RISKWEAVE: ALLOCATING EVIDENCE BY OMISSION COST IN MEDICATION-SAFETY RAG

Anonymous authors
Paper under double-blind review

ABSTRACT

Medication-safety RAG is governed by its most consequential omission: under a ten-item evidence budget, nine relevant passages cannot compensate for a missing contraindication, severe drug–drug interaction, or organ-specific dose rule. RiskWeave turns context construction into clinical-priority support allocation. It infers required medication–patient facets, scales their aggregate omission cost by pre-retrieval query risk, and greedily maximizes a monotone-submodular objective over provenance-bearing assertions; conflict reservation, facet-targeted re-exposure, and residual-support thresholds carry the same support state into citation and abstention. A 200-question frozen-pool control holds candidate access, passage retrieval and packing, prompt, generator, and budgets constant; a matched ablation removes priority weights while retaining the other policy stages. On 400 annotated questions, Facet Coverage@10 is 0.731, versus 0.689 for a flat cross-encoder (paired difference 0.042, 95% CI 0.018–0.067) and 0.612 for uniform submodular coverage. Across 847 questions, blinded grading gives RiskWeave 87.8% acceptability, versus 84.1% and 81.7%, respectively. In the frozen pool, unsupported high-priority claims are 8/200 versus 18/200 for uniform coverage; removing priority weights lowers high-priority coverage from 0.782 to 0.641. With graph-free exposure, RiskWeave reaches 0.715 coverage versus 0.674 for the cross-encoder; with Llama-3.1-70B-Instruct, acceptability is 83.8% versus 80.0%. On 423 eICU questions from 208 hospitals, it reaches 85.3% acceptability versus 78.7% for the strongest tested comparator. At the clinical interface, complete RiskWeave assistance lowers clinician major errors from 88/992 to 23/992 and median task time from 10.4 to 5.8 minutes.

1 INTRODUCTION

Medication recommendations are conjunctive evidence decisions: a context can be topically relevant overall yet unsafe when one required contraindication, severe interaction, or organ-specific dose rule is absent. Relevance rewards useful passages and diversification rewards different topics; neither distinguishes routine background from the missing fact that controls the recommendation under scarce, position-sensitive context (Lewis et al., 2020; Gao et al., 2023; Liu et al., 2024).

RiskWeave represents that decision as asymmetric omission loss. Medication and patient context activate facets such as severe DDI, contraindication, black-box warning, and organ-specific dosing; clinical-priority bases assign relative costs, and query risk scales coverage against relevance and provenance. Greedy selection operates on evidence items retaining assertion, facet, source, timestamp, and conflict identity. The same support state reserves material conflicts, targets missing facets, and triggers abstention.

The objective's first term is classical weighted maximum coverage, so pure greedy selection inherits the $(1 - 1/e)$ guarantee. The technical construction defines medication–patient facets and carries consequence-weighted missing support through provenance, conflict handling, and residual-support decisions. That continuity yields testable differences from intent diversification, uniform coverage, pointwise reranking, and graph-based exposure.

In a 200-question frozen-pool control, candidate access, passage retrieval and packing, and generation settings are common across policies. On separate 400-question annotated and 847-question blinded evaluations, RiskWeave reaches 0.731 Facet Coverage@10 and 87.8% acceptability; Flat-CE reaches

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

0.689 and 84.1%, and SubmodCov 0.612 and 81.7%. The paired RiskWeave–Flat-CE differences are 0.042 coverage (95% CI 0.018–0.067) and 3.8 acceptability points (95% CI 1.4–6.1). Removing priority weights lowers high-priority coverage from 0.782 to 0.641, and the largest coverage advantage appears at four slots. Under graph-free exposure, coverage is 0.715 versus 0.674 for Flat-CE; with Llama-3.1-70B-Instruct, acceptability is 83.8% versus 80.0%.

Frozen-pool comparisons locate support changes before generation, blinded grading measures answer quality, and a randomized crossover evaluates the complete interface. There, major errors fall from 88/992 to 23/992 and median task time from 10.4 to 5.8 minutes. When evidence is scarce, allocating slots by omission consequence protects the facts that control the answer.

2 Related Work

MMR trades relevance against novelty (Carbonell & Goldstein, 1998); IA-Select minimizes dissatisfaction under category probabilities (Agrawal et al., 2009); and xQuAD rewards uncovered query aspects (Santos et al., 2010). With fixed binary facets, Eq. 2's first term is weighted maximum coverage and inherits its diminishing returns (Nemhauser et al., 1978). RiskWeave applies that form to a clinical support contract: medication–patient requirements carry omission priorities, provenance, and conflict state into expansion and abstention.

RankRAG trains ranking jointly with generation (Yu et al., 2024); SetR identifies information requirements before selecting a sufficient passage set (Lee et al., 2025); and OptiSet and SetCE learn set-level complementarity or compatibility (Jiang et al., 2026; Song & Lee, 2026). These methods learn joint utility; RiskWeave exposes consequence-weighted missing support. Flat-CE tests pointwise relevance, SubmodCov tests uniform set coverage, and full-minus-weight tests clinical priority within the complete policy.

RAGAS and RAGChecker diagnose relevance, faithfulness, and claim-level failures (Es et al., 2024; Ru et al., 2024); Sufficient Context separates absent evidence from failure to use it and supports selective generation (Joren et al., 2025). MedRAG, MedRGB, and ASTRID evaluate medical retrieval, insufficient context, refusal, and answer quality (Xiong et al., 2024; Ngo et al., 2024; He et al., 2025); CrossDDI verifies drug interactions against cross-source evidence (Wang et al., 2026). Facet Coverage instead measures clinical-priority required support before generation.

GraphRAG and biomedical knowledge graphs organize entities, relations, summaries, and source traces (Edge et al., 2024; Bodenreider, 2004; Himmelstein et al., 2017; Nicholson & Greene, 2020). RiskWeave uses its graph for exposure and applies the same allocation over graph-free BM25+BGE-M3 fusion. The crossed and frozen-pool controls separate available evidence from allocation; Appendix Table 4 summarizes the distinctions.

3 Problem Setting

In our formulation, retrieval supplies a generation interface for medication-safety question answering. The system receives a natural-language query q together with structured patient context c and must produce an answer y whose non-trivial claims are each supported by inline citations.

Clinically meaningful facets of medication-safety questions include contraindications, DDIs, and organ-specific dosing, as well as special populations, monitoring, and black-box warnings. An otherwise correct answer can still contain a major error if a single high-priority facet is omitted; retrieval must therefore account for asymmetric omission cost in addition to topical relevance. We designate severe DDIs, contraindications, black-box warnings, and organ-specific dosing requirements as high-priority facets; a facet is critical for a query when it is triggered by the medication–patient context.

Evidence selection returns a set $S \subseteq \mathcal{D}$ subject to $|S| \leq B$. Throughout the experiments, the limit of $B = 10$ evidence items corresponds to a common context budget of 4,096 tokens and a P95 latency target of < 5s. Let a_f be the prespecified clinical-priority cost for facet f, and let m_r be a query-level multiplier for the pre-retrieval risk tier r. We use $\alpha_f(r) = a_f m_r$ and define omission loss as

$$L_{\text{omit}}(S; q, c) = \sum_{f \in \mathcal{F}(q,c)} \alpha_f(r)\, \mathbb{I}[\nexists s \in S : f \in C(s)]\,. \tag{1}$$

Because $\sum_f \alpha_f(r)$ is fixed for a query, minimizing Eq. 1 is equivalent to maximizing its covered weight. The common m_r preserves the relative clinical ordering given by a_f but changes how strongly aggregate coverage competes with relevance and provenance for low-, medium-, and high-risk queries. We combine the reduction in omission loss with those modular utilities:

$$U(S; q, c) = \lambda_{\text{rel}}\text{Rel}(S; q) + \lambda_{\text{cov}}\text{Cov}(S; \mathcal{F}, r) + \lambda_{\text{prov}}\text{Prov}(S),$$

where Rel captures topical relevance, Cov captures omission-cost coverage, and Prov rewards provenance quality. For the common evidence budget, Section 4 realizes this decomposition by setting $\lambda_{\text{cov}} = 1$, $\lambda_{\text{rel}} = \beta$, and $\lambda_{\text{prov}} = \gamma$. Either flat retrieval or graph traversal can expose $\mathcal{D}$ without changing the formulation. The primary system path uses graph-derived evidence items because they make provenance traces and contradictions explicit; the crossed evaluation also runs the same selector over an independently retrieved flat-corpus pool.

