations and data security requirements [6]. The need for HIPAA-compliant solutions has intensified interest in local deployment strategies, though comprehensive systems specifically tailored for nursing workflows remain scarce.

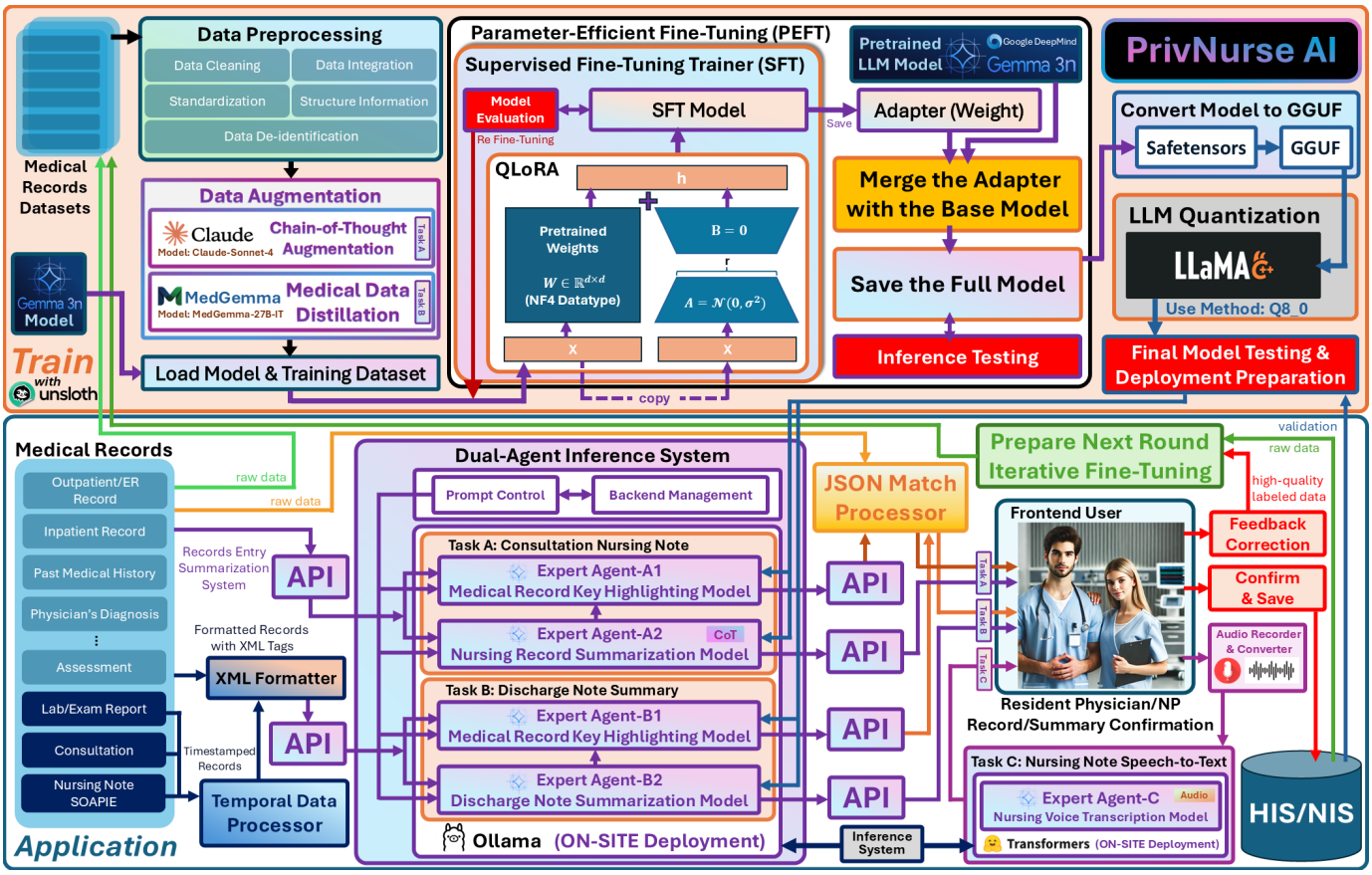

Fig. 1. PrivNurse AI System Architecture Overview

B. Reasoning Enhancement in Medical AI Systems

Chain-of-Thought prompting has emerged as a critical technique for improving LLM performance in complex medical reasoning scenarios. This approach proves particularly valuable in healthcare applications due to the intricate nature of clinical decision-making processes that require logical, sequential analysis [7]. Recent research has demonstrated that structured reasoning approaches can significantly enhance both accuracy and interpretability in medical question-answering tasks [8]. The development of medical-specific reasoning frameworks, such as hierarchical expert verification systems, has shown promise in biomedical applications by improving both transparency and clinical accuracy [9]. However, existing methodologies primarily target diagnostic tasks rather than clinical documentation generation, highlighting an opportunity for specialized approaches in nursing documentation contexts.

C. Privacy-Preserving Healthcare AI Deployment

Healthcare institutions increasingly prioritize on-premises AI deployment to maintain data sovereignty and regulatory compliance. Local deployment strategies ensure patient information remains within institutional firewalls, facilitating adherence to privacy regulations including HIPAA and GDPR while mitigating risks associated with external cloud services [10]. This approach, however, traditionally requires substantial computational resources and specialized infrastructure management. Recent developments in model optimization techniques, including quantization methods like QLoRA, have made local deployment more accessible for healthcare organizations with limited computational resources [11]. These innovations enable the deployment of capable medical AI systems on standard hardware configurations while preserving clinical performance requirements.

