

ARMED: Adaptive Retrieval-Augmented Medication Recommendation

^{1st} Kyungkeon Chung
UX Factory, Inc.

Seongnam-si, South Korea
kunnyjung@uxf.ai

^{2nd} Se Un Park
UX Factory, Inc.

Seongnam-si, South Korea
s4park123@uxf.ai

^{3rd} Hyejin Lee
UX Factory, Inc.

Seongnam-si, South Korea
hyejin@uxf.ai

^{4th} Junyoung Park
UX Factory, Inc.

Seongnam-si, South Korea
junyoung@uxf.ai

Abstract—Medication recommendation requires jointly considering diagnoses, comorbidities, longitudinal patient history, and treatment interactions. Although recent retrieval-augmented generation approaches using language models have shown promise for clinical recommendation, most existing systems rely on fixed retrieval pipelines that cannot adapt to diverse clinical retrieval scenarios and may introduce noisy or irrelevant evidence. We propose ARMED, an adaptive clinical RAG framework that dynamically expands retrieval according to retrieval sufficiency and current-admission relevance. ARMED first observes the initial retrieval outcome and selectively performs supplementary or fallback multi-query retrieval when additional evidence is needed. The framework also incorporates a relevance gate that activates only clinically relevant same-patient history and organizes retrieved evidence into structured query-level clinical evidence. Experiments on MIMIC-IV demonstrate that ARMED consistently outperforms previous static retrieval baselines and improves overall medication recommendation quality. By combining locally runnable small language models with adaptive retrieval, ARMED supports medication recommendation without relying on external language model APIs or task-specific fine-tuning.

Index Terms—Retrieval-Augmented Generation, Medication Recommendation, Adaptive Retrieval, Clinical Decision Support, Small Language Models

I. INTRODUCTION

Medication recommendation is an important task in electronic health record (EHR)-based clinical decision support [1], [2]. Since multiple medication classes are often prescribed simultaneously during hospital admissions, medication recommendation is inherently a complex multi-label prediction problem [3]. Effective recommendation requires consideration of clinical information such as diagnoses, comorbidities, longitudinal patient history, prior treatments, and treatment interactions [4]. Previous studies have therefore proposed various deep learning-based methods using structured EHR information [1], [2], [3], [5], [6].

Recently, large language models (LLMs) have been explored for clinical reasoning and medical recommendation tasks [7], [8]. However, medication recommendation requires clinically grounded and label-constrained outputs, making LLM-only approaches vulnerable to unsupported medication label generation and hallucinated recommendations [9]. Retrieval-augmented generation (RAG)-based clinical recommendation frameworks, such as MEDPLAN [10], PACE-RAG [11], and TreatRAG [12], address this limitation by

grounding generated outputs in retrieved clinical evidence. Such retrieval-grounded systems are also relevant to hospital environments where privacy requirements, internal-network deployment, and restrictions on external API use limit reliance on cloud-based proprietary models [13], [14].

Despite these advances, many clinical RAG methods still rely on static or pre-defined retrieval pipelines. Static retrieval may introduce practical failure modes such as evidence dilution, majority bias, and history noise, where loosely related retrieved cases or irrelevant previous visits negatively affect the final recommendation. Although PACE-RAG progressively refines evidence through a multi-stage pipeline, its retrieval flow remains largely pre-defined rather than dynamically controlled according to patient-specific retrieval sufficiency or history relevance [11].

In this work, we propose **ARMED**, an adaptive clinical RAG framework for medication recommendation. ARMED dynamically adjusts retrieval behavior according to patient-specific clinical context, selectively utilizes clinically relevant same-patient history, and generates recommendations based on dynamically retrieved evidence. Inspired by recent adaptive RAG studies [15], [16], [17], [18], [19], [20], our framework incorporates retrieval sufficiency and patient history relevance to mitigate noisy evidence injection and unsupported recommendations.

The main contributions of this work are as follows:

- We propose **ARMED**, an adaptive clinical RAG framework that dynamically adjusts retrieval strategies according to patient-specific clinical context.
- We introduce a retrieval sufficiency-based controller that determines whether to expand retrieval based on the initial retrieval outcomes.
- We mitigate history noise by selectively utilizing clinically relevant same-patient history based on current admission relevance.
- We design an adaptive retrieval-based medication recommendation framework with consideration for compact and locally deployable language model environments.

The remainder of this paper presents related work in Section II, the proposed framework in Section III, the experimental setup in Section IV, the results in Section V, and the conclusion in Section VI.

II. RELATED WORK

Medication recommendation has been widely studied as an EHR-based multi-label prediction task [1], [3]. Conventional deep learning approaches learn relationships among diagnoses, procedures, medications, and longitudinal patient histories to predict subsequent prescriptions [1], [6]. Although these methods have improved prescription prediction, they typically require task-specific training and do not explicitly retrieve patient-level clinical evidence during inference. This limits their flexibility when medication label spaces, patient populations, or clinical workflows change.

Beyond task-specific learning from EHR data, prior clinical cases can provide evidence for new decisions, an approach that predates modern RAG systems. Case-based reasoning has been studied in clinical decision support as a way to reuse previous cases [21], while structure-mapping theory explains analogical reasoning through relational correspondences between situations [22]. Clinical RAG follows a related case-reuse perspective when retrieving patient records to support recommendations. TreatRAG [12] retrieves similar historical admissions, MEDPLAN [10] incorporates patient evidence into clinical planning, and PACE-RAG [11] refines evidence and prescription decisions through a multi-stage pipeline.

