

RAG-Based LLM Supported by Clinically Structured Re-Ranking, RL-Tuned Retrieval, and Agentic Workflow for ED Triage Prediction

Stanford CS224N Custom Project

Isha Arora
Department of Immunology
Stanford University
iarora02@stanford.edu

Charlotte Louise Kramer
Department of Computer Science
Stanford University
clkramer@stanford.edu

Nino Triandafilidis
Department of Management Science and Engineering
Stanford University
ninot@stanford.edu

Abstract

We present an open-source clinical RAG pipeline for ESI triage prediction on a MIMIC-IV-ED-derived benchmark. On held-out evaluation, broad literature retrieval does not consistently improve over direct LLM prompting. We then test a stronger version of the retrieval hypothesis by replacing broad biomedical retrieval with more task-targeted external guidance, including handbook-based tool use. This targeted access yields a better accuracy-safety trade-off than literature retrieval, but still does not produce a consistent improvement over the strongest few-shot baseline. These results suggest that the limitation is not simply that general-purpose RAG is too broad; rather, making external knowledge useful for triage requires more than substituting a more targeted source. We also report preliminary experiments on alternative retrieval improvements, including reinforcement-learning-based retrieval tuning and re-ranking. RL-tuned retrieval provides only marginal benefit over vanilla RAG and no-retrieval, though performance may improve with further hyperparameter search and reduced top-k retrieved documents.

1 Key Information to include

- Mentor: Nevin George
- External Collaborators and Sharing project: N/A

Team member	Late days remaining
Isha Arora	4
Charlotte Kramer	3
Nino Triandafilidis	2
Total pooled	9
Team size	3
Late days applied for the report	3

2 Introduction

Clinical decision-making in emergency department (ED) triage is cognitively demanding: clinicians must rapidly integrate symptoms, vitals, and patient medical history, all under time pressure. Modern health systems are increasingly saturated and face growing data complexity and patient demand against limited clinical staffing, making early acuity decisions both higher-stakes and more error-prone [1]. Under-triage risks delayed care and over-triage wastes scarce resources. Large language models (LLMs) are positioned as a potential bridge for both patients and clinicians. In particular,

Gaber et al. [1] showed that retrieval-augmented generation (RAG)-augmented LLMs can approach expert-level triage accuracy on MIMIC-IV-ED data, reporting 64.4% exact accuracy for LLM-only and 65.75% for RAG-assisted prediction with Claude 3.5 Sonnet [2]. Our preliminary experiments with Gemini 2.0 Flash, however, show vanilla RAG (top-k=5) yields lower exact accuracy than direct prompting (0.582 vs. 0.610) and virtually identical quadratic-weighted Cohen’s κ (0.293 vs. 0.292), suggesting retrieved evidence may add noise rather than signal. This prompts the question: **can retrieval quality be optimised** so that it genuinely improves triage prediction, and how should this improvement be measured in a clinically meaningful way? We investigate three complementary retrieval improvements, each targeting a different lever in the RAG pipeline (thus operating on different splits). **Clinical Re-ranking:** a clinically structured re-ranker that conditions query embeddings toward inferred clinical categories before re-scoring retrieved articles. **RL-Tuned Retrieval:** a DeepRetrieval-style query rewriter for PubMed literature and an s3-style multi-turn patient similarity searcher [3, 4]. **Agentic Workflow:** workflow chaining together specialised agents in a critique-refinement loop until a retrieval quality threshold is met. A secondary goal is to **identify evaluation metrics better suited to clinical safety**. ESI labels are heavily skewed toward levels 2 and 3, so exact accuracy rewards predicting the mode rather than the clinically safe value. This motivates alternatives such as linearly-weighted Gwet’s AC2 over standard exact accuracy. Together, these aim to produce a stronger triage model and a more honest evaluation framework.

3 Related Work

Evaluating LLM Workflows in Clinical Decision Support for Triage and Referral and Diagnosis:

Gaber et al. [1] define the MIMIC-IV-Ext triage benchmark and provide the natural starting point for this work, but their evaluation centers on exact and range accuracy rather than the design of reproducible agentic retrieval systems. We move beyond that metric frame because triage evaluation must also capture asymmetric safety errors, class imbalance, and prevalence-sensitive chance correction, which motivates our emphasis on under-/over-triage, macro recall, and AC2 alongside benchmark accuracy [5, 6, 7, 8, 9, 10]. More broadly, prior LLM-triage studies establish the clinical relevance of automated triage [11, 12, 13], while the RAG and tool-use literature motivates explicit external evidence access [14, 15, 16]; taken together, these lines of work support shifting attention away from iterative query refinement as the main design question and toward selective tool-mediated evidence access, yet they still do not provide a reproducible agentic retrieval pipeline or a packaged article corpus for follow-on experimentation. **RL-Tuned Retrieval:** There are two relevant papers that influenced the RL strategies adopted in this project. **DeepRetrieval.** Jiang et al. train a large language model to rewrite free-text queries to maximise downstream retrieval quality from RAGs [3]. The model utilizes RL with a reward based on retrieval metrics such as Normalised Discounted Cumulative Gain (NDCG) between the retrieved documents and the ground truth relevant documents. Thus, the model is trained using the HybridFlow RLHF framework to produce queries whose retrieved sets rank relevant documents higher. The model generates its query inside structured `<think>` and `<answer>` tags, encouraging chain of thought reasoning before developing a relevant rewritten query. **s3.** Jiang et al. address a core limitation of existing RL retrieval methods, claiming that retrieval-centric rewards are disconnected from downstream answer quality [4]. s3 addresses this by training just the retriever using RL in an iterative cycle of query generation, evidence accumulation, and assessing whether evidence is sufficient, leaving the generator LLM frozen. The key contribution is the GBR reward which calculates the improvement in generator accuracy when utilizing s3’s retrieved context versus naïve top- k retrieval from the question itself.

Inspired by DeepRetrieval’s query rewriting and s3’s GBR reward, we design two RL-trained retrievers for PubMed articles and similar patient profiles, adapting both to the triage setting where no ground-truth relevance annotations exist.

4 Approach

Data

We use MIMIC-IV-Ext, a dataset of 23,249 de-identified ED cases combining triage scores, demographics, vitals, and discharge notes, to predict ESI triage level (1–5) from a patient’s HPI, demographics, and initial vitals [2, 17, 18]. The data is partitioned into a test set ($n=2,200$, `test.csv`) replicating Gaber et al.’s evaluation slice, a validation set ($n=2,000$, `val.csv`), a development pool

(19,000 rows, `dev.csv`) for prompt engineering, and a 100-row scratch subset for rapid pipeline checks [1]. All splits are disjoint and reproducible via a fixed random seed.

Clinical Re-ranking

After initial FAISS retrieval returns the top- k PubMed articles, a clinically structured re-ranker re-scores them. The method has three steps. (i) A probability distribution over eight clinical categories is inferred from a patient case using keyword matching on the chief complaint, HPI and vitals string, supplemented by numeric vital-sign thresholds. (ii) The original query embedding $\mathbf{q}$ is shifted towards category-representative anchor embeddings weighted by

$$\mathbf{q}_{\text{cond}} = \frac{\mathbf{q} + \alpha \sum_i P(i) \mathbf{a}_i}{\|\mathbf{q} + \alpha \sum_i P(i) \mathbf{a}_i\|},$$

where $\mathbf{a}_i$ is the pre-computed, L2-normalised embedding of a representative clinical phrase for category i and α controls conditioning strength. (iii) A FAISS nearest-neighbour search is performed with $\mathbf{q}_{\text{cond}}$, returning a new top- k candidate set. The anchor embeddings (seven categories $\times$ 768 dimensions, Google text-embedding-005) are computed once and cached. We sweep $\alpha \in \{0.0, 0.25, 0.5, 0.75, 1.0\}$, where $\alpha = 0$ recovers the identity baseline.

