Modeling and Simulation 建模与仿真, 2026, 15(5), 185-196
Published Online May 2026 in Hans. https://www.hanspub.org/journal/mos
https://doi.org/10.12677/mos.2026.155082

Hans汉斯

面向医疗大模型检索增强生成的循证装箱上下文调度算法

田行健[1]，王俊翔[2]，张　勉[3*]

[1]北京建筑大学理学院，北京
[2]中国科学院计算技术研究所，北京
[3]北京建筑大学智能科学与技术学院，北京

收稿日期：2026年4月27日；录用日期：2026年5月20日；发布日期：2026年5月26日

摘　要

在医疗大模型辅助诊断场景中，检索增强生成(RAG)技术被广泛用于融合多模态患者数据。然而，受限于大语言模型(LLM)的上下文窗口，传统截断或均匀压缩策略易导致关键的高临床价值信息(如过敏史)丢失，引发医疗安全隐患。针对此痛点，本文提出一种基于循证医学(EBM)的装箱调度算法(EBM-Pack)。该算法将运筹学中的多重选择背包问题(MCKP)引入上下文调度，首先通过预分类器为不同模态特征赋予基于诊断客观性的临床权重，并允许高危特征扩展出“全量”与“保真压缩”互斥状态；随后采用带互斥检查的线性贪心算法实现全局效用最大化。实验结果表明，在200道MedQA题目的溢出场景及256 Token极限窗口下，EBM-Pack的高危证据保留率(RRCE)达98.4%，端到端诊断准确率较基线提升18.5个百分点。此外，多LLM泛化实验证实了该算法在异构模型上的稳定性；在高信噪比(N = 200)环境与消融实验中，算法展现出优异的抗噪鲁棒性与参数寻优机制的有效性，为医疗大模型的安全落地提供了可靠方案。

关键词

医疗大模型，检索增强生成(RAG)，上下文调度，多重选择背包问题(MCKP)，循证医学

Evidence-Based Packing Context Scheduling Algorithm for Retrieval-Augmented Generation in Medical Large Language Models

Xingjian Tian[1], Junxiang Wang[2], Mian Zhang[3*]

[*]通讯作者。

文章引用: 田行健, 王俊翔, 张勉. 面向医疗大模型检索增强生成的循证装箱上下文调度算法[J]. 建模与仿真, 2026, 15(5): 185-196. DOI: 10.12677/mos.2026.155082

[1]School of Science, Beijing University of Civil Engineering and Architecture, Beijing
[2]Institute of Computing Technology, Chinese Academy of Sciences, Beijing
[3]School of Intelligent Science and Technology, Beijing University of Civil Engineering and Architecture, Beijing

Received: April 27, 2026; accepted: May 20, 2026; published: May 26, 2026

Abstract

In the context of medical large language model (LLM) assisted diagnosis, Retrieval-Augmented Generation (RAG) is widely employed to integrate multimodal patient data. However, constrained by the limited context window of LLMs, traditional truncation or uniform compression strategies often lead to the loss of critical, high-clinical-value information (such as allergy histories), thereby posing significant medical safety risks. To address this critical issue, this paper proposes an Evidence-Based Medicine (EBM) packing scheduling algorithm, termed EBM-Pack. This algorithm introduces the Multiple Choice Knapsack Problem (MCKP) from operations research into context scheduling. It first utilizes a pre-classifier to assign clinical weights to different modality features based on diagnostic objectivity, allowing high-risk features to expand into mutually exclusive states: "full-text" and "high-fidelity compression". Subsequently, a linear greedy algorithm with mutual exclusion checking is employed to maximize the global clinical utility. Experimental results demonstrate that in an overflow scenario involving 200 MedQA questions with an extreme window limit of 256 tokens, EBM-Pack achieves a Retention Rate of Critical Evidence (RRCE) of 98.4% and improves end-to-end diagnostic accuracy by 18.5 percentage points compared to baseline methods. Furthermore, multi-LLM generalization experiments confirm the algorithm's stability across heterogeneous models. In high noise-to-signal ratio environments (N = 200) and ablation studies, the algorithm exhibits excellent noise robustness and validates the effectiveness of its parameter optimization mechanism, providing a reliable solution for the safe deployment of medical LLMs.

Keywords

Medical Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), Context Scheduling, Multiple Choice Knapsack Problem (MCKP), Evidence-Based Medicine (EBM)

1. 引言

近年来，深度学习技术在医学辅助诊断领域取得了突破性进展，业界逐渐形成了一种全新的智慧医疗 AI 范式：采用“领域大语言模型(LLM) + 检索增强生成(RAG)”的混合架构[1]-[3]。这种架构能够将患者实时的多模态检查结果(如病理、影像分析结果)与历史病历相结合，通过大语言模型生成个体化的诊断建议。在现代医疗大模型辅助诊断系统中，面对患者的个体化问诊，系统往往需要调用底层的多个异构专科深度学习模型(如放射科 CT 影像重建模型、病理切片分析模型、心电图时序分类模型等)。由于这些模型的算力消耗与执行时间差异巨大，系统通常采用基于消息中间件的异步架构。各个专科模型推理完成后，会将结构化的诊断结果异步推入消息队列，等待与患者的长期电子病历进行融合，最终提交给大语言模型进行综合判断。然而，将这种 RAG 架构落地于真实的复杂临床环境时，系统面临着严峻的“上下文窗口溢出”与“中间遗忘效应”挑战。当患者具有复杂的既往病史，或在单次会话中触发了大

量多模态检查，所有异步返回的诊断报告与召回的历史病历汇聚到大模型网关层时，其总 Token 长度常常超过当前主流模型的有效上下文窗口。此时，系统必须做出上下文管理决策。

