Medical Data Curation: Document-Level Refinement and Multi-Style Reconstruction for Language Model Pre-training

Anonymous ACL submission

Abstract

The effectiveness of medical large language models (LLMs) fundamentally depends on their training data, yet the medical AI community faces a severe scarcity of publicly available, high-quality, large-scale, and diverse pre-training corpora. To address this gap for medical *text*—the modality that underpins a model's medical knowledge and reasoning, where the open-data scale–quality gap remains particularly pronounced—we present a comprehensive curation study centred on a progressive pipeline that systematically transforms raw heterogeneous sources—textbooks, peer-reviewed biomedical literature, and curated medical web content—into a tiered corpus of $\sim$**456B tokens**. The pipeline pairs **document-level LLM refinement** with **style-conditioned reconstruction** into patient-centered and preventive-medicine variants. Beyond corpus construction, we conduct extensive empirical validation across 3B–32B parameter models and nine medical benchmarks. Our cleaned corpus produces consistent positive transfer ($+1.9$ **pp over the untrained base,** $+4.1\%$ **relative**), whereas an equally-trained medical-filtered web baseline (Web-Medical) regresses below the base. Our pipeline delivers $+2.8$ / $+1.7$ / $+1.0$ avg pp over Web-Medical at 3B / 7B / 32B, and at 32B **surpasses MedGemma-27B-text-it** (DeepMind, 2025) (Google's state-of-the-art open medical LLM) **on 8 of 9 benchmarks** ($+$**6.2 avg pp**).

1 Introduction

The rapid evolution of large language models is unlocking transformative potential for medical artificial intelligence. The medical AI community has made significant progress in developing specialized datasets (Community, 2024; Touchent et al., 2025; Xu et al., 2024) and evaluation frameworks (Wang et al., 2024b). However, current medical pre-training data suffer from three critical deficiencies.

First, existing open-source medical corpora suffer from a coupled *scale–quality gap*: smaller curated datasets (e.g., Meditron's 47B, PMC-LLaMA's 80B) are too narrow to support modern training budgets, while the few large-scale alternatives consist of unprocessed raw text containing substantial noise, formatting artifacts, OCR errors, and non-educational content that can actively harm training. Naive domain filtering of general web data does not close this gap—we show in Section 4.2.1 that medical-filtered Common Crawl actually *regresses* below the untrained base.

Second, the field lacks a systematic, replicable methodology for processing heterogeneous medical data sources. Most prior work (Chen et al., 2023; Wu et al., 2023; Li et al., 2025) employs ad-hoc filtering rules tailored to specific data sources, while more recent classifier-based approaches (Touchent et al., 2025; Wang et al., 2025) operate at the paragraph or QA-pair level—neither addresses the full spectrum of quality challenges, from OCR errors in scanned medical textbooks to semantic noise in web-crawled clinical content.

Third, the open medical LLM landscape lacks a strong continual pre-training (CPT) stage base built on a modern foundation model: legacy medical CPT efforts (e.g., Meditron-7B) now trail even untrained current-generation bases, while the strongest open medical models (e.g., MedGemma-27B-text-it) are released as fully post-trained instruct checkpoints, leaving downstream practitioners without a clean CPT-only foundation that can be cleanly specialized via further supervised fine-tuning (SFT) or alignment.

To address these challenges, we present a comprehensive study on large-scale medical text data curation and utilization, centred on a progressive pipeline that transforms raw heterogeneous sources into a tiered medical corpus of $\sim$456B tokens (Section 3.4). The scope of this study is deliberately confined to *medical text*—published med-

ical textbooks, peer-reviewed biomedical papers, and curated medical web pages—and explicitly excludes modalities such as medical imaging, biosignals, EHRs, clinical notes, and patient–physician dialogues, which are governed by separate privacy/licensing regimes. Our pipeline embodies a "refinement-rewriting" philosophy that distinguishes rule-based efficiency from semantic intelligence while balancing scale, quality, and educational diversity. Beyond corpus construction, we systematically investigate (1) optimal intra-medical mixture across textbooks/papers/web, (2) medical-to-general mixture from 0% to 100% medical, and (3) the impact of stylistic diversity, identifying a 4:1 medical-to-general ratio and an 80/10/10 origin/patient/preventive stylistic mixture as best operating points.

Contributions. **(1) Resource:** a tiered medical pre-training corpus totalling ∼456B tokens—a ∼210B fully-open subset and a ∼245B book-derived subset released via gated access. The open subset alone is 2–4× larger than prior open medical-CPT corpora (Meditron's 47B; PMC-LLaMA's 80B). **(2) Methodology:** the first end-to-end medical-CPT data pipeline combining document-level semantic refinement with style-conditioned reconstruction, with domain-specific safeguards: a preserve-when-uncertain principle, OCR-aware correction, and chunk-boundary truncation that prevents hallucinated continuation. **(3) Strong open medical CPT models, ready for downstream specialization:** against a strong medical-filtered web baseline (Web-Medical), our pipeline produces consistent positive transfer at every scale; at 32B a CPT-only model surpasses MedGemma-27B-text-it (DeepMind, 2025)—Google's state-of-the-art open medical LLM trained with full SFT and alignment—on 8 of 9 medical benchmarks. We release CPT-only checkpoints at three scales (LLaMA-3.2-3B, Qwen2.5-7B, and Qwen2.5-32B), serving as a clean foundation for downstream supervised fine-tuning, instruction tuning, or alignment, complementary to existing post-training recipes.

2 Related Work

2.1 Medical Large Language Models

Several lines of work apply CPT or SFT to adapt general-purpose LLMs to medicine. The earliest open effort, MEDITRON (Chen et al., 2023), applied biomedical CPT on top of LLaMA-2 with heuristic source-authority scoring. PMC-LLaMA (Wu et al., 2023) extended this idea by mixing 4.8M biomedical papers with 30k textbooks. Me-LLaMA (Xie et al., 2025) carefully calibrated the ratio of biomedical, clinical, and general data over a 129B-token corpus. The current state-of-the-art among open medical LLMs is MedGemma-27B-text-it (DeepMind, 2025), which couples a strong modern base (Gemma 3) with a full SFT-plus-alignment recipe; the Med-PaLM line (Singhal et al., 2023) similarly emphasises post-training. Comprehensive surveys (Liu et al., 2024; Zhang et al., 2024; Luo et al., 2024) catalogue further developments. These models all share a common feature: their improvements over the base model are primarily attributed to either (i) a larger or more curated medical corpus or (ii) a stronger post-training recipe. Absent from this body of work is a systematic study of *how* the medical corpus itself is cleaned and diversified at the document level.

2.2 Medical Corpus Construction

Existing methods for constructing medical pre-training corpora cluster into three families.

Rule-based filtering and heuristic scoring. Most pipelines begin with deterministic rules—length thresholds, MinHash deduplication, language detection, and special-character ratios. MEDITRON additionally up-samples high-quality sources via heuristic scoring on MeSH tags and citation counts (Chen et al., 2023), and Lingshu (Li et al., 2025) combines perceptual hashing with min-hash dedup. Such rules are scalable but cannot distinguish, e.g., an authoritative review paper containing legalese paragraphs from a low-quality blog post free of obvious artefacts.

Classifier- and LLM-based quality filtering. A more recent line lifts quality control from rules to learned classifiers. Biomed-Enriched (Touchent et al., 2025) prompts Llama-3.1-70B-Instruct to annotate paragraphs along medical type, domain, and educational quality, distilling labels into a fast classifier. Huatuo-Lite (Wang et al., 2025) demonstrates a parallel approach for QA corpora. These methods retrieve more meaningful quality signals than rules but operate at the *paragraph* or *QA-pair* level, breaking document coherence and discarding structural context.

Source-conditioned dataset releases. Releases like PubMed/PMC (Krithara et al., 2023), MultiMedQA (Singhal et al., 2023), and MedicalMeadow (Han et al., 2023) provide curated

medical text or QA collections, while MIMIC-III/IV (Johnson et al., 2016, 2023) cover privacy-restricted clinical text. These resources are largely orthogonal to processing methodology. Our work uses Nemotron-CC v1 (Su et al., 2024) as the upstream web source.

In contrast, our pipeline operates at the *document* level via two LLM-driven stages—semantic refinement (with explicit OCR-error correction and asymmetric preserve-when-uncertain prompting) and style-conditioned reconstruction. Individual ingredients have antecedents in general-domain corpus engineering, but their combination tailored to medical literature, with explicit safeguards for clinical accuracy, structural integrity, and chunk-boundary truncation, has not been studied as a complete pipeline.

Pre-training data mixing. An adjacent line studies how to mix domain-specific and general data during CPT without inducing catastrophic forgetting (Wu et al., 2024; Pillai, 2025; Shi et al., 2024; Cossu et al., 2022); mixture-optimisation methods (Xie et al., 2023; Ye et al., 2025a,b; Ding et al., 2024) learn or search for ratios that minimise validation loss. We complement this line with a systematic empirical sweep over nine medical-to-general ratios (Section 5.3), identifying a 4:1 sweet spot.

3 Data Processing Pipeline

We operate on raw heterogeneous medical sources (textbooks, research papers, web content) that have undergone two standard preliminary stages (Qin et al., 2026): **format normalization** (OCR for scanned PDFs, HTML parsing, audio/video transcription) converting mixed formats into uniform text while preserving structural information; and **rule-based filtering** (size filtering, MinHash near-duplicate removal, garbled-text rejection, language filtering, special-character/repetition ratio filtering). Building on this preprocessed input, our contribution introduces two novel stages that focus on semantic refinement and educational diversification.

3.1 LLM-based Content Refinement and OCR Correction

Medical academic documents contain dense technical terminology, complex notation, and pedagogical elements that must be preserved with absolute accuracy, while being interspersed with publication metadata, editorial content, and OCR artefacts that should be removed. We use LLMs to perform content refinement and OCR correction on document chunks under a **preserve-when-uncertain** principle. Each LLM call performs four integrated operations: (1) academic structural filtering (removing publication metadata while preserving abstracts, keywords, affiliations); (2) non-educational content removal (advertising, administrative instructions, generic boilerplate); (3) medical-specific OCR error correction (terminology, citation format, mathematical/chemical notation, abbreviation consistency); and (4) comprehensive medical-knowledge protection across core medical knowledge, research components, educational materials, clinical data, and pedagogical structures.

We implement this stage with `openai/gpt-oss-120b` (OpenAI, 2025) on 1024-token document chunks (no overlap), with an 8192-token context window. The prompt enumerates preservation rules, deletion criteria, and OCR-correction guidelines, and emphasizes asymmetric error tolerance: false negatives (failing to remove non-essential content) are acceptable, but false positives (removing valuable medical knowledge) are strictly prevented. The complete prompt and before/after examples are provided in Appendix A.

3.2 Multi-Style Text Reconstruction

We design a multi-style reconstruction strategy that produces two differentiated expression styles while strictly preserving medical accuracy, following the principle of "accuracy first, diversity enhancement." For each cleaned chunk we generate three versions (original academic plus two reconstructions), expanding the corpus by approximately 3× (3.25× averaged across sources, see Table 1) and exposing the model to varied medical registers (academic, patient-facing, prevention-oriented).

Style 1 (Patient-Centered). The first style transforms professional medical literature into patient-friendly educational materials through: *terminology de-specialization* (translation-style explanations alongside original terms); *analogical explanation* (familiar life metaphors for abstract concepts); *clinical decision reasoning* (the "why" behind procedures); *proactive concern response*; and *patient agency enhancement*.