Definition 1 (Omission-cost coverage)**.** *Given facet set $\mathcal{F}(q, c)$, evidence set S, and risk tier r, omission-cost coverage is*

$$\text{Cov}(S; \mathcal{F}, r) = \sum_{f \in \mathcal{F}} \alpha_f(r) \cdot \mathbb{I}\Big[\exists s \in S : f \in C(s)\Big],$$

where $C(s)$ is the set of facets covered by evidence item s. The facet bases assign larger costs to contraindications, black-box warnings, and severe DDIs than to routine background facets; m_r then scales their aggregate contribution against the other objective terms.

For an evidence budget B, evaluation-time Facet Coverage@B is the covered fraction of frozen, manually annotated clinical-priority weight,

$$\text{FCov@}B(q) = \frac{\sum_{f \in \mathcal{F}^\star(q,c)} a_f \mathbb{I}[\exists s \in S_B : f \in C^\star(s)]}{\sum_{f \in \mathcal{F}^\star(q,c)} a_f},$$

where $\mathcal{F}^\star$ and $C^\star$ are the same adjudicated gold requirements and mappings for every method; unstarred $\mathcal{F}$ and C denote the policy's runtime inference. The reported score macro-averages over queries. This is the normalized form obtained with $\alpha_f(r)$ because query-level m_r cancels; FCov therefore isolates priority-weighted facet support, while ECE and Brier score separately assess the risk predictor. FCov is computed on selected evidence before generation, so an eventual abstention remains in the retrieval denominator.

The evidence in S is supplied to GPT-4o (2024-05-13 checkpoint, temperature 0, max_tokens 1024), which generates an answer citing elements of S inline. When targeted expansion still leaves critical facets without support, the system provides a structured explanation for abstaining instead of generating a recommendation that could be harmful.

4 RISKWEAVE

RiskWeave protects the support chain for high-priority clinical facets. Its selection unit is an evidence item: a normalized clinical assertion, its facet mapping, provenance record, source timestamp, extraction confidence, and conflict-set identifier. A supporting passage is source text that realizes an item; up to two passages per selected item may enter the generator context. This separation lets the selector reason over clinical coverage while citations remain anchored to source text.

4.1 CLINICAL SUPPORT CONTRACT AND RISK-SCALED COSTS

Entity linking normalizes drugs, diagnoses, laboratory state, age, pregnancy status, and organ function. Medication-label and guideline triggers then construct the required facet set $\mathcal{F}(q, c)$; for example, a severe interaction creates a DDI facet and a renal threshold can activate an organ-dosing facet. Evidence-side mappings $C(s)$ are assigned from normalized assertion type and source section, with the assertion-link confidence retained in the evidence item. The annotated retrieval set supplies evaluation labels rather than facet- or risk-model supervision.

A logistic model predicts query-level major-error probability before candidate exposure. Its supervised development label is one when an answer's medication recommendation is consequentially incorrect

or omits an action requiring intervention, and zero otherwise; this endpoint is distinct from retrieval coverage. All inputs are computed from (q, c) before exposure: medication count, inferred facet count, a severe-DDI trigger, CKD$\geq$ 4, Child–Pugh B/C, and a rule-based conflict trigger. The model is fitted on 2,700 development queries and isotonic-calibrated on 400 further queries; risk-tier cut points and the complete policy operating point are frozen on a 100-query validation set. These three sets are mutually disjoint and separate from every reported evaluation.

Risk is low below 0.30, medium from 0.30 to below 0.65, and high at or above 0.65. The prespecified facet bases (a_f) are 1.0 for monitoring, 1.3 for routine dosing, 1.8 for organ-specific dosing, 2.3 for severe DDI, 2.5 for contraindication, and 2.7 for black-box warning. These values encode an ordinal clinical-priority policy rather than patient-specific outcome utilities. Risk multipliers (m_r) are 0.75, 1.0, and 1.40, giving $\alpha_f(r) = a_f m_r$. Patient context determines which requirements exist; a_f orders their omission priority; and m_r changes the leverage of total required support against relevance and provenance without changing the within-query facet ordering. The matched full-minus-weight ablation sets every $a_f = 1$ while leaving m_r and all later policy stages unchanged. No cost, threshold, or coefficient is re-estimated on a reported evaluation cohort.

4.2 Budgeted selection

Let $\mathcal{C}$ denote exposed evidence items, $\mathcal{F}$ inferred facets, and B the item budget. We select

$$S^* = \arg \max_{S \subseteq \mathcal{C}, |S| \leq B} \Big[\sum_{f \in \mathcal{F}} \alpha_f(r)\, \mathbb{I}[f \in \cup_{s \in S} C(s)] + \sum_{s \in S} \beta\, \mathrm{sim}(q, s) + \sum_{s \in S} \gamma\, w(s) \Big], \qquad (2)$$

where $\mathrm{sim}(q, s)$ is min–max-normalized BGE-M3 similarity (Chen et al., 2024). The provenance score is

$$w(s) = 0.45 w_{\mathrm{source}}(s) + 0.35 w_{\mathrm{extract}}(s) + 0.20 \exp(-\mathrm{age}(s)/1825),$$

where the fixed source-class scores are 1.00 for regulatory labels, 0.90 for specialty guidelines, 0.82 for curated DDI monographs, and 0.68 for peer-reviewed reports. They encode the source hierarchy used by the policy; extraction confidence is the calibrated assertion-link probability, and recency is unit bounded by the exponential term. We set $\beta = 1.0$ and $\gamma = 0.5$.

Proposition 1. *If $\alpha_f(r), \beta, \gamma \geq 0$, similarity and provenance scores are nonnegative, and coverage is binary, the objective in Eq. 2 is normalized, monotone, and submodular. Greedy marginal-gain selection under $|S| \leq B$ therefore achieves at least $(1 - 1/e)$ of the optimal objective.*

The proof appears in Appendix A.1. The guarantee applies to selection; conflict reservation, expansion, and abstention are explicit policy stages.

4.3 Exposure, conflict preservation, and targeted expansion

The primary path expands two graph hops around linked query entities, then retains the top 20,000 items by $0.80\,\mathrm{sim} + 0.20\,w$. The graph-free path independently retrieves from the same source corpus with BM25 and BGE-M3, fuses ranks with reciprocal-rank constant $k = 60$, and applies the same 20,000-item cap. Both paths then execute the same selector.

An opposing assertion is clinically material when it reverses recommendation polarity, contraindication status, dose action, monitoring threshold, or severity for the same medication–patient facet. Whenever a chosen assertion has such an opponent with provenance score at least 0.65 and a slot remains, RiskWeave assigns the next slot to the highest-objective opposing item before resuming greedy selection. A pair discovered only after the item budget is exhausted leaves its facet unsupported for the answer-or-abstain rule. Greedy ties resolve by higher provenance, newer source date, and finally stable document identifier.

If a critical facet remains uncovered, targeted expansion concatenates the normalized medication, patient trigger, facet name, and missing-support polarity into a facet-specific query. At round t, newly exposed items are added to the current candidate union $\mathcal{C}^{(t)}$, and selection restarts from the empty set over that union. The resulting $S^{(t)}$ replaces the previous selection rather than appending to it, so $|S^{(t)}| \leq B$ in every round and conflict reservations remain inside the same item budget. Expansion stops after two rounds, when all critical facets are supported, or when the best normalized marginal

Algorithm 1 RiskWeave evidence allocation

infer facets $\mathcal{F}$, risk r, and costs $\{\alpha_f(r)\}$
expose and pre-prune $\mathcal{C}$; initialize $S \leftarrow \varnothing$
greedily select S under B, reserving qualifying conflicts inside B
for $t = 1, 2$ while critical support is missing **do**
    expose with the facet-specific query and update $\mathcal{C}$
    reselect S from $\varnothing$ under the same item budget B
end for
retrieve, deduplicate, and budget-pack passages for final S
answer with citations, or abstain under the risk-tier rule

gain is below 0.05. In frozen-pool controls, $\mathcal{C}^{(t)}$ cannot grow: a facet-specific pass may only reorder the original pool.

