MEDPLAN: A Two-Stage RAG-Based System for Personalized Medical Plan Generation

Hsin-Ling Hsu[1]*, Cong-Tinh Dao[2,3]*, Luning Wang[4], Zitao Shuai[4], Nguyen Minh Thao Phan[2,3], Jun-En Ding [5], Chun-Chieh Liao[5], Pengfei Hu [5], Xiaoxue Han [5], Chih-Ho Hsu [6], Dongsheng Luo [7], Wen-Chih Peng [2], Feng Liu [5], Fang-Ming Hung [6], Chenwei Wu [4]

[1]National Chengchi University, [2]National Yang Ming Chiao Tung University,
[3]Can Tho University, [4]University of Michigan, [5]Stevens Institute of Technology,
[6]Far Eastern Memorial Hospital [7]Florida International University

Correspondence: chenweiwu99@gmail.com

Abstract

Despite recent success in applying large language models (LLMs) to electronic health records (EHR), most systems focus primarily on assessment rather than treatment planning. We identify three critical limitations in current approaches: they generate treatment plans in a single pass rather than following the sequential reasoning process used by clinicians; they rarely incorporate patient-specific historical context; and they fail to effectively distinguish between subjective and objective clinical information. Motivated by the SOAP methodology (Subjective, Objective, Assessment, Plan), we introduce MEDPLAN, a novel framework that structures LLM reasoning to align with real-life clinician workflows. Our approach employs a two-stage architecture that first generates a clinical assessment based on patient symptoms and objective data, then formulates a structured treatment plan informed by this assessment and enriched with patient-specific information through retrieval-augmented generation. Comprehensive evaluation demonstrates that our method significantly outperforms baseline approaches in both assessment accuracy and treatment plan quality. Our demo system and code are available at https://github.com/JustinHsu1019/MedPlan.

1 Introduction

Deploying large language models (LLMs) for electronic health records (EHR) (Evans, 2016) analysis in high-stakes medical environments presents significant opportunities for enhancing patient care through automation and improved clinical decision support (Yang et al., 2022; Zhang et al., 2024; Sakai and Lam, 2025; Ding et al., 2024). Despite recent progress in adapting LLM to medical domain (Tang et al., 2025; Jiang et al., 2025; Restrepo et al., 2025), most existing LLM systems (Palepu et al., 2025; Fan and Tao, 2024) for EHR focus largely on diagnostic assessment tasks, neglecting the crucial subsequent step of structured, patient-specific treatment planning (Sarker et al., 2021; Curtis et al., 2017). Effective LLM-based planning could significantly reduce physician cognitive load, standardize care protocols, decrease treatment variability, and enable more personalized interventions.

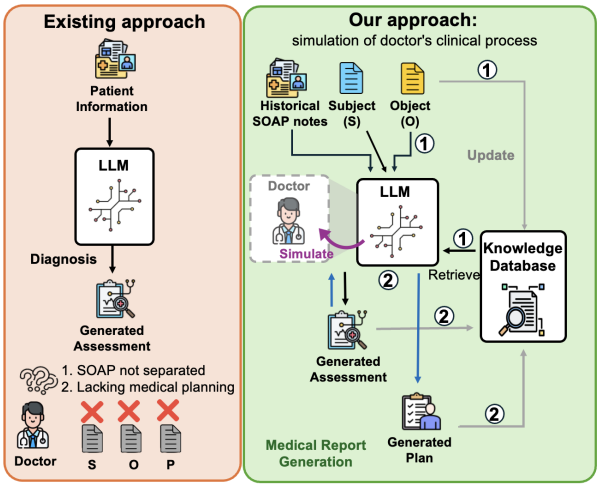

Figure 1: Compare the existing approach (left) with our proposed MEDPLAN (right). We adopt the SOAP protocol and simulate the doctor diagnosis process with LLM for medical plan generation.

Enabling LLM with trustworthy and personalized treatment planning capabilities introduces unique challenges—models must generate medically sound interventions, tailor recommendations to individual patient needs, and maintain a clear rationale connecting diagnosis to treatment (Qiu et al., 2025). Ideally, these systems should align with real-life clinical reasoning processes employed by healthcare professionals. The SOAP methodology (Subjective, Objective, Assessment, Plan) represents one of medicine's fundamental cognitive frameworks (Sorgente et al., 2005; Shechtman, 2002), systematically organizing clinical information into a structured sequen-

*Equal contribution

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track), pages 1072–1082
July 28-30, 2025

tial decision-making process. Under this protocol, clinicians first gather subjective patient-reported symptoms (S) and objective clinical data such as laboratory tests and physical examination findings (O). These elements provide the basis for a clinical assessment (A), subsequently informing a structured treatment plan (P).

However, our analysis identifies several critical limitations in current approaches. First, the few existing works on medical treatment planning with LLMs (Liu et al., 2024; Chen et al., 2025) attempt to generate treatment plans directly from clinical data in a single pass, failing to mirror the sequential cognitive process physicians adopt, where clinicians first reach diagnostic conclusions before developing actionable interventions tailored to each patient's unique circumstances. This collapsed reasoning process risks producing treatment recommendations disconnected from their diagnostic foundations—a critical failure in medical decision-making where transparent causal relationships between findings and interventions are essential.

Second, current approaches rarely incorporate patient-specific historical context—such as medical history, previous treatment responses, and longitudinal trends—that physicians naturally consider when making treatment decisions. This neglect of personalized context leads to generic treatment recommendations that fail to account for individual patient factors crucial to treatment success. Finally, most systems don't effectively distinguish between subjective patient narratives and objective clinical measurements, despite this distinction being fundamental to clinical practice where a patient's subjective experience ("my chest hurts when I breathe") is weighed differently from objective findings (elevated troponin levels) in formulating both diagnoses and treatment plans.