Agentic workflow

Our initial goal was to reproduce RAG setup reported by Gaber et al. (i.e., multi-role agentic pipeline focused on query refinement) and then test whether its main ingredients improved ESI triage prediction on our corpus. Early experiments, however, suggested that iterative query refinement was not the most useful lever: changing how retrieval queries were rewritten or decomposed did not consistently improve prediction quality on our development split (see Appendix Table 5). We, therefore, pivoted to a structured search over the RAG and prediction pipeline, starting from a simple baseline that retrieves the top- k articles from the corpus using similarity search over concatenated patient features. From there, we varied individual components of the pipeline, including query construction, retrieval backend, number of retrieved articles, excerpt length, prompting strategy, passive versus active evidence access, and the number and roles of agents. The full experiment catalog is maintained in the repository’s experiment archive; in the paper, we summarize only the main configuration families. From the best-performing `dev_tune` configurations, we selected a small final set for test-time evaluation: one standalone LLM baselines using few-shot prompting, and four agentic pipelines, split between retrieval-augmented and lookup-tool-based designs (for flowcharts, see Appendix Figure 3). We selected these configurations using a multilevel metric stack rather than a single headline score, combining benchmark accuracy with clinical safety, class-imbalance diagnostics, and chance-corrected agreement. This lets the final comparison reflect not only raw prediction performance, but also whether a pipeline is clinically safer and robust to class imbalances.

RL-Tuned Retrieval

Both retrievers were implemented from scratch, yet the ideas were informed by DeepRetrieval and s3, using Gemini 2.0 Flash as the generator LLM. The *PubMedRetriever* adapts DeepRetrieval’s query-rewriting framework by training Qwen2.5-3B-Instruct to generate effective PubMed queries for ED triage. Since no gold-standard PubMed query annotations exist for triage, NDCG was replaced with a GBR reward, which is the reduction in QWK when Gemini 2.0 Flash predicts triage using RL-retrieved articles versus vanilla RAG-retrieved articles. For the *PatientRetriever*, the structure of s3 was closely followed in that a multi-turn model was implemented to retrieve similar patient cases to the query patient. The `dev.csv` was stratify split into a `retrieval.csv/train.csv` (50/50) to create a retrieval database not used in training with a FAISS index generated on it. Each patient record was represented as a structured text block containing key triage-relevant fields (see Patient Profile Formatting in D). Each turn incorporates a compact summary of prior queries and retrieved subject IDs alongside the query patient profile. Again the GBR reward was used in training this retriever. After retrieval, patient cases and PubMed articles are assembled such that the most relevant items appear at the start and end of the context window, mitigating the lost-in-the-middle effect. While training both retrievers, the top 10 results were retrieved. Due to compute constraints, GRPO was used (500 steps for PatientRetriever, 150 for PubMedRetriever) with an 8-bit AdamW optimizer, GRPO group size scaling from 2 to 8 with training step, KL coefficient 0.05, and learning rate 3e-6.

5 Experiments

5.1 Data

For reader’s convenience, we include the full split pipeline in Appendix Figure 2 and add Table 4, which maps each split to its canonical path and variable name in the repository.

Agentic workflow

We evaluate on the MIMIC-IV-Ext-Triage benchmark ($n=2,200$) from Gaber et al. [1], derived from MIMIC-IV-ED 2.2 [2], using their dataset construction pipeline and “clinical user” input setting (chief complaint, HPI, vitals, demographics). Unlike Gaber et al., we retain the full 2,200 cases—including 117 rows (5.3%) with missing vitals¹. Missing vitals are rendered as “Missing/not recorded” in the prompt rather than passed through as NaN. Gaber et al. augmented their prompts with PubMed article retrieval (30M article abstracts) but evaluated only that single evidence source and did not compare it against task-aligned alternatives. We build a separate retrieval corpus from the PMC Open Access commercial subset (2.31M full articles), drawn from the same upstream PubMed Central repository but a distinct corpus with potentially no overlap, and pair it with a case bank of 60 expert-classified ESI vignettes from the ESI v4 Implementation Handbook [19], enabling a direct comparison between broad biomedical literature and task-aligned clinical evidence. Fast iteration used a 150-case subset of `dev_tune` for rapid strategy screening; pre-test validation used the full `dev_tune` split ($n=1,000$).

Clinical Re-Ranking

Re-ranking experiments use a 150-case stratified subsample `dev_tune.csv` drawn from the development pool. The smaller size reflects the cost of LLM-based triage evaluation: each α sweep requires 150 Gemini 2.5 Flash Lite calls at a 5-RPM rate limit (~ 35 mins per sweep, five sweeps total). Cases are routed to one of seven clinical categories via keyword matching on the chief medical complaint, HPI prefix and vital signs, supplemented by numeric vital-sign thresholds. For cases with a mappable ICD discharge code, the ICD-to-category lookup provides a ground-truth label for routing accuracy. Across all α values the category distribution is stable, at higher α , a small number of cases fail at generation or parsing.

RL-Tuned Retrieval

Two main experiments were run: (1) quantifying performance of models checkpointed at different steps during the training process as a proxy for testing if the models truly had improved performance over time (2) quantifying final model performance in comparison to simple LLM prompting without retrieval. For the first experiment, an evaluation was done on the validation split ($n = 2,000$), which is held out from all training and prompt engineering decisions. For the second experiment, evaluation was done on both the validation split ($n = 2,000$, `val.csv`) and the held-out test split ($n = 2,200$, `test.csv`), the latter of which will enable comparison with Gaber et al. [1]. Moreover, for retrieval by the checkpointed and final trained models, the 9,500 cases in `retrieval.csv` were used.

5.2 Evaluation method

Clinical Re-ranking

Re-ranking evaluation proceeds in two tiers. **(i)** Routing accuracy (retrieval-only): fraction of classifiable cases (those with an ICD code mapping to one of the seven categories) where the inferred top category matches the ground-truth specialty. Measures how re-ranking changes the retrieved article set. **(ii)** Anchor recall: fraction of re-ranked top-10 articles that fall within the 200 nearest FAISS neighbours of the inferred category’s anchor embedding, serving as a retrieval-space relevance proxy. We also report the delta relative to the $\alpha=0$ baseline. **(iii)** Rank displacement: mean absolute rank change for articles appearing in both the original and re-ranked top-10. **(iv)** Overlap@ k : Jaccard

¹We hypothesized that missing vitals would be disproportionately concentrated in the highest-acuity cases, since ESI-1 patients often bypass full triage recording for immediate resuscitation, and that dropping them would introduce a class-inclusion bias against the rarest and most critical category. Post-hoc analysis confirmed the hypothesis, with 75% of missing-vitals rows being ESI-1.

overlap between the original and re-ranked top-10 sets (1.0 at $\alpha=0$ by construction). **(v)** Category precision@ k : fraction of re-ranked top-10 articles containing keyword matches for the inferred top category. The second tier is triage prediction, which passes re-ranked context to Gemini 2.5 Flash Lite (temperature 0) and evaluates the predicted ESI level against the ground-truth label. We report exact accuracy, ± 1 accuracy, MAE, quadratic-weighted Cohen’s κ ($\Delta AC2$) and linear-weighted Gwet’s AC2. We designate $\Delta AC2$ as the primary triage metric because ESI is ordinal and $\Delta AC2$ penalises large disagreements quadratically; AC2 serves as a secondary measure, less sensitive to marginal class distributions.

Agentic workflow

Gaber et al. [1] evaluate triage systems using only exact accuracy and range accuracy. While sufficient for benchmark ranking, these two metrics are blind to failure modes that matter clinically. Range accuracy forgives over-triage by one level, so a model that systematically escalates patients scores well despite wasting emergency department resources. Neither metric distinguishes the *direction* of error: under-triage—assigning a patient lower acuity than warranted—can delay time-sensitive evaluation and treatment [10], whereas over-triage is costly but generally safer. Moreover, ESI labels are heavily imbalanced (ESI-2 and ESI-3 account for roughly 88% of cases), so a model that collapses all predictions into these majority classes achieves attractive headline accuracy while remaining blind to the rarer but clinically critical ESI-1, ESI-4, and ESI-5 levels. To surface these failure modes, we adopt the multilevel metric stack summarized in Table 1.