Although PACE-RAG progressively refines evidence through a pre-defined pipeline [11], ARMED determines whether retrieval expansion is needed based on the initial retrieval outcome and selectively incorporates same-patient history according to current-admission relevance.

Recent general-domain RAG studies have increasingly treated retrieval as a dynamically controllable process rather than a fixed preprocessing step. FLARE [18], Self-RAG [15], CRAG [16], Adaptive-RAG [17], and RQ-RAG [19] have explored retrieval necessity estimation, retrieval quality assessment, and query- or context-dependent retrieval control. In this work, we apply this adaptive retrieval perspective to medication recommendation.

III. METHOD

ARMED combines same-patient history from previous visits with cross-patient evidence from similar admissions in the retrieval database. Given the current admission diagnoses, a history relevance gate selects relevant previous visits (Section III-A), while adaptive retrieval starts with a single query and expands when the retrieved evidence is insufficient (Section III-B). Both evidence sources are structured to guide the final LLM-based medication recommendation (Section III-C). Figs. 1 and 2 illustrate the adaptive retrieval workflow and overall framework, respectively.

A. Relevant Patient History Selection

Same-patient history can provide useful signals for medication recommendation, particularly for chronic diseases or recurrent admissions. However, prior visits may contain diagnoses and prescriptions unrelated to the current admission, introducing irrelevant context when included indiscriminately.

Require: admission x , visits V_x , retrieval DB $\mathcal{D}$, similarity thresholds T, T_{low} , minimum hits K

Require: $R(\mathcal{D}, q)$: retrieve candidate c and score s from $\mathcal{D}$, maximum rewrites R_{max}

```

1:  $H \leftarrow \text{HISTORYGATE}(x, V_x)$ 
2:  $E \leftarrow \text{ADAPTIVERETRIEVE}(x, \mathcal{D})$ 
3:  $E_q \leftarrow \text{BUILDQUERYEVIDENCE}(E, H)$ 
4: return  $\text{LLMRECOMMENDATION}(x, E_q)$ 

5: procedure  $\text{ADAPTIVERETRIEVE}(x, \mathcal{D})$ 
6:    $(C_0, S_0) \leftarrow \{(c, s) \in R(\mathcal{D}, x) : s \geq T\}$ 
7:   if  $|C_0| \geq K$  then
8:     if top and mean of  $S_0$  are high then
9:       return  $C_0$  ▷ Single-query sufficient
10:    else
11:      return  $C_0 \cup \text{MULTIQUERY}(x, \mathcal{D})$  ▷ Supplement
12:   else
13:     return  $\text{MULTIQUERY}(x, \mathcal{D})$  ▷ Fallback

14: procedure  $\text{MULTIQUERY}(x, \mathcal{D})$ 
15:    $\{q_1^0, \dots, q_M^0\} \leftarrow \text{SLMDECOMPOSE}(x)$ 
16:    $E \leftarrow \emptyset$ 
17:   for each decomposed query  $q_i^0$  do
18:     for  $r \leftarrow 0$  to  $R_{\text{max}}$  do
19:        $C \leftarrow \{(c, s) \in R(\mathcal{D}, q_i^r), s \geq T\}$ 
20:       if  $|C| \geq K$  then
21:          $E \leftarrow E \cup C$ 
22:         break
23:       else if  $r < R_{\text{max}}$  then ▷ Rewrite
24:          $q_i^{r+1} \leftarrow \text{SLMREWRITE}(q_i^r)$ 
25:       else ▷ Rescue
26:          $E \leftarrow E \cup \{(c, s) \in R(\mathcal{D}, q_i^0), s > T_{\text{low}}\}$ 
27:   return  $E$ 

```

Fig. 1. Adaptive retrieval workflow.

ARMED therefore employs an SLM-based history relevance gate that jointly considers the current admission diagnoses and the diagnoses and medications from each prior visit to perform binary relevance classification. Only relevant visits are retained to reduce irrelevant historical context in the final LLM prompt. Their medications are incorporated as separate history evidence alongside cross-patient retrieval evidence for the final recommendation.

B. Adaptive Retrieval Policy

Retrieval sufficiency cannot be reliably predicted from query complexity alone, as it also depends on the cases available in the retrieval database. Complex admissions may have strongly matched cases, whereas seemingly simple admissions may retrieve insufficient evidence. ARMED therefore performs an initial search and uses the observed results to determine whether retrieval expansion is needed.

As illustrated in Fig. 3, ARMED first retrieves using the current admission's full diagnosis list as a single query.