These gaps motivate our research questions:

- **How can we structure LLM reasoning processes to mirror the sequential SOAP protocol used by clinicians, and does this improve treatment plan generation?**
- **How can we incorporate patient-specific contexts to better support individualized care decisions?**

To address these challenges, we introduced MEDPLAN, a novel framework that explicitly structures LLM reasoning to mirror the SOAP clinical workflow. Our approach operates in two clinically-grounded stages that parallel physician cognitive processes: (1) a diagnostic phase where we generate an assessment (A) based on patient symptoms and clinical data (S and O), completing the diagnostic reasoning before proceeding, and (2) a therapeutic phase where we formulate a structured treatment plan (P) directly informed by the assessment and tailored to patient-specific factors. This two-stage architecture faithfully replicates how clinicians reason—first establishing what is happening before determining what should be done. We enhanced the planning phase through patient-specific retrieval-augmented generation (RAG) (Lewis et al., 2020), allowing the model to consider longitudinal patient information—mirroring how physicians integrate medical history into their treatment decisions.

Our contributions are three-fold:

- We introduced MEDPLAN, a novel SOAP-inspired two-stage LLM framework for EHR data that structures clinical reasoning to match physician workflows, providing reliable patient-specific assessments and plans.
- We conducted a comprehensive evaluation showing our method significantly outperformed baseline methods on various metrics in both clinical assessment and treatment plan generation.
- We released a fully functional system that tests our approach in a real clinical environment, allowing physicians to efficiently generate structured, patient-specific plans integrated with existing EHR workflows.

2 Related Work

The SOAP framework has been widely recognized as a standard for clinical documentation and reasoning (Cameron and Turtle-Song, 2002). Several computational approaches have attempted to structure medical notes according to SOAP elements (Castillo et al., 2019), but they typically treat these elements as documentation categories rather than as steps in a diagnosis reasoning process. Due to the success of LLMs, such as GPT-4, LLaMA, and Mistral-7B, these models have significantly impacted healthcare, particularly in medical documentation, clinical summarization, and decision support. Studies have demonstrated LLMs' potential in automating discharge note generation, extracting key clinical information from EHRs, and

summarizing medical evidence, though challenges such as factual inconsistency and hallucinations remain (Alkhalaf et al., 2024; Tang et al., 2023).

Recent research used patient physical information and examination results as input to make ChatGPT generate a series of initial diagnostic information, examination results, and recommended measures to create reports (Zhou, 2023). Additionally, RAG was used to improve the efficiency of medical document retrieval and integration of external knowledge (Alkhalaf et al., 2024) or enhance the accuracy of LLMs in EHR summaries and medical note generation (Yang et al., 2025). However, current RAG applications primarily focus on data retrieval and aggregation without truly enhancing the internal generation process of LLMs, particularly when processing complex and large quantities of diagnostic reports to generate personalized diagnostic report plans. In this work, we provide a structured LLM retrieval process that incorporates multiple clinical text information while addressing past patient historical records using a two-stage pipeline for medical planning generation.

3 Methodology

To obtain accurate and personalized clinical plannings that align with physician workflow, we present MEDPLAN, a trustworthy clinical decision support system that employs a two-stage generation pipeline, mirroring the natural progression of clinical planning. To get high-quality planning, we propose to first generate an assessment based on the patient data, then create the treatment plan based on both the patient data and the generated assessment. This separation follows the established SOAP protocol, where clinicians first analyze symptoms and findings to form a diagnosis before determining appropriate interventions. We also explicitly separate S and O components in our prompts (see Appendix C), allowing the model to distinctly process patient-reported symptoms versus clinical observations—a key distinction that enhances clinical relevance. To enhance the personalization and accuracy of the generated plannings, we further leverage two types of references during generation: (1) self-history references—the patient's previous SOAP records, and (2) cross-patient references—similar cases from other patients. Specifically, for the i-th patient, we retrieve their latest N_{hist} SOAP records as self-history references, formulated as $\mathcal{R}^{\text{hist}}i = (Sj, O_j, A_j, P_j) \mid j \in 1, 2, ..., N_{\text{hist}}$. Furthermore, to better align with the clinical reasoning patterns, we incorporate instruction tuning on the models that generate A and P before deploying our two-stage pipeline. Figure 2 illustrates the overall architecture of our inference workflow.

3.1 Assessment Generation Stage

In the Assessment Generation Stage, we integrated the patient's current S and O information with both self-history references $\mathcal{R}_{\text{hist}}$ and cross-patient references $\mathcal{R}^{SOA} = \{(S_j, O_j, A_j)\}_{j=1}^{N_{\text{ref}}}$. To identify the most relevant cross-patient references, we employ a two-step retrieval process. First, we retrieve N_{sim} candidate references $\mathcal{R}_{\text{sim}}^{\text{SOA}}$ via hybrid retrieval (Ma et al., 2020; Bruch et al., 2023; Hsu and Tzeng, 2025) combining BM25 (Robertson et al., 1995) and bi-encoder semantic search (Karpukhin et al., 2020), leveraging both keyword matching and semantic similarity. Then, we refined this selection using a more computationally intensive but more accurate cross-encoder re-ranking model (Nogueira and Cho, 2020) that evaluates the fine-grained clinical relevance by jointly encoding the query and each candidate:

$$\mathcal{R}^{SOA} = \text{Top-}N_{\text{ref}}\Big(\text{ReR}\big(\{S, O\}, \mathcal{R}_{\text{sim}}^{\text{SOA}}\big)\Big),$$

where $\text{ReR}(\{S, O\}, \mathcal{R}_{\text{sim}}^{\text{SOA}})$ represents the cross-encoder re-ranking function that scores each reference in $\mathcal{R}_{\text{sim}}^{\text{SOA}}$ based on its relevance to the current case $\{S, O\}$. After obtaining the refined references, we combine the current (S, O) with both $\mathcal{R}_{SOA}$ and $\mathcal{R}_{\text{hist}}$ to generate the assessment:

$$A_{\text{gen}} = f_{\theta_A}(S, O, \mathcal{R}^{SOA}, \mathcal{R}^{\text{hist}}),$$

where A_{gen} denotes the generated assessment and f_{θ_A} represents the medical language model for assessment generation.