Table 1: Multilevel metric stack. Each tier captures a failure mode invisible to the benchmark metrics of [1].

Tier	Metric(s)	Measures	Value add
Benchmark	Exact accuracy, range accuracy	Match / +1 of true ESI	Retained for comparability with [1]
Safety	Under-triage (%) , over-triage (%)	Directional error rates	Separates harmful under-triage from costly over-triage [10]
Imbalance	Macro recall, ESI- k recall, MA-MAE	Class-balanced accuracy & error magnitude	Detects collapse into dominant ESI-2/3 classes
Chance-adj. agreement	Gwet AC2 (linear); QWK (quadratic)	Chance-corrected ordinal agreement	AC2 stable under skew [6, 9]; QWK retained as collapse alarm [7]

RL-Tuned Retrieval

For both experiments, the trained models are evaluated on the datasets listed in the data section above. The top $k(=10)$ retrieved articles and patient profiles from both the trained PubmedRetriever and PatientRetriever respectively were assembled into a single prompt with the formatted query patient profile and instructions to predict triage. This prompt was then passed into the Gemini-2.0-Flash model(temperature=0) which generated predictions regarding the query patient’s ESI level on which evaluation metrics were applied to quantify model performance. These evaluation metrics used in both experiments include exact accuracy, ± 1 accuracy, MAE, and quadratic-weighted Cohen’s κ , and linear-weighted Gwet’s AC2 metric. Additionally, for the second experiment, we additionally report GBR — the per-metric improvement of RL RAG over the retrieval from a vanilla/baseline RAG.

5.3 Experimental details

Agentic workflow

Each architectural choice is first screened on a deterministic 150-row pilot from the 1,000-row dev_tune split (seed 42); surviving configurations are then run on the full 1,000 rows. We do not use

dev_holdout as an intermediate pruning stage, so the pilot is used to screen discrete architectural variants rather than to tune continuous hyperparameters. Table 6 (in Appendix) summarizes the search space and fixed inference controls; not every combination is exhaustively crossed, because only pilot winners are scaled to decisive runs. Each configuration was scored on the four-tier metric stack described above.

Clinical Re-ranking

For each value of α , the conditioned query q_{cond} is used for an inner-product search over the FAISS index, with an over-fetch of 200 candidates deduplicated to a top-10. Retrieved article text is batch-fetched and truncated. Triage predictions are generated by Gemini-2.5-Flash-Lite at temperature 0. The model is prompted in a system role as an emergency triage nurse and returns a JSON response {"ESI": 1-5, "reasoning": "..."}.

RL-Tuned Retrieval

For the first experiment: During GRPO training, model checkpoints are saved every 50 steps for both the PatientRetriever and PubMedRetriever (Qwen2.5-3B-Instruct backbone). To post-hoc test whether the training led to successively better models, a combinatorial evaluation across checkpoints [50, 100, 150, final] was done and reported. All runs use top- $k=5$ for both retrievers, with Gemini 2.0 Flash as the frozen answer generator at temperature 0. *For the second experiment:* Three conditions are compared based on their performance on val.csv and test.csv. (1) No retrieval: Gemini 2.0 Flash prompted with patient profile only (2) Baseline RAG: patient data concatenation as query with no RL (3) RL RAG: RAG with trained retrievers. All conditions use the same frozen Gemini 2.0 Flash generator at temperature 0. Retrieval uses top- $k=5$ for both patient cases and PubMed articles. Retrieved context/data is assembled with the token-budget-aware interleaving strategy described in the Approach section.

5.4 Results

Clinical Re-Ranking

Table 2 reports the full α -sweep results on dev_tune.csv.

Table 2: Re-ranking α -sweep on dev_tune ($n=150$).

α	Retrieval-only					Triage prediction				
	Routing Acc.	Anchor Recall	Rank Disp.	Overlap @ k	Cat. Prec. @ k	Exact Acc.	± 1 Acc.	MAE $\downarrow$	Δ AC2	AC2 (lin.)
0.00	0.375	0.086	0.000	1.000	0.621	0.527	0.973	0.507	0.249	0.816
0.25	0.375	0.127	1.764	0.724	0.593	0.514	0.980	0.514	0.288	0.811
0.50	0.375	0.188	2.417	0.523	0.558	0.536	0.971	0.500	0.275	0.816
0.75	0.375	0.230	2.598	0.375	0.542	0.556	0.963	0.489	0.282	0.821
1.00	0.375	0.278	2.475	0.276	0.522	0.540	0.978	0.489	0.261	0.822

Re-ranking mechanically replaces the retrieved set: overlap@ k drops from 1.0 to 0.276 at $\alpha = 1.0$, and anchor recall increases monotonically with α from 0.086 to 0.278 (+191%). On the triage side, κ_{qw} peaks at 0.288 ($\alpha = 0.25$, +0.039 over baseline) but is non-monotonic, falling to 0.261 at $\alpha = 1.0$. AC2 varies between 0.811 and 0.822. Best accuracy is 0.556 at $\alpha = 0.75$.

Agentic workflow

Table 7 (in Appendix), Figures 1 below, and 5 (in Appendix) organize the final agentic-workflow results around a clear accuracy–safety trade-off. For exact specification of each run/pipeline, see Table 9 in Appendix.

RL-Tuned Retrieval

For experiment 1: Table 8 (in Appendix) reports performance on the validation split for selected checkpoint combinations. For a graphical version with more datapoints see figure 6 in the Appendix.

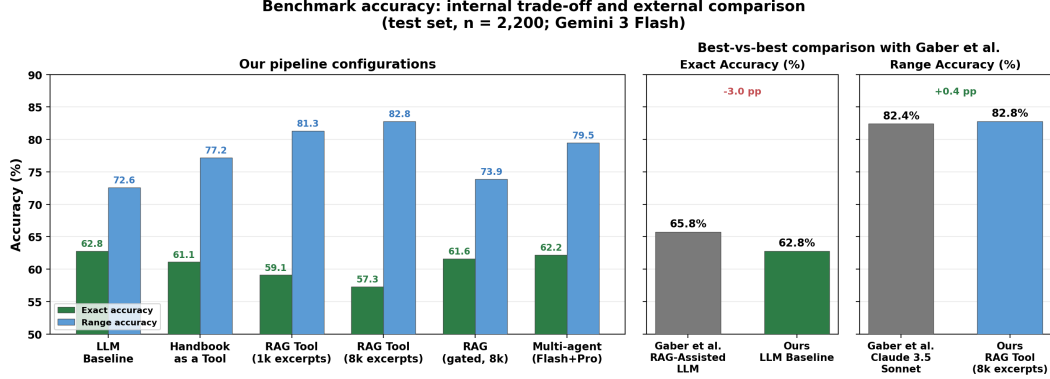


Figure 1: Benchmark accuracy across six pipelines (left) and comparison with Gaber et al. (right).

These results show that the models did train over time yet the improvement in predictions was marginal.

For experiment 2: Table 3 reports validation and test results.

Table 3: Validation ($n = 2,000$) and test ($n = 2,200$) results. GBR is computed relative to retrieval from a vanilla/baseline RAG.

Metric	Validation			Test		
	No Retrieval	Baseline RAG	RL RAG	No Retrieval	Baseline RAG	RL RAG
Exact Acc.	0.584	0.62	0.614	0.571	0.598	0.584
± 1 Acc.	0.984	0.983	0.981	0.983	0.983	0.980
MAE	0.435	0.399	0.406	0.447	0.419	0.436
κ	0.408	0.412	0.410	0.435	0.424	0.415
Linear AC2	0.835	0.849	0.846	0.830	0.841	0.834
GBR with AC2	—	—	-0.003	—	—	-0.006
<i>Gaber et al. (RAG best)</i>		—		<i>Exact Acc. 0.658</i>		

6 Analysis

Clinical Re-ranking

The two retrieval-quality proxies diverge: anchor recall (embedding-space proximity) increases monotonically with α while category precision@k (keyword co-occurrence) decreases monotonically, revealing a misalignment between the embedding-space and keyword-defined relevance. Routing accuracy is flat because α controls conditioning strength, not category inference.