After the final item set is chosen, hybrid BM25+BGE-M3 retrieval obtains at most two passages per item. Passages with cosine similarity above 0.92 are deduplicated; the rest are packed by item marginal gain into 4,096 tokens, preserving one passage per selected item before allocating second passages. Only the final passage may be truncated. High-risk queries abstain if any critical facet remains unsupported; medium-risk queries abstain with at least two unsupported high-priority facets; low-risk queries abstain with at least three unsupported facets. These thresholds are fixed on the 100-query validation set and reused unchanged. On the 847-question internal cohort, they yield 3.7% abstention (8.7% high, 3.2% medium, 0.8% low).

With $N = |\mathcal{C}|$, $F = |\mathcal{F}|$, budget B, and R expansion rounds, lazy greedy selection is $O(R(NF + NB \log N))$ and stores $O(NF)$ facet incidence. Candidate encoding and passage retrieval are shared across baseline comparisons.

4.4 Evidence substrate

The provenance-bearing Neo4j 5.12 substrate contains FDA labels, clinical guidelines, Lexicomp and Micromedex DDI resources, and PubMed-derived snippets. A fine-tuned PubMedBERT-style encoder performs entity extraction with UMLS/RxNorm normalization (Lee et al., 2020; Alsentzer et al., 2019; Aronson & Lang, 2010; Donnelly, 2006). Rule parsing extracts labeled FDA sections, few-shot extraction handles free-text guidelines, and curated imports retain their source identifiers. Contradictory assertions are stored as separate edges in one conflict set, preserving the source and timestamp needed by RiskWeave (Neo4j, Inc., 2023; National Library of Medicine, National Center for Biotechnology Information, 2026).

5 Experimental Setup

RiskWeave is evaluated along three questions: whether omission-sensitive allocation changes support under common candidate access, whether its advantage follows the predicted scarcity mechanism, and whether the resulting policy improves answers and time-bounded clinical decisions.

5.1 Datasets and blinded assessment

The internal benchmark contains 847 clinician-authored medication-safety questions: DDIs (31%), contraindications (25%), dosing (22%), monitoring (13%), and black-box warnings (9%). Twenty-three pharmacists and physicians at seven academic centers authored questions paired with de-identified MIMIC-III contexts (Johnson et al., 2016); the evidence corpus is frozen to sources published before 2023-06-01. The population-replication set contains 423 eICU medication-safety questions spanning 208 hospitals (Pollard et al., 2018).

Seven experts (4 PharmDs, 3 MDs; mean 12.4 years of experience) graded correctness, completeness, safety, and utility. Each answer receives 2 independent ratings in a balanced incomplete-block schedule that equalizes reviewer–system and reviewer–question-family counts; system identity is masked and answer order randomized. Dimension means define acceptability (all dimensions at least 4.0) and harmfulness (aggregate safety at most 2.0). Classification disagreement or a two-point

dimension gap triggers a third rating for 11.6% of answers, after which the median is used. Agreement is 0.81 ICC(2,k) (95% CI 0.78–0.84) and 0.76 Fleiss' κ for acceptability.

Two annotators label 400 internal queries with required facets, gold snippets, and facet–evidence mappings (Cohen's $\kappa = 0.79$; senior-pharmacist adjudication). A disjoint 200-query frozen-candidate control supplies the query-level checker audit over the five frozen system outputs for each query, with a 50-query pilot nested inside. The 3,200-query risk-model pool (2,700 fit/400 calibration/100 operating point), 423 eICU set, and 160-query generator-swap set are separate from both internal evaluation slices. Questions, rather than passages or individual facets, are the sampling and resampling unit. Appendix Table 3 maps every set.

5.2 Systems and parity controls

Every method uses the same corpus freeze, eligible sources, 20,000-item cap, passage retriever, GPT-4o 2024-05-13 generator (temperature 0, max_tokens 1024), prompt, $B = 10$, 4,096-token context, and blinded rubric. Dense RAG uses BGE-M3 similarity. xQuAD treats safety facets as query aspects and applies its probabilistic aspect weighting to reward uncovered facets. SubmodCov assigns one equal gain to every newly covered clinical facet, providing the set-aware uniform-cost control. GraphRAG follows the Microsoft community-summary pipeline with identical truncation. Flat-CE is the zero-shot `BAAI/bge-reranker-v2-m3` cross-encoder, scoring concatenated query/context and normalized evidence assertions at maximum sequence length 1024, without task labels or benchmark tuning. Dense and Flat-CE test pointwise relevance, xQuAD and SubmodCov test aspect- and set-level coverage, and GraphRAG tests community-summary exposure.

The fixed-pool control runs graph exposure once and gives every ranker identical items. No later pass can add a candidate: a facet-targeted pass can only rescore the original pool, after which RiskWeave reselects from the empty set under the same $B = 10$ item and 4,096-token passage budgets. Passage retrieval and packing procedures, generator settings, budgets, and scoring denominators remain identical. SubmodCov and Flat-CE evaluate their stated allocation rules; the complete RiskWeave row retains provenance scoring, conflict reservation, and its answer-or-abstain decision. This comparison estimates the complete policy effect under common exposure. The matched full-minus-weight ablation then sets every $a_f = 1$, isolating clinical-priority ordering while retaining m_r and the other RiskWeave stages.

The exposure–allocation decomposition adds a graph-free path: BM25 and BGE-M3 independently retrieve from the same corpus, reciprocal-rank fusion ($k = 60$) produces the same 20,000-item cap, and each ranker receives that frozen flat pool. This 2×3 design separates exposure (graph versus flat fusion) from allocation (SubmodCov, Flat-CE, or RiskWeave), rather than attributing every policy-stage difference to the omission-cost coefficient alone.

5.3 Metrics and inference

Facet Coverage@B is defined in Section 3; high-priority coverage restricts it to severe DDI, contraindication, black-box warning, and organ-specific dosing facets. Exposure recall@20k is the fraction of gold high-priority facets present before reranking. A claim violation asserts a required high-priority facet without a cited supporting snippet. Major error denotes a consequential incorrect or omitted recommendation requiring intervention; ranking-control, benchmark, and workflow rates apply that endpoint to different decision units. Acceptability and harmfulness follow the blinded rules above. Abstention is unanswered for acceptability and remains in every retrieval denominator. ECE and Brier score target the risk model's major-error probability.