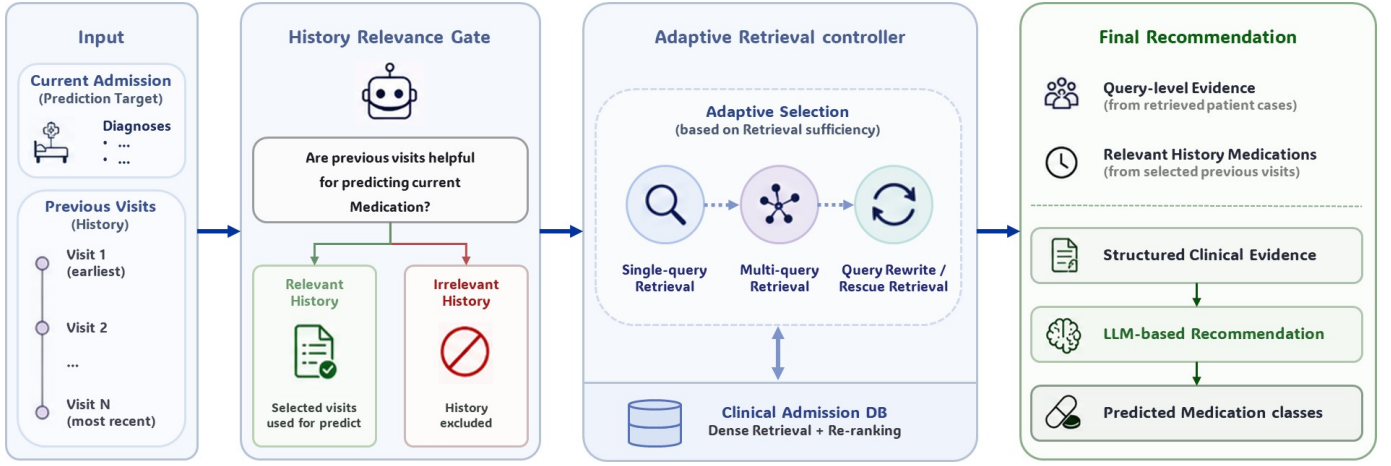


Fig. 2. Overview of ARMED.

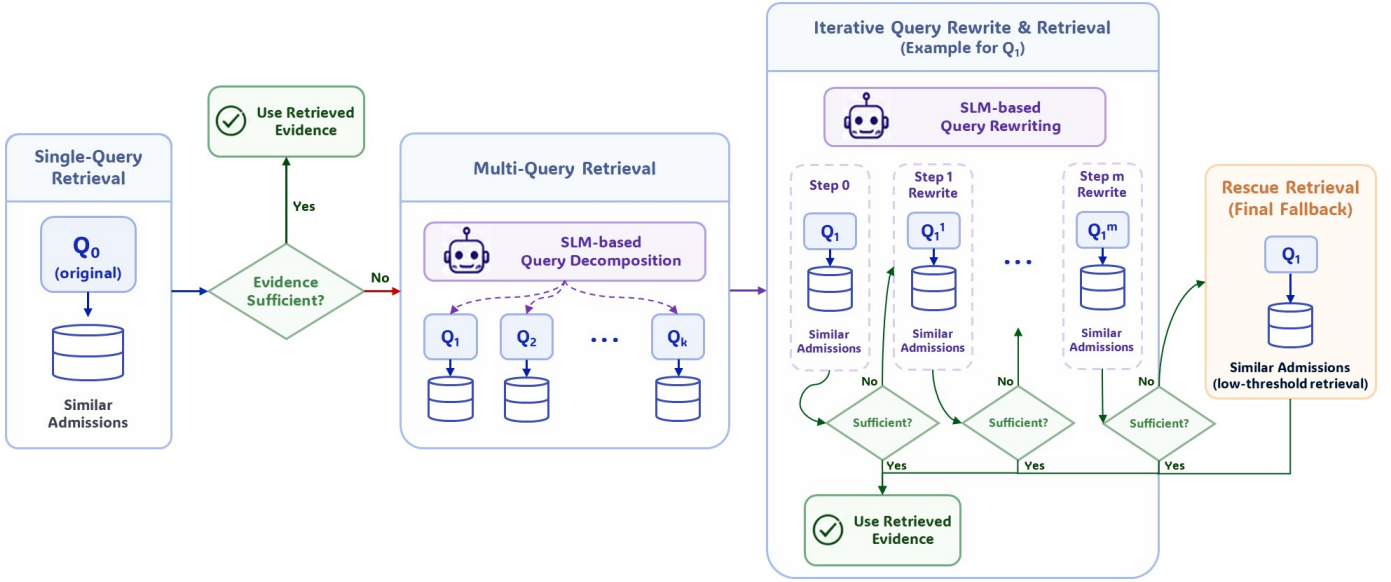


Fig. 3. Adaptive retrieval with single-query selection, multi-query expansion, rewriting, and rescue.

Diagnosis texts are encoded as dense sentence embeddings and compared using cosine similarity. Cases with similarity scores of at least the threshold are accepted. Retrieval sufficiency is assessed using the number of accepted hits and their maximum and mean similarity scores.

High-confidence results are used without expansion (“single-query sufficient”). If the minimum hit requirement is met but the high-confidence criteria are not, ARMED retains the initial results and adds multi-query evidence (“supplement”). Otherwise, multi-query evidence replaces the initial results (“fallback”). For both routes, an SLM decomposes the diagnosis list into focused clinical queries, each retrieved independently.

Sub-queries with insufficient evidence are rewritten by an SLM and retried up to the maximum rewrite count (“rewrite”). If evidence remains insufficient, a rescue pass uses the original sub-query with a lower similarity threshold (“rescue”). Sub-

[System / task description ...]

Given the target patient’s current diagnoses **{CURRENT_DIAGNOSES}**, query-level retrieval evidence **{QUERY_LEVEL_EVIDENCE}**, and optional same-patient history **{HISTORY_EVIDENCE}**, recommend medication class labels for the target admission.

Select labels only from the provided candidate set; do not introduce unsupported labels or alter label names. ...

Fig. 4. Abstracted final recommendation prompt. Ellipses indicate prompt content omitted for brevity.

queries with no results meeting the applicable threshold contribute no evidence. This policy expands retrieval only when the initial evidence is insufficient.

C. Final Recommendation

Retrieved cases are not directly inserted into the LLM prompt as raw retrieval outputs. Instead, ARMED organizes retrieved diagnosis-prescription evidence into a structured query-level evidence format for the final recommendation stage.