现有的处理策略大致可归为三类，但在医疗场景下均存在显著的安全缺陷：(1) 截断流派：按数据生成的先后顺序依次装入直到窗口满。这意味着患者十年前致命的“青霉素过敏休克”记录，可能会被刚刚生成的“轻微咳嗽”主诉挤出上下文。该策略完全无视临床价值，是最朴素的工程兜底方案。(2) 语义重排流派[4]：利用预训练嵌入模型计算各片段与查询的语义相似度，按降序装入窗口。该策略虽然引入了“相关性”，但语义相似度并不等价于临床重要性，一段与主诉高度相关的冗余随访记录可能排在具有“一票否决”属性的过敏史之前。(3) 均匀压缩流派[5]：按窗口与总量的比例对所有片段进行等比截断。该策略虽然保留了全局覆盖性，但等比压缩对医疗数据中关键的数值信息(如“肌酐 2.8 mg/dL”、“肿块直径 8 mm × 6 mm”)具有毁灭性影响。上述三类策略的共同缺陷在于，均未将医学证据的客观性等级纳入上下文分配的决策框架[6]-[8]。近期研究表明[9]，引入循证医学(EBM)理念能显著提升大模型诊断的可靠性。不同模态的特征具有绝对的证据等级差异：一份简短的“病理活检阳性”报告，其临床权重远高于一篇长达千字的“患者主诉”。因此，如何在动态、异步的数据流中，感知模型的剩余窗口容量，并以循证医学证据等级为依据主动筛选出“临床价值最高”的特征片段，成为了解决医疗大模型“幻觉”与“漏诊”问题的关键。

为此，本文提出了一种专为医疗大模型 RAG 定制的基于循证医学的装箱调度算法(EBM-Pack)。其中，“EBM”体现了算法以循证医学证据等级作为特征优先级的核心设计理念，“Pack”则对应将异构医疗特征在有限窗口内进行最优装箱的数学建模思路(MCKP)。本文的主要贡献如下：(1) 提出基于循证医学的效用评价模型：扩展了循证医学在文本生成阶段的应用，将不同医疗模态(病理、影像、检验、主诉等)的证据等级量化为绝对临床权重，建立了上下文片段的“临床效用–Token 代价”数学模型。(2) 设计基于多重选择背包问题(MCKP)的调度算法：将异构诊断特征的上下文融合升维为高频实时求解的 MCKP 问题。摒弃了被动的工程截断补救，允许高危特征扩展出“全量”与“保真压缩”等多个互斥状态。(3) 引入状态展平与带互斥检查的线性贪心算法：通过多态展平与全局密度排序，在满足严苛物理窗口约束的同时，实现了医疗特征的“优雅降级”，以极低的时间复杂度确保了核心临床效用的最大化。

2. 基于循证医学的装箱调度算法(EBM-Pack)

2.1. 临床客观性证据权重(EBM Weight)的量化

在 EBM-Pack 算法中，每条待进入 LLM 上下文的医疗文本片段(如一条病史、一份报告)被定义为一个特征块 q_i。为了打破传统的“文本长度”或“时间远近”偏见，本文引入了循证医学(EBM)中“客观证据优于主观描述”的理念。我们构建了一个基于专家规则的预分类器(Rule-based Pre-classifier)，根据医学检验的诊断客观性(Diagnostic Objectivity)，为不同来源的特征块赋予相对的临床证据权重 W_{EBM}。

该分类器无需复杂的机器学习打标，而是基于医疗信息系统(HIS)中标准化的电子病历(EMR)组件类型进行确定性映射(在理想部署中对应 EMR 组件映射；本文实验中以规则对病例句段进行 Tier 标注)。具体而言，我们定义四层偏序(Partial Order)分级 T1 > T2 > T3 > T4：T1 (金标准/确诊级，$W_{EBM}=\alpha$)：病理活检报告、基因测序结果、明确的高危过敏史(如青霉素过敏)。此类特征具有最高客观性和致命风险。T2 (客观检查级，$W_{EBM}=\beta$)：CT/MRI 等放射科影像诊断结论、生化血液检验异常指标、心电图等功能检查结果。T3 (专业观察级，$W_{EBM}=\gamma$)：首诊医生的体格检查记录、既往病历中的门诊体征记录。T4 (主观描述级，$W_{EBM}=1$，基准)：患者口述主诉、未确诊的疑似症状描述。偏序约束要求 $\alpha \geq \beta \geq \gamma \geq 1$，但不预设各层之间的具体数值差距。其中层级标号 T1 > … > T4 表示序关系；数值权重满足 $\alpha \geq \beta \geq \gamma \geq 1$，以允许

在网格搜索中出现 $\alpha=\beta$ 等权重打平情形。最优系数 $\left(\alpha^*,\beta^*,\gamma^*\right)$ 通过在验证集上的网格搜索自动确定，避免了硬编码权重的主观性问题。

2.2. 基于多重选择背包问题(MCKP)的模型升维

传统工程实践中的“软截断”通常表现为一种被动的错误处理(即“溢出后拦截并压缩”)，这种充满嵌套判断的逻辑不仅在系统高并发时容易引发时延抖动，而且无法保证全局最优。为了从根本上解决这一问题，本文将医疗特征的调度过程在数学上严格重构为多重选择背包问题(Multiple Choice Knapsack Problem, MCKP)。

在每次 RAG 会话中，对于候选队列中的每一个原始医疗特征块 q_i，系统不再将其视为单一实体，而是将其映射为一个包含多个互斥状态(Mutually Exclusive States)的集合 Q_i。状态集 S_i 的定义依赖于该特征的客观性权重 $W_{EBM}(i)$：状态 A (全量态)：包含所有修饰词的原始完整文本，记为 $q_{i,A}$。其 Token 开销为 $C_{i,A}$，由于信息最完整，其临床效用 $U_{i,A}$ 为满分状态。状态 B (保真压缩态)：仅对 T1 级与 T2 级的高危客观证据开放。通过轻量级命名实体识别(NER)强制提取核心数值与阴阳性结论并剔除冗余文本后形成，记为 $q_{i,B}$。其体积发生骤减($C_{i,B}\ll C_{i,A}$)。由于是“实体保真”的压缩，其临床效用仅发生微小折损。对于 T3 级与 T4 级的主观描述，不提供该状态。其中当片段不属于 T1/T2 时，$S_i=\{A\}$ (仅全量态)；仅当为 T1/T2 时 $S_i=\{A,B\}$。基于上述多态扩展，特征 q_i 的第 j 种状态的效用 $U_{i,j}$ 定义为：

$$U_{i,j}=\mathcal{F}_{decay}(j)\times W_{EBM}(i)\times Sim\left(q_{query},q_{i,j}\right)\times \mathrm{e}^{-\lambda\Delta t},\ j\in\{A,B\} \tag{1}$$