3.2 Plan Generation Stage

In the Plan Generation Stage, we utilized the generated assessment A_{gen} along with the original S and O to retrieve and generate an appropriate treatment plan. Mirroring the clinical practice where physicians formulate treatment plans based on their diagnostic assessment and patient information, we employed another retrieval process to find relevant plan references $\mathcal{R}^{SOAP} = \{(S_j, O_j, A_j, P_j)\}_{j=1}^{N_{\text{ref}}}$. Similar to the previous stage, we use a two-step retrieval approach. First, we retrieve N_{sim} candidate references $\mathcal{R}_{\text{sim}}^{\text{SOAP}}$ via hybrid retrieval combining BM25 and bi-encoder semantic search. Then,

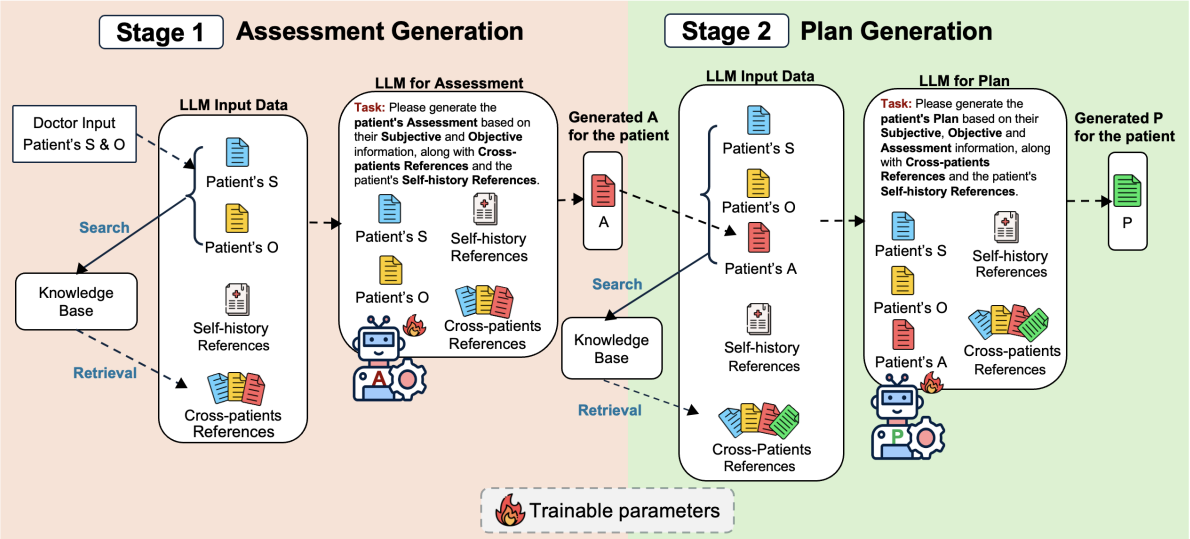

Figure 2: Overall architecture of the proposed MEDPLAN framework.

we refined this selection using a cross-encoder re-ranking model:

$$\mathcal{R}^{SOAP} = \text{Top-}N_{\text{ref}}\Big(\text{ReR}\big(\{S, O, A_{\text{gen}}\}, \mathcal{R}^{\text{SOAP}}_{\text{sim}}\big)\Big),$$

where $\text{ReR}(\{S, O, A_{\text{gen}}\}, \mathcal{R}^{\text{SOAP}}_{\text{sim}})$ represents the cross-encoder re-ranking function that evaluates each reference in $\mathcal{R}^{\text{SOAP}}_{\text{sim}}$ based on its relevance to the current case with the generated assessment. After obtaining the refined references, we combined the current (S, O, A_{gen}) with both $\mathcal{R}^{SOAP}$ and $\mathcal{R}^{\text{hist}}$ to generate the treatment plan:

$$P_{\text{gen}} = f_{\theta_P}(S, O, A_{\text{gen}}, \mathcal{R}_{SOAP}, \mathcal{R}_{\text{hist}}),$$

where P_{gen} denotes the generated plan and f_{θ_P} represents the medical language model for plan generation.

3.3 Information Alignment

To align the models with the clinical reasoning pattern of our dataset, we instruction-tuned both the assessment generation model and plan generation model using the following objectives:

$$\theta_A = \underset{\theta}{\text{argmin}} \sum_{i=1}^{N} \mathcal{L}(f_\theta(S_i, O_i, \mathcal{R}_i^{SOA}, \mathcal{R}_i^{\text{hist}}), A_i),$$

$$\theta_P = \underset{\theta}{\text{argmin}} \sum_{i=1}^{N} \mathcal{L}(f_\theta(S_i, O_i, A_i, \mathcal{R}_i^{SOAP}, \mathcal{R}_i^{\text{hist}}), P_i),$$

where $\mathcal{L}$ is the loss function, N is the number of training samples, and A_i and P_i are the ground truth assessment and plan, respectively. This training process ensures that our models can properly interpret and utilize the medical context specific to our dataset.

4 Experiments

4.1 Datasets

This study utilized 350,684 outpatient and emergency EHR SOAP notes from 55,890 patients collected at Far Eastern Memorial Hospital (FEMH) in 2021. All data were de-identified prior to analysis. We preprocessed all SOAP notes by removing records shorter than two characters and normalizing text (eliminating newlines, redundant spaces, and consecutive punctuation).

Unlike disease-specific approaches, our dataset encompasses general cases, ensuring broader applicability across clinical scenarios. To achieve this, we selected patients with three or more visits and employed a patient-centric sampling strategy. Specifically, records from 6,000 patients constituted our RAG knowledge base embedding, while an additional 3,000 randomly selected patient records were allocated into training and testing sets.