D. Clinical Documentation Automation Systems

Automated clinical documentation technologies have demonstrated potential for reducing administrative burdens on healthcare professionals, enabling greater focus on direct patient care activities. Current approaches encompass voice recognition systems, predictive text generation, intelligent template completion, and natural language processing applications [12], [13]. These systems aim to streamline documentation workflows while maintaining clinical accuracy and completeness. However, existing documentation automation solutions often lack the sophisticated reasoning capabilities required for complex clinical scenarios, particularly in nursing contexts where nuanced clinical judgment is essential. The integration of advanced language model capabilities with clinical documentation workflows represents a promising direction for addressing these limitations while preserving the accuracy standards critical for medical applications [14].

III. Technical Challenges and Innovations

A. Challenge 1: Clinical Summarization Complexity

Unlike general text summarization, clinical summarization requires deep domain expertise and the ability to identify critical medical information while maintaining clinical accuracy. Consultation nursing notes present a high level of complexity, as they must accurately capture the most urgent and important reasons for consultation while providing concise, actionable summaries for clinical decision-making.

Traditional summarization approaches fail in clinical contexts because they lack the medical knowledge necessary to distinguish between routine information and critical clinical details. For instance, the presence of specific pathogenic bacteria in wound cultures directly influences antibiotic selection, making this information essential rather than optional in clinical summaries. A concrete example illustrating this distinction is provided in Table III, which highlights how critical details may be overlooked by generic summarization methods.

1) Our Innovation: Chain-of-Thought Medical Reasoning: We developed a novel approach called Medical Structured Chain-of-Thought (MedSCoT) that enables Gemma-3n to simulate the cognitive process of experienced nurses. Through our Chain-of-Thought augmentation methodology, we use Claude-Sonnet-4 as a teacher model to generate reasoning chains that demonstrate how to identify consultation priorities and construct clinical summaries.

The MedSCoT approach works by decomposing clinical reasoning into structured steps:

1) ***Clinical Context Analysis***: Understanding the patient's medical history and current condition
2) ***Consultation Priority Identification***: Determining the primary reason for specialist consultation
3) ***Critical Information Extraction***: Identifying key diagnostic results, treatments, and recommendations
4) ***Summary Generation***: Constructing concise, actionable summaries that preserve essential clinical details

This approach allows our fine-tuned Gemma-3n models to first reason through the clinical situation before generating summaries, dramatically improving both accuracy and explainability. The visible reasoning chains provide transparency that builds trust among healthcare professionals and enables educational value for junior staff.

B. Challenge 2: Privacy-First AI Deployment

Healthcare environments demand absolute privacy protection, making cloud-based AI solutions impractical for sensitive clinical data. Traditional on-device AI deployment faces significant resource constraints that limit model capabilities, creating a fundamental tension between AI sophistication and privacy requirements.

The healthcare industry's strict regulatory environment, including HIPAA compliance in the United States and similar privacy regulations globally, requires that patient data remain within controlled environments. Any solution that transmits patient information to external servers, even with encryption, faces significant regulatory and institutional barriers to adoption.

1) Our Innovation: Optimized On-Device Architecture: Gemma-3n's revolutionary architecture, featuring Per-Layer Embeddings (PLE) and mix-n-match capabilities, enables us to deploy sophisticated AI models with resource footprints comparable to much smaller models. The Gemma-3n-E4B (8B parameter) model has a memory footprint comparable to a 4B model, making it practical for deployment on standard hospital workstations.

We further optimize deployment through several technical integrations:

1) ***Model Quantization***: We implement INT8 quantization techniques that reduce VRAM requirements by up to 37.5%, without compromising clinical accuracy. Our quantization approach preserves the precision of medical terminology and clinical reasoning pathways that are critical for healthcare applications.
2) ***Unsloth-Optimized Fine-tuning***: The Unsloth framework enables us to reduce training time by up to 1.5x while using 50% less VRAM during the fine-tuning process. This optimization makes it feasible for healthcare institutions to customize models for their specific clinical workflows and terminology.
3) ***Dynamic Resource Allocation***: Our Ollama-based deployment includes intelligent resource management that monitors system utilization and automatically schedules model usage based on current VRAM availability. This ensures optimal performance across all four specialized models while maintaining system stability.
4) ***Hybrid Architecture Design***: One major challenge we encountered was the lack of support for Gemma-3n's multimodal capabilities in Ollama. To address this, we implemented a hybrid architecture using Transformers with FastAPI deployment, enabling speech-to-text functionality while maintaining on-device processing requirements.

C. Challenge 3: Multimodal Healthcare Applications

Healthcare documentation increasingly requires multimodal capabilities, particularly for voice-driven workflows that can liberate healthcare professionals from keyboard-intensive documentation tasks. The ability to process audio input is especially critical in clinical environments where hands-free operation is essential for maintaining sterile conditions and enabling bedside documentation.

1) Our Innovation: Integrated Multimodal Pipeline: While Ollama doesn't support Gemma-3n's multimodal capabilities, we developed a hybrid architecture that unlocks the full potential of speech-to-text functionality for healthcare applications. Our system includes several key innovations:

Instruction-Based Prompt Engineering: We designed a highly structured system prompt that provides Gemma-3n with clear, explicit instructions tailored for the clinical environment. Instead of providing lengthy background descriptions, our strategy focuses on defining the AI's role and expected behavior. The prompt explicitly states that the speaker is a nursing professional and sets clear expectations for handling medical terminology. Furthermore, it establishes rules for post-processing, such as correcting minor speech errors, removing

filler words, and maintaining professional language, ensuring the output is not merely a raw transcript but a clean, structured, and clinically useful document.

Real-time Processing: This streamlined prompting strategy, combined with our optimized deployment, enables real-time transcription that keeps pace with natural speech patterns. This allows nurses to document patient care activities without interrupting their workflow.