Despite these retrieval changes, downstream triage metrics are largely insensitive to re-ranking. The best κ_{qw} gain is +0.039 at $\alpha = 0.25$, while stronger conditioning falls below baseline ($\kappa_{\text{qw}} = 0.261$ at $\alpha = 1.0$), suggesting that aggressive conditioning may introduce topic drift.

Agentic workflow

Compared with Gaber et al., we show that our best held-out exact accuracy remains below their strongest reported result, whereas our best held-out range accuracy is slightly higher. Specifically, LLM Baseline attains 62.8% exact accuracy on the test set, while RAG Tool (8k excerpts) attains 82.8% range accuracy, marginally exceeding Gaber et al.’s best reported range score. This advantage on range accuracy should not, however, be interpreted as a uniform performance gain. In our experiments, improvements in range accuracy are closely associated with more aggressive up-triage, which lowers under-triage but often reduces exact-match performance. Figure 8 makes this trade-off explicit: systems that improve safety generally do so by shifting predictions toward higher acuity. Within this frontier, Multi-agent (Flash+Pro) provides the strongest compromise, preserving near-baseline exact accuracy while reducing under-triage to 20.3%.

We further show that this pattern cannot be explained solely by the mismatch with our PMC-derived corpus. A natural hypothesis was that literature retrieval underperformed because the retrieved articles were semantically similar yet not sufficiently useful for ESI prediction; if that were the case, then a more targeted external source should help. Handbook as a Tool served as a direct test of this hypothesis, because it replaced broad biomedical retrieval with triage-specific guidance that the model could consult selectively. Although this condition slightly outperformed LLM Baseline on the development set, the advantage did not generalize to the held-out test set, where LLM Baseline reached 62.8% exact accuracy and Handbook as a Tool reached 61.1%. The same pattern extends to RAG Tool (1k excerpts), RAG Tool (8k excerpts), and RAG (gated, 8k): making external knowledge smaller, more selective, or more targeted improved safety primarily by shifting predictions upward, rather than by consistently improving core acuity discrimination. Taken together, these results show that the limitation is not simply that PMC retrieval is too broad; making external knowledge useful at inference time requires more than substituting a curated or task-specific source.

RL-Tuned Retrieval

From the results, RL RAG does not consistently improve over vanilla RAG or no-retrieval, likely because state-of-the-art LLMs like Gemini 2.0 Flash are already strong enough that marginal gains from better retrieval are small. Moreover, both retrievers frequently fall back to baseline FAISS retrieval due to invalid outputs where the PubMed retriever omits <answer> tags and the patient retriever hallucinates subject IDs, negating the RL benefit. Wider hyperparameter search would address this. Finally, supplying up to 10 retrieved results likely causes context overload, and reducing top-k with stricter re-ranking may improve signal-to-noise.

7 Conclusion

We investigated whether retrieval quality is the binding constraint on RAG-based ESI triage prediction, evaluating three strategies: clinical re-ranking, RL-tuned retrieval and agentic workflow design, on a MIMIC-IV-ED benchmark. All three converge on the same answer: it is not. Clinical re-ranking replaced up to 72% of retrieved articles and nearly tripled anchor recall, yet κ_{qw} improved by at most +0.039 ($\alpha = 0.25$) and degraded under stronger conditioning. RL-tuned retrieval yielded a GBR Δ AC2 of -0.003 on validation and -0.006 on test, with frequent fallback to baseline FAISS due to malformed model outputs. Among the agentic configurations, the LLM baseline without any retrieval achieved the highest κ_{qw} (0.500) and exact accuracy (62.8%), while retrieval-heavy pipelines traded exact accuracy for reduced under-triage; a trade-off no configuration resolved. The one partial exception was the multi-agent pipeline Flash nurse + Pro critic which achieved near baseline accuracy (62.2%) at substantially lower under-triage (20.3%, $-7.0pp$) by selectively correcting over-triage rather than uniformly shifting predictions.

Team contributions (Required for multi-person team)

Isha: I contributed to the ideation of the project, was in charge of the research and implementation of the RL-backed RAG pipeline, and helped with the generation of the FAISS indices for our datasets.

Charlotte: I contributed to the ideation of the project, was in charge of the research and implementation of the clinical re-ranker.

Nino: Everything else (e.g. agentic workflow, codebase)

References

- [1] F. Gaber, M. Shaik, F. Allega, et al. Evaluating large language model workflows in clinical decision support for triage and referral and diagnosis. *npj Digital Medicine*, 8:263, 2025.
- [2] A. Johnson, L. Bulgarelli, T. Pollard, S. Horng, L.A. Celi, and R. Mark. MIMIC-IV-ED (version 2.2), 2024.
- [3] Pengcheng Jiang, Junyang Lin, Lang Cao, Rui Tian, Shizhe Kang, Zifeng Wang, Jimeng Sun, and Jiawei Han. DeepRetrieval: Hacking real search engines and retrievers with large language models via reinforcement learning. *arXiv preprint arXiv:2503.00223*, 2025.

- [4] Pengcheng Jiang, Xuan Xu, Junyang Lin, Jing Xiao, Zifeng Wang, Jimeng Sun, and Jiawei Han. s3: You don’t need that much data to train a search agent via RL. *arXiv preprint arXiv:2505.14146*, 2025.
- [5] Ahmad Mirhaghi, Abbas Heydari, S. Razieh Mazlom, and Fariba Hasanzadeh. Reliability of the emergency severity index: Meta-analysis. *Sultan Qaboos University Medical Journal*, 15(1):e71–e77, 2015.
- [6] A. R. Feinstein and D. V. Cicchetti. High agreement but low kappa: I. the problems of two paradoxes. *Journal of Clinical Epidemiology*, 43(6):543–549, 1990.
- [7] Matthijs J. Warrens. Some paradoxical results for the quadratically weighted kappa. *Psychometrika*, 77(2):315–323, 2012.
- [8] Kilem L. Gwet. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48, 2008.
- [9] David Quarfoot and Richard A. Levine. How robust are multirater interrater reliability indices to changes in frequency distribution? *The American Statistician*, 70(4):373–384, 2016.
- [10] D. R. Sax, E. M. Warton, D. G. Mark, and M. E. Reed. Emergency department triage accuracy and delays in care for high-risk conditions. *JAMA Network Open*, 8(5):e258498, 2025.
- [11] J. H. Kim, S. K. Kim, J. Choi, and Y. Lee. Reliability of chatgpt for performing triage task in the emergency department using the korean triage and acuity scale. *Digital Health*, 10, 2024.
- [12] C. Colakca, M. Ergin, H. S. Ozensoy, A. Sener, S. Guru, and A. Ozhasenekler. Emergency department triaging using chatgpt based on emergency severity index principles: a cross-sectional study. *Scientific Reports*, 14:22106, 2024.
- [13] B. Arslan, C. Nuhoglu, M. O. Satıcı, and E. Altınbilek. Evaluating llm-based generative ai tools in emergency triage: A comparative study of chatgpt plus, copilot pro, and triage nurses. *American Journal of Emergency Medicine*, 2025.
- [14] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33, 2020.
- [15] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.
- [16] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36, 2023.
- [17] A. Johnson, L. Bulgarelli, T. Pollard, S. Horng, L.A. Celi, and R. Mark. MIMIC-IV (version 3.0), 2024.
- [18] A. Johnson, T. Pollard, T. Fulmer, M. Newberry, I. Ersoy, L. Shi, O. Badawi, and R. Mark. MIMIC-IV-Note: Deidentified free-text clinical notes (version 2.2), 2023.
- [19] Nicki Gilboy, Paula Tanabe, Debbie Travers, and Alexander M. Rosenau. *Emergency Severity Index (ESI): A Triage Tool for Emergency Department Care, Version 4*. Agency for Healthcare Research and Quality, Rockville, MD, 2012. AHRQ Publication No. 12-0014.
- [20] National Library of Medicine. PubMed Central Open Access Subset (commercial use collection), 2025. Accessed via Google BigQuery: `bigquery-public-data.pmc_open_access_commercial.articles`; ~2.3M articles.