For each retrieval query, the framework summarizes retrieved patient cases together with their associated diagnoses and prescription patterns. Repeatedly retrieved medications are additionally aggregated at the query level. This query-specific organization helps preserve the clinical context associated with each retrieval query rather than relying solely on simple frequency-based medication aggregation.

When relevant same-patient history is activated, selected history medications are incorporated as separate history evidence. The final LLM recommendation is then generated using the current diagnosis information together with structured query-level retrieval evidence and selected history evidence (Fig. 4).

IV. EXPERIMENTAL SETUP

A. Dataset and Evaluation Splits

In this study, we constructed an admission-level medication recommendation dataset using MIMIC-IV v3.1 [23], [24]. MIMIC-IV contains real-world clinical records deidentified by the dataset providers before release, including removal of patient identifiers and patient-specific date shifting that preserves within-patient temporal intervals. Access was obtained through PhysioNet [24] after completing the required training and credentialing process and agreeing to the data use agreement. We selected MIMIC-IV because its linked diagnosis, prescription, and longitudinal admission records support evaluation of both cross-patient retrieval and same-patient history utilization.

Diagnosis information was represented using ICD diagnosis titles, while prescription information was converted into ATC level 3 medication class names through RxNorm [25] and RxClass [26] mapping based on the April 2026 NIH RxNorm Prescribable release. The final prediction target was defined as the medication class label set for each admission, formulating the task as a multi-label recommendation problem in which multiple medication classes may exist within a single admission.

After removing admissions without diagnosis information or mapped medication labels, the final dataset consisted of 462,084 admissions from 196,194 unique patients. The dataset contained 390 unique medication labels in total, with an average of 12.8 diagnoses and 12.4 medication labels per admission.

The retrieval corpus comprised 461,084 admissions from 195,818 patients with 390 unique medication labels, while the inference set comprised 1,000 admissions from 376 patients with 248 labels. The two sets were patient-disjoint by `subject_id`, with no patient or admission overlap. All inference labels were present in the retrieval corpus; thus, the

evaluation focused on recommending medication labels supported by existing prescription evidence rather than generating unseen labels.

We used two evaluation settings to separate overall performance evaluation from history selection analysis. In the admission-level setting, we evaluated 1,000 admissions from 376 patients. These comprised 376 first admissions without prior history and 624 subsequent admissions with prior same-patient history.

The patient-latest setting was used only for the history selection analysis (Section V-B), evaluating the latest admissions of the 167 patients with prior history and excluding the 209 patients without history. All other MIMIC-IV evaluations used the admission-level setting. In both settings, only admissions preceding the evaluated admission were available as history candidates.

B. Implementation Details

For final recommendation generation in ARMED, we used Qwen3-8B [27]. In addition, lightweight control tasks such as diagnosis decomposition, query rewriting, and history relevance gating were performed using Qwen2.5-3B [28]. All LLM/SLM inference was conducted using Ollama [29].

For dense retrieval embedding, we used the all-MiniLM-L6-v2 sentence-transformer model [30], [31] and constructed a FAISS [32] vector database using the full retrieval corpus. In addition, a MiniLM-based cross-encoder reranker trained on MS MARCO [33] was used to improve retrieved case ranking quality.

The initial single-query retrieval accepted cases with similarity scores of at least 0.80. With at least two accepted cases, retrieval terminated without expansion if both the maximum and mean scores reached 0.88 and 0.84, respectively; otherwise, supplementary multi-query retrieval was performed. Fewer than two accepted cases triggered multi-query fallback. We allowed up to three alternative rewrites per decomposed query. These thresholds were fixed experimental settings, without a claim of optimality.

We evaluated final medication recommendations as a multi-label prediction task using admission-averaged precision, recall, and F1. These metrics assess agreement between predicted and recorded medication-label sets, rather than the relevance of retrieved documents.

For admission i , let Y_i and $\hat{Y}_i$ denote the ground-truth and predicted medication-label sets, respectively. Labels were matched exactly after lowercasing and normalizing punctuation and whitespace. We computed

$$P_i = \frac{|Y_i \cap \hat{Y}_i|}{|\hat{Y}_i|}, \quad R_i = \frac{|Y_i \cap \hat{Y}_i|}{|Y_i|}, \quad F1_i = \frac{2|Y_i \cap \hat{Y}_i|}{|Y_i| + |\hat{Y}_i|}. \quad (1)$$

Ratios with zero denominators were assigned zero. Each metric was averaged over admissions, not medication classes. Reported F1 is the mean of per-admission F1 scores, not the harmonic mean of averaged precision and recall.

All experiments were conducted on a workstation equipped with an AMD Ryzen 7 9800X3D CPU, an NVIDIA GeForce

TABLE I
OVERALL MEDICATION RECOMMENDATION PERFORMANCE.

Method	F1	95% CI	Precision	Recall
History	0.2926	[0.255, 0.326]	0.2859	0.3352
Static RAG	0.4093	[0.397, 0.422]	0.3841	0.5084
Static RAG + History	0.4472	[0.432, 0.463]	0.4121	0.5669
TreatRAG	0.4009	[0.378, 0.425]	0.4002	0.4712
PACE-RAG	0.3557	[0.312, 0.397]	0.4157	0.3553
ARMED	0.5103	[0.491, 0.531]	0.4910	0.6263

TABLE II
STATISTICAL SIGNIFICANCE OF F1 IMPROVEMENTS.