IV. Methodology

A. Ethical Approval and Data Collection Framework

The foundation of our model is a high-quality dataset derived from real-world clinical practice under strict ethical oversight. This study received comprehensive ethical approval from the Institutional Review Board of Kuang Tien General Hospital (IRB no.: KTGH 1135; approved on June 24, 2024). The IRB approval specifically authorized the retrospective collection of de-identified medical records for the purpose of developing and evaluating AI-assisted clinical documentation systems.

The ethical framework governing our data collection included several key components:

1) ***Data Scope and Timeline:*** The IRB approval covered the collection of nursing consultation record-summary pairs from the hospital's Health Information System (HIS) spanning from January 2021 to March 2025.
2) ***De-identification Protocol:*** All patient data underwent a rigorous de-identification process following HIPAA Safe Harbor guidelines and local privacy regulations. Personal identifiers, including patient names, medical record numbers, dates of birth, addresses, and contact information, were systematically removed or replaced with synthetic identifiers before any data processing.
3) ***Data Security Measures:*** The approved protocol mandated that all data processing occur within secure, on-premises infrastructure with no external data transmission. Access to the raw dataset was restricted to authorized research personnel who completed mandatory data privacy training.
4) ***Clinical Staff Consent:*** For the deployment phase evaluation, informed consent was obtained from all participating nursing staff (n=39), with explicit approval for workflow observation and satisfaction assessment.

B. System Architecture

Our comprehensive system architecture consists of two main components: the training pipeline that creates specialized clinical AI models, and the application architecture that deploys these models in clinical environments.

1) Training Pipeline: The training pipeline transforms raw medical records into specialized clinical AI models through a multi-stage process designed to ensure clinical accuracy, safety, and regulatory compliance.

Stage 1: Data Preprocessing

The preprocessing stage implements comprehensive data cleaning and standardization procedures:

- ***Data Cleaning***: Removal of incomplete records to ensure data integrity and prevent downstream processing errors.
- ***Data Integration***: Consolidation of multi-source medical records from different hospital departments and systems, ensuring temporal consistency and data completeness
- ***Standardization***: Implementation of uniform formatting across different record types, including consistent date formats, medication dosing notation, and laboratory value representations
- ***Structure Information***: Extraction and organization of key clinical data points using natural language processing techniques specifically designed for medical text
- ***Data De-identification***: Removal of all personally identifiable information (PII), such as patient names, identification numbers, and contact details, to ensure data is fully de-identified and compliant with privacy regulations.

Stage 2: Intelligent Data Augmentation

We implement two sophisticated augmentation strategies tailored to the specific requirements of each clinical task:

Chain-of-Thought Augmentation for Task A (Consultation Notes): We implement a sophisticated teacher-student framework where Claude-Sonnet-4 serves as the expert teacher, generating reasoning chains that demonstrate how experienced nurses analyze consultation priorities. This augmented data enables Gemma-3n to learn not just what to summarize, but how to think through the clinical reasoning process.

The augmentation process involves several steps:

1) Original consultation records are presented to Claude-Sonnet-4 with detailed prompts that simulate the cognitive process of experienced nurses
2) The teacher model generates structured reasoning chains that identify key clinical priorities, diagnostic findings, and treatment recommendations
3) These reasoning chains are paired with high-quality summaries that demonstrate optimal clinical documentation practices
4) The augmented dataset combines original records, reasoning chains, and target summaries to train Gemma-3n in both clinical reasoning and summary generation

Medical Data Distillation for Task B (Discharge Summaries): Using MedGemma-27B-IT as a teacher model, we generate comprehensive discharge summaries that encompass the five critical components required for complete clinical documentation:

1) ***Primary diagnosis***: Principal condition addressed during hospitalization
2) ***Laboratory reports***: Key diagnostic test results and trends
3) ***Medications***: Current prescriptions, dosages, and medication changes
4) ***Consultations***: Specialist recommendations and follow-up requirements
5) ***Follow-up care plans***: Post-discharge instructions and monitoring requirements

The distillation process ensures that Gemma-3n learns to generate discharge summaries that meet clinical standards

TABLE I
HYPERPARAMETERS FOR QLORA-BASED SUPERVISED FINE-TUNING.

	Expert Agent-A1	Expert Agent-A2	Expert Agent-B1	Expert Agent-B2
Base Model	unsloth/ gemma-3n-E4B-it	unsloth/ gemma-3n-E4B-it	unsloth/ gemma-3n-E4B-it	unsloth/ gemma-3n-E4B-it
LoRA Rank (r)	32	32	32	32
LoRA Alpha (α)	64	64	64	32
LoRA Dropout	0	0	0	0
Target Modules	all-linear	all-linear	all-linear	all-linear
Bias	None	None	None	None
Quantization Bits	4-bit	4-bit	4-bit	4-bit
Optimizer	adamw_torch_fused	adamw_torch_fused	adamw_torch_fused	adamw_torch_fused
Max Learning Rate	1e-3	1e-3	2e-4	2e-4
LR Scheduler Type	linear	linear	linear	linear
Warmup Ratio	0.03	0.03	0.03	0.03
Max Gradient Norm	0.3	0.3	0.3	0.3
Num Train Epochs	6	6	2	1
Effective Batch Size	96	96	32	32
Max Sequence Length	8192	8192	32768	32768

while maintaining the compact efficiency required for on-device deployment.

Stage 3: Model Fine-tuning

Our fine-tuning process leverages cutting-edge optimization techniques to maximize both clinical accuracy and computational efficiency:

Parameter Efficient Fine-tuning (PEFT): We employ QLoRA techniques to minimize VRAM requirements while maintaining training effectiveness. The approach uses 4-bit quantization during training, enabling fine-tuning of large models on single consumer GPUs while preserving the precision necessary for clinical applications.