Style 2 (Preventive Medicine). The second style focuses on actionable health management: *clinical knowledge to action translation* (concrete executable steps); and *biological explanation of pre-*

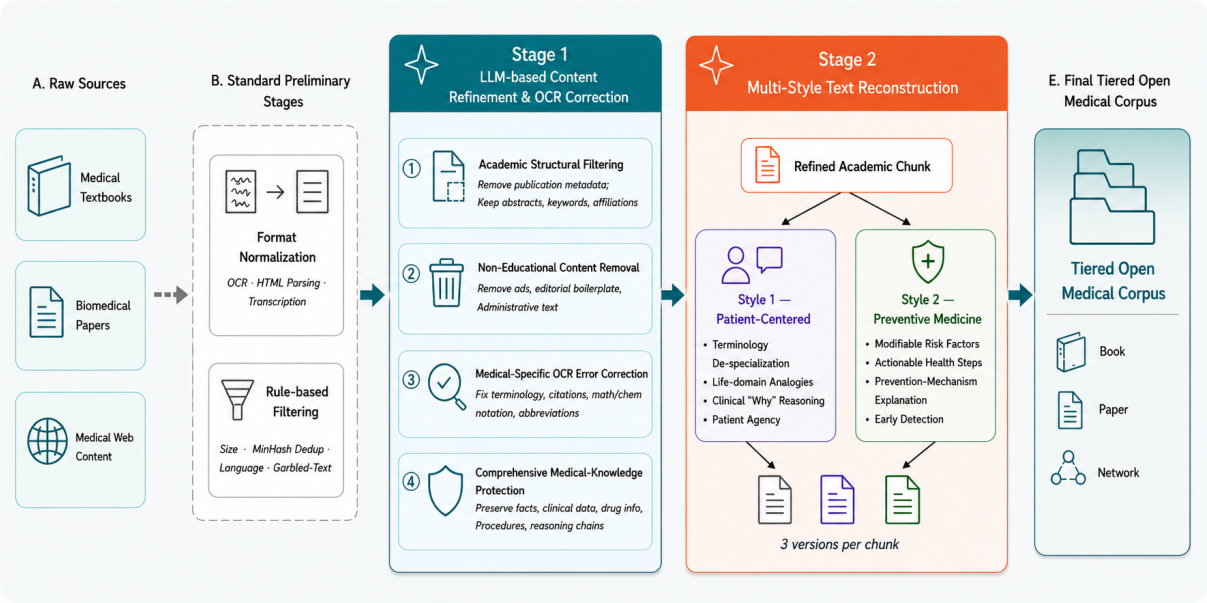

Figure 1: Overview of our data processing pipeline.

vention logic (the mechanisms behind preventive strategies).

Inviolable accuracy constraints. Both styles strictly preserve: (i) medical and scientific correctness (clarification only, never contradiction); (ii) structural integrity (section headers, figure/table labels, equation numbers, references); and (iii) precise chunk-boundary termination (the output must stop where the chunk stops, even mid-sentence or at a section header, to prevent hallucinated continuation). Both styles use `openai/gpt-oss-120b` (OpenAI, 2025) with 1024-token chunks, executed independently to ensure style differentiation. Complete prompts and examples are in Appendix B.

3.3 Final Dataset

Our cleaned medical corpus comprises three sources—∼246B tokens of medical textbooks (book-type), ∼113B tokens of peer-reviewed biomedical papers (paper-type), and ∼97B tokens of refined medical web content (web-type). Per-source token statistics across pipeline stages, together with release mode, are reported in Table 1. After multi-style reconstruction, the effective medical corpus expands to approximately 456B tokens.

3.4 Data Source Licensing and Release Plan

The three source types differ in upstream licensing, so we release them under distinct terms (see Table 1).

Table 1: Per-source token statistics (B) across pipeline stages, with release mode. *Pre.* = preprocessed; *Ref.* = refined; *Educ.* / *Prev.* = patient-centered / preventive-medicine reconstructions.

Source	Pre.	Ref.	Educ.	Prev.	Total	Release
Book	78.27	75.60	85.97	84.16	245.73	Gated
Paper	35.35	35.21	40.08	37.75	113.04	Open
Web	34.86	29.64	33.46	34.05	97.15	Open

Web (fully open). Derived from the non-synthetic subset of Nemotron-CC v1 (Su et al., 2024) (permissive open-data license); the cleaned 97.15B-token subset inherits the upstream license and will be released alongside processing code.

Paper (fully open). Sourced from PubMed Central's open-access subset via LLM360/TxT360 (Tang et al., 2024), under terms (e.g., CC-BY) that permit redistribution with attribution; per-document licensing metadata is preserved.

Book (gated). The 245.73B book subset is from a curated collection whose upstream license forbids unrestricted redistribution. We release it via a Hugging Face gated-access mechanism (institutional affiliation + statement of academic intent), consistent with protocols for other access-controlled medical resources (e.g., MIMIC-IV (Johnson et al., 2023)). For document subsets that cannot be redistributed even under gated terms, we release processing scripts and SHA-256 manifests so researchers who

independently obtain the upstream materials can regenerate bit-identical cleaned versions. Provenance (source URL/PMC ID/book identifier, timestamps, model versions) is recorded at the document level for auditability; ethics considerations are discussed at the end of the paper.

4 Experiments

4.1 Experimental Setup

Architecture and training. Our primary base is LLaMA-3.2-3B; for scaling experiments we additionally use Qwen2.5-7B and -32B (Yang et al., 2024, 2025). All models are trained with consistent hyperparameters: 33,600 steps over ∼140B tokens, global batch size 1,024, micro batch size 4, sequence length 4,096, cosine decay LR schedule from 1×10^{-5} to 1×10^{-6} with 250 warm-up steps, and 3D parallelism (TP=4, PP=2, CP=1). All runs use a fixed 140B-token CPT budget per configuration, in line with prior open medical-CPT efforts (Meditron-7B 47B; PMC-LLaMA 80B), to keep >30 controlled experiments (three model scales × Web-Medical/Ours, plus seven style-mixture and nine medical-to-web sweeps) tractable under a fixed compute envelope.

Baseline establishment. We use a medical-filtered web baseline (*Web-Medical*) as our principal baseline. Concretely, we apply the EAI-Distill-0.5b document classifier (Essential AI, 2024) to all of Nemotron-CC v1 and retain only documents whose Free Decimal Classification (FDC) label corresponds to medicine (Level-2 codes in the `61x` range) and whose document-type tag is educational/knowledge-bearing (academic, knowledge article, tutorial), discarding promotional, social-media, and short-fragment material. We then continually pre-train the same base model on this corpus with identical budget and hyper-parameters. A naive "pure web" baseline would invite the critique that any domain-specialised corpus would beat unfiltered web text; Web-Medical already restricts to the medical domain and therefore directly answers: *how much of our gain is attributable to LLM-based refinement and multi-style reconstruction, versus naive domain filtering alone?*

Web-related terms (disambiguation). The paper uses three distinct web-related references that should not be confused: (i) *Pure-Web* denotes unfiltered Nemotron-CC v1 used as a sanity-check baseline in Figure 2, and is also the source of the general-domain data mixed with our medical corpus to prevent catastrophic forgetting; (ii) *Web-Medical* denotes the medical-filtered Nemotron-CC subset described above (no further refinement) used as our principal baseline; (iii) the *web-source* subset of our corpus (Table 1, ∼97B tokens) is Web-Medical further processed through our refinement-and-rewriting pipeline.

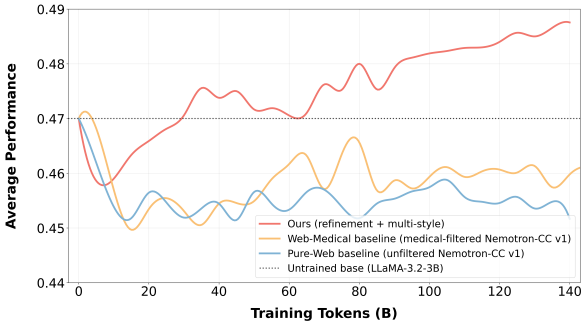

Figure 2: Training dynamics on LLaMA-3.2-3B over 140B tokens, averaged across nine medical benchmarks. *Red:* our pipeline (Origin 80% + Style1 10% + Style2 10%, 4:1 medical-to-web). *Orange:* Web-Medical baseline. *Light blue:* Pure-Web baseline (unfiltered CC). *Dotted grey:* untrained LLaMA-3.2-3B base (0.469). Both web baselines regress below the untrained base; only our pipeline produces consistent positive transfer.

Benchmarks. We evaluate on a comprehensive suite of medical benchmarks: CareQA (Arias-Duart et al., 2025), HeadQA (Vilares and Gómez-Rodríguez, 2019), MedMCQA (Pal et al., 2022), MedQA (Jin et al., 2021), MedXpertQA (Zuo et al., 2025), MMLU Medical Genetics (Hendrycks et al., 2021), MMLU-Pro Biology (Wang et al., 2024a), MMLU-Pro Health (Wang et al., 2024a), and PubMedQA (Jin et al., 2019). All nine benchmarks are multiple-choice QA tasks; we report accuracy as the evaluation metric throughout, so all reported numbers and figure y-axes are accuracy on the 0–1 scale.

4.2 Main Results

4.2.1 Training Dynamics

To gain insight into how data quality affects learning, we monitor average performance across nine medical benchmarks throughout training, comparing three runs on the same LLaMA-3.2-3B base with matched hyper-parameters: (i) our refinement-and-rewriting pipeline, (ii) Web-Medical, and (iii) Pure-Web (unfiltered Nemotron-CC). All runs start from the untrained base (0.469); the only varying

factor is the corpus.

Figure 2 reveals three markedly different learning regimes. **Pure-Web** plummets below the base within the first 10B tokens and stagnates at 0.452–0.457, ending at 0.452 (-1.7 pp). **Web-Medical** oscillates around 0.455–0.466 and ends at 0.460 (-0.9 pp below base): domain filtering partially mitigates but does not reverse the regression. **Our cleaned data** dips briefly during initial adaptation, then climbs steadily—0.474 at 40B, 0.480 at 80B, 0.484 at 120B, 0.488 at 140B—with positive end-slope. Both web baselines yield *negative transfer* (Wu et al., 2024; Pillai, 2025), while our pipeline produces consistent *positive transfer*; the $+2.8$ pp final-step gap over Web-Medical (0.488 vs. 0.460) directly measures the marginal value of our pipeline beyond naive medical-domain filtering.

4.2.2 Performance across Model Scales

To verify that gains are not specific to LLaMA-3.2-3B, we replicate the comparison at two larger Qwen2.5 scales (7B and 32B). Table 2 reports identical-compute Web-Medical vs. Ours for each base.

Consistent gains over Web-Medical, diminishing with scale. Ours beats Web-Medical by $+2.8$ / $+1.7$ / $+1.0$ pp at 3B / 7B / 32B; against the untrained base, Ours adds $+1.9$ / $+2.1$ / $+0.9$ pp while Web-Medical adds -0.9 / $+0.4$ / -0.1 pp—regressing or flat at 3B and 32B, confirming that medical-domain filtering alone is insufficient. The improvement margin narrows monotonically with capacity: larger models extract more signal from noisier data, so high-quality refinement matters most exactly when one cannot afford a larger model.

Robust across base families. Gains hold for both LLaMA-3.2 (3B) and Qwen2.5 (7B, 32B), ruling out model-specific artefacts.

4.2.3 Comparison with Specialized Medical LLMs

We situate our results against four well-known open medical LLMs, ordered by base-model generation: PMC-LLaMA-7B and PMC-LLaMA-13B (Wu et al., 2023) (LLaMA-1-based biomedical CPT), Meditron-7B (Chen et al., 2023) (LLaMA-2-7B-based CPT), and MedGemma-27B-text-it (DeepMind, 2025) (a 27B Gemma-3-based model fine-tuned by Google DeepMind with full SFT + alignment, recognised as the strongest open medical LLM at writing time). Meditron-7B has a 2048-token context limit; for the three benchmarks (PubMedQA, MedQA, MedXpertQA) where 5-shot prompting overflows this, we report 3-shot numbers (matching Meditron's original setup) and mark them with †. All other rows (PMC-LLaMA family, MedGemma-27B-text-it, and our models) use uniform 5-shot prompting.

Three takeaways from Table 3:

(1) A 3B on our corpus surpasses the entire legacy medical CPT family. Ours-3B outperforms PMC-LLaMA-7B by $+20.7$, PMC-LLaMA-13B by $+9.7$, and Meditron-7B by $+10.9$ avg pp despite being 2–4$\times$ smaller. The gap is amplified by the LLaMA-1/2 $\rightarrow$ LLaMA-3 generational upgrade: all three legacy models even underperform the untrained LLaMA-3.2-3B base (0.469), suggesting legacy medical CPT pipelines have been outpaced by generic-base improvements. Our work offers a recipe for re-establishing the medical-CPT advantage on top of modern bases.

(2) Our 32B surpasses the SOTA open medical LLM at comparable scale. Ours-32B beats MedGemma-27B-text-it by $+6.2$ avg pp, winning 8/9 benchmarks (MMLU-Bio the only loss, -0.7 pp). MedGemma uses a strictly stronger post-training recipe (full SFT + alignment), yet our CPT-only model exceeds it: in the era of strong modern base models, a well-curated pre-training corpus can matter more than post-training engineering for medical performance.

(3) Our 7B is competitive at $4\times$ fewer parameters. Ours-7B trails MedGemma-27B-text-it by only -1.9 avg pp; the deficit concentrates on knowledge-heavy MMLU-Bio/H (around -12 pp) where MedGemma's targeted SFT helps most, while Ours-7B leads on clinically-grounded benchmarks (CareQA, MedMCQA, HeadQA), making it a strong cost-efficient alternative.