Comparison	$\Delta F1$	95% CI	p_{Holm}
vs. Static RAG + History	+0.0631	[0.0522, 0.0735]	< 0.001
vs. TreatRAG	+0.1094	[0.0916, 0.1279]	< 0.001
vs. PACE-RAG	+0.1546	[0.1271, 0.1855]	< 0.001

RTX 5080 GPU with 16 GB of VRAM, and 64 GB of DDR5 memory, running 64-bit Windows 10.

V. RESULTS

A. Overall Performance

Table I reports overall medication recommendation performance using the same Qwen3-8B backbone and 1,000 inference admissions from 376 patients for all methods. For F1, we report 95% patient-cluster bootstrap confidence intervals computed from 10,000 bootstrap resamples.

The History baseline uses the current admission’s diagnoses and available same-patient history without external retrieval. Static RAG uses the complete diagnosis list of the current admission as a single query for fixed dense retrieval and incorporates the retrieved cases into the LLM context, without same-patient history. Static RAG + History combines this retrieved evidence with same-patient history. Both history-using baselines include available history without relevance filtering. Static RAG achieved higher F1 than History, and adding same-patient history further improved F1.

TreatRAG [12] was reimplemented following its original lexical n-gram Jaccard retrieval pipeline and achieved slightly higher precision but slightly lower F1 than Static RAG. We evaluated PACE-RAG [11] using the authors’ public implementation. Consistent with the history-dependent performance reported in that study, PACE-RAG achieved F1 scores of 0.5269 with prior same-patient history and 0.0716 without history in our evaluation, compared with 0.5462 and 0.4508 for ARMED, respectively. Overall, ARMED achieved the highest F1, precision, and recall among all evaluated methods.

Paired patient-cluster bootstrap tests using 10,000 resamples showed that ARMED achieved significantly higher F1 than Static RAG + History, TreatRAG, and PACE-RAG (Table II), with Holm-adjusted $p < 0.001$ across all three comparisons.

Table III shows that ARMED predicted fewer medication labels than Static RAG + History while increasing true positives (TP) and reducing false positives (FP) and false negatives

TABLE III
MEAN MEDICATION PREDICTION COUNTS PER ADMISSION.

Method	Predicted	TP	FP	FN
Static RAG + History	15.834	6.490	9.344	6.136
ARMED	14.652	7.146	7.506	5.480

TABLE IV
EFFECT OF SAME-PATIENT HISTORY SELECTION.

Method	F1	Precision	Recall
w/o History	0.4387	0.4279	0.5293
Always-History	0.4783	0.4629	0.5757
ARMED	0.4792	0.4635	0.5777

(FN), indicating more accurate label selection rather than a larger prediction set.

B. Relevant History Selection Analysis

Table IV evaluates how same-patient history selection affects final medication recommendation on the latest admissions of the 167 test patients with prior history. The *w/o History* condition excludes history, *Always-History* incorporates recent prior visits without relevance filtering, and ARMED incorporates only visits retained by its history relevance gate. All three conditions use the same ARMED pipeline, differing only in same-patient history usage. Including history improved recommendation performance, while relevance filtering yielded a small additional gain.

The gate removed 6 visits from 5 patients. The 342 retained visits shared an average of 6.62 medication labels with the current admission’s ground truth, compared with 1.00 for removed visits. Among the 5 affected patients, mean F1 increased from 0.3632 to 0.3908, with no decrease in F1 for any patient. Together, these findings suggest that most available prior visits provided relevant medication context in our test set, while selective filtering helped exclude less relevant history.

C. Retrieval Policy Ablation

Table V compares retrieval policies with same-patient history disabled. All policies share the same backbone, retrieval corpus, and no-history ARMED pipeline, differing only in retrieval policy. Unlike the simple Static RAG baseline, which directly provides densely retrieved cases to the LLM, these ablations retain ARMED’s reranking, structured evidence aggregation, and label-constrained generation. GT Coverage measures the fraction of ground-truth medication labels covered by retrieved evidence, while Evidence Precision measures the fraction of retrieved medication labels matching the ground truth. Both metrics are averaged over admissions using the union of retrieved medication labels, with empty retrievals scored as zero. They assess retrieved evidence, not final predictions.

The *single-query only* baseline retrieves using the complete diagnosis list as one query. The *multi-query only* baseline

TABLE V
RETRIEVAL POLICY ABLATION WITHOUT SAME-PATIENT HISTORY.

Method	F1	GT Coverage	Evidence Precision
Single-query only	0.4178	0.7174	0.2406
Multi-query only	0.3943	0.6329	0.2690
Adaptive (ARMED)	0.4417	0.7825	0.2437

TABLE VI
PERFORMANCE AND RUNTIME COMPARISON WITHOUT SAME-PATIENT HISTORY.

Metric	Adaptive	Multi-query only
F1	0.4417	0.3943
Mean latency (s)	14.686	16.865
P50 latency (s)	12.088	13.466
P95 latency (s)	28.630	33.934
Throughput (/h)	245.1	213.5

always decomposes the diagnosis list into focused queries and retrieves evidence for each. The *adaptive retrieval* policy proposed in ARMED retains the initial single-query results when sufficient and otherwise invokes multi-query retrieval to supplement or replace them.

Always applying multi-query retrieval yielded the highest Evidence Precision but the lowest GT Coverage and F1, suggesting that more focused evidence alone does not ensure better recommendations. Adaptive retrieval achieved the highest GT Coverage and F1 while maintaining Evidence Precision comparable to single-query retrieval. These results support selectively expanding insufficient retrieval to improve evidence coverage and recommendation performance.