Appendix

A Data

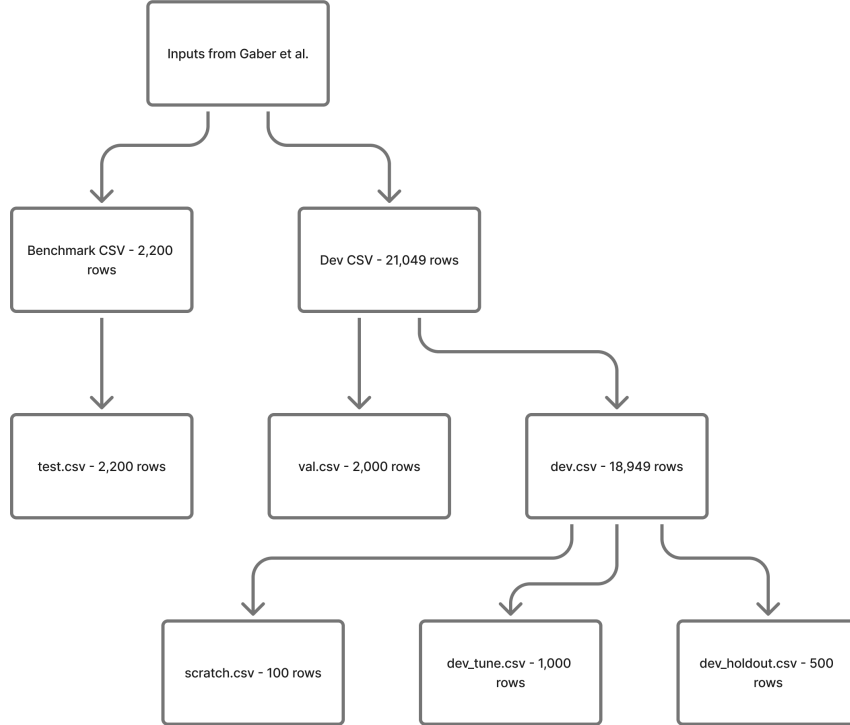


Figure 2: Data splits pipeline from upstream Gaber et al. benchmark CSVs to canonical project splits.

Table 4: Data splits and their canonical variable names.

Split	Path	Rows	Variable
Benchmark	MIMIC-IV-Ext-Triage-....csv	2,200	_SRC_TEST
Dev pool	MIMIC-IV-Ext-Dev.csv	21,049	_SRC_DEV
test	splits/test.csv	2,200	TEST_CSV
val	splits/val.csv	2,000	VAL_CSV
dev	splits/dev.csv	18,949	DEV_CSV
dev_tune	splits/dev_tune.csv	1,000	DEV_TUNE_CSV
dev_holdout	splits/dev_holdout.csv	500	DEV_HOLDOUT_CSV
scratch	splits/scratch.csv	100	SCRATCH_CSV

B Pipeline Diagrams

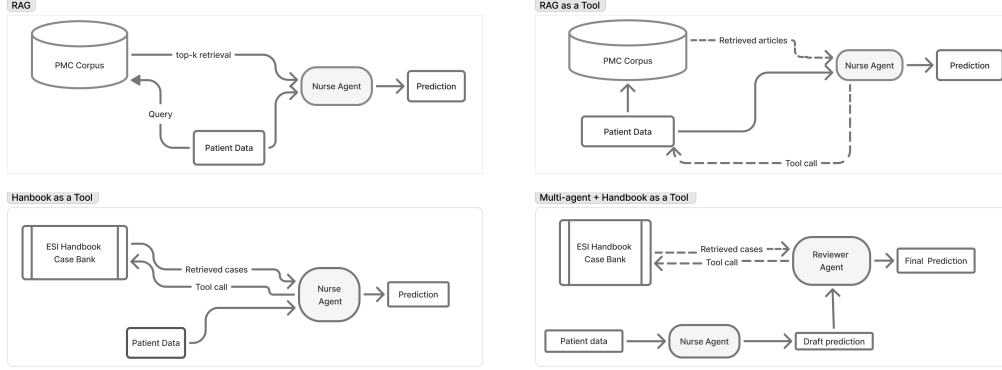


Figure 3: Framework variants evaluated for ESI triage prediction.

C Query Refinement Evidence

Table 5: Curated query-refinement experiments on the development split. Metrics are taken from the experiment archive entries for E02a/E02b, E09, and E12.

Experiment	n	Exact Acc.	QWK	Result
E02a concat	150	56.0%	0.232	Baseline RAG query; below LLM-only ($\kappa = 0.338$)
E02a hpi_only	150	50.7%	0.184	Worse than concat
E02b rewrite	150	54.7%	0.236	Flat vs concat on QWK, lower exact accuracy
E09 multifacet	150	54.0%	0.218	Worse than single-query retrieval
E12 adaptive	150	58.0%	0.292	Worse than ungated RAG and distance gating

D Prompt Templates

This appendix reproduces the core system prompt for each strategy listed in Table 6. All templates share three patient-level placeholders (`{hpi}`, `{patient_info}`, `{initial_vitals}`) and the same five-level ESI definition block. We show only the structural differences below; the shared ESI definition (“1: Requires immediate lifesaving intervention ... 5: No diagnostic/therapeutic interventions needed beyond physical exam”) is abbreviated as `[ESI-DEFS]`.

Zero-shot

The baseline prompt provides the triage nurse persona, ESI definitions, and patient data with no examples or tool access.

```
You are a nurse with emergency and triage experience. Using the
patient's history of present illness, their information and initial
vitals, determine the triage level based on the Emergency Severity
Index (ESI), ranging from ESI level 1 (highest acuity) to ESI level
5 (lowest acuity): [ESI-DEFS] History of present illness: {hpi}
Patient info: {patient_info}
Initial vitals: {initial_vitals}
```

Few-shot

Prepends five gold cases from the ESI Implementation Handbook [19], one per ESI level, each with a clinical vignette, correct label, and rationale. We show one example; the remaining four follow the same format.

...[ESI-DEFS] Below are five reference triage cases from the ESI Handbook with their correct ESI levels and clinical rationales. Use them to calibrate your judgement. -- Example 1 --
Case: A 76-year-old male is brought to the ED because of severe abdominal pain. He tells you “it feels like someone is ripping me apart.” ... His skin is cool and diaphoretic. VS: HR 122, BP 88/68, RR 24, SpO2 94%.
ESI: 1
Rationale: Signs of shock -- hypotensive, tachycardic, decreased peripheral perfusion. Needs immediate IV access, aggressive fluid resuscitation. *[Examples 2–5 omitted for space; one case per ESI level.]* Now classify this patient:
History of present illness: {hpi} ...

Decision-tree

Replaces the flat ESI definitions with a structured step-by-step algorithm (Steps A–D) derived from the ESI v4 handbook. Combined with tool-use access to the case bank.

...determine the triage level by following this algorithm: Step A -- Does this patient require an immediate lifesaving intervention? If YES → ESI-1.
Step B -- Is this a high-risk situation? Is the patient confused, lethargic, disoriented, or in severe distress? If YES → ESI-2.
Step C -- How many resources will this patient need? 0 → ESI-5; 1 → ESI-4; ≥2 → go to Step D.
Step D -- Vital-sign danger zone check. Any in danger zone → consider ESI-2. All stable → ESI-3.