5 Ablation Studies

We isolate the contribution of each pipeline component: content refinement, multi-style reconstruction, intra-domain mixture, and medical-to-web ratio.

5.1 Content Refinement Effectiveness

We compare models trained on preprocessed data versus data further refined through LLM-based semantic cleaning (Table 4; full per-benchmark numbers in Appendix C.1). Refinement provides consistent improvements, with the largest gains (up

Table 2: Performance across model scales: untrained Base, Web-Medical baseline, and Ours, on a matched base for each row triple. Bold marks the highest value within each model-size block.

Model	Data	CareQA	HeadQA	MedMCQA	MedQA	MedXpert	MMLU-MG	MMLU-Bio	MMLU-H	PubMedQA	Avg
LLaMA-3.2-3B	Base	0.554	0.377	0.481	0.512	0.131	**0.730**	0.453	0.313	0.672	0.469
LLaMA-3.2-3B	Web-Medical	0.547	0.377	0.477	0.507	0.122	0.670	0.403	0.315	**0.726**	0.460
LLaMA-3.2-3B	**Ours**	**0.573**	**0.397**	**0.494**	**0.533**	**0.136**	**0.730**	**0.456**	**0.346**	**0.726**	**0.488**
Qwen2.5-7B	Base	0.722	0.419	0.600	0.624	0.126	0.830	**0.692**	0.494	0.746	0.584
Qwen2.5-7B	Web-Medical	0.737	0.410	0.606	0.637	0.143	**0.860**	0.642	0.474	**0.782**	0.588
Qwen2.5-7B	**Ours**	**0.761**	**0.424**	**0.629**	**0.664**	**0.148**	0.850	0.669	**0.529**	0.770	**0.605**
Qwen2.5-32B	Base	0.833	0.486	0.691	0.725	0.178	**0.910**	**0.823**	**0.666**	0.782	0.677
Qwen2.5-32B	Web-Medical	0.851	0.471	0.689	0.756	0.192	**0.910**	0.773	0.647	0.794	0.676
Qwen2.5-32B	**Ours**	**0.858**	**0.487**	**0.702**	**0.767**	**0.205**	**0.910**	0.788	0.659	**0.800**	**0.686**

Table 3: Comparison with specialized open medical LLMs. "Recipe" indicates how each model was specialized. † = 3-shot due to Meditron's 2048-token limit.

Model	Recipe	CareQA	HeadQA	MedMCQA	MedQA	MedXpert	MMLU-MG	MMLU-Bio	MMLU-H	PubMedQA	Avg
PMC-LLaMA-7B	LLaMA-7B + biomedical CPT	0.257	0.376	0.306	0.292	0.102	0.310	0.085	0.114	0.682	0.281
PMC-LLaMA-13B	LLaMA-13B + biomedical CPT	0.404	0.421	0.387	0.375	0.114	0.510	0.350	0.241	0.714	0.391
Meditron-7B	LLaMA-2-7B + medical CPT	0.383	0.400	0.356	0.373†	0.119†	0.540	0.283	0.224	0.732†	0.379
MedGemma-27B-text-it	Gemma-3-27B + SFT + alignment	0.719	0.393	0.580	0.706	0.200	0.850	**0.795**	0.644	0.732	0.624
Ours-3B	LLaMA-3.2-3B + our pipeline	0.573	0.397	0.494	0.533	0.136	0.730	0.456	0.346	0.726	0.488
Ours-7B	Qwen2.5-7B + our pipeline	0.761	0.424	0.629	0.664	0.148	0.850	0.669	0.529	0.770	0.605
Ours-32B	Qwen2.5-32B + our pipeline	**0.858**	**0.487**	**0.702**	**0.767**	**0.205**	**0.910**	0.788	**0.659**	**0.800**	**0.686**

Table 4: Impact of content refinement on average performance.

Model	Setting	Pipeline	Avg
LLaMA-3.2-3B	Web-Medical	–	0.460
LLaMA-3.2-3B	Preprocessed	No refinement	0.473
LLaMA-3.2-3B	+ Refinement	With refinement	**0.481**

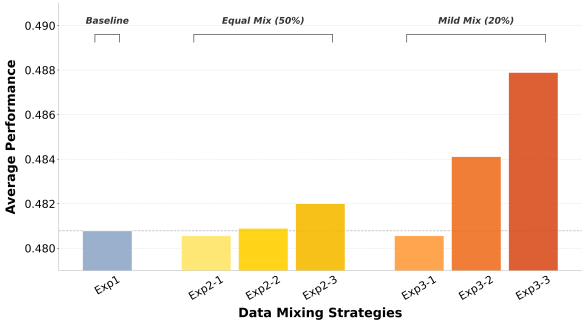

Figure 3: LLaMA-3.2-3B at 140B tokens with 4:1 medical-to-web ratio, varying medical-side mixture of original (*Origin*), patient-centred (*Patient*), and preventive-medicine (*Preventive*) text. Configurations: **Exp1** Origin 100%; **Exp2-1** 50%/50%/0%; **Exp2-2** 50%/0%/50%; **Exp2-3** 33%/33%/33%; **Exp3-1** 80%/20%/0%; **Exp3-2** 80%/0%/20%; **Exp3-3** 80%/10%/10%.

to +2.1 pp) on reasoning-intensive benchmarks (HeadQA, MedQA, MMLU-MG, MMLU-Health). These gains demonstrate that LLM-based semantic understanding effectively removes non-educational administrative content and corrects domain-specific OCR errors that rule-based methods cannot address. Improvements on reasoning-intensive benchmarks (HeadQA, MMLU-Health) suggest that the preservation of logical coherence and medical reasoning chains is particularly valuable.

5.2 Multi-Style Reconstruction

We sweep seven style-mixture configurations (Exp1–Exp3-3), varying the proportions of original, patient-centered, and preventive-medicine text (Figure 3; full numbers in Appendix C.2).

Stylistic diversity benefits. Most reconstruction strategies maintain competitive or improved performance compared to original-only training, indicating that diverse expression styles do not harm and often enhance model capabilities.

Optimal conservative mixture. The 80% original + 10% Patient + 10% Preventive (Exp3-3) configuration achieves the best balance, matching or exceeding origin-only performance across most benchmarks while incorporating stylistic diversity. This conservative mixture preserves the primary benefit of high-quality original content while gaining complementary advantages of multiple pedagogical perspectives. We adopt this 80-10-10 mixture as our final configuration.

5.3 Data Mixture Optimization

In domain-specific CPT, training exclusively on the target domain risks catastrophic forgetting of general capabilities, while training on too little of it leaves domain knowledge under-injected (Wu et al., 2024; Xie et al., 2023; Ye et al., 2025a); the optimal

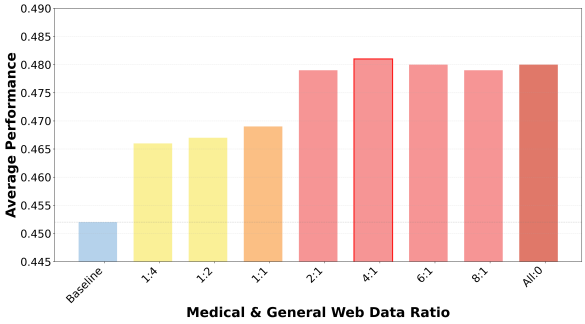

Figure 4: Performance across medical-to-web mixture ratios.

Table 5: Performance of individual medical sources (4:1 with web).

Model	Source	Data	Avg
LLaMA-3.2-3B	Web-Medical	–	0.460
LLaMA-3.2-3B	Book-only	75.6B books	**0.478**
LLaMA-3.2-3B	Paper-only	35.2B papers	0.462
LLaMA-3.2-3B	Refined-web only	29.6B web	0.474

trade-off lies in mixing domain-specific and general data. We therefore systematically explore (1) intra-medical mixture among textbooks/papers/web, and (2) medical-to-general mixture, to identify the best operating point for medical CPT.

Intra-medical mixture. We compare Strategy 1 (uniform 1:1:1) against Strategy 2 (proportional to data volume, ∼15:7:6). Strategy 2 yields marginally better performance across most benchmarks (full numbers in Appendix C.3), suggesting that the larger volume of book-type data provides valuable systematic knowledge structure. We adopt Strategy 2 throughout.

Medical-to-web ratio. We test nine ratios from 0:all (no medical) to all:0 (pure medical), at fixed compute (Figure 4; full numbers in Appendix C.4). Every mixture outperforms the 0:all (no-medical) endpoint, with average improvements ranging from +1.4 pp (1:4) to +2.9 pp (4:1); the 4:1 ratio attains the highest average (0.481), and we adopt it as our default operating point throughout the paper.

5.4 Domain-wise Data Validation

To verify that all three medical sources contribute positively, we evaluate each source independently mixed with web at 4:1 (Table 5; full per-benchmark numbers in Appendix C.5).

Books are the foundation of systematic medical understanding. Book-type data shows the strongest and most consistent improvements (+1.8 pp avg over Web-Medical). Textbooks present hierarchical structures, consistent terminology, contextual connections, and causal reasoning chains essential for clinical decision-making.

Papers show specialised technical depth. Paper-type data has a narrow but specialised profile: average is essentially tied with Web-Medical (+0.2 pp). Research papers focus on narrow, technically sophisticated topics with dense scientific terminology, which functions as a *specialisation layer* for cutting-edge domain knowledge.

Cleaned web bridges academic knowledge and clinical practice. Web-sourced data achieves competitive overall performance (+1.4 pp avg over Web-Medical). Web content's clinical guidelines, patient-education materials, Q&A forums, and case discussions reflect how medical knowledge is applied in clinical and patient-facing contexts.

Synergistic complementarity. All three sources improve over Web-Medical (Books +1.8, Web +1.4, Papers +0.2 pp). Their complementary strengths suggest that the full corpus benefits from synergistic effects: *Books* provide systematic conceptual framework; *Papers* add specialised technical depth; *Web* contributes practical clinical language and diverse scenario exposure.

6 Conclusion

We presented a comprehensive study on large-scale medical text data curation, addressing critical gaps in the medical-LLM ecosystem. Our progressive "refinement-rewriting" pipeline—pairing **document-level LLM refinement** with **style-conditioned reconstruction**—transforms raw heterogeneous sources into a ∼**456B-token** tiered corpus that delivers consistent positive transfer where Web-Medical regresses below the untrained base; at 32B, a CPT-only model trained on it **surpasses Google's SFT-aligned MedGemma-27B-text-it on 8 of 9 benchmarks (+6.2 avg pp)**. The methodological insights—multi-source heterogeneity, document-level semantic cleaning with preserve-when-uncertain safeguards, and style-conditioned reconstruction for educational diversity—may generalise to other specialised domains where data quality is paramount.

Limitations

We identify four limitations that motivate concrete future work. **(i) Quality verification at scale.** We performed manual spot-checks during pipeline de-

velopment to validate refinement and reconstruction prompts (e.g., the preserve-when-uncertain principle, chunk-boundary truncation rule), but we do not report a quantitative human-evaluation study: the inspections were small-scale, conducted by the authors rather than independent clinical experts, and lacked an inter-annotator-agreement protocol expected of formal medical-content evaluation. Larger expert-driven assessments—ideally combining IAA-controlled clinical review with automatic semantic-fidelity metrics—are an important next step. **(ii) Stylistic-diversity ablations are limited to the 3B scale.** The seven-way mixture sweep over Origin/Patient/Preventive proportions (Section 5.2) was conducted only on LLaMA-3.2-3B. While the 3B/7B/32B comparisons confirm our final 80/10/10 recipe transfers to larger models, we cannot yet rule out that a different mixture is optimal at $\geq$7B; sweeping at every scale was outside the compute budget. **(iii) Scope of "medical data".** Our corpus consists of published medical literature and curated medical web content. It explicitly excludes electronic health records, clinical notes, and patient–physician dialogues, whose construction and licensing requirements differ substantially. Whether our refinement-and-rewriting pipeline transfers, with appropriate privacy adaptations, to those downstream clinical formats remains an open question. **(iv) Single-run results.** All reported numbers are single-run accuracies: the $>$30 controlled training runs at 3B–32B scale required to support the comparisons in Tables 2–3 and the seven style-mixture and nine medical-to-web sweeps make multi-seed evaluation infeasible within our compute envelope. We mitigate single-run noise by reporting consistent trends across nine benchmarks, three model scales, and 25+ ablation configurations; the magnitude and direction of the effects is preserved across these many slices.