4.2 Metrics

For evaluation metrics, we used BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), ROUGE (Lin, 2004), and BERTScore (Zhang et al., 2019) using an independent inference script. Lexical similarity is evaluated using METEOR (Metric for Evaluation of Translation with Explicit Ordering) and BLEU (Bilingual Evaluation Understudy), with METEOR considering stemming and synonyms. ROUGE, which is the abbreviation of Recall-Oriented Understudy for Gisting Evaluation scores, compares the produced and reference summaries for the longest common subse-

quence (ROUGE-L) and n-gram overlaps (ROUGE-1, ROUGE-2). In order to properly evaluate text coherence and meaning, BERTScore balances recall and accuracy by using contextual embeddings to estimate semantic similarity beyond precise matches.

4.3 Implementation Details

We utilized prompt engineering techniques and applied LoRA for parameter-efficient fine-tuning. Specifically, we instruction-tuned several open-source LLMs—Medical-Llama3-8B (Vsevolodovna, 2024a), Medical-Mixtral-7B-v2k (Vsevolodovna, 2024b), and Bio-Medical-Llama3-8B (ContactDoctor, 2024)—using the Unsloth framework (Daniel Han and team, 2023). To support long-context retrieval in our RAG-based design, we adopted OpenAI's text-embedding-3-large model (OpenAI, 2024) for semantic similarity search, and used VoyageAI Reranker-2 (VoyageAI, 2024) as a cross-encoder model to re-rank the retrieved candidates. For baseline comparison, we additionally evaluated two general-purpose models: o1 (OpenAI, 2024b) and GPT-4o (OpenAI, 2024a), without domain-specific adaptation.

We set $N_{\text{hist}} = 20$ and $N_{\text{ref}} = 10$ for our RAG module, retrieving $N_{\text{sim}} = 80$ initial candidates based on semantic similarity. To evaluate MEDPLAN, we simulated clinical diagnostic processes by using the first $N-2$ visits as history $\mathcal{R}_{\text{hist}}$ and the second-to-last visit as the training target for patients with N visits, while the first $N-1$ visits and the most recent visit were used as history and evaluation target respectively during testing. We conducted ablation experiments with various configurations by selectively enabling components in our pipeline, including: **Self-history**, **Instruction Tuning**, **Cross-patient References**, **Direct Plan Generation**, and a **Two-step Approach with Pre-plan Assessment**. Additional implementation details, including training environment and hyperparameter settings, are provided in Appendix A.1.

4.4 Results

Does MEDPLAN help improve clinical planning? In Table 1, our SOAP-inspired MEDPLAN (S+O→A→P) outperforms the baseline approach (S+O→P) across all backbone models and evaluation metrics. For example, on the Medical-Llama3-8B model, MEDPLAN increases BLEU from 0.307 to 0.315 and METEOR from 0.501 to 0.516. This is likely because MEDPLAN structures LLM reasoning in a manner that mirrors real-world clinical workflows, leading to more reliable planning.

Does MEDPLAN help improve clinical assessment? In Table 2, MEDPLAN method integrates historical cross-patient assessments records, and consistently promotes base versions of all backbones on all metrics. In particular, on the Medical-Llama3-8B backbone, MEDPLAN improves METEOR by 2%, with ROUGE1 and ROUGE2 by 2% and 1.5%, respectively. Similar gains are also observed in other models. This improvement likely results from the inference-time knowledge augmentation provided by the cross-patient information, which enriches the contextual input and helps the model generate more accurate and trustworthy assessments.

How do we better support personalized planning? As shown in Table 1, integrating patient history and cross-patient information via RAG enables our MEDPLAN to significantly enhance plan generation across all evaluated models. For instance, adding RAG in the instruction-tuned Medical-Llama3-8B model raises BLEU from 0.052 to 0.307 and METEOR from 0.173 to 0.501. This might due to the enriched contextual input brought by the RAG, which augments the knowledge in the inference time and help the model to generate more trustworthy clinical plans.

How do our generated treatment plans compare qualitatively to baseline approaches? Figure 3 illustrates the qualitative improvement in clinical decision support capabilities. When presented with a complex patient case featuring multiple cardiovascular risk factors (hyperlipidemia, hypertension, metabolic syndrome, and pre-diabetes), the baseline Medical-Mixtral-7B-v2k model produced only a simplistic "Keep current Rx" recommendation—missing critical diagnostic and treatment components necessary for evidence-based care. In contrast, our approach generated a comprehensive clinical recommendation: "Cardiac catheterization. If symptoms persist, keep Kerlone, Cozaar, and encourage exercise and diet control." This output demonstrates enhanced capabilities to: (1) prioritize appropriate diagnostic procedures, (2) implement condition-based medication management, and (3) incorporate preventive lifestyle interventions for modifiable risk factors.

Table 1: Performance Comparison of Different Models and Settings for Plan Generation

Planning Method	Model	Self-history	Instruction Tuning	Cross-patient	BLEU ↑	METEOR ↑	ROUGE1 ↑	ROUGE2 ↑	ROUGE_L ↑	Bertscore_F1 ↑
S+O→P	o1	✓			0.016399	0.140358	0.125431	0.046444	0.107900	0.817148
	GPT-4o	✓			0.028817	0.166348	0.154136	0.070183	0.139563	0.827025
	Medical-Llama3-8B		✓		0.052796	0.173414	0.220035	0.129617	0.214548	0.847451
		✓			0.178594	0.306591	0.343440	0.274914	0.340154	0.867276
		✓	✓		0.291157	0.477312	0.535286	0.434203	0.531056	0.907823
		✓	✓	✓	0.307380	0.501418	0.559243	0.456576	0.554414	0.911653
	Bio-Medical-Llama3-8B		✓		0.061325	0.188050	0.235100	0.148139	0.228682	0.850004
		✓			0.112796	0.217000	0.235758	0.174116	0.230855	0.848391
		✓	✓		0.299377	0.486631	0.544217	0.441678	0.539558	0.908784
		✓	✓	✓	0.309457	0.501485	0.557870	0.456750	0.553876	0.911572
	Medical-Mixtral-7B-v2k		✓		0.067164	0.196569	0.249694	0.156125	0.243456	0.852184
		✓			0.170338	0.311579	0.365305	0.285245	0.360484	0.869952
		✓	✓		0.298256	0.482994	0.541785	0.442677	0.537791	0.910507
		✓	✓	✓	0.312393	0.510814	0.570339	0.464942	0.565761	0.914185
S+O→A→P (MEDPLAN)	**Bio-Medical-Llama3-8B**	✓	✓	✓	**0.312238**	**0.516716**	**0.574780**	**0.467528**	**0.569738**	**0.915024**
	Medical-Llama3-8B	✓	✓	✓	**0.314718**	**0.516189**	**0.576113**	**0.469581**	**0.571199**	**0.915500**
	Medical-Mixtral-7B-v2k	✓	✓	✓	**0.318286**	**0.521312**	**0.581657**	**0.475762**	**0.577055**	**0.917194**