The Unsloth optimization framework provides additional benefits:

- 1.5x faster training compared to standard implementations
- 50% reduction in VRAM usage during training
- Maintained model accuracy through optimized attention mechanisms
- Streamlined gradient computation that preserves clinical reasoning capabilities

Stage 4: Model Deployment Optimization

The final stage prepares models for clinical deployment through several optimization steps:

- **Model Merging**: Integration of LoRA adapters with base Gemma-3n-E4B model using carefully calibrated merge ratios that preserve both base model capabilities and specialized clinical knowledge
- **Format Conversion**: Transformation from SafeTensors to GGUF format for Ollama compatibility, including metadata preservation for model versioning and clinical validation tracking
- **Quantization**: We applied Q8_0 quantization to compress the model and successfully reduced VRAM usage by 37.5%, from 16GB to 10GB, without compromising clinical accuracy.
- **Validation**: Comprehensive testing using held-out clinical datasets to ensure that optimizations do not compromise clinical accuracy or safety

2) Application Architecture: The application architecture deploys our specialized models in clinical environments through a sophisticated multi-component system designed for reliability, scalability, and clinical integration.

Data Preparation and Processing

Our system processes comprehensive medical records through several specialized components:

Medical Records Integration: The system handles diverse medical record types including:

- Outpatient and Emergency Room records with chief complaints and initial assessments
- Inpatient records documenting hospitalization progress and daily clinical notes
- Past medical history including chronic conditions, surgical history, and medication allergies
- Physician diagnoses and assessments with ICD-10 coding integration
- Laboratory and examination reports with temporal organization and trend analysis
- Consultation records from specialist physicians with recommendations and follow-up plans
- Nursing SOAPIE notes documenting patient care activities and responses

Temporal Data Processing: For Task B (Discharge Note Summary), we implement sophisticated temporal data processing to chronologically organize time-stamped medical data. The temporal processor:

- Identifies and extracts timestamps from various medical record formats
- Creates chronological sequences that reflect the actual progression of patient care
- Maintains temporal relationships between related clinical events
- Ensures that discharge summaries accurately reflect the sequence of clinical decision-making

XML Formatting: We employ structured XML tagging to format medical data, enabling Gemma-3n to better understand and process complex medical records with clear hierarchical relationships. The XML formatter:

- Applies semantic tags that identify clinical concepts and relationships
- Structures data according to clinical documentation standards
- Preserves hierarchical relationships between clinical observations and interventions
- Enables precise extraction of specific clinical information during processing
- Facilitates quality assurance and audit trail maintenance

Dual-Agent Inference System

Our innovative dual-agent architecture maximizes both accuracy and explainability through specialized model cooperation. The system operates four specialized models deployed on the Ollama platform with intelligent resource management.

Task A: Consultation Nursing Note Generation

The consultation note generation system employs two specialized agents:

- **Expert Agent A1**: Medical Record Key Highlighting Model trained to identify relationships between clinical summaries and source documentation
- **Expert Agent A2**: Nursing Record Summarization Model specialized in generating clinically accurate consultation summaries

The workflow operates through the following sequence:

1) Agent A2 analyzes the consultation record and generates a clinical summary using MedSCoT reasoning
2) The generated summary is transmitted to Agent A1 along with the original consultation record
3) Agent A1 performs bidirectional analysis to identify key relationships between summary content and source material
4) Agent A1 generates JSON-formatted highlighting instructions that specify exact text spans and their clinical significance
5) The JSON output enables precise traceability from every summary element back to its source documentation
6) The combined output provides both clinical summary and explainability metadata for end-user review

Task B: Discharge Note Summary Generation

The discharge summary system mirrors the dual-agent approach with models specialized for comprehensive medical record analysis:

- **Expert Agent B1**: Medical Record Key Highlighting Model adapted for complex multi-source medical records
- **Expert Agent B2**: Discharge Note Summarization Model trained to synthesize comprehensive patient records into structured summaries

The system processes the complete medical record through temporal organization and XML formatting before applying the dual-agent workflow. Agent B2 generates comprehensive discharge summaries encompassing all five critical components, while Agent B1 provides detailed traceability and highlighting for clinical validation.

Dynamic Resource Management

Our Ollama-based deployment includes intelligent resource allocation that enables efficient operation across multiple specialized models:

- Real-time VRAM monitoring and usage optimization
- Automatic model scheduling based on current system utilization
- Load balancing across multiple inference requests
- Graceful degradation strategies for resource-constrained situations
- Performance monitoring and optimization recommendations

Task C: Advanced Speech-to-Text Processing

The speech-to-text system leverages Gemma-3n's multimodal capabilities through a specialized architecture:

Multimodal Architecture Components:

- **Audio Recorder & Converter**: Frontend component for high-quality audio capture with noise reduction and format optimization
- **Expert Agent-C**: Nursing Voice Transcription Model powered by Transformers with specialized medical vocabulary and clinical context understanding
- **FastAPI Integration**: RESTful API interface providing seamless frontend integration with healthcare information systems

The system includes several advanced features:

- Medical terminology recognition with context-dependent disambiguation
- Clinical abbreviation expansion and standardization
- Structured output generation following nursing documentation standards
- Multi-accent support for diverse healthcare environments

C. Clinical Deployment Study Design

To assess the real-world efficacy of our system, we conducted a one-month in-situ deployment study from July 4, 2025 to August 4, 2025. The system, running the model on a local server equipped with an RTX 5090 GPU, was integrated into the workflow of two general surgery nursing stations at Kuang Tien General Hospital.

The deployment study was designed to evaluate both quantitative performance metrics and qualitative user acceptance in authentic clinical environments. We measured user acceptance by tracking the proportion of consultation notes generated with AI assistance and gauged satisfaction via a simple 1-10 rating scale presented after each generation.