Retrieval means, paired deltas, confidence intervals, and two-sided p values use 10,000 paired bootstrap resamples; binary rates receive exact-binomial intervals, and paired binary contrasts use exact McNemar tests with Holm correction within the graph-exposure family reported in Appendix Table 8 (Holm, 1979). The declared primary graph-exposure contrasts are RiskWeave versus Flat-CE and SubmodCov for FCov and acceptability. Claim violations and harmful outcomes are supportive sparse-event endpoints; the graph-free crossing, eICU cohort, generator substitution, and slices test stability and mechanism. Workflow outcomes use mixed-effects logistic regression with clinician random intercepts plus period and carryover terms; completion times use Wilcoxon signed-rank tests under the crossover schedule's within-clinician pairing. eICU intervals use questions as units for

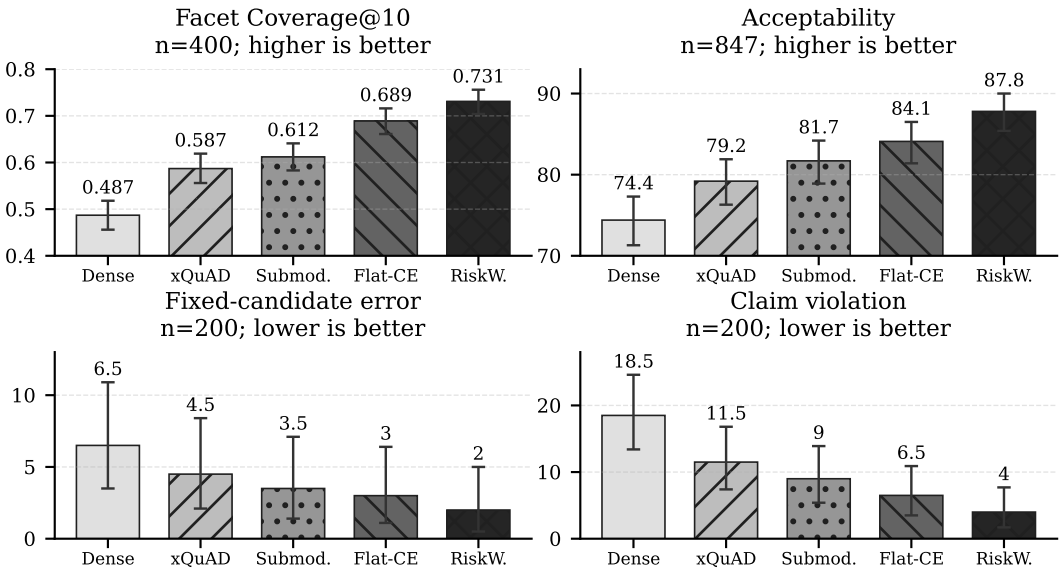


Figure 1: Evidence support and answer outcomes under the shared 4,096-token budget. The panel denominators are 400, 847, 200, and 200; bars show paired-bootstrap intervals for FCov and exact-binomial intervals for rates.

patient-context cohort replication rather than hospital-level effects. Retrieval runs use 5 fixed seeds and report pooled query-level estimates. The generator swap holds retrieval, prompt skeleton, budget, and rubric fixed while substituting Llama-3.1-70B-Instruct.

Experiments run on 4×A100 GPUs with Neo4j 5.12, BGE-M3 embeddings, and the frozen settings in Section 4. In the balanced crossover, 248 clinicians each complete eight trials—four unassisted and four with the complete RiskWeave answer, citations, conflict indicators, and abstention state—under clinician-level AB/BA assignment; all 1,984 trials enter the intention-to-treat analysis. Separate evidence comprises 1,440 three-arm task trials, 1,847 shadow-mode decisions, and 384 paired decisions. Appendix Table 15 specifies units, assignment, endpoints, and inference under crossover-trial and clinical-AI reporting guidance (Dwan et al., 2019; Liu et al., 2020).

6 Results

6.1 Omission-sensitive retrieval

Under the common ten-item budget, RiskWeave reaches 0.731 Facet Coverage@10 on 400 annotated questions and 87.8% clinical acceptability across 847 questions per system under blinded expert grading. Flat-CE reaches 0.689 and 84.1%, while SubmodCov reaches 0.612 and 81.7% (Figure 1 and Table 1). Relative to SubmodCov, the paired FCov gain is 0.119 (95% CI 0.087–0.151; $p = 0.002$). In the separate 200-query frozen-pool control, the complete allocation policy records 4 major errors and 8 unsupported high-priority claims, versus 7 and 18 for SubmodCov and 6 and 13 for Flat-CE.

Table 1 summarizes the broad internal evaluation. Table 2 then crosses candidate exposure and allocation while holding corpus, passage retrieval, generator, prompt, and budget fixed. This design separates exposure from complete-policy differences; the full-minus-weight comparison separately targets clinical-priority ordering. The 200-query checker audit has 87.8% precision, 90.0% recall, 88.9% F1, and 95.5% query-level agreement (Appendix Table 14).

6.2 Exposure versus allocation

Flat fusion lowers high-priority exposure recall from 0.963 to 0.925; within that pool, RiskWeave yields 0.715 FCov versus 0.674 for Flat-CE and records 9/200 rather than 14/200 violations. Under

Table 1: Broad internal uncertainty summary. Denominators appear in the headers; cells report an estimate or event count followed by rate and 95% CI.

System	FCov@10 [95% CI]	Acceptable / 847 [95% CI]
Dense	0.487 [0.456, 0.518]	630 (74.4% [71.3,77.3])
xQuAD	0.587 [0.556, 0.619]	671 (79.2% [76.3,81.9])
SubmodCov	0.612 [0.583, 0.641]	692 (81.7% [78.9,84.2])
Flat-CE	0.689 [0.661,0.716]	712 (84.1% [81.4,86.5])
RiskWeave	0.731 [0.704, 0.756]	744 (87.8% [85.4,90.0])
System	Fixed error / 200 [95% CI]	Claim violation / 200 [95% CI]
Dense	13 (6.5% [3.5,10.9])	37 (18.5% [13.4,24.6])
xQuAD	9 (4.5% [2.1,8.4])	23 (11.5% [7.4,16.8])
SubmodCov	7 (3.5% [1.4,7.1])	18 (9.0% [5.4,13.9])
Flat-CE	6 (3.0% [1.1,6.4])	13 (6.5% [3.5,10.9])
RiskWeave	4 (2.0% [0.5,5.0])	8 (4.0% [1.7,7.7])

Table 2: Exposure–allocation decomposition. The primary allocation comparison is RiskWeave versus Flat-CE within each path. Recall and FCov use $n = 400$; acceptability uses $n = 847$; violations and harmful outcomes show exact counts.

Exposure	Ranker	Recall@20k	FCov@10	Accept. (%)	Viol./200	Harm/248
Graph	SubmodCov	0.963	0.612	81.7	18	11
Graph	Flat-CE	0.963	0.689	84.1	13	9
Graph	RiskWeave	0.963	0.731	87.8	8	6
Flat fusion	SubmodCov	0.925	0.596	80.3	20	13
Flat fusion	Flat-CE	0.925	0.674	83.1	14	10
Flat fusion	RiskWeave	0.925	0.715	86.3	9	7

graph exposure, RiskWeave exceeds Flat-CE by 0.042 FCov (95% CI 0.018–0.067; $p = 0.001$) and 3.8 acceptability points (95% CI 1.4–6.1; Holm-adjusted McNemar $p = 0.012$). The sparse claim-violation contrast is directionally aligned but individually uncertain: -2.5 points (95% CI -5.4–0.4; adjusted $p = 0.111$). Appendix Tables 6 and 7 report intervals, fixed-candidate errors, safety events, and P50/P95 latency.

6.3 Mechanisms and operating point

The matched full-minus-weight ablation directly tests asymmetric clinical priorities inside the complete policy: setting every $a_f = 1$ while retaining m_r lowers FCov from 0.731 to 0.667, high-priority coverage from 0.782 to 0.641, and acceptability from 87.8% to 84.3%, while violations rise from 8/200 to 14/200 (Appendix Table 9). Other removals locate complementary functions: provenance and conflict retention chiefly affect unsupported claims, while abstention exchanges answer rate for fewer harmful outcomes. The declared paired tests remain the inferential basis for the primary system contrasts; the component rows identify where the observed policy gains enter.

The budget response supplies an independent mechanism check. At $B = 4$, RiskWeave exceeds Flat-CE by 0.102 FCov and records 16 rather than 29 violations; at $B = 10$, the differences narrow to 0.042 and 8 versus 13 (Appendix Table 12). Risk/facet point estimates concentrate in high-risk, severe-DDI, contraindication, and black-box-warning questions, while low-risk and monitoring slices show no systematic gain (Appendix Table 11). The aggregate paired contrasts, full-minus-weight ablation, and prespecified budget response carry the mechanism conclusion; the sparse slices localize the pattern.