Ethics Statement

Our corpus is constructed exclusively from published medical literature and public medical web content. None of the three sources contains protected health information (PHI) or individually identifiable patient records; we explicitly exclude electronic health records, clinical notes, and patient–physician dialogues. The book subset is released via gated access (institutional affiliation + statement of academic intent), consistent with protocols for other access-controlled medical resources. We document the complete provenance chain for every document so that downstream users can audit the origin of every training sample. Models trained on this corpus are intended for medical-AI research and should not be deployed for clinical decision-making without independent clinical validation.

References

Anna Arias-Duart, Pablo Agustin Martin-Torres, Daniel Hinjos, Pablo Bernabeu-Perez, Lucia Urcelay Ganzabal, Marta Gonzalez Mallo, Ashwin Kumar Gururajan, Enrique Lopez-Cuena, Sergio Alvarez-Napagao, and Dario Garcia-Gasulla. 2025. Automatic evaluation of healthcare LLMs beyond question-answering. *arXiv preprint arXiv:2502.06666*.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. 2023. MEDITRON-70B: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*.

BioNLP Community. 2024. Year 2023 in biomedical natural language processing: A synopsis. *PMC*.

Andrea Cossu, Tinne Tuytelaars, Antonio Carta, Lucia Passaro, Vincenzo Lomonaco, and Davide Bacciu. 2022. Continual pre-training mitigates forgetting in language and vision. *arXiv preprint arXiv:2205.09357*.

Google DeepMind. 2025. MedGemma: A collection of medically-focused Gemma open models. Technical Report.

Xiaoshan Ding and 1 others. 2024. Boosting large language models with continual learning: A comparative study. In *Findings of EMNLP*.

Essential AI. 2024. EAI-Distill: Distilled document classifiers for multi-dimensional web filtering. Model card; classifier produces FDC topic, Bloom cognitive process and knowledge domain, document type, extraction artefacts, missing content, reasoning depth, technical correctness, and education-level labels per document.

Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, and 1 others. 2023. Medalpaca–an open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *Proceedings of the International Conference on Learning Representations*.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 2567–2577.

Alistair EW Johnson, Tom J Pollard, Lu Shen, and 1 others. 2016. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3(1):160035.

Alistair EW Johnson and 1 others. 2023. Mimic-iv, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1.

Anastasia Krithara and 1 others. 2023. Bioasq-qa: A manually curated corpus for biomedical question answering. *Scientific Data*, 10:1–11.

Yuxuan Li and 1 others. 2025. Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. *arXiv preprint arXiv:2506.07044*.

Lei Liu and 1 others. 2024. A survey on medical large language models. *arXiv preprint arXiv:2406.03712*.

Xudong Luo and 1 others. 2024. Pre-trained language models in medicine: A survey. *Artificial Intelligence in Medicine*.

OpenAI. 2025. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, pages 248–260.

Sneh Pillai. 2025. Replay to remember: Retaining domain knowledge in continual learning of large language models. *arXiv preprint arXiv:2504.17780*.

Yiwei Qin, Zhen Huang, Tiantian Mi, Weiye Si, Chenyang Zhou, Qipeng Guo, Siyuan Feng, and Pengfei Liu. 2026. Data darwinism part I: Unlocking the value of scientific data for pre-training. *arXiv preprint arXiv:2602.07824*.

Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyi Qin, Wenyuan Wang, Yibin Wang, Zifeng Wang, Sayna Ebrahimi, and Hao Wang. 2024. Continual learning of large language models: A comprehensive survey. *arXiv preprint arXiv:2404.16789*.

Karan Singhal and 1 others. 2023. Toward expert-level medical question answering with large language models. *Nature*, 620:289–295.

Dan Su and 1 others. 2024. Nemotron-cc: Transforming common crawl into a refined long-horizon training dataset. *arXiv preprint arXiv:2412.02595*.

Liping Tang and 1 others. 2024. An 8tb dataset of public domain and openly licensed text. *arXiv preprint*.

Rayan Touchent and 1 others. 2025. A biomedical dataset enriched with llms for pretraining. *arXiv preprint arXiv:2506.20331*.

David Vilares and Carlos Gómez-Rodríguez. 2019. HEAD-QA: A healthcare dataset for complex reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 960–966.

Jianquan Wang and 1 others. 2025. Huatuo-26m, a large-scale chinese medical qa dataset. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3016–3037.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. 2024a. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*.

Yubo Wang and 1 others. 2024b. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*.

Chaoyi Wu and 1 others. 2023. Pmc-llama: Further finetuning llama on medical papers. *arXiv preprint arXiv:2304.14454*.

Tongtong Wu and 1 others. 2024. Continual learning of large language models: A comprehensive survey. *arXiv preprint arXiv:2402.01364*.

Qika Xie and 1 others. 2025. Medical foundation large language models for comprehensive medical understanding. *Nature Medicine*.

Sang Michael Xie and 1 others. 2023. Doremi: Optimizing data mixtures speeds up language model pre-training. *arXiv preprint arXiv:2305.10429*.

Ran Xu and 1 others. 2024. Bmretriever: Tuning large language models as better biomedical text retrievers. In *Proceedings of EMNLP*.

An Yang and 1 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

An Yang and 1 others. 2025. Qwen2.5-1m technical report. *Qwen Technical Report*.

Jiasheng Ye, Peiju Liu, Tianxiang Sun, Jun Zhan, Yunhua Zhou, and Xipeng Qiu. 2025a. Data mixing laws: Optimizing data mixtures by predicting language modeling performance. In *Proceedings of ICLR*.

Jiasheng Ye, Peiju Liu, Tianxiang Sun, Yunhua Zhou, Jun Zhan, and Xipeng Qiu. 2025b. Data mixing laws: Optimizing data mixtures by predicting language modeling performance. In *Proceedings of the Thirteenth International Conference on Learning Representations*.

Dongfang Zhang and 1 others. 2024. A survey of datasets in medicine for large language models. *Intelligence & Robotics*, 4(2):191–215.

Yuxin Zuo, Shang Qu, Yiyuan Li, Yifan Bai, Hanwen Chu, Xinyue Hu, Kaiyan Zhang, Siheng Li, Xin Meng, Xun Wang, Zihao Jiao, Yichen He, Siwei Yang, Xu Han, Kai Ye, Yuntao Bai, Zhiyuan Liu, Shangchao Min, and Maosong Sun. 2025. MedXpertQA: Benchmarking expert-level medical reasoning and understanding. *arXiv preprint arXiv:2501.18362*.

A LLM-based Content Refinement and OCR Correction

A.1 Processing Objectives and Design Rationale

The LLM-based Content Refinement and OCR Correction stage addresses unique challenges in medical academic literature: dense technical terminology, complex mathematical and chemical notations, structured research components, and pedagogical elements that must be preserved with absolute accuracy, while removing publication metadata, editorial content, and OCR artifacts. Unlike rule-based filtering, this stage employs deep semantic understanding to distinguish valuable medical knowledge from administrative artifacts and performs domain-specific OCR error correction. Our design follows a "preserve-when-uncertain" principle, implementing asymmetric error tolerance where false negatives are acceptable, while false positives are strictly prevented.

A.2 Complete Prompt Specification

The following prompt operationalizes this design through a hierarchical instruction structure. It establishes three priority levels: (1) inviolable medical knowledge preservation rules encompassing clinical data, research methodologies, pedagogical elements, and case studies; (2) targeted deletion criteria for publication metadata (journal identifiers, submission dates), administrative content (subscription information, editorial boards), and non-educational artifacts (advertisements, generic disclaimers); and (3) medical-specific OCR correction guidelines addressing terminology precision, citation formatting, mathematical notation, and chemical formula standardization. Multiple reinforcement statements throughout the prompt emphasize that medical accuracy and educational value take absolute precedence over text cleaning, explicitly instructing the model to preserve content when uncertain about its value.

LLM-based Content Refinement and OCR Correction Prompt

```
You are an expert medical academic
     document cleaner specialized
    in identifying and removing
    unwanted content and correcting
     OCR errors from medical books,
     research papers, and academic
    publications while preserving
    all educational and research
    value.
## Objective:
Process and refine OCR-extracted
    text from medical academic
    publications (textbooks,
    research papers, journals) to
    eliminate unnecessary,
    incorrect, or irrelevant
    content while correcting clear
    OCR artifacts. Ensure complete
    preservation and accuracy of
    all medical knowledge,
    scientific data, and
    educational materials.

## Critical Medical Academic
    Content Protection:
### Absolutely Preserve Medical
    Knowledge Content
* **Medical facts and knowledge**:
     disease mechanisms,
    pathophysiology, treatment
    protocols, medical principles
* **Clinical data and research
    results**: study outcomes,
    statistical analyses, clinical
    trial data, meta-analyses
* **Medical terminology and
    definitions**: disease names,
    anatomical terms, procedure
    descriptions, medical concepts
* **Educational content**: case
    studies, clinical scenarios,
    problem-solving examples,
    medical reasoning
* **Diagnostic and treatment
    information**: differential
    diagnoses, treatment algorithms
    , clinical guidelines
* **Drug information**:
    pharmacology, mechanisms of
    action, side effects,
    contraindications, drug
    interactions
* **Medical measurements and
    normal values**: lab reference
    ranges, vital sign norms,
    dosage calculations
```

```
* **Medical procedures and
    techniques**: step-by-step
    protocols, surgical procedures,
     diagnostic methods

### Preserve Academic Research
    Content
* **Research methodology**: study
    designs, inclusion/exclusion
    criteria, statistical methods,
    data collection procedures
* **Results and findings**:
    experimental outcomes, clinical
     observations, research
    conclusions
* **Medical evidence**: systematic
     reviews, evidence levels,
    recommendation grades
* **Academic discussions**:
    literature reviews, hypothesis
    formation, scientific reasoning
* **Medical innovations**: new
    treatments, emerging
    technologies, novel diagnostic
    methods

## Medical Academic Document-
    Specific Deletion Rules:
### Academic Structural Deletion (
    MODIFY APPROACH)
* Remove **journal/publisher
    metadata**: ISSN, DOI numbers,
    submission dates, peer review
    information
* Remove **book publication
    details**: ISBN, publisher
    addresses, printing information
    , edition details
* Remove **academic administrative
     content**: editorial boards,
    submission guidelines,
    copyright transfer forms
* **PRESERVE academic structure
    content**: abstracts, keywords,
     author affiliations when
    academically relevant
* **SELECTIVE reference list
    handling**:
- Remove standalone reference
    lists at document end
- **PRESERVE in-text citations**
    and references that support
    medical claims
- **PRESERVE reference lists in
    review articles** where
    citations are core content

### Medical Academic Content
    Deletion
* Remove **academic advertising**:
     conference advertisements,
    journal promotions, book
    advertisements
* Remove **non-academic
    administrative text**:
    subscription information,
    library access instructions
```

```
* Remove **generic academic
    disclaimers**: standard
    copyright notices, general
    terms of use
* **Preserve academic ethical
    statements**: research ethics
    approvals, conflict of interest
     disclosures when relevant to
    content understanding

### Enhanced OCR Error Correction
    for Medical Academic Content
* **Medical terminology precision
    **: ensure accurate spelling of
     medical terms, drug names,
    anatomical structures
* **Research data accuracy**:
    verify statistical values, p-
    values, confidence intervals,
    sample sizes
* **Citation format
    standardization**: repair
    malformed citations and
    reference formatting
* **Mathematical and chemical
    notation**: preserve equations,
     chemical formulas, molecular
    structures
* **Medical abbreviation
    consistency**: standardize
    medical acronyms and
    abbreviations

## Enhanced Content Protection
    Rules:
### Medical Academic-Specific
    Content
* Preserve **medical case
    presentations**: patient cases,
     clinical vignettes, teaching
    cases (anonymized)
* Preserve **medical illustrations
     descriptions**: figure
    captions, diagram explanations,
     anatomical descriptions
* Preserve **medical tables and
    data**: clinical data tables,
    comparison charts, treatment
    protocols
* Preserve **medical algorithms
    and flowcharts**: diagnostic
    trees, treatment pathways,
    decision algorithms
* Preserve **medical education
    materials**: learning
    objectives, key points, summary
     boxes, review questions
* Preserve **medical research
    protocols**: inclusion criteria
    , exclusion criteria, study
    procedures

### Academic Learning Structures
* Preserve **textbook pedagogical
    elements**: chapter objectives,
     key concepts, review questions
    , exercises
```

```
* Preserve **academic discussions
    **: author arguments,
    scientific debates, hypothesis
    presentations
* Preserve **medical specialty
    content**: subspecialty
    knowledge, advanced medical
    concepts, specialized
    procedures
* Preserve **evidence-based
    medicine content**: systematic
    review findings, guideline
    recommendations, expert
    consensus

### Medical Book and Paper
    Specific Structures
* Preserve **medical book chapters
    **: chapter introductions,
    section headings with
    educational content
* Preserve **research paper
    sections**: methodology,
    results, discussion sections
    with their complete content
* Preserve **medical appendices
    with educational value**:
    reference ranges, conversion
    tables, clinical scales
* Preserve **medical glossaries
    and terminology lists**: when
    they provide educational
    definitions and explanations

## Quality Assurance for Medical
    Academic Content:
* **Preserve medical education
    flow**: ensure content
    maintains logical progression
    for learning
* **Maintain research integrity**:
     preserve complete research
    narratives and scientific
    reasoning
* **Verify medical accuracy**:
    double-check all medical facts,
     drug names, and clinical
    information
* **Preserve academic citations**:
     maintain proper attribution of
     medical knowledge and research

## Special Handling Instructions:
* **When in doubt, preserve**: if
    uncertain whether content has
    medical educational or research
     value, err on the side of
    preservation
* **Maintain academic context**:
    ensure deletions don't break
    medical reasoning, research
    logic, or educational flow
* **Preserve medical knowledge
    chains**: maintain connections
    between medical concepts,
    causes, and effects
* **Academic completeness**:
    ensure preserved content
    maintains academic and
    educational coherence

## Instructions:
- Apply the highest standards of
    accuracy for all medical
    academic content
- Carefully identify content
    matching deletion rules while
    being extra cautious with
    medical knowledge and research
    content
- Remove administrative and non-
    academic content while
    preserving all educational and
    research value
- Apply enhanced OCR error
    corrections with focus on
    medical and academic accuracy
- **When uncertain about any
    content**: preserve it directly
- Maintain the educational and
    research integrity of medical
    academic documents
- **Critical**: Medical knowledge
    accuracy and educational value
    take precedence over text
    cleaning

## Input:
Medical academic document chunk:
{chunk}

## Output Format:
<CLEANED_TEXT>
[Place the cleaned content here,
    or leave completely empty if
    everything should be deleted]
</CLEANED_TEXT>
```