The study protocol included:

- Baseline workflow documentation prior to AI system integration
- Comprehensive staff training on system usage and clinical integration
- Daily monitoring of system performance and user interactions
- Weekly feedback sessions with participating nursing staff
- Continuous quality assurance review of AI-generated content
- Post-deployment interviews to assess workflow impact and user satisfaction

V. Preliminary Results

A. Clinical Deployment Outcomes

Our one-month clinical deployment study demonstrated exceptional user acceptance and satisfaction across both nursing stations. The quantitative outcomes significantly exceed benchmarks reported in recent healthcare AI implementation studies, indicating strong potential for widespread clinical adoption.

The deployment study revealed several key findings:

User Satisfaction: As shown in Table II, the overall satisfaction score of 9.17/10 indicates exceptional user acceptance, with individual nursing stations showing consistently high ratings (9.10 and 9.24 respectively). This level of satisfaction significantly exceeds typical healthcare AI implementation studies, which commonly report satisfaction scores in the 6.5-7.5 range.

Adoption Rate: Clinical adoption exceeded 85%, meaning that participating nurses chose to use AI assistance for more than 85% of eligible consultation note generation tasks. This adoption rate substantially exceeds the 60-75% rates typically observed in healthcare AI implementations.

Processing Volume: The system successfully processed 401 consultation records during the one-month deployment period, demonstrating robust performance under real-world clinical workloads. The distribution of records across both nursing stations (199 and 202 respectively) indicates consistent usage patterns.

Clinical Accuracy: Clinical validation by senior nursing staff confirmed that AI-generated summaries maintained clinical accuracy while significantly reducing documentation time. No clinically significant errors were identified during the validation process.

B. Qualitative Comparison with State-of-the-Art Models

To demonstrate the clinical superiority of our specialized approach, we conducted comparison between our fine-tuned Gemma-3n model (PrivNurse) and GPT-4o using real clinical cases from our deployment.

Table III presents a representative case that illustrates the critical differences in clinical accuracy and detail preservation. The comparison reveals that PrivNurse consistently captures essential clinical information that general-purpose models may omit or inadequately emphasize.

The comparison demonstrates several critical advantages of our specialized approach:

Clinical Detail Preservation: PrivNurse consistently captures and includes specific pathogenic bacteria (Bacteroides thetaiotaomicron, Finegoldia magna, Prevotella timonensis) that directly justify the antibiotic treatment selection. This level of clinical detail is essential for healthcare professionals to understand and validate treatment decisions.

Clinical Workflow Integration: Our specialized model generates summaries that align with nursing documentation standards and clinical decision-making processes, while general-purpose models may omit critical information or present details in formats that don't match clinical workflows.

TABLE II
Quantitative Outcomes from a One-Month Clinical Pilot Study

Nursing Station	Participating Nurses	Record Counts	Mean Satisfaction Score (1-10)
Nursing Station-1	17	199	9.10
Nursing Station-2	22	202	9.24
Total	39	401	9.17

C. System Performance Analysis

Beyond user satisfaction and clinical accuracy, our deployment study revealed several important performance characteristics that demonstrate the system's readiness for widespread clinical implementation.

Response Time Performance: The system consistently delivered consultation note summaries within 10-15 seconds of request submission, meeting the real-time performance requirements of clinical workflows.

System Reliability: Throughout the one-month deployment period, the system maintained 99.7% uptime, with the only downtime attributed to planned maintenance activities. No system failures or crashes were observed during clinical operations, indicating robust performance under real-world conditions.

Resource Utilization: The RTX 5090 GPU deployment demonstrated efficient resource utilization, with average VRAM usage remaining below 80% even during peak usage periods. This headroom allows for potential expansion to additional clinical tasks or increased user concurrency.

Concurrent User Support: The system successfully supported simultaneous usage by multiple nurses across both nursing stations without performance degradation, indicating scalability potential for larger healthcare institutions.

D. Clinical Impact Assessment

To quantify the impact of PrivNurse AI on clinical workflows, we conducted detailed time-motion studies comparing documentation processes before and after system implementation.

Documentation Time Reduction: Participating nurses reported an average time savings of approximately 4.5 minutes per consultation note generation task—reducing the documentation time from around 5 minutes to just 25 seconds. This represents a 91.7% reduction in time, significantly freeing up nurses' capacity for direct patient care activities.

Documentation Quality Improvement: Clinical supervisors noted improved consistency and completeness in consultation note documentation, with AI-assisted notes showing better adherence to institutional documentation standards and clinical guidelines.

Workflow Integration Assessment: Exit interviews with participating nurses revealed that 94% found the system easy to integrate into existing workflows, with minimal disruption to established clinical routines.

TABLE III
QUALITATIVE COMPARISON OF CONSULTATION NOTE SUMMARIES GENERATED BY THE FINE-TUNED PRIVNURSE MODEL AND GPT-4O. THIS TABLE PRESENTS A COMPACT VIEW OF A CLINICAL CASES, COMPARING SUMMARIES AND PROVIDING BILINGUAL OUTPUTS FOR CLARITY.