Table VI additionally compares runtime under the same settings. Relative to *multi-query only*, *adaptive retrieval* reduced mean latency by 12.9% and increased batch-size-1 sequential throughput by 14.8%, while achieving higher F1. Thus, selective retrieval expansion improved both recommendation performance and inference efficiency.

D. Retrieval Strategy Breakdown

Table VII summarizes ARMED’s retrieval routes with same-patient history disabled. *Single-query sufficient* uses the initial retrieval alone when sufficient; *Supplement* retains the initial results and adds diagnosis-focused multi-query evidence; and *Fallback* replaces insufficient initial results with multi-query evidence. *Single-query sufficient* achieved the highest F1, precision, and recall, while *Fallback* was selected least often and had the lowest scores. Overall, 44.9% of admissions required no retrieval expansion, while 44.8% used supplementation and only 10.3% required fallback, illustrating selective rather than uniform multi-query expansion.

E. Generalization Across Backbone Models

Table VIII compares ARMED with Static RAG + History across backbone models under the same evaluation protocol, with $\Delta F1$ denoting ARMED’s improvement. Representative clinical and general-purpose models include MedGemma 1.5

TABLE VII
RETRIEVAL STRATEGY BREAKDOWN WITHOUT SAME-PATIENT HISTORY.

Route	<i>N</i>	F1	Precision	Recall
Single-query sufficient	449	0.4787	0.4587	0.5881
Supplement	448	0.4167	0.4215	0.4962
Fallback	103	0.3888	0.3884	0.4798

TABLE VIII
MEDICATION RECOMMENDATION PERFORMANCE ACROSS BACKBONE MODELS.

Backbone	Static RAG + History	ARMED	$\Delta F1$
MedGemma 1.5 4B	0.4074	0.4891	+0.0817
Qwen3-8B	0.4472	0.5103	+0.0631
Phi-3.5 Mini 3.8B	0.3687	0.4300	+0.0613
Gemma 3 4B	0.4323	0.4825	+0.0502
EXAONE 3.5 2.4B	0.4143	0.4554	+0.0411
Ministral 3 8B	0.4900	0.4921	+0.0021

4B, Gemma 3 4B, Phi-3.5 Mini 3.8B, EXAONE 3.5 2.4B, and Ministral 3 8B [34], [35], [36], [37], [38].

ARMED improved F1 across all six backbones, with gains ranging from 0.0021 for Ministral 3 8B to 0.0817 for MedGemma 1.5 4B. These results suggest that its benefits extend across backbone models, although their magnitude varies.

F. Qualitative Case Study: Static Retrieval Failure and Adaptive Recovery

Table IX illustrates three examples of adaptive recovery, comparing ARMED with Static RAG + History. ARMED supplemented retrieval and retained relevant history in the *Ambiguous Comorbidity Case*, switched to multi-query fallback in the *Weak Retrieval Case*, and excluded weakly relevant prior visits in the *History Noise Control Case*, improving case-level F1 in all three examples.

G. Additional Evaluation on MIMIC-III

We additionally evaluated ARMED on MIMIC-III [39], which contains real-world clinical records de-identified by the dataset providers and was accessed through PhysioNet [24] after completing the required training and credentialing process and agreeing to the applicable data use agreement. We used the same preprocessing procedure and NDC-to-ATC mapping as for MIMIC-IV. All model and retrieval settings were kept unchanged, with no dataset-specific tuning. The retrieval corpus comprised 42,784 admissions from 33,385 patients, and the test set comprised 6,410 admissions from 5,000 patients, with no patient or admission overlap between the two sets. Of the test admissions, 1,410 had prior same-patient history.

We computed 95% confidence intervals and conducted paired method comparisons using 10,000 patient-cluster bootstrap resamples. Table X shows that ARMED improved admission-averaged F1 over Static RAG + History by 0.0247 (95% CI: [0.0220, 0.0275]; $p < 0.001$), with higher precision and similar recall. F1 increased from 0.4617 to 0.4853 on

TABLE IX
CASE STUDIES OF STATIC RETRIEVAL FAILURE AND ADAPTIVE RECOVERY.

Case Type	Ambiguous Comorbidity Case	Weak Retrieval Case	History Noise Control Case
Clinical Description	Advanced pulmonary malignancy with respiratory failure, renal disease, and cardiac ischemic comorbidity	End-stage renal disease with metabolic derangements and complex renal/urologic history	Neurologic/vascular presentation with diabetes and cardiovascular risk factors, but weakly relevant prior history
Baseline Failure	Broad retrieval caused many FP/FN errors	Single-query evidence failed to recover sufficient treatments	Recent history context increased FP burden
ARMED Action	Added multi-query supplement and retained relevant history	Switched to multi-query fallback	Rejected prior visits via history gate
Evidence Change	Recovered focus-specific candidates	Retrieved diagnosis-focused evidence	Suppressed noisy longitudinal context
F1 Change	0.125 $\rightarrow$ 0.667	0.400 $\rightarrow$ 0.812	0.235 $\rightarrow$ 0.500

TABLE X
MEDICATION RECOMMENDATION PERFORMANCE ON MIMIC-III.

Method	F1	95% CI	Precision	Recall
Static RAG + History	0.4631	[0.4590, 0.4671]	0.4144	0.6179
ARMED	0.4878	[0.4836, 0.4918]	0.4601	0.6189

admissions without prior history and from 0.4682 to 0.4967 on those with prior history ($p < 0.001$ in both groups), indicating that the gains do not depend solely on same-patient history.