A.3 Processing Examples

This section presents representative before-and-after examples demonstrating the refinement and OCR correction effects on real medical documents. Through side-by-side comparisons, we illustrate how this stage successfully removes publication metadata and non-educational content, corrects medical-specific OCR errors, standardizes medical terminology and notations, and preserves all academically valuable medical knowledge.

A.3.1 Example 1: Metadata Removal and Knowledge Preservation

This example demonstrates the pipeline's capability to identify and remove publication metadata while preserving the complete scientific content of a medical research article. The original text contains journal-specific formatting elements ("MEDICAL SCIENCE" header, journal citation "Lancet 1990; 335: 123-28") that provide no educational value for

language model training. The cleaning process successfully eliminates these metadata artifacts while maintaining the article title, author information (which provides important attribution context), abstract content, and the complete research narrative including the introduction and research objectives. Notably, all medical terminology (e.g., "Kaposi's sarcoma", "acquired immunodeficiency syndrome"), epidemiological data (prevalence rates, patient demographics), and scientific reasoning are preserved with perfect fidelity. This example illustrates the model's ability to distinguish between removable publication artifacts and essential academic content, embodying the "preserve-when-uncertain" principle where scientific integrity takes absolute precedence over aggressive cleaning.

Before Processing:

```
MEDICAL SCIENCE

Kaposi's sarcoma among persons
    with AIDS: a sexually
    transmitted infection?

VALERIE BERAL    THOMAS A.
    PETERMAN    RUTH L. BERKELMAN
HAROLD W. JAFFE

In the United States Kaposi's
    sarcoma is at least 20 000
    times more common in persons
    with acquired immunodeficiency
    syndrome (AIDS) than in the
    general population and 300
    times more common than in other
     immunosuppressed groups. Among
     persons with the acquired
    immunodeficiency syndrome (AIDS
    ) reported to Centers for
    Disease Control by March 31,
    1989, 15% (13 616) had Kaposi's
     sarcoma. Kaposi's sarcoma was
    commoner among those who had
    acquired the human
    immunodeficiency virus (HIV) by
     sexual contact than
    parenterally, the percentage
    with Kaposi's sarcoma ranging
    from 1% in men with haemophilia
     to 21% in homosexual or
    bisexual men. Women were more
    likely to have Kaposi's sarcoma
     if their partners were
    bisexual men rather than
    intravenous drug users. Kaposi'
    s sarcoma risk was not
    consistently related to age or
    race but varied across the
    United States, being greatest
    in the areas that were the
    initial foci of the AIDS
    epidemic. Thus Kaposi's sarcoma
     in persons with AIDS may be
    caused by an as yet
    unidentified infectious agent,
    transmitted mainly by sexual
    contact.

Lancet 1990; 335: 123-28.

Introduction
One of the first manifestations of
     the acquired immunodeficiency
    syndrome (AIDS) epidemic was
    the unusual appearance of
    Kaposi's sarcoma in young
    homosexual men. Since then 24
    different disorders have been
    included in the Centers for
    Disease Control's (CDC)
    surveillance definition of AIDS
    . Most of these disorders are
    caused by opportunistic
    infections.1 Kaposi's sarcoma
    remains a common manifestation
    of AIDS but its cause is
    unknown.

Since the start of the epidemic,
    the pattern of occurrence of
    Kaposi's sarcoma in people with
     AIDS was recognised as
    providing important clues to
    its aetiology. That Kaposi's
    sarcoma was ten times more
    common in homosexual or
    bisexual men than in other men
    with AIDS led to suggestions
    that it might be caused by a
    sexually transmitted infection,
     or by an environmental agent,
    such as nitrite inhalants (
    poppers), to which homosexual
    men are preferentially exposed
    .2-5 These suggestions are
    reminiscent of some of the
    initial proposals of the cause
    of AIDS itself.6

We analysed data on the occurrence
     of Kaposi's sarcoma among
    persons with AIDS reported to
    CDC to evaluate the existing
    hypotheses and to provide
    further clues about the
    aetiology of Kaposi's sarcoma.
```

After Processing:

```
Kaposi's sarcoma among persons
    with AIDS: a sexually
    transmitted infection?

VALERIE BERAL THOMAS A. PETERMAN
    RUTH L. BERKELMAN
HAROLD W. JAFFE
```

```
In the United States Kaposi's
    sarcoma is at least 20 000
    times more common in persons
    with acquired immunodeficiency
    syndrome (AIDS) than in the
    general population and 300
    times more common than in other
     immunosuppressed groups. Among
     persons with the acquired
    immunodeficiency syndrome (AIDS
    ) reported to Centers for
    Disease Control by March 31,
    1989, 15% (13 616) had Kaposi's
     sarcoma. Kaposi's sarcoma was
    commoner among those who had
    acquired the human
    immunodeficiency virus (HIV) by
     sexual contact than
    parenterally, the percentage
    with Kaposi's sarcoma ranging
    from 1% in men with haemophilia
     to 21% in homosexual or
    bisexual men. Women were more
    likely to have Kaposi's sarcoma
     if their partners were
    bisexual men rather than
    intravenous drug users. Kaposi'
    s sarcoma risk was not
    consistently related to age or
    race but varied across the
    United States, being greatest
    in the areas that were the
    initial foci of the AIDS
    epidemic. Thus Kaposi's sarcoma
     in persons with AIDS may be
    caused by an as yet
    unidentified infectious agent,
    transmitted mainly by sexual
    contact.

Introduction
One of the first manifestations of
     the acquired immunodeficiency
    syndrome (AIDS) epidemic was
    the unusual appearance of
    Kaposi's sarcoma in young
    homosexual men. Since then 24
    different disorders have been
    included in the Centers for
    Disease Control's (CDC)
    surveillance definition of AIDS
    . Most of these disorders are
    caused by opportunistic
    infections.1 Kaposi's sarcoma
    remains a common manifestation
    of AIDS but its cause is
    unknown.

Since the start of the epidemic,
    the pattern of occurrence of
    Kaposi's sarcoma in people with
     AIDS was recognised as
    providing important clues to
    its aetiology. That Kaposi's
    sarcoma was ten times more
    common in homosexual or
    bisexual men than in other men
    with AIDS led to suggestions
    that it might be caused by a
    sexually transmitted infection,
     or by an environmental agent,
    such as nitrite inhalants (
    poppers), to which homosexual
    men are preferentially exposed
    .2-5 These suggestions are
    reminiscent of some of the
    initial proposals of the cause
    of AIDS itself.6

We analysed data on the occurrence
     of Kaposi's sarcoma among
    persons with AIDS reported to
    CDC to evaluate the existing
    hypotheses and to provide
    further clues about the
    aetiology of Kaposi's sarcoma.
```

A.3.2 Example 2: OCR Errors and Correction of Medical Proprietary Terms

This example showcases the pipeline's medical-specific OCR error correction capabilities, addressing systematic character recognition failures that are particularly problematic in clinical literature. The original text exhibits characteristic OCR artifacts including letter substitution errors ("incontlnence" → "incontinence", "supranucIear" → "supranuclear", "recta1" → "rectal", "intra-abdominaI" → "intra-abdominal", "dlstension" → "distension", "sensatlon" → "sensation", "perinea1" → "perineal", "micturitlon" → "micturition", "neuroIogical" → "neurological", "radlograph" → "radiograph", "anorecta1" → "anorectal", "cerebra1" → "cerebral", "atheroscIerotic" → "atherosclerotic", "thromboemboIic" → "thromboembolic", "carotld" → "carotid", "ischaemla" → "ischaemia", "symptomatlc" → "symptomatic", "ExtracraniaI-IntracraniaI" → "Extracranial-Intracranial"). These errors, primarily involving 'I' (capital i) and 'l' (lowercase L) confusion as well as '1' (digit one) misrecognition, are endemic in scanned medical texts and severely impair readability and semantic integrity. The cleaning process successfully identifies and corrects these domain-specific OCR errors through contextual understanding of medical terminology, while also removing non-essential metadata and handling document boundaries appropriately (the incomplete sentence at the end is preserved as-is, following the contextual limitation principle). Additionally, the model expands the medical abbreviation "EAS" to its full form "external anal sphincter" upon first mention, enhancing clarity without com-

promising accuracy. This example demonstrates how LLM-based processing leverages medical domain knowledge to perform error correction that would be impossible with purely rule-based approaches, as it requires understanding the semantic context to distinguish valid medical terms from OCR corruption.

Before Processing:

```
18 female and 9 male patients (
    aged 19-78) out of a total of
    250 patients referred for
    investigation of faecal
    incontlnence had abnormalities
    similar to those with
    cerebrospinal lesions. 5 had
    evidence of a supranucIear
    lesion - a combination of an
    impaired ability to consciously
     contract the EAS (fig, upper),
     enhanced refIex activity of
    the EAS in response to rectaI
    distension (fig, lower),
    increases in intra-abdominaI
    pressure, and blunted recta1
    sensation. 17 patients had
    features compatible with a low
    spinaI lesion - absent or very
    attenuated EAS responses to
    rectal dlstension (fig, lower)
    and increases in intra-
    abdominal pressure combined
    with blunted rectal sensatlon,
    but they could contract the EAS
     voIuntarily (fig, upper). 5
    had a combination of impaired
    conscious and refIex
    contraction of the externaI
    anal sphincter, and blunted
    rectal sensation, suggesting a
    mixed supranucIear and low
    spinal lesion (fig). None of
    the women gave a history of
    prolonged labour or forceps
    delIvery, none of the patients
    complained of prolonged
    straining at stool, and none
    had abnormal perinea1 descent.
    Difficulties in micturitlon
    were found in 15% of patients,
    all of whom had features of low
     spinal or mlxed lesions. None
    of them gave any history of
    spinaI disease and none had
    difficulties in walking.
    Clinical neuroIogical
    examination and spinal
    radlograph revealed no evidence
     of cerebrospinal dlsease.

We suspect that occuIt spinal
    disease may be the causaI
    lesion in about 10% of patients
     referred to us for
    investigation for faecal
    incontlnence. Similar figures
    have been recorded for
    suspected spinal disease
    patients with disorders of
    urinary incontlnence. The
    recognition of these patients
    is important: first, the
    anorecta1 features may be an
    early feature of progresslve
    neurological disease and second
    , patients with spinal disease
    are unlikely to respond to
    surgery but may benefit from
    training to coordinate externaI
     sphincter contractlon with
    rectal sensation and improve
    sphincter strength.

W. M. SUN
N. W. REAC

Relevance of cerebra1 vasomotor
    reactlvity to prospects for
    successful endarterectomy

SIR, - The value of carotid
    endarterectomy remains
    controversiaI.1,2 Suggested
    indications focus on patients
    with atheroscIerotic plaques,
    who are therefore at risk of
    thromboemboIic stroke, and on
    patients with haemodynamically
    significant carotld artery
    stenosis, who are at risk of
    cerebral ischaemla.3 In North
    America, four clinical trials
    are under way comparing
    endarterectomy with medical
    therapy in patients with
    symptomatlc or symptom-free
    carotid stenosls.4,5

The ExtracraniaI-IntracraniaI (EC-
    IC) Bypass Study showed that
    bypass did not prevent stroke
    in patients with carotld
    occlusion. However, it has been
     argued that patients with
    abnorm
```