Risk predictions remain calibrated: internal ECE is 0.031 and Brier score 0.089; external ECE is 0.038 and Brier score 0.094. On the 847-question internal end-to-end cohort, the frozen operating point yields 3.7% abstention and a 2.3% benchmark major-error rate. This cohort-level endpoint is distinct from the 200-query ranking-control error and per-trial clinician error. Moving to 8%

abstention lowers the benchmark rate to 1.1%, exposing a controllable answer-rate–error tradeoff at a fixed retrieval policy.

6.4 Replication across cohort and generator

On 423 eICU questions drawn from 208 hospitals, RiskWeave reaches 85.3% acceptability and 3.1% harmful answers, compared with 78.7% and 5.0% for SubmodCov, the strongest evaluated comparator on this cohort (Appendix Table 5). With Llama-3.1-70B-Instruct, RiskWeave reaches 83.8% acceptability and 0.706 FCov@10, compared with 80.0% and 0.662 for Flat-CE (Appendix Table 17).

The frozen pool measures complete-policy differences under common exposure, full-minus-weight isolates clinical priorities within RiskWeave, and blinded grading measures answer quality. The clinician randomization below instead estimates the complete interface.

6.5 Clinician-task consequences

In the randomized crossover (248 clinicians, eight trials each), complete RiskWeave assistance lowers major errors from 88/992 (8.9%) to 23/992 (2.3%; adjusted OR 0.22, 95% CI 0.14–0.34; $p < 0.001$) and median time from 10.4 to 5.8 minutes (difference -4.6, 95% CI [-4.9, -4.2] minutes; $p < 0.001$). This estimates the complete interface; fixed-pool and component analyses supply retrieval attribution.

Appendix Tables 15 and 16 report the three-arm and paired-decision replications with their units, assignments, and inference.

Removing both citation grounding and abstention raises harmful outcomes from 6/248 to 29/248. RiskWeave operates at 2.24s P50 and 4.18s P95 (Appendix Table 7).

7 Discussion

The budget curve locates the scarcity regime, and full-minus-weight isolates a_f. At $B = 4$ and $B = 6$, routine background can displace a contraindication or severe DDI, producing the largest RiskWeave advantage; the gap narrows as more facets fit. FCov uses the prespecified priorities and cancels m_r, measuring protection of the declared support contract. Blinded acceptability, audited claim violations, and clinician errors provide consequence endpoints outside that score.

Provenance and conflict state connect allocation to answering: expansion, passage packing, citation, and abstention operate on the same assertions. Accordingly, weights chiefly improve high-priority coverage, while provenance, conflicts, grounding, and abstention chiefly affect unsupported or harmful outputs. Graph and flat pools vary exposure, the frozen pool fixes candidate access, full-minus-weight isolates priorities, and randomized tasks measure the complete interface. The systems insight is that a retrieval objective becomes decision-relevant when its missing-support state remains available after ranking.

The evidence establishes curated medication-safety QA under four to ten items. The a_f values are global ordinal priorities, and m_r is not isolated from the operating policy; eICU intervals quantify question-level replication rather than hospital effects; and clinician randomization estimates the complete interface, with selector attribution supplied by frozen-pool and weight controls. Within those units, exposure changes, blinded grading, generator substitution, and time-bounded decisions favor allocation by omission consequence. Transfer requires defining the target specialty's facets, source hierarchy, and operating point while preserving the support contract.

8 Conclusion

RiskWeave allocates by high-priority omission cost, carrying residual support into retrieval and abstention. Under ten slots, Facet Coverage@10 rises from 0.612 to 0.731 and acceptability from 81.7% to 87.8% over uniform coverage, with significant paired gains over Flat-CE. In the randomized crossover, the complete interface cuts errors from 88/992 to 23/992 and time from 10.4 to 5.8 minutes: pricing omitted support protects decision-controlling evidence.

Ethics Statement

The retrospective analyses use de-identified MIMIC-III and eICU records under their credentialed research terms and the anonymous institution's applicable review determination. The clinician-task studies received institutional review approval; participants provided written informed consent and were compensated at approved local rates. No intervention changed patient care, and all reported records and judgments are de-identified. RiskWeave surfaces citations, preserves material conflicts, and abstains on unsupported critical facets; clinicians retain responsibility for every task decision. These safeguards address automation bias while preserving the study's focus on retrieval and decision support (Wiens et al., 2019; Kelly et al., 2019; Obermeyer et al., 2019; Chen et al., 2023).

Reproducibility Statement

The data construction, frozen split map, model checkpoints, selection rule, operating point, baseline parity controls, metrics, and uncertainty procedures are specified in Sections 3–5. The appendices provide the optimization proof, algorithmic details, exposure–allocation decomposition, mechanism and slice analyses, grader-assignment protocol, raw workflow counts, and second-generator evaluation needed to reproduce every reported comparison.

AI Use Statement

Generative AI was used in the documented manuscript-preparation workflow to edit prose and presentation, refine the title and abstract, normalize formatting, verify bibliography metadata, and conduct submission-readiness checks. This statement describes documented preparation assistance and does not make claims about undocumented uses.

References

Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, and Samuel Ieong. Diversifying search results. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pp. 5–14. ACM, 2009. doi: 10.1145/1498759.1498766.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pp. 72–78. Association for Computational Linguistics, 2019. doi: 10.18653/v1/W19-1909. URL https://aclanthology.org/W19-1909/.

Alan R Aronson and François-Michel Lang. An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236, 2010. doi: 10.1136/jamia.2009.002733.

Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270, 2004. doi: 10.1093/nar/gkh061.

Jaime Carbonell and Jade Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 335–336. ACM, 1998. doi: 10.1145/290941.291025.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. M3-Embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 2318–2335. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.137. URL https://aclanthology.org/2024.findings-acl.137/.

Richard J. Chen, Judy J. Wang, Drew F. K. Williamson, Tiffany Y. Chen, Jana Lipkova, Ming Y. Lu, Sharifa Sahai, and Faisal Mahmood. Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nature Biomedical Engineering*, 7(6):719–742, 2023. doi: 10.1038/s41551-023-01056-8.

Kevin Donnelly. Snomed-ct: The advanced terminology and coding system for ehealth. *Studies in health technology and informatics*, 121:279–290, 2006. URL https://pubmed.ncbi.nlm.nih.gov/17095826/.

Kerry Dwan, Tianjing Li, Douglas G Altman, and Diana Elbourne. Consort 2010 statement: extension to randomised crossover trials. *BMJ*, 366:l4378, 2019. doi: 10.1136/bmj.l4378.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024. URL https://arxiv.org/abs/2404.16130.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. RAGAs: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pp. 150–158. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.eacl-demo.16.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023. URL https://arxiv.org/abs/2312.10997.

Yajie Vera He, Mohita Chowdhury, Jared Joselowitz, Aisling Higham, and Ernest Lim. ASTRID—an automated and scalable TRIaD for the evaluation of RAG-based clinical question answering systems. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 16700–16716. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.857.

Daniel S Himmelstein, Antoine Lizee, Christine Hessler, Leo Brueggeman, Sabrina L Chen, Dexter Hadley, Ari Green, Pouya Khankhanian, and Sergio E Baranzini. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *eLife*, 6:e26726, 2017. doi: 10.7554/eLife.26726.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. doi: 10.2307/4615733. URL https://doi.org/10.2307/4615733.

Yi Jiang, Sendong Zhao, Jianbo Li, Bairui Hu, Yanrui Du, Haochun Wang, and Bing Qin. OptiSet: Unified optimizing set selection and ranking for retrieval-augmented generation. *arXiv preprint arXiv:2601.05027*, 2026. doi: 10.48550/arXiv.2601.05027. URL https://arxiv.org/abs/2601.05027.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):160035, 2016. doi: 10.1038/sdata.2016.35.

Hailey Joren, Jianyi Zhang, Chun-Sung Ferng, Da-Cheng Juan, Ankur Taly, and Cyrus Rashtchian. Sufficient context: A new lens on retrieval augmented generation systems. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Jjr2Odj8DJ.

Christopher J Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine*, 17(1):195, 2019. doi: 10.1186/s12916-019-1426-2.