Table 2: Comparison Performance in Patient-Specific Assessments Generation

Model	Self-history	Instruction Tuning	Cross-patient	BLEU ↑	METEOR ↑	ROUGE1 ↑	ROUGE2 ↑	ROUGE_L ↑	Bertscore_F1 ↑
Medical-Mixtral-7B-v2k	✓			0.302052	0.469219	0.535851	0.437234	0.532359	0.905538
	✓	✓		0.484695	0.653686	0.704872	0.606026	0.700879	0.940547
	✓	✓	✓	**0.493051**	**0.665725**	**0.715743**	**0.616415**	**0.712651**	**0.942709**
Bio-Medical-Llama3-8B	✓			0.234989	0.35864	0.378168	0.310427	0.372989	0.872104
	✓	✓		0.479665	0.645509	0.697491	0.596622	0.693297	0.938073
	✓	✓	✓	**0.490539**	**0.664329**	**0.717387**	**0.61274**	**0.713025**	**0.942353**
Medical-Llama3-8B	✓			0.303517	0.431265	0.466276	0.401507	0.463519	0.889349
	✓	✓		0.474254	0.641288	0.692784	0.594512	0.68923	0.937197
	✓	✓	✓	**0.487554**	**0.658435**	**0.713324**	**0.610607**	**0.710027**	**0.941513**

5 Clinical Application Demo and System Design

To demonstrate the real-world applicability of our Plan generation system, we developed a clinical prototype that has been reviewed by practicing physicians for viability in actual healthcare settings. An overview of the clinical interface is shown in Figure 4. Our system works as follows: The physician first inputs the patient's S and O, and the system generates A and P based on these inputs. At the same time, physicians can modify A according to their clinical judgment and regenerate P, while our system can update retrievals through RAG, which leverages a knowledge base of patient SOAP notes. The more specific technical architecture of the backend system is shown in Figure 2. The frontend is developed using React, the backend is based on FastAPI service, and communication between frontend and backend is conducted through RESTful API. The core of the system includes two specialized LLMs, responsible for generating A and P respectively. The system uses Microsoft SQL (MSSQL) database to store patient historical data, and enhances semantic retrieval and case matching through vector embedding using Weaviate database.

The detailed system architecture is provided in Appendix A.

6 Conclusion

In this study, we introduced MEDPLAN, a novel approach leveraging LLMs with RAG to produce personalized treatment plans following the SOAP methodology. By structuring LLM reasoning into a two-stage process mirroring physician workflows, MEDPLAN generates assessments before formulating plans informed by patient-specific context. Empirical evaluation on an in-house dataset demonstrated promising outcomes and potential for future LLM diagnostic generation research work.

References

Mohammad Alkhalaf, Ping Yu, Mengyang Yin, and Chao Deng. 2024. Applying generative ai with retrieval augmented generation to summarize and extract key clinical information from electronic health records. *Journal of biomedical informatics*, 156:104662.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Sebastian Bruch, Siyu Gai, and Amir Ingber. 2023. An analysis of fusion functions for hybrid retrieval. *ACM Transactions on Information Systems*, 42(1):1–35.

Susan Cameron and Imani Turtle-Song. 2002. Learning to write case notes using the soap format. *Journal of Counseling & Development*, 80(3):286–292.

Human Diagnosis		Model Generation	
Real Diagnosis input	**Human Doctor Planning**	**(Baseline) Medical-Mixtral-7B-v2k**	**(Ours) Base Model: Medical-Mixtral-7B-v2k**
Subjective: R/O CAD, hyperlipidemia, rheumatoid arthritis, hypertension, hyperlipidemia, metabolic syndrome, GERD (?), and pre-DM. No chest pain, no radiation pain, no cold sweating. No orthopnea, no PND. HBP is not monitored. Prescription from other Hospital includes Lipanthyl, folic acid, Cozaar 1# qd, Kerlone (20) 0.5# qd, Prednisolone 1# qd, Salazine 1# bid, Methotrexate 1# TIW, and Orencia (Abatacept) 250mg qd per month. Smoking and drinking are denied. In November 2017, TXT was positive. **Objective:** BP: 122/81 mmHg, PR: 85/min. Heart: Regular heartbeat, no gallop. Chest: Symmetric expansion, clear breath sounds. No leg edema. 110/03/23: creatinine was 0.96, ALT (SGPT) 13, glucose AC 108, total cholesterol 180, triglyceride 191, HDL cholesterol 42, LDL cholesterol 102, and HGB 14.1. 110/08/04: GPT was 17, creatinine 1.0, total cholesterol/triglyceride was 158/245, LDL/HDL was 82/37, and Hb was 13.8. 110/09/07: potassium was 4.3, creatinine 0.97, ALT (SGPT) 21, glucose AC 99, total cholesterol 154, triglyceride 176, and LDL cholesterol 92.	**Planning:** Cardiac catheterization. if symptoms persisted Keep Kerlone, Cozaar Encourage exercise and diet control	Keep current Rx	**Cardiac catheterization. If symptoms persisted, keep Kerlone, Cozaar, and encourage exercise and diet control.**

Figure 3: Plan Generation Results: Human Doctor, Baseline LLM, and MEDPLAN

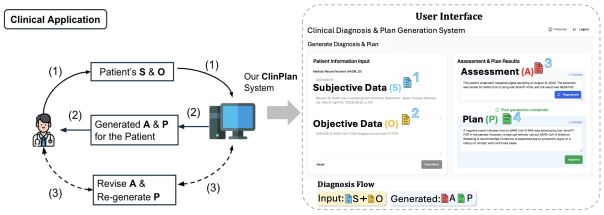

Figure 4: Overview of the Clinical Application of the MEDPLAN System

Víctor H Castillo, Ana I Martínez-García, Leonel Soriano-Equigua, Fermín Marcelo Maciel-Mendoza, José Luis Álvarez-Flores, and Reyes Juárez-Ramírez. 2019. An interaction framework for supporting the adoption of ehrs by physicians. *Universal Access in the Information Society*, 18(2):399–412.