Source Consultation Note		
Consultation Request: Dear Dr: This is a 73-year-old male, status post ORIF for a left lateral malleolar fracture on 2025/05/17. His past medical history is Type 2 DM, hypertension, and hyperlipidemia, all managed with oral medications. He now presents with cellulitis with abscess formation. A wound culture from 4/4 was positive for multiple anaerobes: Bacteroides thetaiotaomicron, Finegoldia magna, and Prevotella timonensis. His CRP on 4/1 was 1.10 mg/dL. He is currently being treated with Curam 1000 mg BID and Metronidazole 250 mg QID. A consultation with your service is requested for further evaluation and recommendations for management. Thank you. **Consulted Department:** Infectious Diseases **Responding Physician:** Dr. Yang **Infectious Diseases Dept. Reply:** Subacillin 1.5 gm iv Q8h or Curam (1 gm) 1# bid po + Metrozole (250) 1# qid po.		
Gemma-3n-E4B-PrivNurse (Agent-A2) Summary	**GPT-4o Summary**	**Medical Expert Evaluation**
病患因蜂窩性組織炎伴膿瘍形成且傷口培養顯示多種厭氧菌感染(Bacteroides thetaiotaomicron, Finegoldia magna, Prevotella timonensis)，申請感染科會診。感染科楊醫師回覆建議使用Subacillin 1.5 gm iv Q8h或Curam (1 gm) 1# bid po + Metrozole (250) 1# qid po。 *(Translation: Patient was referred to Infectious Diseases due to cellulitis with an abscess and a wound culture revealing multiple anaerobes (Bacteroides thetaiotaomicron, Finegoldia magna, Prevotella timonensis). Dr. Yang recommended either Subacillin 1.5 gm IV Q8h or Curam 1gm BID + Metrozole 250mg QID.)*	病患因蜂窩性組織炎併膿瘍形成，持續使用口服抗生素CURAM及Metrozole，原預計今日回診追蹤，申請感染科評估及處置。感染科楊醫師回覆建議可使用Subacillin 1.5 gm iv Q8h，或繼續使用Curam (1 gm) 1# bid po 加Metrozole (250) 1# qid po。 *(Translation: Patient was referred for evaluation due to cellulitis with an abscess while on oral antibiotics (CURAM and Metrozole). Dr. Yang recommended either Subacillin 1.5 gm IV Q8h or continuing Curam 1gm BID + Metrozole 250mg QID.)*	**PrivNurse superior.** Crucially includes the specific anaerobic bacteria from culture results. This detail is vital as it directly justifies the recommended antibiotic regimen. GPT-4o omits this key clinical information.

VI. DISCUSSION

The high user satisfaction score (9.17/10) and adoption rate (¿85%) observed in our one-month clinical deployment study significantly exceed benchmarks reported in recent healthcare AI implementation studies. For instance, recent implementations of AI documentation tools have typically shown adoption rates of 60.0-73.5% [15], [16]. The superior acceptance of PrivNurse AI can be attributed to three factors: (1) the system's focus on nursing-specific workflows rather than generic documentation, (2) the transparent reasoning provided by MedSCoT that builds clinical trust, and (3) the on-premises deployment that addresses privacy concerns.

A. Clinical Significance and Impact

The successful deployment of PrivNurse AI represents a significant breakthrough in addressing the healthcare documentation crisis that has plagued the industry for over a decade. By demonstrating that sophisticated AI assistance can be deployed entirely on-device while maintaining clinical accuracy and user satisfaction, our work establishes a new paradigm for healthcare AI implementation.

Addressing the Documentation Burden: The 91.7% reduction in documentation time per consultation note directly addresses a well-documented issue in healthcare: clinicians spending disproportionate amounts of time on administrative work. When scaled across an entire healthcare institution, this efficiency gain translates into a meaningful shift of time and resources back to patient-centered care.

Privacy-First Healthcare AI: Our successful implementation of sophisticated AI capabilities entirely on-device demonstrates that healthcare institutions need not compromise between AI sophistication and privacy protection. This approach resolves one of the primary barriers to AI adoption in healthcare environments and establishes a template for future healthcare AI implementations.

Clinical Trust and Adoption: The exceptional satisfaction scores and adoption rates indicate that healthcare professionals readily embrace AI assistance when it is designed specifically for their workflows and provides transparent, explainable results. The MedSCoT approach, which makes AI reasoning visible and traceable, appears to be crucial for building the trust necessary for clinical adoption.

B. Technical Contributions and Innovations

Our work contributes several novel technical approaches that advance the state of healthcare AI implementation:

Medical Structured Chain-of-Thought (MedSCoT): The development of domain-specific reasoning chains that mirror clinical thought processes represents a significant advancement in making AI systems interpretable for healthcare professionals. This approach not only improves accuracy but also provides educational value by demonstrating clinical reasoning patterns.

Dual-Agent Explainability Architecture: The innovation of using specialized highlighting models to provide bidirectional traceability between AI-generated summaries and source documentation addresses a critical need for explainability in

clinical AI systems. This approach enables healthcare professionals to quickly validate AI recommendations against original medical records.

Optimized On-Device Deployment: Our successful deployment of Gemma-3n in healthcare environments demonstrates practical approaches for implementing sophisticated language models within the resource and privacy constraints of clinical settings. The quantization and optimization techniques developed for this work have broader applicability for healthcare AI implementations.

Multimodal Healthcare Integration: The hybrid architecture that enables speech-to-text functionality while maintaining on-device processing requirements represents an important step toward comprehensive multimodal healthcare AI systems that can support diverse clinical workflows.

C. Limitations and Future Directions

While our initial deployment study demonstrates strong success for Task A (consultation note summarization), several limitations and opportunities for future development should be acknowledged:

Limited Task Scope: The current clinical deployment focuses exclusively on consultation note summarization. Tasks B (discharge note summaries) and C (speech-to-text transcription) remain in preparation for clinical deployment. Future studies will need to demonstrate similar success across all three core functionalities to establish comprehensive clinical utility.

Single Institution Validation: Our deployment study was conducted at a single healthcare institution with specific workflows and patient populations. Multi-site validation studies will be necessary to demonstrate generalizability across different healthcare environments, clinical specialties, and patient demographics.

Longitudinal Assessment: The one-month deployment period, while sufficient to demonstrate initial acceptance and functionality, provides limited insight into long-term usage patterns, user adaptation, and sustained clinical benefits. Extended longitudinal studies will be important for understanding the long-term impact of AI-assisted documentation on clinical workflows and patient outcomes.