其中，$\mathcal{F}_{decay}(j)$ 为状态保真折损系数，定义为：

$$\mathcal{F}_{decay}(j)=\begin{cases}1, & j=A\left(\text{全量态}\right)\\ \min\left(1,\dfrac{N_{entity}}{N_{total}}+\epsilon\right), & j=B\left(\text{保真压缩态}\right)\end{cases} \tag{2}$$

N_{entity} 为状态 B 下保留的核心医疗实体(如指标数值、阴阳性结论)数量，N_{total} 为原始文本中对应关键实体的总数，比值 N_{entity}/N_{total} 用作保真度的计数型代理，ϵ 为作用于该代理的标量校准项(亦称平滑常数)，用于缓解粗粒度实体计数与互斥选段排序需求之间的偏差，其取值由验证集敏感性分析确定(本文最优为 0)。Δt 为特征生成距今的物理时间间隔，引入自然指数衰减以强化病历时效性，其中 $\lambda>0$ 为时间衰减强度系数，控制远期证据在总效用中的削弱程度(与 ϵ 一同由实验确定)。q_{query} 代表当前患者的问诊查询或系统的主 Prompt，$Sim\left(q_{query},q_{i,j}\right)$ 为查询向量与特征片段向量之间的语义相似度函数。本文采用预训练的 Sentence-Transformers 模型计算两者的余弦相似度，以确保与主诉高度相关的特征能够获得基础的效用加成。

由此，LLM 上下文的融合过程被优雅地转化为一个标准的多态约束优化问题：

$$\max\sum_{i=1}^{n}\sum_{j\in S_i}x_{i,j}\cdot U_{i,j} \tag{3}$$

受到以下三个核心约束的严格限制：

$$\text{s.t.}\sum_{i=1}^{n}\sum_{j\in S_i}x_{i,j}\cdot C_{i,j}\le W_{\max}\quad\text{(窗口容量约束)} \tag{4}$$

$$\sum_{j\in S_i}x_{i,j}\le 1,\forall i\in\{1,\cdots,n\}\quad\text{(互斥选择约束)} \tag{5}$$

$$\forall t\in T_{patient},\ \sum_{\{i\mid q_i\in t\}}\sum_{j\in S_i}x_{i,j}\ge\min\left(k_{\min},N_t\right)\quad\text{(最低层级表征约束)} \tag{6}$$

上述公式表明：对于同一个患者的同一份病理报告，算法最多只能选择将其“全量态”或“压缩态”其一装入窗口(或完全丢弃)，绝不能同时装入。此外，为了防止纯密度贪心策略在极端情况下导致某些低密度但关键的循证层级(如包含过敏史的 T1 层级)被完全忽略，公式(6)引入了最低层级表征约束(Minimum Tier Representation Constraint)，记为 $k_{\min}$。其中，设当前患者所具备的循证医学层级集合为 $T_{patient}$；对于任意存在的层级 $t \in T_{patient}$，设候选队列中属于该层级的特征总数为 N_t。该约束相当于为不同层级的医学证据设定了“保底名额”，即保证对于每一个存在的层级，最终装入上下文的特征块数量不得少于 $k_{\min}$ (若该层级原始特征数不足 $k_{\min}$，则需全部保留)。在满足上下文窗口限制的前提下，此机制强制保障了大模型输入信息的跨层级多样性，从而守住诊断安全的底线。

2.3. 状态展平与带互斥检查的线性贪心算法

由于严格求解 MCKP 具有 NP-Hard 的时间复杂度，本文设计了一种带有互斥检查机制的线性启发式贪心算法。该算法摒弃了传统业务代码中“装不下再尝试压缩”的嵌套逻辑，通过“状态预展平”实现了极简的线性流水线处理。第一步：多态展平与密度计算(Flatting & Scoring)。系统首先计算所有特征块可用状态的“性价比密度”(Utility-Cost Ratio)：

$$\rho_{i,j} = U_{i,j} / C_{i,j}$$

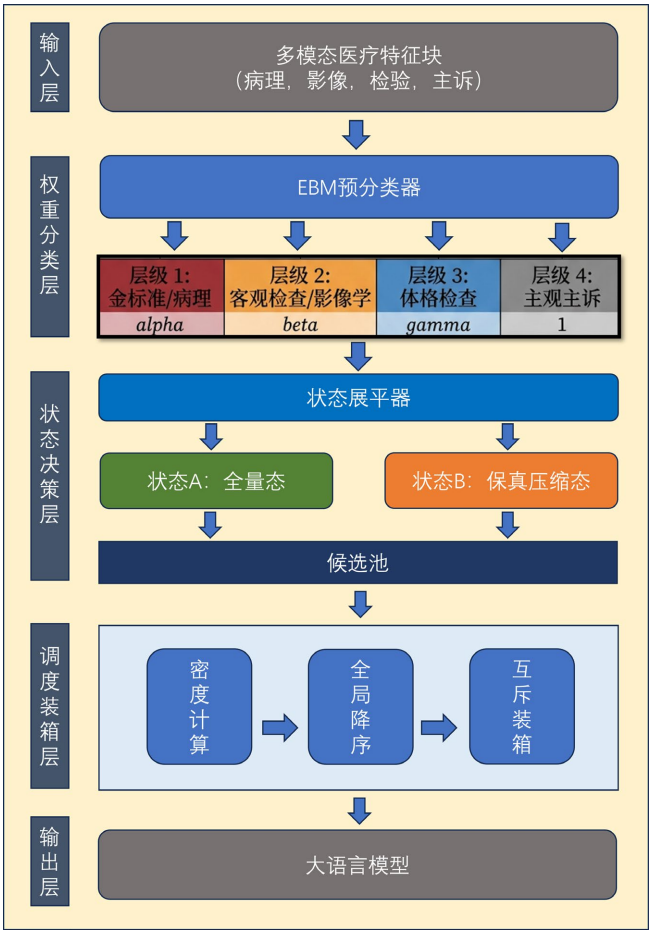

Figure 1. Flowchart of the EBM-Pack scheduling algorithm based on the classification of evidence-based medical evidence
图 1. 基于循证医学证据分级的 EBM-Pack 调度算法流程图