Zhen Chen, Zhihao Peng, Xusheng Liang, Cheng Wang, Peigan Liang, Linsheng Zeng, Minjie Ju, and Yixuan Yuan. 2025. Map: Evaluation and multi-agent enhancement of large language models for inpatient pathways. *arXiv preprint arXiv:2503.13205*.

ContactDoctor. 2024. Bio-medical: A high-performance biomedical language model. https://huggingface.co/ContactDoctor/Bio-Medical-Llama-3-8B.

Kate Curtis, Margaret Fry, Ramon Z Shaban, and Julie Considine. 2017. Translating research findings to clinical nursing practice. *Journal of clinical nursing*, 26(5-6):862–872.

Michael Han Daniel Han and Unsloth team. 2023. Unsloth.

Jun-En Ding, Phan Nguyen Minh Thao, Wen-Chih Peng, Jian-Zhe Wang, Chun-Cheng Chug, Min-Chen Hsieh, Yun-Chien Tseng, Ling Chen, Dongsheng Luo, Chenwei Wu, et al. 2024. Large language multimodal models for new-onset type 2 diabetes prediction using five-year cohort electronic health records. *Scientific Reports*, 14(1):20774.

R Scott Evans. 2016. Electronic health records: then, now, and in the future. *Yearbook of medical informatics*, 25(S 01):S48–S61.

Xiaojing Fan and Chunliang Tao. 2024. Towards resilient and efficient llms: A comparative study of efficiency, performance, and adversarial robustness. *arXiv preprint arXiv:2408.04585*.

Hsin-Ling Hsu and Jengnan Tzeng. 2025. Dat: Dynamic alpha tuning for hybrid retrieval in retrieval-augmented generation. *arXiv preprint arXiv:2503.23013*.

Yixing Jiang, Kameron C Black, Gloria Geng, Danny Park, Andrew Y Ng, and Jonathan H Chen. 2025. Medagentbench: Dataset for benchmarking llms as agents in medical applications. *arXiv preprint arXiv:2501.14654*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Sheng Liu, Oscar Pastor-Serrano, Yizheng Chen, Matthew Gopaulchan, Weixing Liang, Mark Buyyounouski, Erqi Pollom, Quynh-Thu Le, Michael Gensheimer, Peng Dong, et al. 2024. Automated radiotherapy treatment planning guided by gpt-4vision. *arXiv preprint arXiv:2406.15609*.

Ji Ma, Ivan Korotkov, Keith B. Hall, and Ryan T. McDonald. 2020. Hybrid first-stage retrieval models for biomedical literature. In *Conference and Labs of the Evaluation Forum*.

Rodrigo Nogueira and Kyunghyun Cho. 2020. Passage re-ranking with bert. *arXiv preprint arXiv:1901.04085*.

OpenAI. 2024a. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

OpenAI. 2024b. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

OpenAI. 2024. text-embedding-3-large. https://platform.openai.com/docs/models/text-embedding-3-large.

Anil Palepu, Valentin Liévin, Wei-Hung Weng, Khaled Saab, David Stutz, Yong Cheng, Kavita Kulkarni, S Sara Mahdavi, Joëlle Barral, Dale R Webster, et al. 2025. Towards conversational ai for disease management. *arXiv preprint arXiv:2503.06074*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Pengcheng Qiu, Chaoyi Wu, Shuyu Liu, Weike Zhao, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2025. Quantifying the reasoning abilities of llms on real-world clinical cases. *arXiv preprint arXiv:2503.04691*.

David Restrepo, Chenwei Wu, Zhengxu Tang, Zitao Shuai, Thao Nguyen Minh Phan, Jun-En Ding, Cong-Tinh Dao, Jack Gallifant, Robyn Gayle Dychiao, Jose Carlo Artiaga, et al. 2025. Multi-ophthalingua: A multilingual benchmark for assessing and debiasing llm ophthalmological qa in lmics. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 28321–28330.

Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1995. Okapi at trec-3. *Nist Special Publication Sp*, 109:109.

Hajar Sakai and Sarah S Lam. 2025. Large language models for healthcare text classification: A systematic review. *arXiv preprint arXiv:2503.01159*.

Abeed Sarker, Mohammed Ali Al-Garadi, Yuan-Chi Yang, Jinho Choi, Arshed A Quyyumi, Greg S Martin, et al. 2021. Defining patient-oriented natural language processing: a new paradigm for research and development to facilitate adoption and use by medical experts. *JMIR Medical Informatics*, 9(9):e18471.

Zipora Shechtman. 2002. Child group psychotherapy in the school at the threshold of a new millennium. *Journal of Counseling & Development*, 80(3):293–299.

Tami Sorgente, Eduardo B Fernandez, and MM Larrondo Petrie. 2005. The soap pattern for medical charts. In *Proceedings of PLoP*, volume 2005.

Liyan Tang, Zhaoyi Sun, Betina Idnay, Jordan G Nestor, Ali Soroush, Pierre A Elias, Ziyang Xu, Ying Ding, Greg Durrett, Justin F Rousseau, et al. 2023. Evaluating large language models on medical evidence summarization. *NPJ digital medicine*, 6(1):158.