Dahyun Lee, Yongrae Jo, Haeju Park, and Moontae Lee. Shifting from ranking to set selection for retrieval augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 17606–17619. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.861. URL https://aclanthology.org/2025.acl-long.861/.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020. doi: 10.1093/bioinformatics/btz682.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33: 9459–9474, 2020. URL https://arxiv.org/abs/2005.11401.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638.

Xiaoxuan Liu, Samantha Cruz Rivera, David Moher, Melanie J. Calvert, Alastair K. Denniston, and SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the consort-ai extension. *The Lancet Digital Health*, 2(10):e537–e548, 2020. doi: 10.1016/S2589-7500(20)30218-1.

National Library of Medicine, National Center for Biotechnology Information. NIH NLM NCBI PubMed Central (PMC) Article Datasets—Full-Text Biomedical and Life Sciences Journal Articles on AWS. https://registry.opendata.aws/ncbi-pmc/, 2026. Accessed 2026-08-30.

George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. An analysis of approximations for maximizing submodular set functions—i. *Mathematical programming*, 14(1):265–294, 1978. doi: 10.1007/BF01588971.

Neo4j, Inc. Neo4j Graph Database, Version 5.12. Computer software, https://neo4j.com/docs/operations-manual/5/, 2023. Accessed 2026-08-30.

Nghia Trung Ngo, Chien Van Nguyen, Franck Dernoncourt, and Thien Huu Nguyen. Comprehensive and practical evaluation of retrieval-augmented generation systems for medical question answering. *arXiv preprint arXiv:2411.09213*, 2024. URL https://arxiv.org/abs/2411.09213.

David N Nicholson and Casey S Greene. Constructing knowledge graphs and their biomedical applications. *Computational and Structural Biotechnology Journal*, 18:1414–1428, 2020. doi: 10.1016/j.csbj.2020.05.017.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019. doi: 10.1126/science.aax2342.

Tom J. Pollard, Alistair E. W. Johnson, Jesse D. Raffa, Leo A. Celi, Roger G. Mark, and Omar Badawi. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific data*, 5(1):180178, 2018. doi: 10.1038/sdata.2018.178.

Dongyu Ru, Lin Qiu, Xiangkun Hu, Tianhang Zhang, Peng Shi, Shuaichen Chang, Cheng Jiayang, Cunxiang Wang, Shichao Sun, Huanyu Li, Zizhao Zhang, Binjie Wang, Jiarong Jiang, Tong He, Zhiguo Wang, Pengfei Liu, Yue Zhang, and Zheng Zhang. RAGChecker: A fine-grained framework for diagnosing retrieval-augmented generation. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 37, pp. 21999–22027, 2024. doi: 10.52202/079017-0692. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/27245589131d17368cccdfa990cbf16e-Abstract-Datasets_and_Benchmarks_Track.html.

Rodrygo LT Santos, Craig Macdonald, and Iadh Ounis. Exploiting query reformulations for web search result diversification. In *Proceedings of the 19th international conference on World wide web*, pp. 881–890. ACM, 2010. doi: 10.1145/1772690.1772780.

Mooho Song and Jay-Yoon Lee. Retrieving a set, not independent passages: Set-level compatibility learning for efficient set exploration. *arXiv preprint arXiv:2607.05712*, 2026. doi: 10.48550/arXiv.2607.05712. URL https://arxiv.org/abs/2607.05712.

Hui Wang, Bohao Chu, and Norbert Fuhr. CrossDDI: Cross-source evidence-grounded drug–drug interaction verification. In *BioNLP 2026*, pp. 911–919. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.bionlp-1.73. URL https://aclanthology.org/2026.bionlp-1.73/.

Jenna Wiens, Suchi Saria, Mark Sendak, Marzyeh Ghassemi, Vincent X Liu, Finale Doshi-Velez, Kenneth Jung, Katherine Heller, David Kale, Mohammed Saeed, et al. Do no harm: a roadmap for responsible machine learning for health care. *Nature Medicine*, 25(9):1337–1340, 2019. doi: 10.1038/s41591-019-0548-6.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. Benchmarking retrieval-augmented generation for medicine. *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 6233–6251, 2024. doi: 10.18653/v1/2024.findings-acl.372. URL https://aclanthology.org/2024.findings-acl.372/.

Yue Yu, Wei Ping, Zihan Liu, Boxin Wang, Jiaxuan You, Chao Zhang, Mohammad Shoeybi, and Bryan Catanzaro. RankRAG: Unifying context ranking with retrieval-augmented generation in LLMs. In *Advances in Neural Information Processing Systems*, volume 37, pp. 121156–121184, 2024. doi: 10.52202/079017-3850. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/db93ccb6cf392f352570dd5af0a223d3-Abstract-Conference.html.

A Supplementary Methods and Results

A.1 Optimization property

Proof of Proposition 1. For facet f, define $g_f(S) = \alpha_f(r)\mathbb{I}[\exists s \in S : f \in C(s)]$. With $\alpha_f(r) \geq 0$, g_f is normalized and monotone. Its marginal gain is $\alpha_f(r)$ until f is covered and zero thereafter, so the marginal gain cannot increase as the selected set grows. Thus g_f is submodular. The coverage term is a nonnegative sum of such functions. Relevance and provenance are nonnegative modular functions, which are also monotone submodular; their nonnegative sum with coverage preserves these properties. The standard greedy bound for a cardinality constraint therefore gives $U(S_{\text{greedy}}) \geq (1 - 1/e)U(S^*)$ (Nemhauser et al., 1978). □

A.2 Evaluation-set map

Table 3: Dataset and split relations. Cross-cutting subsets may overlap their named parent but not another disjoint evaluation slice.

Set	n	Relation	Role
Risk development	3,200	separate; 2,700/400/100	fit / isotonic calibration / operating point
Internal benchmark	847	parent evaluation set	blinded end-to-end grading
Annotated retrieval	400	internal; disjoint from control	facet and evidence labels
Fixed-candidate control	200	internal; disjoint from annotated	held-out allocation mechanism
Checker pilot	50	nested in control	initial checker validation
High-risk cohort	248	cross-cutting internal subset	harmful-outcome analysis
eICU	423	external, 208 hospitals	descriptive cohort replication
Generator swap	160	separate held-out set	Llama-3.1-70B-Instruct test

A.3 Technical positioning and external cohort

Table 4 compares the allocation object and downstream decision state, which are the relevant distinctions for the method. Table 5 then reports the complete external cohort results summarized in Section 6.4.

Table 4: Closest technical neighbors and the evidence-allocation decision they implement.

Family	Allocation objective and unit	Missing-support treatment and decision coupling
MMR	passage-level relevance and pairwise novelty	no explicit facet requirement; ranking only
IA-Select / xQuAD	weighted taxonomy categories / query aspects	nonuniform intent priors; ranking only
SubmodCov	set-level uniform clinical-facet coverage	one gain per newly covered facet; ranking only
Flat-CE	pointwise query–assertion relevance	no selected-set state or explicit missing-support cost
GraphRAG / bio-KG	graph-neighborhood context exposure	organizes provenance without an omission-loss allocation rule
SetR / OptiSet / SetCE	learned set-level sufficiency, complementarity, or compatibility	jointly values passages without an explicit clinical omission-cost contract
RiskWeave	set-level marginal clinical omission-cost reduction	required-facet support with provenance, conflict, expansion, and abstention

Table 5: External eICU benchmark ($n = 423$, 208 hospitals) under the shared evidence budget. Intervals use questions as units.

System	Accept. (%)	Harmful (%)
Dense RAG	71.6 (67.1–75.9)	8.3 (5.8–11.3)
GraphRAG	75.4 (71.0–79.4)	6.4 (4.2–9.2)
xQuAD	76.1 (71.8–80.1)	5.9 (3.9–8.6)
SubmodCov	78.7 (74.5–82.5)	5.0 (3.1–7.5)
RiskWeave	85.3 (81.6–88.6)	3.1 (1.6–5.2)

Across the evaluated eICU systems, RiskWeave leads SubmodCov, the strongest comparator, by 6.6 acceptability points and 1.9 harmful-answer points. The question-level intervals quantify replication across patient contexts drawn from 208 hospitals.