Tool-use (case bank)

The model is given a function-calling tool `search_esi_case_bank` and decides autonomously when to query it.

You are a nurse with emergency and triage experience. You have access to a tool called `search_esi_case_bank` that lets you search a bank of 60 expert-classified ESI triage cases from the ESI Implementation Handbook. Each case includes a clinical vignette, the correct ESI level, and the clinical rationale. Before making your triage decision, consider whether consulting similar reference cases would help you classify this patient more accurately. You may call the tool zero or more times with different search filters. Once you have enough information, determine the triage level ...
[ESI-DEFS] History of present illness: {hpi} ...

Two-role (nurse + critic)

A two-stage pipeline. The nurse agent sees only patient data and produces a provisional ESI with a confidence tag. The critic agent then receives the patient data, the nurse draft, and case-bank tool access.

Nurse prompt:

You are the triage nurse in a two-role emergency triage workflow. Review the patient presentation only and make a provisional ESI decision before any tool use or evidence review. ... Output ONLY a JSON object with keys: “acuity” (int 1-5), “confidence” (high/medium/low), “rationale” (string).

Critic prompt:

You are the critic in a two-role emergency triage workflow. Review the patient’s presentation and the nurse draft. You have access to

the ``search_esi_case_bank`` tool with expert-classified ESI examples. Use it if it helps you challenge or refine the nurse's decision, then return the final ESI level. ... Nurse draft: `{nurse_draft}`
Output ONLY a JSON object with exactly one key: "acuity" (int 1-5).

Reinforcement Learning: PubMed Retriever Prompt:

You are an expert medical librarian specialising in PubMed search. Given a patient's clinical presentation, generate an effective PubMed search query to retrieve academic articles most relevant for emergency triage level predictions.

Focus on:

- Relevant triage or acuity assessment criteria
- Key differentials that affect triage level
- Relevant triage or acuity assessment criteria
- Use clinical MeSH-style terminology where appropriate

Show your reasoning inside `<think>` `</think>` tags. Your final PubMed query must be inside `<answer>` `</answer>` tags. Output nothing outside these tags.

Reinforcement Learning: Patient Retriever Prompt:

You are a medical search assistant for an emergency department triage system.

You will iteratively search for similar past patients whose outcomes help predict the current patient's Emergency Severity Index (ESI) level.

Each turn you must output exactly three XML tags:

- `<query>`your search query (max 100 words)`</query>`
- `<important_docs>`[list of subject IDs from current results that are useful, e.g. [32796544, 45432895]]`</important_docs>`
- `<search_complete>`True or False`</search_complete>`

Focus on triage-predictive features:

- Chief complaint and primary symptoms
- Vital sign abnormalities (high/low HR, BP, SpO2, RR, temp)
- Key risk factors (age, comorbidities)
- Acuity indicators (severity, onset, progression, pain score)

Rules:

- Select at most 3 documents per turn in `<important_docs>`.
- Set `<search_complete>`True`</search_complete>` when you have enough cases to confidently predict ESI level.
- Set `<search_complete>`False`</search_complete>` to issue another query.
- Output ONLY the three tags - no preamble, no explanation.

Reinforcement Learning: Generator Prompt:

You are an experienced emergency department triage nurse. Your task is to assign an Emergency Severity Index (ESI) triage level to a patient.

ESI levels:

- 1 - Immediate: life-threatening, requires immediate intervention
- 2 - Emergent: high risk, severe pain/distress, should be seen very quickly
- 3 - Urgent: vitals are not in danger zone but needs multiple resources (labs, imaging, IV fluids, etc.) to stabilize patient

- 4 - Less Urgent: needs one resource to stabilize patient
- 5 - Non-Urgent: stable, no resources needed

Respond with ONLY a single integer: 1, 2, 3, 4, or 5. Do not explain your reasoning. Do not add any other text.

Reinforcement Learning: Patient Profile Formatting:

Patient Presentation:

- Subject ID: `subject_id`
- Chief Complaint: `chiefcomplaint`
- Vitals: `initial_vitals`
- Patient Info: `patient_info`
- Pain Score: `pain`
- Arrival Transport: `arrival_trans`
- History of Present Illness: `HPI`
- Past Medical History: `past_medical`
- Past Medications: `past_medication`

E Additional Results Tables

Table 6: Architectural search space and fixed inference controls in the staged pilot-to-decisive workflow.

Component	Variants / settings
LLM backbone	Gemini 2.5 Flash, Gemini 3 Flash, Gemini 3.1 Pro
Reasoning / orchestration ^a	Zero-shot, few-shot (5 gold ESI v4 cases), decision-tree, tool-use (case bank), two-role (nurse + critic)
Retrieval backend	None, FAISS ^b IVF4096 (dense), BM25S ^c (sparse), hybrid ^d
Retrieval activation	Passive (always-on), distance-gated ($d < 0.25$), tool-based (on-demand PMC search)
Query formulation ^e	Full triage fields, HPI-only, model-rewritten
Context budget	Top- $k \in \{3, 5, 10\}$; characters per article $\in \{1000, 8000\}$
Tool budget	Up to 3 tool turns per case; up to 10 results per call
Corpus / index	Approximately 2.3M PMC open-access articles [20]; Vertex AI ^f <code>text-embedding-005</code> ; local FAISS + SQLite
Fixed decoding controls	Temperature 0.0; structured JSON output

^a Full prompt templates for each strategy are provided in Appendix D.

^b FAISS (Facebook AI Similarity Search) is a library for efficient dense vector nearest-neighbor retrieval; IVF4096 denotes an inverted-file index with 4,096 Voronoi cells.

^c BM25S is a sparse lexical retrieval model that ranks documents by term-frequency and inverse document-frequency overlap with the query.

^d Reciprocal rank fusion (RRF); fusion constant $k = 60$, equal sparse/dense weights.

^e Query formulation strategies with examples are given in Appendix C. HPI = history of present illness.

^f Vertex AI is Google Cloud’s managed ML platform; we use its `text-embedding-005` model to produce 768-dimensional dense vectors.

Table 7: Test-set results ($n=2,200$) across all metric tiers. Bold = best per metric; $\downarrow$ = lower is better. All pipelines use Gemini 3 Flash except Multi-agent (Gemini 3.1 Pro critic). ESI-5 recall omitted (no test cases).

Tier	Metric	LLM Baseline	Handbook as a Tool	RAG Tool (1k chars)	RAG Tool (8k chars)	RAG (gated, 8k chars)	Multi-agent (Flash+Pro)
<i>Benchmark</i>	Exact accuracy (%)	62.8	61.1	59.1	57.3	61.6	62.2
	Range accuracy (%)	72.6	77.2	81.3	82.8	73.9	79.5
<i>Safety</i>	Under-triage (%) $\downarrow$	27.3	22.7	18.5	16.8	25.9	20.3
	Over-triage (%) $\downarrow$	10.0	16.1	22.4	25.9	12.4	17.5
<i>Imbalance</i>	Macro recall	0.394	0.374	0.365	0.354	0.386	0.384
	ESI-1 recall	0.186	0.140	0.159	0.159	0.182	0.165
	ESI-2 recall	0.607	0.734	0.807	0.844	0.638	0.770
	ESI-3 recall	0.781	0.624	0.494	0.414	0.724	0.601
	ESI-4 recall	0.000	0.000	0.000	0.000	0.000	0.000
	MA-MAE $\downarrow$	0.667	0.683	0.724	0.796	0.674	0.670
<i>Chance-adj. agreement</i>	Gwet AC2 (linear)	0.771	0.764	0.757	0.748	0.765	0.772
	QWK (quadratic)	0.500	0.461	0.414	0.370	0.479	0.469

Table 8: Checkpointed RL Model Performance on validation set ($n = 2,000$). Patient and PubMed columns indicate training step of each retriever. Best performance for each metric in **bold**.