Xiangru Tang, Daniel Shao, Jiwoong Sohn, Jiapeng Chen, Jiayi Zhang, Jinyu Xiang, Fang Wu, Yilun Zhao, Chenglin Wu, Wenqi Shi, et al. 2025. Medagentsbench: Benchmarking thinking models and agent frameworks for complex medical reasoning. *arXiv preprint arXiv:2503.07459*.

VoyageAI. 2024. Reranker-2. https://docs.voyageai.com/docs/reranker.

Ruslan Magana Vsevolodovna. 2024a. Medical-llama3-8b-16bit: Fine-tuned llama3 for medical q&a. https://huggingface.co/ruslanmv/Medical-Llama3-8B.

Ruslan Magana Vsevolodovna. 2024b. Medical-mixtral-7b-v2k. https://huggingface.co/ruslanmv/Medical-Mixtral-7B-v2k.

Rui Yang, Yilin Ning, Emilia Keppo, Mingxuan Liu, Chuan Hong, Danielle S Bitterman, Jasmine Chiat Ling Ong, Daniel Shu Wei Ting, and Nan Liu. 2025. Retrieval-augmented generation for generative artificial intelligence in health care. *npj Health Systems*, 2(1):2.

Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony B Costa, Mona G Flores, et al. 2022. A large language model for electronic health records. *NPJ digital medicine*, 5(1):194.

Jingqing Zhang, Kai Sun, Akshay Jagadeesh, Parastoo Falakaflaki, Elena Kayayan, Guanyu Tao, Mahta Haghighat Ghahfarokhi, Deepa Gupta, Ashok Gupta, Vibhor Gupta, et al. 2024. The potential and pitfalls of using a large language model such as chatgpt, gpt-4, or llama as a clinical assistant. *Journal of the American Medical Informatics Association*, 31(9):1884–1891.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Zeyu Zhou. 2023. Evaluation of chatgpt's capabilities in medical report generation. *Cureus*, 15(4).

A System Architecture

Our system architecture is designed for real-world deployment, ensuring robustness and efficiency when handling large-scale requests in the future. As illustrated in Figure 5, the backend is implemented using FastAPI, designed for high concurrency and efficient request handling. Instead of synchronous API calls, which may lead to memory overload or timeouts, we adopt an asynchronous task management approach. Upon receiving input, the backend assigns a unique task ID and forwards the request to the LLM. Once processing is completed, the system returns the results alongside the task ID, ensuring a seamless experience without blocking other requests.

MEDPLAN integrates two databases to support its functionality. Microsoft SQL Server stores structured patient data, allowing efficient retrieval of the latest consultation records using MRN (Medical Record Number) as a key. Additionally, Weaviate, a vector database, is employed to store a large repository of past patient records. These enable retrieval-augmented generation (RAG), allowing the system to identify cross-patient similar cases and provide physicians with relevant contextual information.

The user interface is developed using React, providing an intuitive web-based platform for physicians to interact with the system. The underlying LLM is deployed on our GPU server, which is equipped with NVIDIA hardware, ensuring efficient real-time inference and responsiveness.

A.1 Implementation Details

We instruction-tuned three domain-specific LLMs—Medical-Llama3-8B (Vsevolodovna, 2024a), Medical-Mixtral-7B-v2k (Vsevolodovna, 2024b), and Bio-Medical-Llama3-8B (ContactDoctor, 2024)—using the Unsloth framework (Daniel Han and team, 2023) for efficient adaptation with long-context support. All models were trained on NVIDIA RTX 6000 Ada Generation GPUs with Low-Rank Adaptation (LoRA), dynamically adjusted for each model's architecture. A maximum sequence length of 65,536 tokens was used to accommodate extended patient histories and cross-patient references. The training employed the AdamW optimizer in 8-bit precision, along with a cosine learning rate scheduler and a warm-up phase equal to 1.6% of the total steps.

For semantic retrieval, we used OpenAI's text-embedding-3-large model (OpenAI, 2024), which supports high-dimensional dense representations suitable for medical content. As our cross-encoder model, we employed the VoyageAI Reranker-2 (VoyageAI, 2024), which was used to re-rank the semantically retrieved candidates in our RAG pipeline. All experiments were conducted under consistent hardware and software configurations to ensure comparability.

B Generation Samples

Figure 3 demonstrates a significant improvement in clinical decision support capabilities between the best baseline Medical-Mixtral-7B-v2k model and MEDPLAN with the Medical-Mixtral-7B-v2k model as the base model. The baseline model only produced the simple result, "Keep current Rx", while dealing with a complicated patient scenario that included several cardiovascular risk factors, such as hyperlipidemia, hypertension, metabolic syndrome, and pre-diabetes. This result indicates a troubling missing core diagnostic and treatment components necessary for evidence-based treatment.

In contrast, our approach produced a comprehensive, clinically sound recommendation that aligns remarkably with expert human physician judgment. Our model's output "Cardiac catheterization. If symptoms persist, keep Kerlone, Cozaar, and encourage exercise and diet control" demonstrates the model's enhanced capacity to (1) prioritize appropriate diagnostic procedures for suspected coronary artery disease, (2) implement condition-based medication management strategies, and (3) incorporate preventive lifestyle interventions addressing modifiable risk factors.

When a subset of the generated samples was presented to physicians at Far Eastern Memorial Hospital (FEMH) for evaluation, the proposed method demonstrated approximately 66% improvement in clinical assessments compared to the baseline approach.