VI. CONCLUSION

We proposed ARMED, an adaptive clinical RAG framework that controls retrieval depth based on retrieval sufficiency and selectively incorporates relevant same-patient history. ARMED outperformed the evaluated RAG baselines, including TreatRAG and PACE-RAG. Ablation analyses further supported the benefits of adaptive retrieval and relevance-aware same-patient history selection.

By combining small language models with adaptive retrieval, ARMED supports local inference without external language model APIs or task-specific fine-tuning. Application to a new hospital would require local data preparation and validation.

Although ARMED improved agreement with recorded medication labels, this study did not establish the clinical appropriateness or safety of its recommendations. We did not evaluate drug-drug interactions (DDIs), validate recommendations against clinical guidelines, or conduct an expert clinical assessment. Moreover, the outputs are ATC level 3 medication classes rather than specific drugs with doses and administration routes, limiting their direct applicability to prescribing decisions. Performance on long-tail medication classes and generalization to other institutions remain to be evaluated.

Future work could replace the current heuristic-based retrieval controller with a trainable lightweight SLM agent capable of learning more robust retrieval decisions from clinical context. Another important direction is to integrate medical guideline retrieval into the framework in order to refine retrieved evidence in a guideline-aware manner. Furthermore, future studies may evaluate ARMED on disease-specific EHR

cohorts, such as diabetes cohorts, to assess its potential utility for prescription decision support in real hospital environments.

ACKNOWLEDGMENT

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2025-02263669, Development of server-level DRAM-stacked PIM solution for accelerating ultra-large AI models).

REFERENCES

- [1] J. Shang, C. Xiao, T. Ma, H. Li, and J. Sun, "Gamenet: Graph augmented memory networks for recommending medication combination," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, p. 1126–1133, July 2019. [Online]. Available: <http://dx.doi.org/10.1609/aaai.v33i01.33011126>
- [2] Y. An, L. Zhang, M. You, X. Tian, B. Jin, and X. Wei, "Mesin: Multilevel selective and interactive network for medication recommendation," *Knowledge-Based Systems*, vol. 233, p. 107534, Dec. 2021. [Online]. Available: <http://dx.doi.org/10.1016/j.knosys.2021.107534>
- [3] R. Wu, Z. Qiu, J. Jiang, G. Qi, and X. Wu, "Conditional generation net for medication recommendation," in *Proceedings of the ACM Web Conference 2022*, ser. WWW '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 935–945. [Online]. Available: <https://doi.org/10.1145/3485447.3511936>
- [4] J. Mi, Y. Zu, Z. Wang, and J. He, "Acndnet: Attention-guided collaborative decision network for effective medication recommendation," *Journal of Biomedical Informatics*, vol. 149, p. 104570, Jan. 2024. [Online]. Available: <http://dx.doi.org/10.1016/j.jbi.2023.104570>
- [5] Y. Zhang, R. Chen, J. Tang, W. F. Stewart, and J. Sun, "Leap: Learning to prescribe effective and safe treatment combinations for multimorbidity," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 1315–1324. [Online]. Available: <https://doi.org/10.1145/3097983.3098109>
- [6] C. Yang, C. Xiao, F. Ma, L. Glass, and J. Sun, "Safedrug: Dual molecular graph encoders for recommending effective and safe drug combinations," 2021. [Online]. Available: <https://arxiv.org/abs/2105.02711>
- [7] C. Wu, W. Lin, X. Zhang, Y. Zhang, W. Xie, and Y. Wang, "Pmc-llama: toward building open-source language models for medicine," *Journal of the American Medical Informatics Association*, vol. 31, no. 9, p. 1833–1843, Apr. 2024. [Online]. Available: <http://dx.doi.org/10.1093/jamia/ocae045>
- [8] K. Singhal *et al.*, "Toward expert-level medical question answering with large language models," *Nature Medicine*, vol. 31, no. 3, p. 943–950, Jan. 2025. [Online]. Available: <http://dx.doi.org/10.1038/s41591-024-03423-7>
- [9] Y. Kim *et al.*, "Medical hallucinations in foundation models and their impact on healthcare," 2025. [Online]. Available: <https://arxiv.org/abs/2503.05777>