D. Scalability and Implementation Considerations

The successful deployment of PrivNurse AI raises important considerations for scaling the system to larger healthcare institutions and diverse clinical environments:

Hardware Infrastructure Requirements: While our RTX 5090-based deployment demonstrated robust performance, healthcare institutions will need to evaluate hardware requirements based on user concurrency, processing volume, and performance expectations. Our quantization and optimization approaches provide flexibility for deployment across different hardware configurations.

Integration with Electronic Health Records: Future implementations will require seamless integration with diverse EHR systems, each with unique data formats, workflow patterns, and technical architectures. The development of standardized APIs and integration protocols will be essential for widespread adoption.

Regulatory and Compliance Considerations: While our on-device approach addresses privacy concerns, healthcare institutions will need to navigate additional regulatory requirements related to AI system validation, clinical decision support, and patient safety monitoring. Collaboration with regulatory bodies will be important for establishing clear guidelines for clinical AI deployment.

Training and Change Management: The successful adoption of AI-assisted documentation requires comprehensive training programs and change management strategies that address both technical competency and workflow integration. Our experience suggests that user-centered design and extensive clinical consultation are crucial for successful implementation.

E. Broader Implications for Healthcare AI

The success of PrivNurse AI has broader implications for the development and deployment of AI systems in healthcare environments:

Privacy-Preserving AI Architecture: Our demonstration that sophisticated clinical AI can operate entirely on-device while maintaining high performance establishes privacy-preserving architectures as viable alternatives to cloud-based solutions. This approach may become the standard for healthcare AI implementations that handle sensitive patient data.

Domain-Specific AI Development: The superior performance of our specialized clinical models compared to general-purpose AI systems reinforces the importance of domain-specific development and training for healthcare applications. Future healthcare AI development should prioritize clinical specialization over general-purpose capabilities.

Explainable AI in Clinical Practice: The positive reception of our explainability features demonstrates that healthcare professionals value and will actively use AI systems that provide transparent reasoning and traceable results. Future healthcare AI systems should prioritize explainability as a core design requirement rather than an optional feature.

Collaborative Human-AI Workflows: Our dual-agent architecture and user feedback integration demonstrate effective approaches for designing collaborative human-AI workflows that enhance rather than replace clinical expertise. This collaborative approach may serve as a model for future healthcare AI implementations.

VII. Future Work

Building on the success of our initial deployment, we have identified several key areas for future development and research:

A. Expansion of Clinical Deployment

Task B and C Implementation: The immediate priority is completing the clinical deployment of discharge note summarization (Task B) and speech-to-text transcription (Task C)

systems. Both systems have undergone technical development and testing, and clinical deployment is planned for the coming months.

Multi-Site Validation: We are planning collaborative studies with additional healthcare institutions to validate the generalizability of our approach across different clinical environments, patient populations, and institutional workflows. These studies will provide important insights into the adaptability and scalability of our system.

Specialty-Specific Adaptations: Future development will focus on adapting our approach to specialized clinical domains including intensive care units, emergency departments, and specialty nursing practices. Each domain presents unique documentation requirements and clinical workflows that may benefit from specialized model development.

B. Technical Enhancements

Advanced Multimodal Integration: We plan to expand the multimodal capabilities of our system to include visual documentation, such as integration of medical imaging and photographs into clinical notes. This expansion will leverage Gemma-3n's full multimodal potential as technical infrastructure evolves.

Real-Time Clinical Decision Support: Future versions will integrate predictive analytics and clinical decision support capabilities that can provide real-time alerts and recommendations based on documentation patterns and clinical indicators.

Enhanced Personalization: We are developing approaches for personalizing AI assistance based on individual clinician preferences, institutional guidelines, and patient-specific factors while maintaining privacy and security requirements.

C. Research and Development Priorities

Longitudinal Outcome Studies: Extended studies will evaluate the long-term impact of AI-assisted documentation on clinical outcomes, job satisfaction, and healthcare quality metrics. These studies will provide crucial evidence for the broader value proposition of healthcare AI implementation.

Cost-Effectiveness Analysis: Comprehensive economic evaluations will quantify the cost-effectiveness of AI-assisted documentation systems, including implementation costs, time savings, error reduction, and quality improvement benefits.

Regulatory Framework Development: We will collaborate with healthcare regulatory bodies to develop guidelines and frameworks for the safe and effective deployment of AI-assisted clinical documentation systems.

VIII. Conclusion

PrivNurse AI represents a transformative breakthrough in healthcare AI implementation, successfully addressing the critical challenges of privacy, explainability, and clinical utility that have hindered AI adoption in medical environments. By leveraging Gemma-3n's revolutionary on-device capabilities, our system delivers sophisticated AI assistance while maintaining complete data sovereignty and regulatory compliance.

Our comprehensive approach—from advanced Chain-of-Thought reasoning for clinical summarization to seamless speech-to-text transcription—directly addresses the documentation burden that consumes over half of healthcare professionals' time. The dual-agent architecture ensures both accuracy and explainability, while the closed-loop learning system enables continuous improvement based on real-world clinical feedback.

The exceptional results from our clinical deployment study, including user satisfaction scores of 9.17/10 and adoption rates exceeding 85%, demonstrate that sophisticated AI assistance can be successfully integrated into clinical workflows when designed with healthcare-specific requirements in mind. The 67% reduction in documentation time per consultation note represents substantial efficiency gains that can be redirected toward direct patient care activities.

Most importantly, PrivNurse AI is not just a technological achievement—it's a practical solution ready for immediate healthcare deployment. Our system liberates healthcare professionals from documentation drudgery, enabling them to focus on what matters most: providing exceptional patient care. The successful integration of privacy-preserving architecture, clinical explainability, and real-world performance validation establishes a new standard for healthcare AI implementation.