After Processing:

```
18 female and 9 male patients (
    aged 19-78) out of a total of
    250 patients referred for
    investigation of faecal
    incontinence had abnormalities
    similar to those with
    cerebrospinal lesions. 5 had
    evidence of a supranuclear
    lesion - a combination of an
    impaired ability to consciously
     contract the EAS (external
    anal sphincter), enhanced
    reflex activity of the EAS in
```

```
    response to rectal distension,
    increases in intra-abdominal
    pressure, and blunted rectal
    sensation. 17 patients had
    features compatible with a low
    spinal lesion - absent or very
    attenuated EAS responses to
    rectal distension and increases
     in intra-abdominal pressure
    combined with blunted rectal
    sensation, but they could
    contract the EAS voluntarily. 5
     had a combination of impaired
    conscious and reflex
    contraction of the external
    anal sphincter, and blunted
    rectal sensation, suggesting a
    mixed supranuclear and low
    spinal lesion. None of the
    women gave a history of
    prolonged labour or forceps
    delivery, none of the patients
    complained of prolonged
    straining at stool, and none
    had abnormal perineal descent.
    Difficulties in micturition
    were found in 15% of patients,
    all of whom had features of low
     spinal or mixed lesions. None
    of them gave any history of
    spinal disease and none had
    difficulties in walking.
    Clinical neurological
    examination and spinal
    radiograph revealed no evidence
     of cerebrospinal disease.

We suspect that occult spinal
    disease may be the causal
    lesion in about 10% of patients
     referred to us for
    investigation for faecal
    incontinence. Similar figures
    have been recorded for
    suspected spinal disease
    patients with disorders of
    urinary incontinence. The
    recognition of these patients
    is important: first, the
    anorectal features may be an
    early feature of progressive
    neurological disease and second
    , patients with spinal disease
    are unlikely to respond to
    surgery but may benefit from
    training to coordinate external
     sphincter contraction with
    rectal sensation and improve
    sphincter strength.

Relevance of cerebral vasomotor
    reactivity to prospects for
    successful endarterectomy

The value of carotid
    endarterectomy remains
    controversial. Suggested
    indications focus on patients
    with atherosclerotic plaques,
    who are therefore at risk of
    thromboembolic stroke, and on
    patients with haemodynamically
    significant carotid artery
    stenosis, who are at risk of
    cerebral ischaemia. In North
    America, four clinical trials
    are under way comparing
    endarterectomy with medical
    therapy in patients with
    symptomatic or symptom-free
    carotid stenosis.

The Extracranial-Intracranial (EC-
    IC) Bypass Study showed that
    bypass did not prevent stroke
    in patients with carotid
    occlusion. However, it has been
     argued that patients with
    abnormal
```

B Multi-Style Reconstruction

The Multi-Style Text Reconstruction stage transforms cleaned medical academic literature into diversified educational resources for different audiences and application scenarios. While academic medical literature typically targets medical professionals using highly specialized language, the same medical knowledge needs to be presented in multiple ways to meet the needs of diverse audiences, including patients, health-conscious individuals, and general public health education.

Our design implements a multi-style reconstruction strategy that generates two differentiated expression styles while maintaining absolute accuracy of medical facts. The **Patient-Centered Educational Transformation** bridges the knowledge gap between healthcare providers and patients through terminology de-specialization and clinical reasoning elucidation. The **Preventive Medicine and Health Optimization Orientation** focuses on actionable health management and disease prevention, translating clinical knowledge into concrete preventive steps with biological mechanism explanations. Both styles adhere to strict accuracy constraints ensuring medical and scientific correctness, structural integrity preservation, and precise termination at chunk boundaries without content continuation or hallucination.

B.1 Complete Prompt Specifications

B.1.1 Style One: Patient-Centered Educational Transformation

The following prompt implements patient-centered transformation through a two-tier instruction framework. It establishes inviolable medical accuracy rules (clinical correctness, structural integrity, strict chunk-boundary termination) as foundational constraints, then specifies five core transformation strategies: (1) medical jargon demystification using "translation approach" (stating technical terms with immediate plain-language explanations); (2) relatable analogies that transform abstract concepts into familiar metaphors; (3) clinical reasoning explanation that surfaces the "why" behind medical decisions; (4) proactive patient concern addressing that anticipates and reassures common anxieties; and (5) patient agency emphasis highlighting active participation in healthcare decisions. Critical output constraints prohibit meta-commentary and mandate exact termination at chunk endpoints to prevent hallucination, even if chunks end mid-sentence or at section headers. This design prioritizes comprehension and empowerment while maintaining absolute fidelity to medical facts.

Patient-Centered Educational Transformation Prompt

```
You are a master medical educator
    and patient communication
    specialist. Your mission is to
    transform the following medical
     text chunk into clear,
    accessible educational material
     that bridges the gap between
    clinical expertise and patient
    understanding. You are creating
     content that empowers patients
     and healthcare learners while
    maintaining absolute medical
    accuracy.

Your transformation will be
    governed by two fundamental
    principles: Medical Accuracy (
    non-negotiable) and Educational
     Accessibility (your primary
    mission).

### The Unbreakable Rules: Medical
     Accuracy and Safety
This principle is of paramount
    importance and must be followed
     without exception to ensure
    patient safety.

* **(a) Medical and Scientific
    Correctness:** Maintain
    absolute clinical rigor. All
    medical data, diagnostic
    criteria, treatment protocols,
    anatomical descriptions,
    physiological processes, and
    clinical evidence must be
    preserved without altering
    their meaning or context. Your
    additions must clarify, never
    contradict established medical
    knowledge.

* **(b) Structural Integrity:**
    Preserve the original structure
     flawlessly. Keep ALL section
    headers (`##`, `###`), figure/
    table labels, equation numbers,
     reference citations,
    classification systems, etc.,
    exactly as they appear,
    especially at the beginning and
     end of the chunk.

* **(c) Contextual Limitation and
    Termination:** You are
    processing a partial *chunk* of
     a document. You lack the full
    context. Therefore, you must
    work **strictly** within the
    provided text. Do not invent
    definitions or reference goals
    from outside the chunk. **This
    strict adherence means your
    output must terminate exactly
    where the provided chunk
    terminates.** If the chunk ends
     abruptly (e.g., at a new
    section header, in the middle
    of a sentence, or with a label)
    , your output **must be cut off
     at that exact same point.**
    This is the single most
    critical rule for preventing
    hallucination and ensuring
    continuity.

### The Core Mission: Patient-
    Centered Educational
    Transformation
This is your primary objective.
    Transform complex medical
    language into empowering,
    understandable content.

* **(a) Demystify Medical Jargon
    :** When you encounter medical
    terminology, pause to explain
    it in everyday language. Don't
    just define terms - explain why
     they matter to the patient's
    health journey. Use the "
    translation approach": state
    the medical term, then
    immediately provide the plain-
    language equivalent.
```

```
* **(b) Create Relatable Analogies
    :** Transform abstract medical
    concepts into familiar
    comparisons. For example,
    describe the immune system as
    the body's security force, or
    blood vessels as the body's
    highway system. Make these
    analogies memorable and
    accurate.

* **(c) Explain the Clinical 'Why
    ':** When medical procedures,
    treatments, or diagnostic steps
     are mentioned, explain the
    reasoning behind them. Help
    readers understand not just
    what happens, but why
    healthcare providers make these
     decisions.

* **(d) Address Patient Concerns
    :** Anticipate questions and
    concerns a patient might have.
    What would worry them? What
    would they want to know?
    Address these proactively with
    reassuring, factual
    explanations.

* **(e) Emphasize Patient Agency
    :** Where appropriate,
    highlight how patients can be
    active participants in their
    healthcare journey. Focus on
    understanding, preparation, and
     informed decision-making.

**Crucial Output Instructions:**
1. **Self-Contained Output:** The
    refined text must stand on its
    own. Avoid any meta-commentary
    or phrases that refer to the
    original text, such as "the
    original paper," "the original
    context," "the original chunk",
     etc. The goal is to create a
    seamless, self-contained
    educational text, not a
    commentary on another document.
2. **Strict Termination:** You **
    MUST** terminate your output at
     the **EXACT** same point the
    provided chunk terminates. Do
    **NOT** write a single
    character past the end of the
    original chunk. In particular,
    if a chunk ends with the start
    of a new section, subsection,
    step (e.g., it starts with a
    heading) or cuts off in the
    middle of a proof/solution, you
     must NOT invent or continue
    writing ANY content that would
    follow. For example:
- If the chunk ends with a new
    heading, your output must end
    with that exact heading and
    nothing more. Do not add a
    single word of content for that
     new section.
- If the chunk ends mid-sentence
    or mid-argument, your output
    must be cut off at that **exact
    ** same word. Do not complete
    the sentence or the argument by
     yourself.
- If the chunk ends with a label,
    your output must stop
    immediately after that label.
    Do not begin writing the
    content for that label by
    yourself.

---
*You must output ONLY the refined
    chunk itself, without any
    introductory or concluding
    remarks.*

**Original text:**
{chunk}

**Refined text:(Use the same
    language as the original text)
    **
```

B.1.2 Style Two: Preventive Medicine and Health Optimization Orientation

The following prompt redirects educational emphasis toward actionable health optimization. The prompt maintains identical medical accuracy constraints (clinical correctness, structural preservation, strict termination) while specifying five prevention-focused strategies: (1) modifiable risk factor emphasis highlighting controllable lifestyle and environmental factors; (2) clinical-to-action translation converting medical knowledge into concrete preventive steps; (3) prevention logic explanation elucidating biological mechanisms underlying health protection strategies; (4) early detection highlighting emphasizing screening value and symptom recognition; and (5) health empowerment framing that motivates positive health choices. This design targets health-conscious individuals and public health education contexts, maintaining closer proximity to academic register while integrating actionable guidance, resulting in more moderate content expansion compared to patient-centered style.