由于压缩态(B)的体积$C_{i,B}$极小且保留了核心效用，其密度$\rho_{i,B}$往往极高。系统将所有合法的状态项$\{q_{i,j}\}$放入一个全局统一的候选池中。第二步：全局降序(Global Sorting)。在$O(M\log M)$的时间复杂度下(M为展平后的总状态数)，将候选池中的所有项按$\rho_{i,j}$进行降序排列。第三步：互斥装箱(Mutually Exclusive Packing)。系统从队首开始进行线性遍历。在正式按密度装箱前，算法会优先扫描候选池，为每个存在的层级(Tier)强制挑出密度最高的$k_{\min}$个特征装入上下文，以满足保底配额。完成保底分配后，对于剩余的项$q_{i,j}$，算法继续执行两次轻量级检查：(1) 互斥检查：查询特征索引i是否已被标记为“已装入”。若已装入，则直接跳过该项(例如，如果某病理报告的压缩态因密度极高已被提前装入，后续即便遇到其全量态也会被优雅拦截)。(2) 容量检查：判断当前剩余窗口$W_{\max}$是否大于$C_{i,j}$。

若同时满足上述两点，则将$q_{i,j}$装入上下文批次，扣减$W_{\max}$，并标记特征i为已处理。

这一设计将复杂的医疗特征“降级”逻辑(Graceful Degradation)隐藏在客观的密度排序与互斥验证中，使得整个调度过程极具工业级并发处理美感。EBM-Pack 的调度流程图如图 1 所示。

3. 实验设计与分析

3.1. 实验环境与数据集构建

数据集。本文基于公开的 MedQA-USMLE [10]数据集构建实验。MedQA 包含来自美国医师执照考试(USMLE)的四选一多选题，每道题附带完整的临床病例描述(clinical vignette)和标准答案。我们从测试集中筛选出包含至少两种模态(如实验室检查 + 影像)且具有高客观性证据(Tier 1 或 Tier 2)的题目，随机抽取 200 道作为实验样本。

多模态溢出场景构造。为模拟真实临床场景中电子病历的信息冗余与跨科室数据混杂，对于每道题目，我们将其临床描述解析为多个模态片段(主诉、体检、检验、影像等)，并按照 1:12 的信噪比注入异构噪声片段。噪声包含三种类型：(1) 40%为医疗系统常见的模板化冗余文本(如“患者遵医嘱按时服药”等无诊断价值的随访记录)；(2) 30%为其他病例的断句碎片；(3) 30%为跨科室病例的交叉拼接文本(模拟多科室会诊数据混杂)。最终每道题的候选上下文总量为原始病例信号量的 10 至 15 倍，确保在受限窗口下产生显著溢出。

嵌入模型。语义相似度计算采用 Sentence-Transformers 框架的 paraphrase-multilingual-MiniLM-L12-v2 模型(384 维)，该模型支持多语言且在医学文本检索任务中表现稳健。

推理模型。端到端准确率测试统一采用 DeepSeek-V3.2 (deepseek-chat)作为基座大语言模型，temperature 设为 0 以确保可复现性。

基线算法。本文选取以下三种代表性基线进行对比：(1) FIFO 截断：按特征在候选队列中的原始顺序，从前向后依次装入直到窗口满。作为传统工程中最朴素的空白对照基线。(2) 语义重排(Semantic Re-ranking)：代表现代 RAG 主流的重排流派。利用嵌入模型计算各片段与问题的余弦相似度，按相似度降序装箱。(3) 均匀压缩(Uniform Compression)：代表提示词压缩流派。按窗口与总量的比例对所有片段进行等比例截断，保留每段的前r比例词汇。

EBM-Pack 权重寻优。本文采用偏序约束下的网格搜索策略确定各模态层的权重系数。定义四层偏序 T1 > T2 > T3 > T4，参数化权重$\theta=(\alpha,\beta,\gamma,1)$，搜索空间为$\alpha\in\{1.5,2,2.5,3\}$，$\beta\in\{1.2,1.5,2,2.5\}$，$\gamma\in\{1,1.2,1.5\}$。共产生 35 组满足偏序约束$\alpha\geq\beta\geq\gamma\geq1$的有效组合，以 RRCE 为指标在验证集上筛选出最优权重配置$\theta^{*}=(2.5,2.5,1.0,1.0)$。

3.2. 不同窗口限制下的高危证据保留率

为验证 EBM-Pack 在极端窗口限制下的关键证据保护能力，实验设定$W_{max}\in\{256,512,1024,2048\}$

Token。评价指标为高危证据保留率(Retention Rate of Critical Evidence, RRCE)，定义为 Tier 1 或 Tier 2 的信号片段在策略输出中未被丢弃的比例。实验结果如表 1 所示。

Table 1. The retention rate of high-risk evidence under different window restrictions (RRCE, %)
表 1. 不同窗口限制下的高危证据保留率(RRCE, %)

算法策略	$W_{max}=2048$	$W_{max}=1024$	$W_{max}=512$	$W_{max}=256$
FIFO 截断	100.0	63.5	28.6	14.1
语义重排	100.0	61.2	41.1	25.7
均匀压缩	100.0	100.0	98.4	65.5
EBM-Pack	**100.0**	**100.0**	**99.0**	**98.4**