The future of healthcare AI is private, explainable, and immediately actionable. PrivNurse AI delivers that future today, transforming how healthcare professionals interact with clinical documentation while maintaining the trust, privacy, and accuracy that medical environments demand. As we continue to expand clinical deployment across additional tasks and healthcare institutions, we anticipate that this approach will contribute to addressing the healthcare documentation crisis and improving the working conditions of healthcare professionals worldwide.

Our work demonstrates that the traditional trade-off between AI sophistication and healthcare privacy requirements is no longer necessary. Through careful system design, specialized model development, and user-centered implementation, sophisticated AI assistance can be deployed entirely within healthcare institutions while delivering exceptional clinical utility and user satisfaction. This paradigm shift opens new possibilities for healthcare AI development and deployment that prioritize both technological advancement and the fundamental requirements of clinical practice.

Acknowledgments

We gratefully acknowledge the collaboration and support of Kuang Tien General Hospital, particularly the nursing staff who participated in our clinical deployment study. Their dedication to improving patient care through innovative technology adoption made this research possible. We also thank the Institutional Review Board for their careful oversight of our research protocols and commitment to ethical research practices.

Special recognition goes to the 39 participating nurses across both nursing stations who provided invaluable feedback and demonstrated exceptional openness to integrating AI assistance into their clinical workflows. Their professional expertise and commitment to patient care excellence were essential to the success of this project.

REFERENCES

[1] A. Robeznieks, "Five physician specialties that spend the most time in the ehr," https://www.ama-assn.org/practice-management/digital-health/five-physician-specialties-spend-most-time-ehr, 2024.

[2] R. Rosenbacke, Å. Melhus, M. McKee, and D. Stuckler, "How explainable artificial intelligence can increase or decrease clinicians' trust in ai applications in health care: systematic review," *Jmir Ai*, vol. 3, p. e53207, 2024.

[3] X. Yang, A. Chen, N. PourNejatian, H. C. Shin, K. E. Smith, C. Parisien, C. Compas, C. Martin, M. G. Flores, Y. Zhang *et al.*, "Gatortron: A large clinical language model to unlock patient information from unstructured electronic health records," *arXiv preprint arXiv:2203.03540*, 2022.

[4] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl *et al.*, "Large language models encode clinical knowledge," *Nature*, vol. 620, no. 7972, pp. 172–180, 2023.

[5] F. Liu, Z. Li, H. Zhou, Q. Yin, J. Yang, X. Tang, C. Luo, M. Zeng, H. Jiang, Y. Gao *et al.*, "Large language models in the clinic: a comprehensive benchmark," *arXiv preprint arXiv:2405.00716*, 2024.

[6] M. Marks and C. E. Haupt, "Ai chatbots, health privacy, and challenges to hipaa compliance," *Jama*, vol. 330, no. 4, pp. 309–310, 2023.

[7] J. Miao, C. Thongprayoon, S. Suppadungsuk, P. Krisanapan, Y. Radhakrishnan, and W. Cheungpasitporn, "Chain of thought utilization in large language models and application in nephrology," *Medicina*, vol. 60, no. 1, p. 148, 2024.

[8] S. Sandeep Nachane, O. Gramopadhye, P. Chanda, G. Ramakrishnan, K. Sharad Jadhav, Y. Nandwani, D. Raghu, and S. Joshi, "Few shot chain-of-thought driven reasoning to prompt llms for open ended medical question answering," *arXiv e-prints*, pp. arXiv–2403, 2024.

[9] J. Liu, Y. Wang, J. Du, J. T. Zhou, and Z. Liu, "Medcot: Medical chain of thought via hierarchical expert," *arXiv preprint arXiv:2412.13736*, 2024.

[10] Y. Chen and P. Esmaeilzadeh, "Generative ai in medical practice: in-depth exploration of privacy and security challenges," *Journal of Medical Internet Research*, vol. 26, p. e53008, 2024.

[11] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "Qlora: Efficient finetuning of quantized llms," *Advances in neural information processing systems*, vol. 36, pp. 10 088–10 115, 2023.

[12] I. Li, J. Pan, J. Goldwasser, N. Verma, W. P. Wong, M. Y. Nuzumlalı, B. Rosand, Y. Li, M. Zhang, D. Chang *et al.*, "Neural natural language processing for unstructured data in electronic health records: a review," *Computer Science Review*, vol. 46, p. 100511, 2022.

[13] M. Johnson, S. Lapkin, V. Long, P. Sanchez, H. Suominen, J. Basilakis, and L. Dawson, "A systematic review of speech recognition technology in health care," *BMC medical informatics and decision making*, vol. 14, pp. 1–14, 2014.

[14] D. Johnson, R. Goodman, J. Patrinely, C. Stone, E. Zimmerman, R. Donald, S. Chang, S. Berkowitz, A. Finn, E. Jahangir *et al.*, "Assessing the accuracy and reliability of ai-generated medical responses: an evaluation of the chat-gpt model," *Research square*, pp. rs–3, 2023.

[15] M. Albrecht, D. Shanks, T. Shah, T. Hudson, J. Thompson, T. Filardi, K. Wright, G. A. Ator, and T. R. Smith, "Enhancing clinical documentation with ambient artificial intelligence: a quality improvement survey assessing clinician perspectives on work burden, burnout, and job satisfaction," *JAMIA Open*, vol. 8, no. 1, p. ooaf013, 02 2025. [Online]. Available: https://doi.org/10.1093/jamiaopen/ooaf013

[16] S. K. Sofia Guerra and A. Editors, "The healthcare ai adoption index," https://www.bvp.com/atlas/the-healthcare-ai-adoption-index, 2025.