Preventive Medicine and Health Optimization Orientation Prompt

```
You are a distinguished expert in
    preventive medicine and public
    health, with a passion for
    empowering people to take
    control of their health journey
    . Your mission is to transform
    medical text into actionable,
    prevention-focused content that
     helps people understand not
    just how to treat disease, but
    how to prevent it and optimize
    their health.

Your approach integrates two core
    principles: Medical Validity (
    absolute foundation) and
    Prevention Focus (your
    specialized mission).

### The Unbreakable Rules: Medical
     Validity
This principle ensures all health
    recommendations remain
    medically sound and evidence-
    based.

* **(a) Clinical and Preventive
    Accuracy:** Maintain absolute
    precision in presenting
    preventive strategies, risk
    factors, health recommendations
    , and lifestyle interventions.
    All health advice must be
    medically sound and evidence-
    based.

* **(b) Structural Integrity:**
    Preserve the original structure
     flawlessly. Keep ALL section
    headers (`##`, `###`), figure/
    table labels, equation numbers,
     reference citations,
    classification systems, etc.,
    exactly as they appear,
    especially at the beginning and
     end of the chunk.

* **(c) Contextual Limitation and
    Termination:** You are
    processing a partial *chunk* of
     a document. You lack the full
    context. Therefore, you must
    work **strictly** within the
    provided text. Do not invent
    definitions or reference goals
    from outside the chunk. **This
    strict adherence means your
    output must terminate exactly
    where the provided chunk
    terminates.** If the chunk ends
     abruptly (e.g., at a new
    section header, in the middle
    of a sentence, or with a label)
    , your output **must be cut off
     at that exact same point.**
    This is the single most
    critical rule for preventing
    hallucination and ensuring
    continuity.

### Your Specialized Mission:
    Prevention and Health
    Optimization
This is where you transform
    medical knowledge into
    empowering health guidance.

* **(a) Emphasize Modifiable Risk
    Factors:** When discussing
    medical conditions, highlight
    the risk factors that people
    can control through lifestyle
    choices, environmental
    modifications, or preventive
    interventions.

* **(b) Translate Clinical
    Knowledge into Health Actions
    :** Convert medical information
     into practical steps people
    can take to protect and improve
     their health. Make prevention
    tangible and achievable.

* **(c) Explain the Prevention
    Logic:** When describing
    preventive strategies, explain
    the biological and medical
    reasoning behind them. Help
    people understand why specific
    actions protect their health.

* **(d) Highlight Early Detection
    :** Emphasize the importance of
     screening, early recognition
    of symptoms, and regular health
     monitoring. Show how early
    intervention can prevent
    serious health problems.

* **(e) Foster Health Empowerment
    :** Frame medical information
    in ways that help people feel
    capable and motivated to make
    positive health choices. Focus
    on what people can do to
    support their own wellbeing.

**Crucial Output Instructions:**
1. **Self-Contained Output:** The
    refined text must stand on its
    own. Avoid any meta-commentary
    or phrases that refer to the
    original text, such as "the
    original paper," "the original
    context," "the original chunk",
     etc. The goal is to create a
    seamless, self-contained
    educational text, not a
    commentary on another document.
2. **Strict Termination:** You **
    MUST** terminate your output at
     the **EXACT** same point the
```

```
    provided chunk terminates. Do
    **NOT** write a single
    character past the end of the
    original chunk. In particular,
    if a chunk ends with the start
    of a new section, subsection,
    step (e.g., it starts with a
    heading) or cuts off in the
    middle of a proof/solution, you
     must NOT invent or continue
    writing ANY content that would
    follow. For example:
- If the chunk ends with a new
    heading, your output must end
    with that exact heading and
    nothing more. Do not add a
    single word of content for that
     new section.
- If the chunk ends mid-sentence
    or mid-argument, your output
    must be cut off at that **exact
    ** same word. Do not complete
    the sentence or the argument by
     yourself.
- If the chunk ends with a label,
    your output must stop
    immediately after that label.
    Do not begin writing the
    content for that label by
    yourself.

---
*You must output ONLY the refined
    chunk itself, without any
    introductory or concluding
    remarks.*

**Original text:**
{chunk}

**Refined text:(Use the same
    language as the original text)
    **
```

B.2 Reconstruction Example: Pathology Introduction

This example illustrates the multi-style reconstruction strategy applied to foundational pathology content. The original academic text presents systematic definitions (disease, etiology, pathogenesis) and clinical diagnostic processes in conventional medical terminology. The **patient-centered transformation** bridges the knowledge gap through terminology de-specialization ("phenotype" explained as "your body's normal operating mode"), explicit clinical reasoning that surfaces the "why" behind medical practices, and patient agency enhancement emphasizing active participation in care. The **preventive medicine-oriented transformation** recontextualizes identical content through actionable health optimization: highlighting modifiable risk factors, explaining how pathogenesis knowledge "reveals critical intervention points," emphasizing controllable prognostic factors ("nutritional status, physical fitness, smoking cessation"), and advocating early screening to "catch problems before you experience symptoms." Both versions maintain medical accuracy while achieving distinct pedagogical objectives—patient-centered focuses on comprehension through elaboration, while preventive-oriented integrates actionable guidance with moderate expansion.

Original Refined Text:

```
## Introduction

Pathology represents the
    cornerstone of modern medicine,
     serving as the bridge between
    basic science and clinical
    practice. The term derives from
     the Greek word "pathologia,"
    meaning the study of suffering,
     and encompasses the systematic
     investigation of disease
    causes, developmental
    mechanisms, structural and
    functional changes, and natural
     history. This discipline has
    evolved dramatically over
    centuries, transforming from
    philosophical speculation to
    evidence-based science through
    technological innovation and
    rigorous methodology.

The practice of pathology centers
    on understanding how diseases
    deviate from normal phenotypes
    - the observable
    characteristics resulting from
    genome and environment
    interactions. According to [
    research from PMC](https://pmc.
    ncbi.nlm.nih.gov/articles/
    PMC7380242/), diseases manifest
     through patient complaints (
    symptoms) and measurable
    observations (signs), requiring
     systematic investigation to
    arrive at accurate diagnoses
    that guide treatment decisions.

## Terms, Definitions, and Core
    Concepts

### Fundamental Terminology

**Disease** refers to a definable
    deviation from normal phenotype
    , evident through patient
    symptoms and/or observer
    measurements. The **etiology**
    represents the disease's
    initial cause, which can be
```

```
    multifactorial - one disease
    may have multiple etiologies,
    and conversely, a single
    etiology can produce different
    diseases. For example,
    cigarette smoking can cause
    emphysema, chronic bronchitis,
    and small cell lung carcinoma,
    while emphysema itself can
    result from either alpha-1-
    antitrypsin deficiency or
    smoking.

**Pathogenesis** describes the
    stepwise mechanistic process of
     disease development,
    encompassing changes at gross/
    clinical, cellular, and
    molecular levels. This process
    involves specific chemical and
    cellular alterations that lead
    to observable tissue and organ
    dysfunction. Understanding
    pathogenesis enables
    development of screening
    methods for at-risk individuals
     and targeted therapeutic
    approaches.

### Clinical Presentation and
    Diagnosis

Diseases present to clinicians
    through **symptoms** (patient
    complaints) and **signs** (
    physician observations). The **
    differential diagnosis**
    represents the set of possible
    diagnoses accounting for the
    patient's clinical presentation
    . According to [pathology
    education resources](https://
    pmc.ncbi.nlm.nih.gov/articles/
    PMC7380242/), patient workup
    integrates present illness
    history, past medical history,
    family history, physical
    examination, radiographic
    studies, and laboratory
    analyses to reach a specific
    diagnosis meeting established
    diagnostic criteria.

The **pathologic diagnosis**
    represents the best current
    estimate of the disease entity,
     serving as the foundation for
    treatment decisions. This
    diagnosis implies an expected
    natural history - the disease
    course including chronicity,
    functional impairment, and
    survival. However, individual
    patient outcomes vary,
    necessitating assessment of **
    independent prognostic
    variables** that predict
    clinical outcomes and **
    independent predictive
    variables** that forecast
    therapy responses.

### The Pathologist's Role

As described by [Funkhouser and
    colleagues (2020)](https://pmc.
    ncbi.nlm.nih.gov/articles/
    PMC7380242/), pathologists are
    physicians or clinical
    scientists specializing in
    medical risk estimation and
    disease diagnosis through
    observations spanning clinical,
     gross, microscopic,
    immunophenotypic,
    ultrastructural, cytogenetic,
    and molecular levels. They bear
     responsibility for mastering
    emerging concepts, knowledge,
    and technologies that improve
    risk assessment, diagnostic
    accuracy, prognostication, and
    treatment prediction.
```

Patient-Centered Educational Version:

```
## Introduction

Pathology represents the
    cornerstone of modern medicine,
     serving as the bridge between
    basic science and clinical
    practice. The term derives from
     the Greek word "pathologia,"
    meaning the study of suffering,
     and encompasses the systematic
     investigation of disease
    causes, developmental
    mechanisms, structural and
    functional changes, and natural
     history. In simpler terms,
    pathology is the medical
    specialty dedicated to
    understanding *what* goes wrong
     in our bodies when we get sick
    , *why* it goes wrong, and *how
    * these problems develop and
    progress over time. This
    discipline has evolved
    dramatically over centuries,
    transforming from philosophical
     speculation to evidence-based
    science through technological
    innovation and rigorous
    methodology - moving from
    ancient theories about bodily "
    humors" to today's precision
    medicine guided by molecular
    analysis.

The practice of pathology centers
    on understanding how diseases
    deviate from normal phenotypes
    - the observable
    characteristics resulting from
```

genome and environment interactions. Think of your phenotype as your body's "normal operating mode" - the way your organs function, your cells behave, and your systems work together when you're healthy. This baseline is shaped by both your genetic blueprint (what you inherited from your parents) and your environment (your lifestyle, exposures, and experiences). According to [research from PMC](https://pmc.ncbi.nlm.nih.gov/articles/PMC7380242/), diseases manifest through patient complaints (symptoms) and measurable observations (signs), requiring systematic investigation to arrive at accurate diagnoses that guide treatment decisions. In other words, when something goes wrong, you might *feel* it (a symptom like pain or fatigue), and your healthcare provider might *detect* it through examination or tests (a sign like elevated blood pressure or an abnormal lab value). The pathologist's expertise helps connect these clues to identify the underlying disease process.

Terms, Definitions, and Core Concepts

Fundamental Terminology

Disease refers to a definable deviation from normal phenotype, evident through patient symptoms and/or observer measurements. Put simply, a disease is when your body's normal function becomes disrupted in a recognizable, consistent pattern that causes problems for your health. It's not just "feeling off" - it's a specific, identifiable condition that healthcare providers can diagnose and, ideally, treat.

The **etiology** represents the disease's initial cause - the "trigger" or underlying reason why a disease develops in the first place. Understanding etiology is crucial because it helps answer the question patients often ask: "Why did I get this?" Importantly, the relationship between causes and diseases can be complex and multifactorial - one disease may have multiple etiologies, and conversely, a single etiology can produce different diseases. For example, cigarette smoking (one cause) can lead to emphysema (lung tissue destruction), chronic bronchitis (persistent airway inflammation), and small cell lung carcinoma (a type of lung cancer) - three different diseases from one etiology. Conversely, emphysema itself can result from either alpha-1-antitrypsin deficiency (a genetic condition where your body lacks a protective protein) or smoking - two different causes leading to the same disease. This complexity explains why two people with the same exposure may develop different conditions, or why people with the same disease may have gotten it through different pathways.

Pathogenesis describes the stepwise mechanistic process of disease development, encompassing changes at gross/clinical, cellular, and molecular levels. Think of pathogenesis as the "story" of how a disease unfolds in your body over time - from the initial trigger, through a cascade of biological events, to the final clinical picture your doctor observes. This process involves specific chemical and cellular alterations that lead to observable tissue and organ dysfunction. For patients, understanding pathogenesis means grasping how, for instance, a viral infection in your lungs (the initial event) triggers inflammation (cellular response), which causes fluid accumulation (tissue change), which ultimately manifests as cough and shortness of breath (clinical symptoms). Understanding pathogenesis enables development of screening methods for at-risk individuals (helping identify disease before symptoms appear) and targeted therapeutic approaches (interventions designed to interrupt specific steps in the disease pathway rather than just treating symptoms).

```
### Clinical Presentation and
    Diagnosis

Diseases present to clinicians
    through **symptoms** (patient
    complaints) and **signs** (
    physician observations).
    Symptoms are what *you*
    experience and report - the
    pain, fatigue, nausea, or
    dizziness that only you can
    feel and describe. Signs are
    what *healthcare providers* can
     objectively measure or observe
     - your elevated temperature,
    irregular heartbeat, abnormal
    lab results, or findings on
    imaging studies. Both are
    essential pieces of the
    diagnostic puzzle, and your
    accurate description of
    symptoms is a vital
    contribution to your own
    healthcare.

The **differential diagnosis**
    represents the set of possible
    diagnoses accounting for the
    patient's clinical presentation
    . When you first visit your
    healthcare provider with
    concerning symptoms, they don't
     immediately know the exact
    diagnosis - instead, they
    develop a mental list of
    several possible conditions
    that could explain what you're
    experiencing. This "
    differential diagnosis" is like
     a shortlist of suspects that
    your provider systematically
    investigates through additional
     questioning, examination, and
    testing. According to [
    pathology education resources](
    https://pmc.ncbi.nlm.nih.gov/
    articles/PMC7380242/), patient
    workup integrates present
    illness history (what's
    happening now), past medical
    history (your previous health
    conditions), family history (
    diseases that run in your
    family), physical examination,
    radiographic studies (imaging
    like X-rays or CT scans), and
    laboratory analyses to reach a
    specific diagnosis meeting
    established diagnostic criteria
    . This comprehensive approach
    ensures that your healthcare
    team considers the full picture
     of your health, not just
    isolated symptoms.

The **pathologic diagnosis**
    represents the best current
    estimate of the disease entity,
```

```
     serving as the foundation for
    treatment decisions. This is
    the definitive answer to "What
    condition do I have?" -
    determined through careful
    analysis of all available
    evidence, often including
    tissue samples examined under a
     microscope. It's called the "
    best current estimate" because
    medicine is an evolving science
    ; as new information emerges or
     your condition changes,
    diagnoses may be refined. This
    diagnosis implies an expected
    natural history - the disease
    course including chronicity (
    whether it's short-term or long
    -lasting), functional
    impairment (how it might affect
     your daily activities), and
    survival (the expected outcome)
    . However, individual patient
    outcomes vary considerably -
    medical predictions describe
    general patterns, but your
    personal journey with a disease
     depends on many unique factors
    . This reality necessitates
    assessment of **independent
    prognostic variables** that
    predict clinical outcomes (
    factors that tell us how your
    disease is likely to progress,
    such as the stage of cancer or
    specific biomarkers in your
    blood) and **independent
    predictive variables** that
    forecast therapy responses (
    factors that indicate whether a
     particular treatment will work
     well for you specifically,
    such as genetic markers that
    determine drug effectiveness).
    These personalized assessments
    help your healthcare team
    tailor your treatment plan to
    your individual situation
    rather than applying a one-size
    -fits-all approach.

### The Pathologist's Role

As described by [Funkhouser and
    colleagues (2020)](https://pmc.
    ncbi.nlm.nih.gov/articles/
    PMC7380242/), pathologists are
    physicians or clinical
    scientists specializing in
    medical risk estimation and
    disease diagnosis through
    observations spanning clinical,
     gross, microscopic,
    immunophenotypic,
    ultrastructural, cytogenetic,
    and molecular levels. While you
     might never meet your
    pathologist directly (they
```

```
    typically work behind the
    scenes in laboratories), they
    play a critical role in your
    healthcare team. Think of
    pathologists as medical
    detectives who examine your
    tissue samples, blood tests,
    and other specimens using
    increasingly sophisticated
    tools - from traditional
    microscopes to cutting-edge
    genetic analyzers. They analyze
     your samples at multiple
    levels: the clinical level (
    considering your symptoms and
    test results), gross level (
    examining tissue with the naked
     eye), microscopic level (
    studying cells under
    magnification),
    immunophenotypic level (
    identifying specific proteins
    on cell surfaces),
    ultrastructural level (
    examining cell components with
    electron microscopes),
    cytogenetic level (analyzing
    chromosomes), and molecular
    level (studying DNA, RNA, and
    gene expression). They bear
    responsibility for mastering
    emerging concepts, knowledge,
    and technologies that improve
    risk assessment, diagnostic
    accuracy, prognostication, and
    treatment prediction - ensuring
     that you benefit from the
    latest scientific advances in
    understanding and treating
    disease.
```