在 2048 Token 的充裕窗口下，各算法均能完整保留高危证据。然而随着窗口收紧，差异急剧拉大：在 256 Token 的极限窗口下，FIFO 截断的 RRCE 跌至 14.1%——意味着超过 85%的关键临床证据被丢弃；语义重排仅为 25.7%，因为大量语义相关度低但具有致命属性的边缘病史(如过敏史、基因突变)被长篇主诉挤出窗口；均匀压缩通过全面缩减将 RRCE 提升至 65.5%，但仍有近三分之一的高危证据因过度压缩而丢失关键数值。相比之下，EBM-Pack 通过 MCKP 多态装箱与最低层级表征约束，在 256 Token 极限下仍保持 98.4%的高危证据保留率，如图 2 所示。

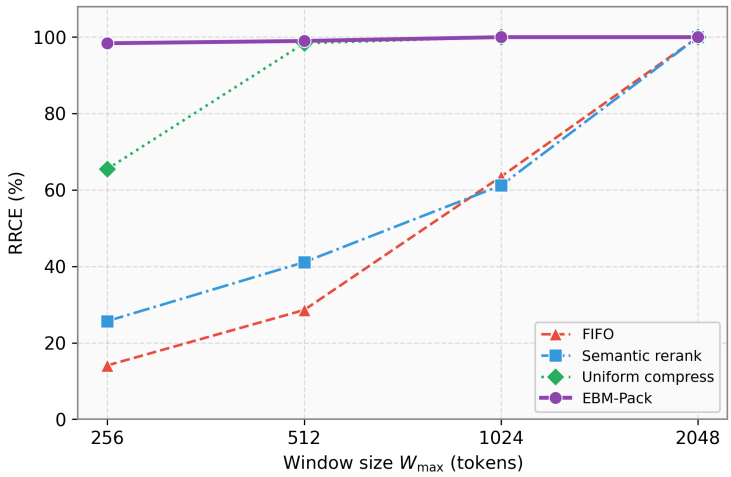

Figure 2. Comparison trend of RRCE under different windows
图 2. 不同窗口下 RRCE 对比趋势

3.3. 端到端诊断准确率与准确率保持率

仅保留证据并不意味着最终诊断正确。为验证 RRCE 的提升是否真正转化为诊断质量的提升，本文进行了端到端的准确率测试。同时，为了消除不同题目难度差异对绝对准确率的影响，本文提出准确率保持率(Accuracy Preservation Ratio, APR)作为上下文管理策略的鲁棒性评价指标：

$$\mathrm{APR}=\frac{\mathrm{Accuracy}_{W_{max}}}{\mathrm{Accuracy}_{oracle}}\times 100\%$$

其中 $\mathrm{Accuracy}_{oracle}$ 为无噪声干扰下 LLM 对原始题目的基准准确率(本实验中为 80.0%)。APR 衡量的是各策略在上下文溢出后“保持诊断能力”的比例。

Table 2. End-to-end diagnostic accuracy rate (Accuracy, %)
表 2. 端到端诊断准确率(Accuracy, %)

算法策略	$W_{max}=2048$	$W_{max}=1024$	$W_{max}=512$	$W_{max}=256$
FIFO 截断	61.0	55.0	46.0	49.2
语义重排	60.0	52.5	47.0	49.5
均匀压缩	62.0	52.0	45.5	43.5
EBM-Pack	**65.0**	**61.0**	**61.5**	**62.0**

如表 2 和图 3 所示，EBM-Pack 在所有窗口下均取得最高的准确率和 APR。尤其值得注意的是：(1) EBM-Pack 在小窗口下准确率保持稳定。在 256 Token 下 EBM-Pack 取得 62.0%的准确率(APR = 77.5%)，而其在 2048 Token 下为 65.0%，仅下降 3 个百分点。相比之下，均匀压缩从 62.0%跌至 43.5% (下降 18.5 个百分点)，退化幅度远大于 EBM-Pack。这说明 EBM-Pack 的权重–密度排序机制具有显著的“窗口鲁棒性”——即便窗口大幅缩减，其核心诊断能力依然稳健。(2) 基线算法在小窗口下急剧退化。均匀压缩在 256 Token 下 APR 仅为 54.4%，意味着近半数诊断能力被上下文溢出所吞噬。FIFO 虽然表面绝对值不低，但其保留的内容高度依赖随机排列顺序，可复现性差。(3) EBM-Pack 在小窗口下反超最强基线 12.8 个百分点。在 256 Token 下，EBM-Pack (62.0%)较语义重排(49.5%)提升 12.5 个百分点，较均匀压缩(43.5%)提升 18.5 个百分点。图 3 直观展示了各策略的 APR 随窗口缩小的衰减趋势。

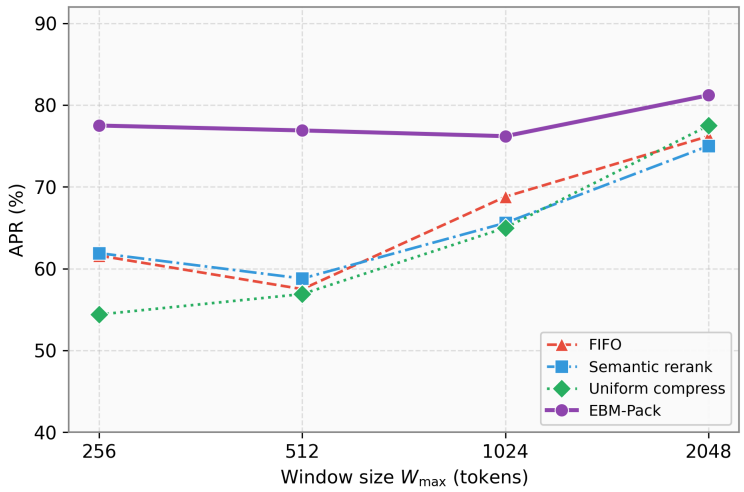

Figure 3. Comparison of Accuracy Preservation Rate (APR) under different windows
图 3. 不同窗口下准确率保持率(APR)对比