A.4 Exposure–allocation decomposition

The decomposition first compares support and answer quality, then reports sparse safety events and latency. Splitting the display keeps the denominators and confidence intervals readable while preserving one row per exposure–ranker combination.

Table 6: Full same-corpus exposure–allocation support results. Recall and FCov show 95% paired-bootstrap intervals; acceptability uses $n = 847$ and fixed error uses $n = 200$.

Exposure	Ranker	Recall@20k	FCov@10	Acceptable	Fixed error
Graph	SubmodCov	0.963 [0.951,0.974]	0.612 [0.583,0.641]	81.7%	3.5%
Graph	Flat-CE	0.963 [0.951,0.974]	0.689 [0.661,0.716]	84.1%	3.0%
Graph	RiskWeave	0.963 [0.951,0.974]	0.731 [0.704,0.756]	87.8%	2.0%
Flat fusion	SubmodCov	0.925 [0.908,0.940]	0.596 [0.566,0.625]	80.3%	4.0%
Flat fusion	Flat-CE	0.925 [0.908,0.940]	0.674 [0.646,0.702]	83.1%	3.5%
Flat fusion	RiskWeave	0.925 [0.908,0.940]	0.715 [0.687,0.741]	86.3%	2.5%

Table 7: Safety events and end-to-end latency for the same exposure–allocation cells. Event intervals are reported with the primary contrasts in Table 8.

Exposure	Ranker	Claim violation	Harmful	P50 / P95 (s)
Graph	SubmodCov	18/200 (9.0%)	11/248 (4.4%)	2.08 / 3.86
Graph	Flat-CE	13/200 (6.5%)	9/248 (3.6%)	2.41 / 4.46
Graph	RiskWeave	8/200 (4.0%)	6/248 (2.4%)	2.24 / 4.18
Flat fusion	SubmodCov	20/200 (10.0%)	13/248 (5.2%)	1.86 / 3.54
Flat fusion	Flat-CE	14/200 (7.0%)	10/248 (4.0%)	2.19 / 4.08
Flat fusion	RiskWeave	9/200 (4.5%)	7/248 (2.8%)	2.04 / 3.91

Graph exposure recall is 0.963 versus 0.925 for flat fusion. RiskWeave exceeds Flat-CE in FCov by 0.042 in the graph pool and 0.041 in the flat pool; safety-event counts favor the same policy but retain the wider uncertainty shown by the paired contrasts.

Table 8: Paired primary contrasts under graph exposure. Deltas are RiskWeave minus comparator in percentage points except FCov; p values are Holm-adjusted for binary outcomes.

Metric	Comparator	Delta	95% CI	p
FCov@10	Flat-CE	0.042	[0.018,0.067]	0.001
Acceptability	Flat-CE	3.8	[1.4,6.1]	0.012
Claim violation	Flat-CE	-2.5	[-5.4,0.4]	0.111
Harmful	Flat-CE	-1.2	[-3.3,0.8]	0.219
FCov@10	SubmodCov	0.119	[0.087,0.151]	0.002
Acceptability	SubmodCov	6.1	[3.6,8.7]	<0.001
Claim violation	SubmodCov	-5.0	[-8.2,-1.8]	0.008
Harmful	SubmodCov	-2.0	[-4.3,0.4]	0.109

A.5 Mechanism ablations

All rows use the same exposed pools, generator, and budget. "Risk weights only" and "provenance only" retain the named selection signal beside common relevance; "risk + provenance" combines them. Each "full −" row removes one named component from the complete policy while retaining targeted expansion and the other stages. The partial policy variants retain the frozen residual-support decision, which accounts for their nonzero abstention rates.

Table 9: Same-pool mechanism ablation for support, acceptability, and answer rate. High-priority coverage (HP cov.) includes severe DDI, contraindication, black-box warning, and organ-specific dosing facets.

Variant	FCov@10	HP cov.	Acceptable	Abstain
SubmodCov	0.612	0.581	81.7%	0.0%
Flat-CE	0.689	0.654	84.1%	0.0%
Risk weights only	0.683	0.724	84.9%	1.4%
Provenance only	0.652	0.633	83.1%	1.1%
Risk + provenance	0.711	0.758	86.2%	2.0%
Full − priority weights	0.667	0.641	84.3%	2.8%
Full − provenance	0.698	0.746	85.1%	3.1%
Full − conflict retention	0.719	0.769	86.4%	3.5%
Full − abstention	0.731	0.782	89.1%	0.0%
Complete RiskWeave	0.731	0.782	87.8%	3.7%

Table 10: Safety outcomes for the same matched mechanism ablations.

Variant	Claim violation	Harmful
SubmodCov	18/200 (9.0%)	11/248 (4.4%)
Flat-CE	13/200 (6.5%)	9/248 (3.6%)
Risk weights only	12/200 (6.0%)	9/248 (3.6%)
Provenance only	15/200 (7.5%)	10/248 (4.0%)
Risk + provenance	10/200 (5.0%)	8/248 (3.2%)
Full − priority weights	14/200 (7.0%)	10/248 (4.0%)
Full − provenance	12/200 (6.0%)	9/248 (3.6%)
Full − conflict retention	11/200 (5.5%)	9/248 (3.6%)
Full − abstention	12/200 (6.0%)	15/248 (6.0%)
Complete RiskWeave	8/200 (4.0%)	6/248 (2.4%)

Clinical-priority weights contribute most to high-priority facet support. Provenance and conflict retention reduce unsupported claims, whereas abstention changes harmful outcomes and answer rate without changing pre-generation coverage.

A.6 RISK-TIER AND FACET-FAMILY SLICES

Table 11: Claim violations in the 200-query control, stratified by risk tier and principal facet. Difference is RiskWeave minus Flat-CE in percentage points; intervals are paired-bootstrap 95% intervals.

Slice	n	Flat-CE	RiskWeave	Difference [95% CI]
High risk	72	10/72 (13.9%)	5/72 (6.9%)	-6.9 [-14.6,0.7]
Medium risk	79	3/79 (3.8%)	2/79 (2.5%)	-1.3 [-5.6,3.1]
Low risk	49	0/49 (0.0%)	1/49 (2.0%)	2.0 [-2.0,6.1]
Severe DDI	62	5/62 (8.1%)	2/62 (3.2%)	-4.8 [-11.3,1.6]
Contraindication	50	4/50 (8.0%)	2/50 (4.0%)	-4.0 [-11.5,3.5]
Black-box warning	18	2/18 (11.1%)	1/18 (5.6%)	-5.6 [-20.2,9.1]
Organ dosing	42	1/42 (2.4%)	1/42 (2.4%)	0.0 [-6.6,6.6]
Monitoring	28	1/28 (3.6%)	2/28 (7.1%)	3.6 [-7.1,14.3]

A.7 EVIDENCE-BUDGET RESPONSE

The budget sweep varies only the number of evidence items. Coverage and high-priority support are shown separately from claim violations, abstention, and latency so that the scarcity response can be compared without a landscape display.

Table 12: Coverage response to evidence budget B on the 400-query annotated set. GF denotes graph-free exposure.

B	Exposure	Ranker	FCov@B	HP coverage
4	Graph	SubmodCov	0.402	0.372
4	Graph	Flat-CE	0.454	0.431
4	GF	Flat-CE	0.438	0.414
4	Graph	RiskWeave	0.556	0.628
6	Graph	SubmodCov	0.493	0.472
6	Graph	Flat-CE	0.565	0.548
6	GF	Flat-CE	0.548	0.529
6	Graph	RiskWeave	0.644	0.701
8	Graph	SubmodCov	0.561	0.548
8	Graph	Flat-CE	0.635	0.619
8	GF	Flat-CE	0.619	0.601
8	Graph	RiskWeave	0.698	0.756
10	Graph	SubmodCov	0.612	0.581
10	Graph	Flat-CE	0.689	0.654
10	GF	Flat-CE	0.674	0.656
10	Graph	RiskWeave	0.731	0.782

Table 13: Claim violations, abstention, and P95 latency across evidence budgets. Violations use the 200-query control; latency is end-to-end seconds.