These findings highlight how combining RAG with two-stage targeted instruction tuning of LLMs can substantially improve AI clinical reasoning capabilities, potentially enhancing model utility in real-world medical decision support systems. Our proposed approach exhibits precise clinical reasoning, addressing both urgent diagnostic needs and long-term illness management concerns, suggest-

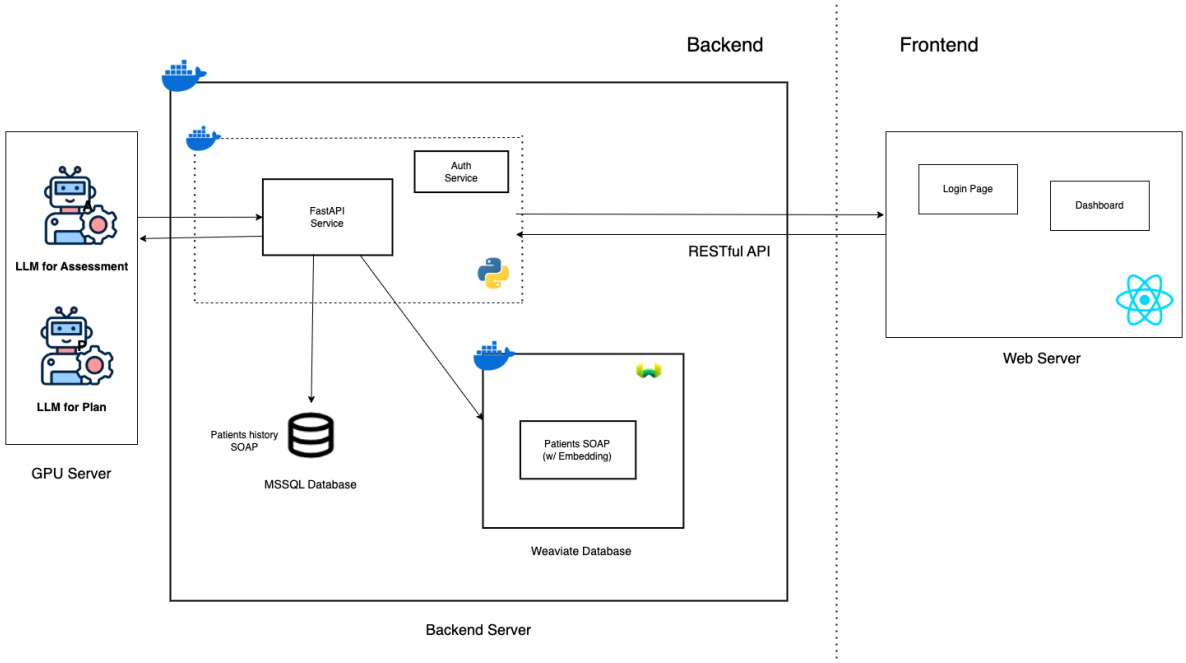

Figure 5: MEDPLAN System Architecture.

ing promising directions for medical AI applications in healthcare settings.

C Prompt Template

We present our prompt template (Figure 6) to guide the generation by the LLMs. The left figure outlines the Assessment Generation template, while the right figure introduces the Plan Generation template. Each template contains three key sections:

- **Role & Instruction**: Directs an AI Medical Assistant to synthesize patient data using chain-of-thought reasoning.
- **User Prompt**: Provides structured query formats with placeholders for patient-specific information.
- **Generation**: Designates space for AI-generated content ([A_latest] or [P_latest]).

D Limitation

The main limitation of this study lies in the data source and applicability. Our models are trained on EHR SOAP records from a specific hospital, which may limit its generalizability to other medical institutions or specialties. Additionally, while MEDPLAN employs retrieval-augmented generation (RAG) to enhance accuracy, it is still subject to inherent biases in language models, potentially leading to generating content that does not fully align with medical standards. These limitations highlight the need for continuous improvements and rigorous evaluation in real-world settings.

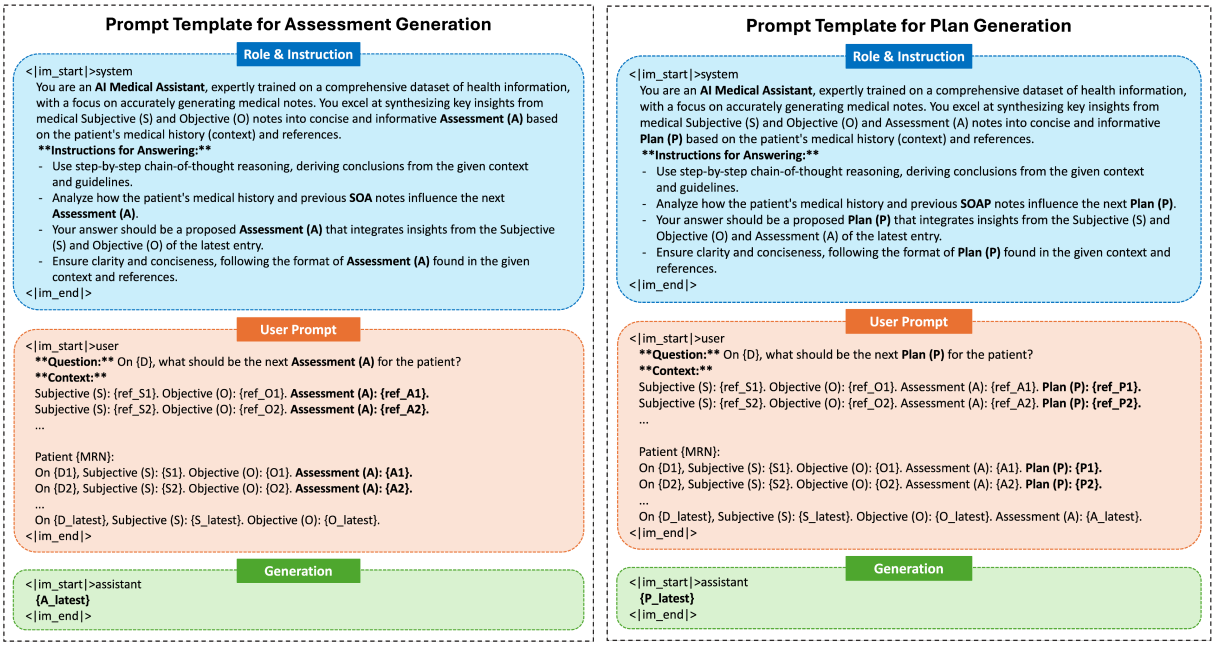

Figure 6: Prompt Template for Generation