3.4. 多 LLM 泛化性验证与噪声鲁棒性分析

为验证 EBM-Pack 的有效性不依赖于特定的大语言模型，本文在三种不同架构、不同厂商的 LLM 上进行了对比实验($n=50$)：DeepSeek-V3.2、qwen3-next-80b-a3b-instruct 和 doubao-seed-2-0-lite-260215。实验仅对比 FIFO (最强传统基线)与 EBM-Pack，结果如图 4 所示。

如图 4 所示，三种 LLM 上 EBM-Pack 均在所有窗口下全面超越 FIFO 基线。尤其是 Qwen3-80B 在 256 Token 下 EBM-Pack (64.0%)较 FIFO (48.0%)提升 16 个百分点，Doubao-Seed 在 512 Token 下 EBM-Pack (76.0%)较 FIFO (56.0%)提升 20 个百分点。这表明 EBM-Pack 的优势来源于上下文质量的结构性提

升，而非对特定 LLM 推理偏好的过拟合。

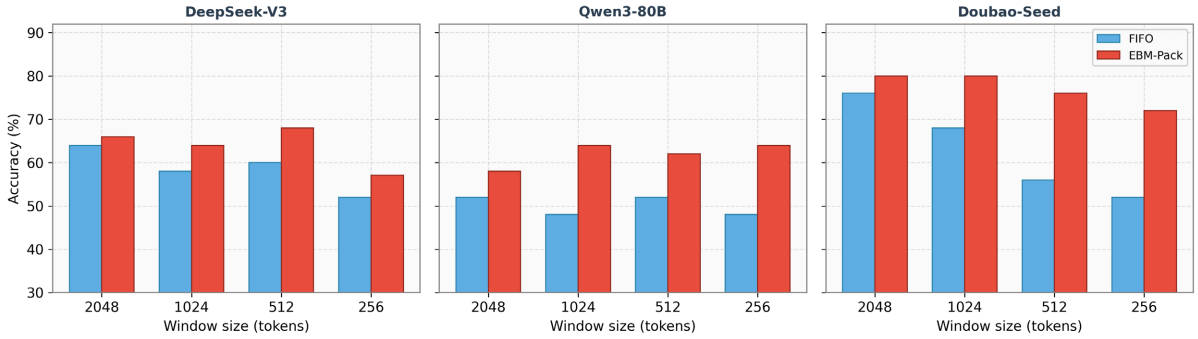

Figure 4. Comparison of generalization abilities of multiple LLMs—EBM-Pack vs FIFO in different window sizes
图 4. 多 LLM 泛化性对比——EBM-Pack vs FIFO 在不同窗口下的准确率

为评估 EBM-Pack 在不同信息冗余程度下的抗噪能力，固定 $W_{\max}=512$，变化每道题的噪声注入量 $N_{noise}\in\{20,40,80,120,200\}$，对比 FIFO 与 EBM-Pack 的表现。结果如表 3 和图 5 所示。

Table 3. Comparison of FIFO and EBM-Pack under different noise levels ($W_{\max}=512$)
表 3. 不同噪声量下 FIFO 与 EBM-Pack 对比($W_{\max}=512$)

N_{noise}	*FIFO-RRCE* (%)	*FIFO-Acc* (%)	*EBM-Pack-RRCE* (%)	*EBM-Pack-Acc* (%)
20	100.0	68.0	100.0	66.0
40	70.8	62.0	98.6	62.0
80	41.7	60.0	98.6	56.0
120	36.1	52.0	97.2	54.0
200	**20.8**	**50.0**	**97.2**	**58.0**

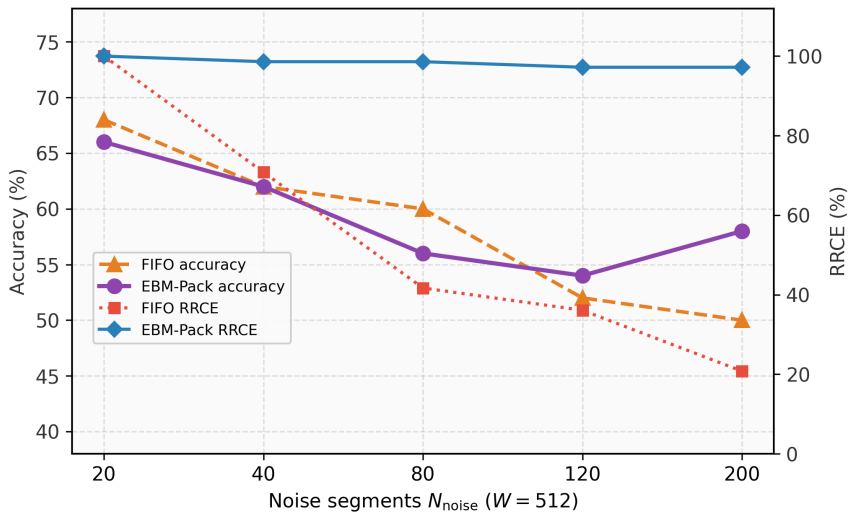

Figure 5. Comparison of FIFO and EBM-Pack under different noise levels: The impact of noise levels on diagnostic accuracy and RRCE ($W_{\max}=512$)
图 5. 噪声量对诊断准确率与 RRCE 的影响($W_{\max}=512$)

在低噪声 $N_{noise}=20$ 时，窗口充裕，两者表现接近。但随着噪声急剧增长：FIFO 的 RRCE 从 100%线性衰减至 20.8%，准确率从 68.0%下滑至 50.0% (接近随机猜测)；而 EBM-Pack 的 RRCE 始终保持在 97%以上，准确率在 $N_{noise}=200$ 时反而回升至 58.0%，展现出显著的抗噪鲁棒性。EBM-Pack 在高噪声下的“准确率反弹”现象可归因于其密度排序机制在噪声越多时越能有效过滤低效用片段，体现了“越乱越强”的特性。