B	Exposure	Ranker	Claim violation	Abstain	P95
4	Graph	SubmodCov	34/200 (17.0%)	0.0%	2.88
4	Graph	Flat-CE	29/200 (14.5%)	0.0%	3.21
4	GF	Flat-CE	31/200 (15.5%)	0.0%	2.97
4	Graph	RiskWeave	16/200 (8.0%)	7.8%	3.35
6	Graph	SubmodCov	27/200 (13.5%)	0.0%	3.26
6	Graph	Flat-CE	22/200 (11.0%)	0.0%	3.68
6	GF	Flat-CE	24/200 (12.0%)	0.0%	3.42
6	Graph	RiskWeave	12/200 (6.0%)	5.9%	3.71
8	Graph	SubmodCov	22/200 (11.0%)	0.0%	3.61
8	Graph	Flat-CE	17/200 (8.5%)	0.0%	4.06
8	GF	Flat-CE	19/200 (9.5%)	0.0%	3.81
8	Graph	RiskWeave	10/200 (5.0%)	4.6%	3.94
10	Graph	SubmodCov	18/200 (9.0%)	0.0%	3.86
10	Graph	Flat-CE	13/200 (6.5%)	0.0%	4.46
10	GF	Flat-CE	14/200 (7.0%)	0.0%	4.08
10	Graph	RiskWeave	8/200 (4.0%)	3.7%	4.18

The advantage is largest when four or six items must carry the answer. As B increases, every method covers more facets and violation rates fall, and the RiskWeave–Flat-CE FCov gap narrows from 0.102 at $B = 4$ to 0.042 at $B = 10$.

A.8 Claim-violation checker audit

Table 14: Automatic claim-violation checker against manual adjudication of 200 frozen control-query output sets.

Sample	TP	FP	FN	TN	Prec.	Rec.	F1	Agreement
200 queries	36	5	4	155	87.8	90.0	88.9	95.5

After all five outputs for each control query were frozen, the checker and manual adjudicators classified whether that output set contained a violation. The confusion counts therefore sum to 200 queries, not 1,000 outputs. Agreement is 95.5% on the complete audit versus the nested 50-query pilot.

A.9 Clinician-study protocol and inference

The randomized crossover is the primary clinician consequence study. Its unit is a clinician–case trial: 248 clinicians each complete eight trials, four under each condition, for 992 assisted and 992 unassisted decisions. The assisted interface presents the complete RiskWeave answer with inline citations, conflict state, and any abstention; the control presents the same task without system assistance. Clinician-level balanced AB/BA assignment defines treatment order, and the mixed model accounts for repeated trials with clinician random intercepts and period and carryover terms. All 1,984 scheduled observations enter the intention-to-treat analysis. The three-arm task study compares control, checklist, and the complete interface over 480 trials per arm; shadow monitoring is observational; and the 384 before/after decisions are analyzed as paired observations. All studies use the major-error endpoint defined in Section 5.

Table 15: Blinded grading and clinician-task analysis protocol.

Component	Prespecified implementation
Answer grading	Balanced assignment to 2 blinded experts per answer; randomized answer order; third rating for classification disagreement or a two-point dimension gap
Aggregation	Dimension means for concordant pairs; median of three after adjudication; acceptability requires every aggregate dimension ≥ 4.0
Reliability	ICC(2,k) 0.81 [0.78,0.84]; acceptability Fleiss' $\kappa = 0.76$
Crossover unit	248 clinicians $\times$ 8 trials; four complete-interface and four unassisted trials per clinician; 1,984 analyzed observations
Crossover assignment	Clinician-level balanced AB/BA randomization; period and carryover terms; intention-to-treat population
Crossover error	Mixed logit OR 0.22 [0.14,0.34], $p < 0.001$
Crossover time	Within-clinician paired median difference -4.6 [-4.9,-4.2] minutes, $p < 0.001$
Three-arm contrasts	RiskWeave vs. control OR 0.23 [0.11,0.45], adjusted $p < 0.001$; vs. checklist OR 0.34 [0.16,0.70], adjusted $p = 0.006$
Decision impact	Paired OR 0.26 [0.13,0.48], $p < 0.001$

Table 16: Clinician-task raw counts and uncertainty.

Study	Arm	n	Errors	Rate / uncertainty	Time / note
Crossover	Assisted	992	23	2.3%	5.8 min
Crossover	Unassisted	992	88	8.9%	10.4 min
Three-arm	Control	480	44	9.2%	reference
Three-arm	Checklist	480	31	6.5%	+2.2 min
Three-arm	RiskWeave	480	11	2.3%	-4.6 min
Shadow mode	Monitoring	1,847	35	1.9% [1.3,2.6]; abstain 76/1,847	12 weeks
Decision impact	Before	384	45	11.7% [8.7,15.4]	paired queries
Decision impact	After	384	13	3.4% [1.8,5.7]	paired queries

The crossover and three-arm results agree on the complete assistance package, while the paired study repeats the direction on a distinct decision set. These studies quantify the consequence of the interface; the fixed-pool and ablation experiments provide the retrieval-policy attribution.

A.10 Second-generator evaluation

The second-generator analysis keeps the selected evidence, prompt skeleton, budget, and blinded rubric fixed while replacing GPT-4o. This intervention tests the observed allocation gaps under a second answer model.

Table 17: Results with Llama-3.1-70B-Instruct. Harmful-rate intervals are wide on the 58-query high-risk subset and are reported in full.

Metric	SubmodCov	Flat-CE	RiskWeave
Acceptable ($n = 160$)	122/160 (76.3%)	128/160 (80.0%)	134/160 (83.8%)
FCov@10	0.595 [0.560,0.631]	0.662 [0.627,0.697]	0.706 [0.671,0.741]
Harmful ($n = 58$)	6/58 (10.3%; 3.9–21.2)	4/58 (6.9%; 1.9–16.7)	3/58 (5.2%; 1.1–14.4)

RiskWeave exceeds Flat-CE by 0.044 FCov and 3.8 acceptability points with Llama-3.1-70B-Instruct. Harmful outcomes are directionally aligned, but their 58-query intervals are too wide to carry the generator comparison.

A.11 EVIDENCE-RISK MAP

Table 18 links each protected element to the intervention that exposes its failure and to the most direct reported endpoint. It also separates evidence exposure, allocation, answer support, and clinician decision effects.

Table 18: How each medication-safety evidence risk is isolated and measured.

Protected element	Stress condition and policy response	Strongest control and evidence
High-priority support	Relevant background crowds out a required facet; minimize risk-weighted omission loss.	SubmodCov and Flat-CE; FCov@10 and fixed-pool violations in Eq. 1 and Tables 2–6.
Cited recommendation	The generator asserts a required facet without supporting text; preserve item–passage links through packing.	No-grounding ablation; harmful outcomes on the 248-query cohort in Section 6.5.
Conflict visibility	One side of a material source disagreement is displaced; reserve a qualifying opposing assertion.	No-conflict ablation; claim violations and harmful outcomes in Table 10.
Source quality	Stale or low-confidence evidence outranks authoritative support; score source class, extraction confidence, and recency.	No-provenance ablation; acceptability and support changes in Eq. 2 and Table 9.
Exposure robustness	Candidate construction changes the available facet set; apply the same allocation policy to graph and flat-fusion pools.	Flat-CE within each pool; exposure recall, FCov, violations, and latency in Tables 2–7.
Clinician decision	Time pressure encourages recommendation over-reliance; expose citations, conflicts, and abstention state.	Unassisted and checklist controls; error and time effects in Tables 15–16.

The map makes the evidentiary hierarchy explicit: weighted coverage is tested before generation, claim support after generation, and workflow effects only at the complete human-facing interface.