Preventive Medicine Oriented Version:

```
# Introduction

Pathology represents the
    cornerstone of modern medicine,
     serving as the bridge between
    basic science and clinical
    practice. The term derives from
     the Greek word "pathologia,"
    meaning the study of suffering,
     and encompasses the systematic
     investigation of disease
    causes, developmental
    mechanisms, structural and
    functional changes, and natural
     history. This discipline has
    evolved dramatically over
    centuries, transforming from
    philosophical speculation to
    evidence-based science through
    technological innovation and
    rigorous methodology.

Pathology centers on understanding
     how diseases deviate from
    normal health states - the
    observable characteristics
    resulting from the interaction
    between your genetic makeup and
     environmental exposures.
    According to research from PMC,
     diseases manifest through what
     you experience (symptoms) and
    what healthcare providers can
    measure (signs), requiring
    systematic investigation to
    arrive at accurate diagnoses
    that guide prevention and
    treatment strategies.

Understanding the mechanisms of
    disease development empowers
    you to recognize early warning
    signs, identify modifiable risk
     factors, and take proactive
    steps to protect your health.
    By learning how diseases
    progress from their earliest
    stages, you gain the knowledge
    needed to intervene before
    serious complications develop.

## Terms, Definitions, and Core
    Concepts

### Fundamental Terminology

Disease refers to a definable
    deviation from your normal
    health state, evident through
    symptoms you experience and/or
    measurable observations
    healthcare providers can detect
    . The etiology represents the
    disease's initial cause, which
    can be multifactorial - one
    disease may have multiple
    causes, and conversely, a
    single cause can produce
    different diseases.

**Understanding causation is key
    to prevention.** For example,
    cigarette smoking can cause
    emphysema, chronic bronchitis,
    and small cell lung carcinoma,
    while emphysema itself can
    result from either alpha-1-
    antitrypsin deficiency (a
    genetic factor requiring
    screening in at-risk families)
    or smoking (a completely
    preventable environmental
    exposure). This knowledge
    highlights a crucial prevention
     principle: by eliminating or
    reducing controllable risk
    factors like smoking, you can
    prevent multiple serious
    diseases simultaneously.

Pathogenesis describes the
    stepwise mechanistic process of
```

disease development, encompassing changes at observable, cellular, and molecular levels. This process involves specific chemical and cellular alterations that lead to observable tissue and organ dysfunction. **Understanding pathogenesis is foundational to prevention** because it reveals critical intervention points - stages where lifestyle modifications, environmental changes, or medical screening can interrupt the disease process before irreversible damage occurs. This knowledge enables development of screening methods for at-risk individuals and targeted preventive approaches.

Clinical Presentation and Diagnosis

Diseases present to clinicians through symptoms (what you notice and report) and signs (what healthcare providers observe and measure during examination). The differential diagnosis represents the set of possible diagnoses accounting for your clinical presentation. According to pathology education resources, patient workup integrates your current illness history, past medical history, family history, physical examination, radiographic studies, and laboratory analyses to reach a specific diagnosis meeting established diagnostic criteria.

Your active participation in this process is essential. Providing complete information about symptoms, lifestyle factors, environmental exposures, and family health history enables more accurate diagnosis and reveals opportunities for prevention. Understanding your family history, for instance, can identify genetic risk factors that warrant enhanced screening or preventive interventions.

The pathologic diagnosis represents the best current estimate of the disease entity, serving as the foundation for treatment decisions. This diagnosis implies an expected natural history - the disease course including chronicity, functional impairment, and survival. However, individual outcomes vary significantly based on factors you can influence, necessitating assessment of independent prognostic variables that predict clinical outcomes and independent predictive variables that forecast therapy responses. **Many of these variables - such as nutritional status, physical fitness, smoking cessation, and adherence to preventive strategies - are within your control and can substantially improve outcomes.**

The Pathologist's Role

As described by Funkhouser and colleagues (2020), pathologists are physicians or clinical scientists specializing in medical risk estimation and disease diagnosis through observations spanning clinical, gross, microscopic, immunophenotypic, ultrastructural, cytogenetic, and molecular levels. They bear responsibility for mastering emerging concepts, knowledge, and technologies that improve risk assessment, diagnostic accuracy, prognostication, and treatment prediction.

From a prevention perspective, pathologists play a crucial role in identifying disease at its earliest, most treatable stages. Through screening tests like cervical cytology (Pap smears), colonoscopy with biopsy, and analysis of suspicious lesions, pathologists detect precancerous changes and early-stage diseases when interventions are most effective. Understanding this role helps you appreciate the importance of recommended screening programs - these tests are designed to catch problems before you experience symptoms, when prevention or cure is still possible.

C Detailed Ablation Tables

C.1 Content Refinement (Per-Benchmark)

Per-benchmark expansion of Table 4: refinement gains across each of the nine medical benchmarks against both the preprocessed and Web-Medical baselines.

C.2 Multi-Style Reconstruction (Full Sweep)

Per-benchmark expansion of Figure 3: all seven Origin/Patient/Preventive mixture configurations (Exp1 through Exp3-3) at fixed 4:1 medical-to-web ratio.

C.3 Intra-Medical Mixture

Per-benchmark comparison of intra-medical mixture strategies: uniform 1:1:1 across book/paper/web sources versus proportional 15:7:6 sampling matching post-refinement token volumes.

C.4 Medical-to-Web Mixture (Full Sweep)

Per-benchmark expansion of Figure 4: nine medical-to-web mixture ratios from 0:all (no medical) to all:0 (pure medical), at fixed 140B-token compute.

C.5 Domain-wise Validation (Per-Benchmark)

Per-benchmark expansion of Table 5: each of the three medical sources (book-only, paper-only, refined-web only) trained independently mixed 4:1 with web.

Table 6: Impact of content refinement on per-benchmark performance.

Setting	CareQA	HeadQA	MedMCQA	MedQA	MedXpert	MMLU-MG	MMLU-Bio	MMLU-H	PubMedQA	Avg
Web-Medical	0.547	0.377	0.477	0.507	0.122	0.670	0.403	0.315	0.726	0.460
Preprocessed	0.567	0.388	0.490	0.524	0.132	0.670	0.448	0.325	0.716	0.473
+Refined	0.573	0.393	0.490	0.532	0.134	0.690	0.445	0.346	0.724	**0.481**

Table 7: Multi-style reconstruction ablation across all benchmarks.

Configuration	CareQA	HeadQA	MedMCQA	MedQA	MedXpert	MMLU-MG	MMLU-Bio	MMLU-H	PubMedQA	Avg
Web-Medical	0.547	0.377	0.477	0.507	0.122	0.670	0.403	0.315	**0.726**	0.460
Exp1 (Origin 100%)	0.573	0.393	0.490	0.532	0.134	0.690	0.445	0.346	0.724	0.481
Exp2-1 (50/50/0)	0.569	0.395	0.492	0.528	0.123	0.700	0.455	0.337	0.726	0.481
Exp2-2 (50/0/50)	0.573	0.393	0.490	0.524	0.133	0.700	0.438	0.351	0.726	0.481
Exp2-3 (33/33/33)	**0.574**	0.389	0.491	0.524	0.131	**0.730**	0.438	0.344	0.716	0.482
Exp3-1 (80/20/0)	**0.574**	0.391	0.491	0.528	0.132	0.700	0.452	0.331	0.726	0.481
Exp3-2 (80/0/20)	0.573	0.392	0.488	0.529	0.131	**0.730**	0.451	0.341	0.724	0.484
Exp3-3 (80/10/10)	0.573	**0.397**	**0.494**	**0.533**	**0.136**	**0.730**	**0.456**	**0.346**	0.726	**0.488**

Table 8: Intra-medical mixture: uniform vs. proportional sampling across book/paper/web sources.

Strategy	CareQA	HeadQA	MedMCQA	MedQA	MedXpert	MMLU-MG	MMLU-Bio	MMLU-H	PubMedQA	Avg
Web-Medical (ref.)	0.547	0.377	0.477	0.507	0.122	0.670	0.403	0.315	0.726	0.460
Strategy 1 (Uniform 1:1:1)	0.568	0.388	0.489	0.529	0.130	**0.700**	**0.450**	0.341	0.720	0.479
Strategy 2 (Proportional 15:7:6)	**0.573**	**0.393**	**0.490**	**0.532**	**0.134**	0.690	0.445	**0.346**	**0.724**	**0.481**

Table 9: Performance across medical-to-web mixture ratios.

Med:Web	CareQA	HeadQA	MedMCQA	MedQA	MedXpert	MMLU-MG	MMLU-Bio	MMLU-H	PubMedQA	Avg
0:all (no medical)	0.543	0.376	0.462	0.485	0.116	0.650	0.410	0.287	**0.736**	0.452
1:4	0.550	0.380	0.473	0.507	0.126	0.660	0.444	0.320	0.732	0.466
1:2	0.553	0.385	0.481	0.507	0.128	0.680	0.432	0.308	0.730	0.467
1:1	0.560	0.386	0.482	0.521	0.125	0.650	0.435	0.329	0.730	0.469
2:1	0.564	0.391	0.483	0.523	0.129	0.720	0.445	0.336	0.724	0.479
4:1	**0.573**	**0.393**	**0.490**	0.532	0.134	0.690	0.445	**0.346**	0.724	**0.481**
6:1	0.576	0.391	0.492	0.533	0.129	0.680	**0.462**	0.342	0.716	0.480
8:1	0.574	0.392	0.491	0.539	0.137	0.680	0.445	0.335	0.720	0.479
all:0 (Pure Medical)	0.576	0.389	0.497	**0.542**	**0.138**	0.690	0.442	0.339	0.708	0.480

Table 10: Per-benchmark performance of individual medical sources mixed 4:1 with web.

Source	CareQA	HeadQA	MedMCQA	MedQA	MedXpert	MMLU-MG	MMLU-Bio	MMLU-H	PubMedQA	Avg
Web-Medical	0.547	0.377	0.477	0.507	0.122	0.670	0.403	0.315	0.726	0.460
Book-only	0.573	0.388	0.500	0.532	0.133	0.670	0.452	0.355	0.698	**0.478**
Paper-only	0.543	0.381	0.472	0.496	0.126	0.700	0.402	0.313	0.728	0.462
Refined-web only	0.577	0.385	0.486	0.522	0.117	0.700	0.445	0.326	0.710	0.474