3.5. 消融实验与参数敏感性分析

为验证最低层级表征约束 $k_{\min}$ 的有效性，固定 $W_{\max}=512$，在 EBM-Pack 中分别设置 $k_{\min}\in\{0,1,2,3\}$，结果如表 4 所示。

Table 4. The impact of $k_{\min}$ on the performance of EBM-Pack ($W_{\max}=512$)
表 4. $k_{\min}$ 对 EBM-Pack 性能的影响 ($W_{\max}=512$)

$k_{\min}$	*RRCE* (%)	*Accuracy* (%)
0	95.7	36.0
1	97.1	52.0
2	100.0	52.0
3	100.0	54.0

当 $k_{\min}=0$ 时，准确率骤降至 36.0%，说明纯密度贪心在特定场景下可能完全忽略某些低密度但关键的循证层级(如包含患者口述过敏史的 T1 层级)，导致 LLM 做出错误推断。引入 $k_{\min}=1$ 后准确率提升 16 个百分点，验证了最低层级表征约束的必要性。继续增大 $k_{\min}$ 至 2 或 3 时，RRCE 略有提升但准确率趋于饱和，表明 $k_{\min}=1$ 是效率与效果的最佳平衡点。

为验证偏序约束下的参数寻优机制的有效性，本文对 EBM-Pack 的层级权重配置进行了消融实验。在 35 组满足偏序约束 $\alpha\geq\beta\geq\gamma\geq1$ 的有效参数组合中，我们比较了三类权重配置下的性能差异，见表 5：

Table 5. Weight configuration ablation experiment ($W_{\max}=512$)
表 5. 权重配置消融实验($W_{\max}=512$)

权重配置	$(\alpha,\beta,\gamma,1)$	*RRCE* (%)	*Accuracy* (%)
Uniform	(1, 1, 1, 1)	41.1	47.0
Hand-picked	(5, 4, 2, 1)	99.0	46.0
Optimized	**(2.5, 2.5, 1, 1)**	**99.0**	**61.5**

结果表明：(1) Uniform 配置退化为纯语义重排，RRCE 大幅下降至 41.1%，验证了偏序权重机制的必要性；(2) Hand-picked 虽然 RRCE 与 Optimized 持平，但 Accuracy 显著偏低(46.0% vs 61.5%)，说明过大的层级差距(5:1)导致低权重层的有用信息被过度丢弃；(3) 数据驱动的扁平化寻优在保持高 RRCE 的同时实现了最高准确率，证明了偏序约束下超参数自动优化策略的优越性。

为验证效用函数中保真度校准系数 ϵ 与时间衰减强度 λ 对上下文排序及端到端诊断的有效性，在固定分层权重 $\theta^{*}=(2.5,2.5,1.0,1.0)$、窗口 $W_{\max}=512$、噪声协议与主实验一致的 MedQA 子集上，对 ϵ 与 λ 进行 5 × 6 全因子网格搜索：$\epsilon\in\{0,0.02,0.05,0.1,0.2\}$，$\lambda\in\{0,0.1,0.5,1,2,5\}$。评价指标为端到端准确率(Accuracy)与高危证据保留率 RRCE；最优格点取准确率最大、并以 RRCE 为高指标平局时的次要准则。

图 6 给出 $\epsilon-\lambda$ 平面上的 Accuracy 与 RRCE 热力图。可见 λ 对两指标均具主效应：当 λ 较小时，RRCE 未饱和，Accuracy 亦处于约 60%~65.5%的低位；当 λ 增大至约 0.5 及以上时，RRCE 达到并维持 100%，Accuracy 则随 λ 继续上升至约 73%~76.5%区间并形成平台。相较之下，ϵ 在所扫区间内对 Accuracy 与 RRCE 的扰动极小：对任一固定 λ，沿 ϵ 方向热力图近似呈竖向条带，列内数值变化通常不超过约 1 个百分点，表明在当前实体计数型保真度代理与装箱协议下，压缩态校准项对排序结果的边际影响有限。

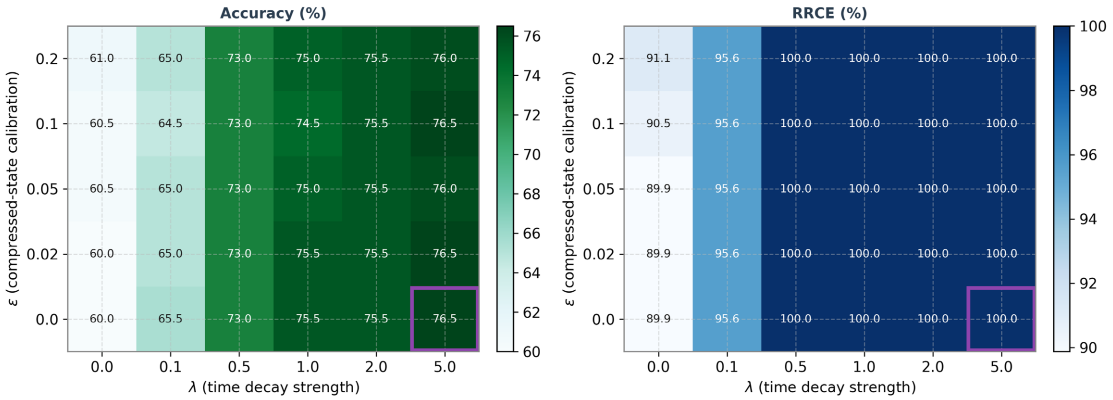

Figure 6. Heatmaps of end-to-end accuracy (left) and RRCE (right) as functions of ϵ and λ