- [10] H.-L. Hsu *et al.*, “MedPlan: A two-stage RAG-based system for personalized medical plan generation,” in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*. Vienna, Austria: Association for Computational Linguistics, 2025, pp. 1072–1082. [Online]. Available: <https://aclanthology.org/2025.acl-industry.76/>
- [11] C. Huh, H. Hwang, J. H. Shin, S. Jo, J. Park, and J. C. Ye, “PACE-RAG: Patient-aware contextual and evidence-constrained RAG for clinical drug recommendation,” *arXiv preprint arXiv:2603.17356*, 2026, version 3. [Online]. Available: <https://arxiv.org/abs/2603.17356v3>
- [12] C.-C. Liu, H.-R. Yao, D.-C. Chang, and O. Frieder, “TreatRAG: A framework for personalized treatment recommendation,” in *Proceedings of the Nineteenth ACM Conference on Recommender Systems*, ser. RecSys ’25. New York, NY, USA: Association for Computing Machinery, 2025, pp. 690–695. [Online]. Available: <https://doi.org/10.1145/3705328.3748022>
- [13] M. A. Tahera, K. S. Sidhu, S. Dass, and S. Saha, “Sok: Privacy-aware llm in healthcare: Threat model, privacy techniques, challenges and recommendations,” 2026. [Online]. Available: <https://arxiv.org/abs/2601.10004>
- [14] F. Dennstädt, J. Hastings, P. M. Putora, M. Schmerder, and N. Cihoric, “Implementing large language models in healthcare while balancing control, collaboration, costs and security,” *npj Digital Medicine*, vol. 8, no. 1, Mar. 2025. [Online]. Available: <http://dx.doi.org/10.1038/s41746-025-01476-7>
- [15] A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi, “Self-rag: Learning to retrieve, generate, and critique through self-reflection,” 2023. [Online]. Available: <https://arxiv.org/abs/2310.11511>
- [16] S.-Q. Yan, J.-C. Gu, Y. Zhu, and Z.-H. Ling, “Corrective retrieval augmented generation,” 2024. [Online]. Available: <https://arxiv.org/abs/2401.15884>
- [17] S. Jeong, J. Baek, S. Cho, S. J. Hwang, and J. C. Park, “Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity,” 2024. [Online]. Available: <https://arxiv.org/abs/2403.14403>
- [18] Z. Jiang, F. F. Xu, L. Gao, Z. Sun, Q. Liu, J. Dwivedi-Yu, Y. Yang, J. Callan, and G. Neubig, “Active retrieval augmented generation,” 2023. [Online]. Available: <https://arxiv.org/abs/2305.06983>
- [19] C.-M. Chan, C. Xu, R. Yuan, H. Luo, W. Xue, Y. Guo, and J. Fu, “RQ-RAG: Learning to refine queries for retrieval augmented generation,” 2024. [Online]. Available: <https://arxiv.org/abs/2404.00610>
- [20] B. Jin, H. Zeng, Z. Yue, J. Yoon, S. Arik, D. Wang, H. Zamani, and J. Han, “Search-rl: Training llms to reason and leverage search engines with reinforcement learning,” 2025. [Online]. Available: <https://arxiv.org/abs/2503.09516>
- [21] I. Bichindaritz and C. Marling, “Case-based reasoning in the health sciences: What’s next?” *Artificial Intelligence in Medicine*, vol. 36, no. 2, pp. 127–135, 2006.
- [22] D. Gentner, “Structure-mapping: A theoretical framework for analogy,” *Cognitive Science*, vol. 7, no. 2, pp. 155–170, 1983.
- [23] A. E. W. Johnson *et al.*, “MIMIC-IV, a freely accessible electronic health record dataset,” *Scientific Data*, vol. 10, no. 1, 2023. [Online]. Available: <https://doi.org/10.1038/s41597-022-01899-x>
- [24] A. L. Goldberger *et al.*, “PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals,” *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.
- [25] U.S. National Library of Medicine, “RxNorm,” <https://www.nlm.nih.gov/research/umls/rxnorm/>, 2026, NIH RxNorm Prescribable release, accessed 2026.
- [26] —, “RxClass,” <https://mor.nlm.nih.gov/RxClass/>, 2026, ATC drug classification mapping, accessed 2026.
- [27] A. Yang *et al.*, “Qwen3 technical report,” 2025. [Online]. Available: <https://arxiv.org/abs/2505.09388>
- [28] Qwen *et al.*, “Qwen2.5 Technical Report,” Jan. 2025, arXiv:2412.15115 [cs]. [Online]. Available: <http://arxiv.org/abs/2412.15115>
- [29] Ollama, “Ollama: Get up and running with large language models locally,” <https://github.com/ollama/ollama>, 2023.
- [30] N. Reimers and I. Gurevych, “Sentence-bert: Sentence embeddings using siamese bert-networks,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, 2019, p. 3980–3990. [Online]. Available: <http://dx.doi.org/10.18653/v1/D19-1410>
- [31] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou, “Minilm: deep self-attention distillation for task-agnostic compression of pre-trained transformers,” in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS ’20. Red Hook, NY, USA: Curran Associates Inc., 2020.
- [32] M. Douze, A. Guzhva, C. Deng, J. Johnson, G. Szilvasy, P.-E. Mazaré, M. Lomeli, L. Hosseini, and H. Jégou, “The Faiss library,” 2024. [Online]. Available: <https://arxiv.org/abs/2401.08281>
- [33] P. Bajaj *et al.*, “MS MARCO: A human generated MACHine reading COMprehension dataset,” 2016. [Online]. Available: <https://arxiv.org/abs/1611.09268>
- [34] A. Sellergren *et al.*, “Medgemma 1.5 technical report,” 2026. [Online]. Available: <https://arxiv.org/abs/2604.05081>
- [35] Gemma Team *et al.*, “Gemma 3 technical report,” 2025. [Online]. Available: <https://arxiv.org/abs/2503.19786>
- [36] M. Abdin *et al.*, “Phi-3 technical report: A highly capable language model locally on your phone,” 2024. [Online]. Available: <https://arxiv.org/abs/2404.14219>
- [37] S. An *et al.*, “Exaone 3.5: Series of large language models for real-world use cases,” 2024. [Online]. Available: <https://arxiv.org/abs/2412.04862>
- [38] A. H. Liu *et al.*, “Minstral 3,” 2026. [Online]. Available: <https://arxiv.org/abs/2601.08584>
- [39] A. E. W. Johnson *et al.*, “MIMIC-III, a freely accessible critical care database,” *Scientific Data*, vol. 3, p. 160035, 2016. [Online]. Available: <https://doi.org/10.1038/sdata.2016.35>