Patient Checkpoint	PubMed Checkpoint	Exact Acc.	± 1 Acc.	MAE	κ	Linear AC2
step-50	step-50	0.603	0.984	0.415	0.391	0.843
step-100	step-100	0.609	0.986	0.407	0.404	0.846
step-150	step-150	0.618	0.984	0.399	0.423	0.849

F Additional Figures

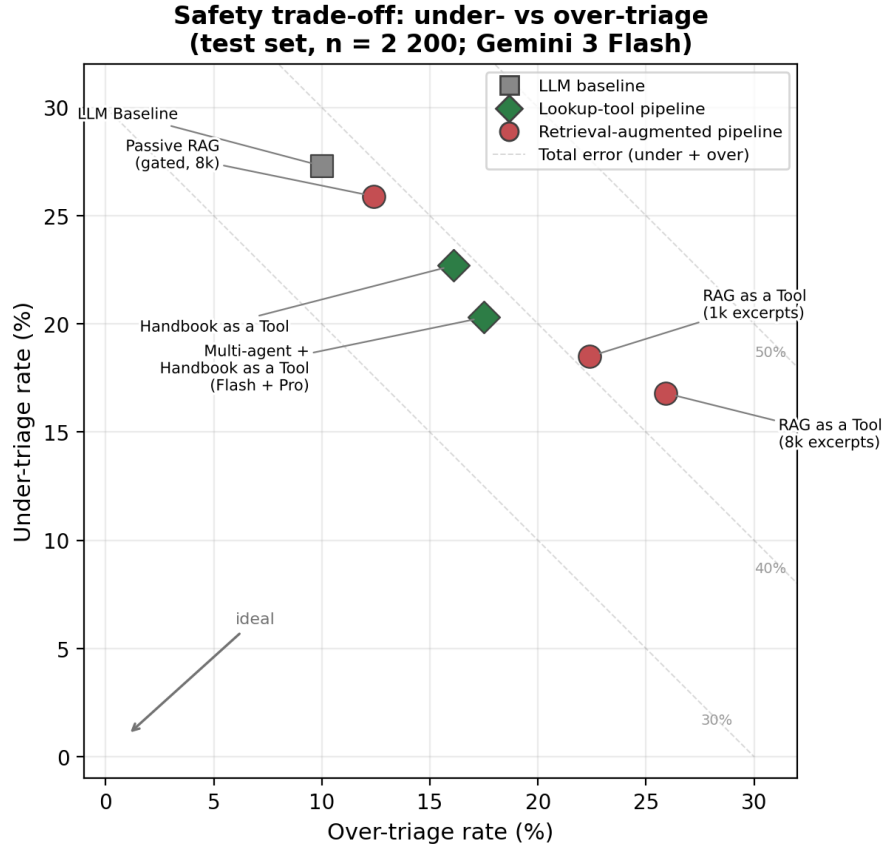


Figure 4: Under- vs over-triage trade-off. Dashed lines are iso-total-error contours. Retrieval shifts errors toward over-triage; the multi-agent pipeline achieves the most balanced profile.

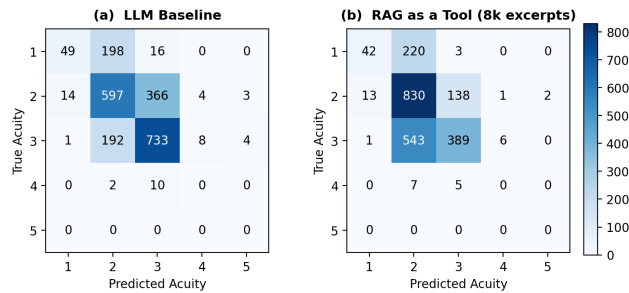


Figure 5: Confusion matrices for the two safety extremes from Fig. 4: LLM Baseline (a) vs RAG Tool 8k (b).

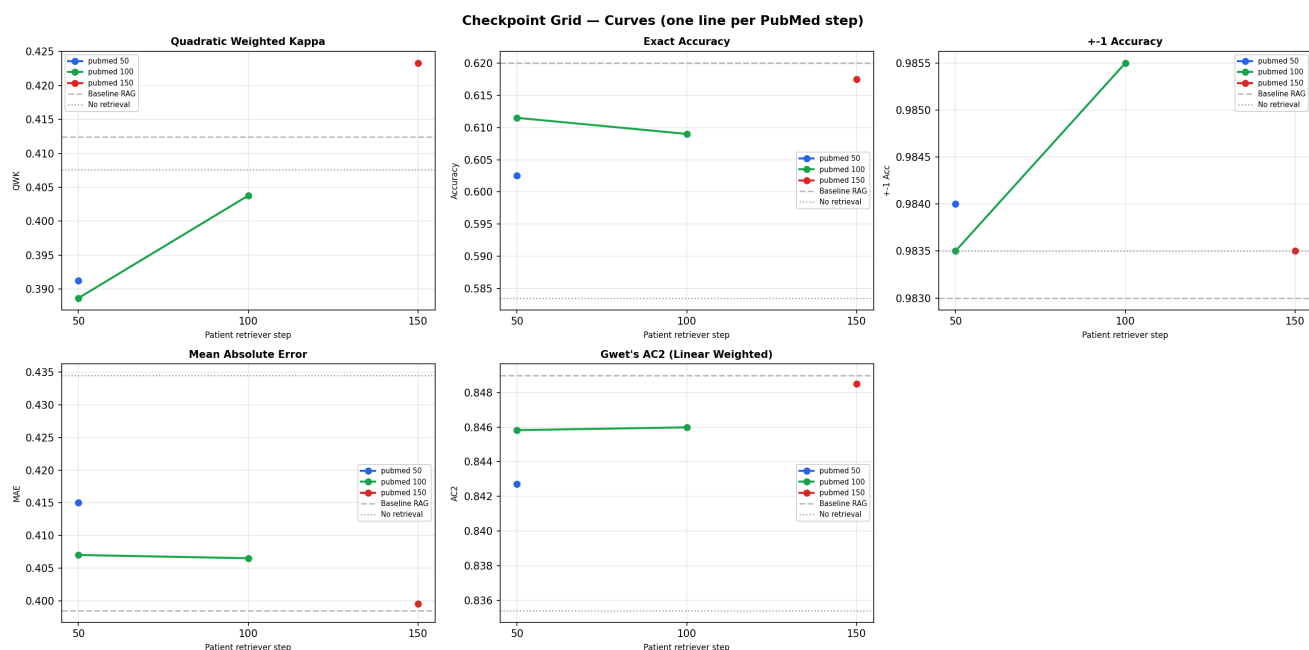


Figure 6: Scatter plots/Curves describing performance on val.csv of trained (checkpointed) models at increasing training steps of 50, 100, and 150.