图 6. ϵ 与 λ 对端到端准确率(左)及高危证据保留率 RRCE (右)的敏感性热力图

图 7 给出过最优点 ϵ^* 与 λ^* 的一维切片以对照上述结论。固定 $\lambda=\lambda^*$ 扫描 ϵ 时，Accuracy 仅在约 76.0%~76.5%间小幅波动，RRCE 恒为 100%；固定 $\epsilon=\epsilon^*$ 扫描 λ 时，Accuracy 自 $\lambda=0$ 的约 60.0%单调抬升至 $\lambda=5.0$ 的 76.5%，RRCE 则在 $\lambda\approx0.5$ 后饱和至 100%。网格搜索得到 $\epsilon^*=0.0$、$\lambda^*=5.0$：(Accuracy = 76.5%，RRCE = 100%)，其中 $\epsilon^*=0.0$ 表示无需通过额外抬高 $F_{decay}(B)$ 来改善下游表现；λ^* 取网格最大档，结合 λ 自 2.0 增至 5.0 时准确率增幅已明显收窄，可解释为时间衰减权重在所设搜索区间内已接近收敛。

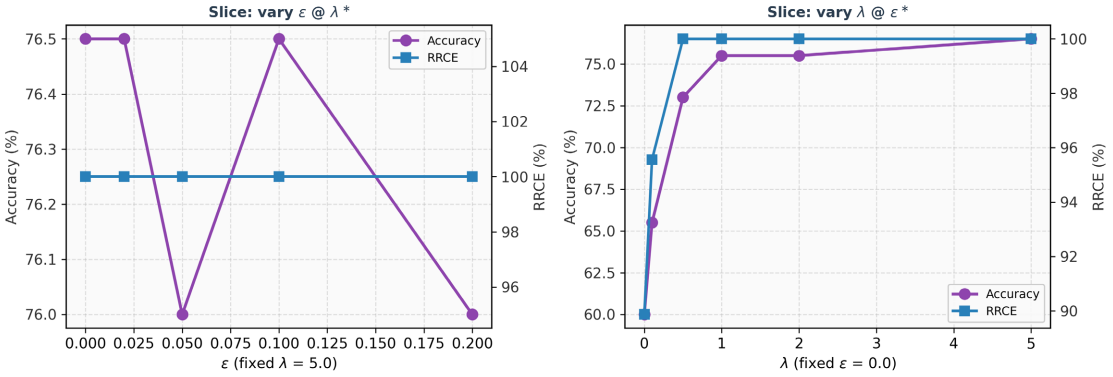

Figure 7. One-dimensional sensitivity slices through (ϵ^*,λ^*) $(W_{\max}=512)$

图 7. 过最优点(ϵ^*,λ^*)的一维敏感性切片($W_{\max}=512$)

4. 结论

本文针对医疗大模型 RAG 场景中的“上下文窗口溢出”难题，创新性地提出 EBM-Pack 调度算法。该算法突破了传统文本压缩的思维局限，将循证医学客观性等级与 MCKP 装箱模型深度结合，通过多态

展平和线性贪心调度，实现了严苛容量约束下的特征优雅降级。全面的对比实验与消融分析表明：(1) 在极限窗口(256 Token)下，EBM-Pack 能以 98.4%的高保留率死守核心医疗证据，显著优于传统截断与重排策略；(2) 在保障证据的基础上，其端到端诊断准确率及准确率保持率(APR)均实现大幅领跑，并在 DeepSeek、Qwen、Doubao 等多款主流大模型上表现出优秀的泛化能力；(3) 面对海量冗余文本的强噪声干扰，密度排序机制赋予了算法极强的抗噪鲁棒性；最低层级表征约束与偏序权重寻优机制则进一步锁定了诊断安全下限。未来的工作将探索结合强化学习的个性化动态权重分配，以更好地应对复杂的多学科会诊场景。

参考文献

[1] Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., *et al*. (2017) Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks. *Nature*, **542**, 115-118. https://doi.org/10.1038/nature21056

[2] Lewis, P., Perez, E., Piktus, A., *et al*. (2020) Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information Processing Systems*, **33**, 9459-9474.

[3] Xiong, G., Jin, Q., Lu, Z. and Zhang, A. (2024) Benchmarking Retrieval-Augmented Generation for Medicine. *Findings of the Association for Computational Linguistics ACL* 2024, Bangkok, 11-16 August 2024, 6233-6251. https://doi.org/10.18653/v1/2024.findings-acl.372

[4] Yue, Z., Wang, Y., Chen, Y., *et al*. (2024) RankRAG: Unifying Context Ranking with Retrieval-Augmented Generation in LLMs. *Advances in Neural Information Processing Systems*, **37**, 121156-121184. https://doi.org/10.52202/079017-3850

[5] Xu, F., Shi, W. and Choi, E. (2024) RECOMP: Improving Retrieval-Augmented LMs with Compression and Selective Augmentation. *The Twelfth International Conference on Learning Representations* (*ICLR* 2024), Vienna, 7-11 May 2024. https://openreview.net/forum?id=mlJLVigNHp

[6] Peng, C., Wang, B., Long, Z. and Sheng, J. (2025) AdaGReS: Adaptive Greedy Context Selection via Redundancy-Aware Scoring for Token-Budgeted RAG. arXiv:2512.25052.

[7] Cao, L., Chen, Q. and Guo, Y. (2026) EHR-RAG: Bridging Long-Horizon Structured Electronic Health Records and Large Language Models via Enhanced Retrieval-Augmented Generation. arXiv:2601.21340.

[8] Nath, S., *et al*. (2025) Less Context, Same Performance: A RAG Framework for Resource-Efficient LLM-Based Clinical NLP. arXiv:2505.20320.

[9] Sun, M., Zhao, S., Chen, J., Wang, H. and Qin, B. (2025) META-RAG: Meta-Analysis-Inspired Evidence-Re-Ranking Method for Retrieval-Augmented Generation in Evidence-Based Medicine. arXiv:2510.24003.

[10] Jin, D., Pan, E., Oufattole, N., *et al*. (2020) What Disease Does This Patient Have? A Large-Scale Open Domain Question Answering Dataset from Medical Exams. arXiv:2009.13081.