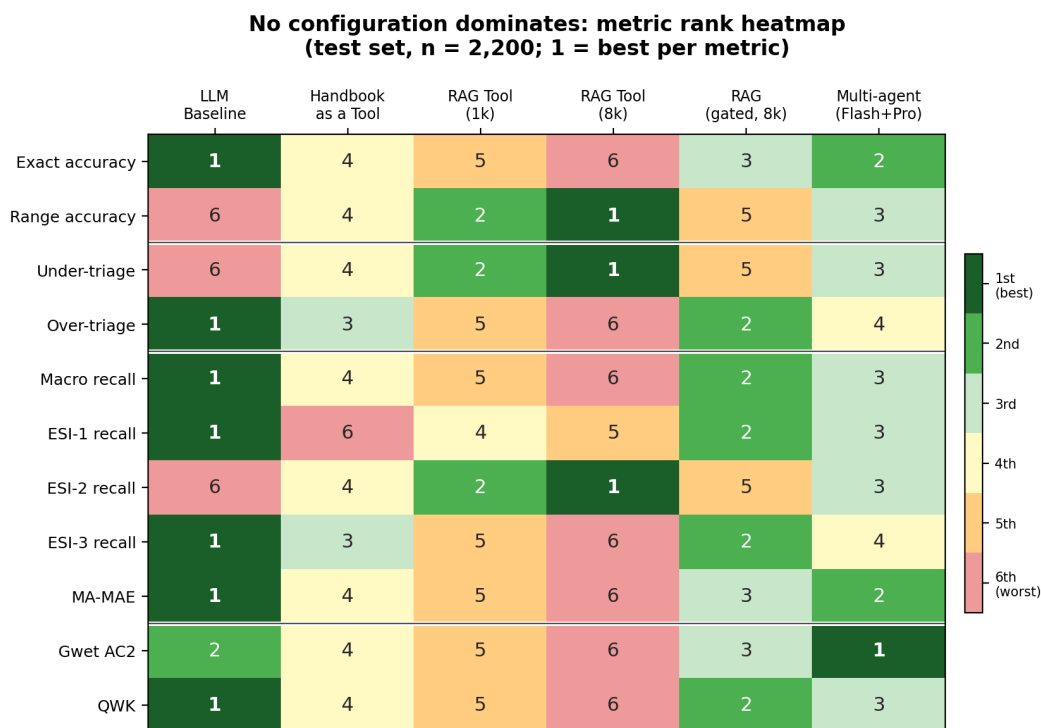


Figure 7: Metric rank heatmap across six configurations on the test set ($n = 2,200$). Rank 1 (dark green) is best per metric. The LLM Baseline ranks first on 7 of 11 metrics but last on under-triage and range accuracy. No configuration dominates across all evaluation tiers.

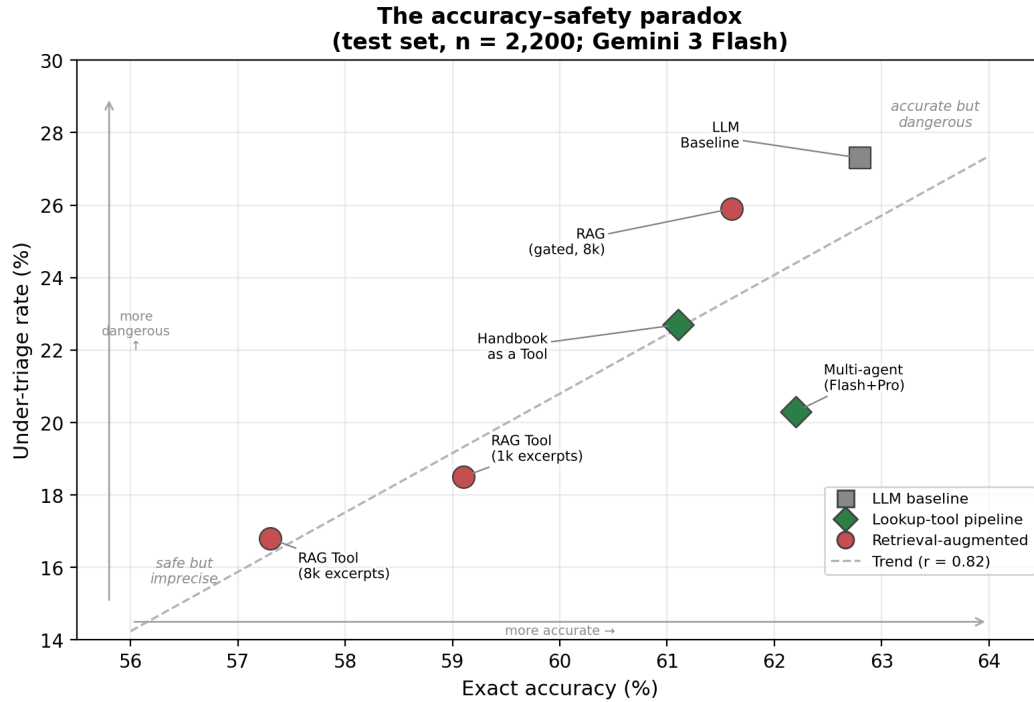


Figure 8: The accuracy–safety paradox: exact accuracy vs. under-triage rate across six configurations ($n = 2,200$, Gemini 3 Flash). Configurations that improve safety do so by shifting predictions toward higher acuity, proportionally destroying accuracy. The multi-agent configuration (Flash+Pro) is the sole positive outlier — it achieves near-baseline accuracy (62.2%) at substantially lower under-triage (20.3%) by selectively correcting the nurse’s over-triage rather than uniformly shifting all predictions.

G Detailed Specifications

Table 9: Exhaustive specification of pipeline configurations evaluated on the test set ($n = 2,200$; Gemini 3 Flash unless noted).

	LLM Baseline	Handbook as a Tool	RAG Tool (1k)	RAG Tool (8k)	RAG (gated, 8k)	Multi-agent (Flash+Pro)
<i>Model & orchestration</i>						
Model	G3 Flash	G3 Flash	G3 Flash	G3 Flash	G3 Flash	Nurse: G3 Flash / Critic: G3.1 Pro
Temperature	0.0	0.0	0.0	0.0	0.0	0.0
Pipeline mode	11m	11m (tool_use)	11m (tool_use)	11m (tool_use)	rag (passive)	11m (two_role)
Orchestration	Single-agent	Single-agent	Single-agent	Single-agent	Single-agent	Two-role: nurse → critic
Prompt template	fewshot	tool_use	tool_use_pmc	tool_use_pmc	fewshot_rag	Nurse / Critic prompts
Few-shot examples	5 (ESI 1–5)	0	0	0	5 (ESI 1–5)	0
<i>Retrieval specification</i>						
Retrieval type	None	On-demand tool	On-demand tool	On-demand tool	Passive injection	On-demand tool (critic only)
Backend	—	Case bank (in-memory)	FAISS	FAISS	FAISS	Case bank (in-memory)
Corpus	—	60 ESI v4 cases	~2.3M PMC articles	~2.3M PMC articles	~2.3M PMC articles	60 ESI v4 cases
Embedding model	—	—	text-embedding-005	text-embedding-005	text-embedding-005	—
Tool name	—	search_esi_case_bank	search_pmc_articles	search_pmc_articles	— (auto-injected)	search_esi_case_bank
Top- k	—	$\leq 10/\text{call}$	5/call	5/call	5 (fixed)	$\leq 10/\text{call}$
Excerpt chars/article	—	~150	1,000	8,000	8,000	~150
Snippet cleaning	—	—	Yes	Yes	Yes	—
Distance gate	—	—	—	—	0.25	—
Gate fallback	—	—	—	—	LLM-only (fewshot)	—
<i>Token usage & cost</i>						
n evaluated	2,197	2,200	2,200	2,200	2,198	2,178
Parse failures	3	0	0	0	2	22 (timeout)
Prompt tokens	2.46M	9.03M	9.74M	32.75M	6.10M	7.78M
Completion tokens	24.2K	169.2K	104.8K	102.3K	24.2K	255.3K
Cost (USD)	\$7.96	\$18.41	\$22.32	\$33.27	\$10.45	N/A
<i>Benchmark metrics</i>						
Exact accuracy (%)	62.8	61.1	59.1	57.3	61.6	62.2
Range accuracy (%)	72.6	77.2	81.3	82.8	73.9	79.5
<i>Safety metrics</i>						
Under-triage (%)	27.3	22.7	18.5	16.8	25.9	20.3
Over-triage (%)	10.0	16.1	22.4	25.9	12.4	17.5
<i>Diagnostics</i>						
κ (QWK)	0.500	0.461	0.414	0.370	0.480	0.469
AC2 (Gwet)	0.771	0.764	0.757	0.748	0.765	0.773
Macro recall	0.394	0.374	0.365	0.354	0.386	0.384
MA-MAE	0.667	0.683	0.724	0.796	0.674	0.670
<i>Per-class recall</i>						
ESI-1	0.186	0.140	0.159	0.159	0.182	0.165
ESI-2	0.607	0.734	0.807	0.844	0.638	0.770
ESI-3	0.781	0.624	0.494	0.414	0.724	0.601
ESI-4	0.000	0.000	0.000	0.000	0.000	0.000