Biomed-Enriched: Data-Efficient Biomedical Pretraining via Paragraph-Level Annotation

Rian Touchent and Nathan Godey* **and Éric de la Clergerie**
Inria, Paris, France
{firstname.lastname}@inria.fr

Abstract

We annotate PubMed Central paragraphs for document type, domain, and educational quality using a two-stage pipeline: Llama-3.1-70B labels 400K paragraphs, then a fine-tuned XLM-RoBERTa propagates annotations to the full corpus. This paragraph-level approach captures content diversity within scientific articles that document-level labels miss. The resulting Biomed-Enriched corpus contains 2M clinical case paragraphs, providing a publicly available alternative to restricted clinical datasets. For decoders, continual pretraining experiments enable targeted improvements, with clinical upsampling boosting performance by 4 points on MMLU ProfMed and educational filtering improving MedQA and MedMCQA by ∼1 point. Combinations of these techniques led to faster convergence, reaching the same performance with a third of training tokens. For encoders, our best recipe matches BioClinical-ModernBERT on 11 tasks (77.3% vs. 77.1% F1) while using 2.5× fewer tokens and only public data.

1 Introduction

Large language models (LLMs) have demonstrated remarkable capabilities across a wide range of general tasks, from question answering to code generation. However, their performance often lags in specialized domains such as biomedical and clinical medicine, which demand domain expertise and precise terminology. This performance gap can be explained by the composition of standard pretraining corpora, which predominantly consist of web-scraped content from CommonCrawl. While diverse, these datasets lack sufficient representation of specialized knowledge required for complex biomedical reasoning. Although pre-training datasets are often supplemented with high-quality domain-specific corpora like PubMed, these curated additions represent only a small fraction compared to the vast amount of general web text (Li et al., 2024). Available clinical text is particularly scarce in public datasets, with hospital records and clinical notes largely inaccessible due to strict privacy regulations. The situation is further complicated for non-English biomedical content, with resources like PMC containing over 98% English articles. A central challenge in developing effective domain-specialized LLMs is therefore identifying strategies for curating, filtering, and upsampling domain-relevant documents to enhance performance on specialized tasks without compromising general capabilities.

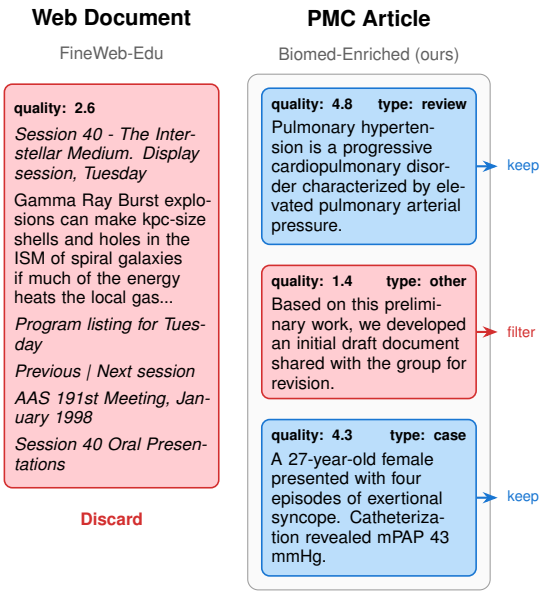

Figure 1: Motivation for paragraph-level filtering. Scientific articles mix high-value content (clinical cases, reviews) with boilerplate. Document-level filtering would discard valuable paragraphs.

2 Related Work

To address the domain gap issue, researchers have employed continual pre-training on large

*Now at Cornell University.

Findings of the Association for Computational Linguistics: ACL 2026, pages 34276–34287

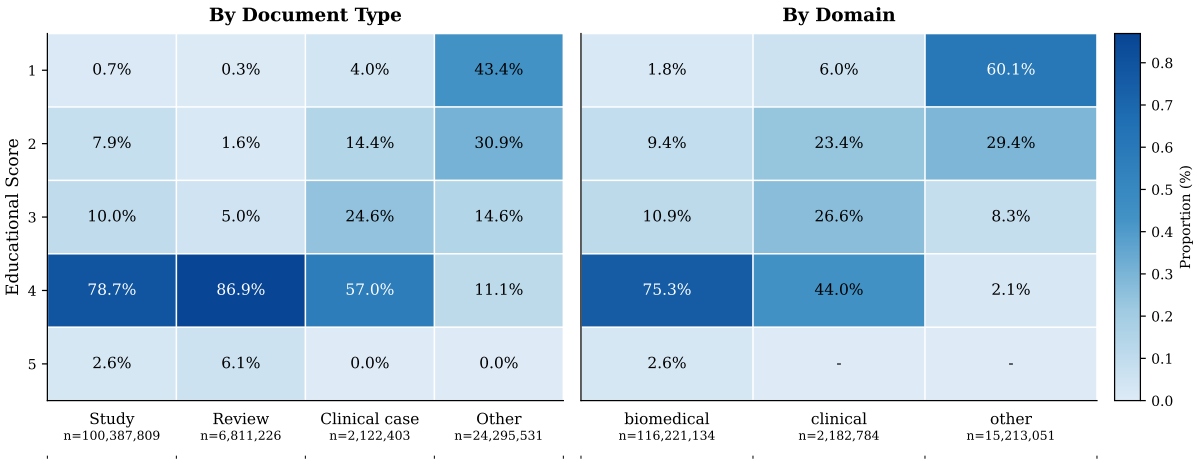

Figure 2: Distribution of educational quality scores by document type and domain. Reviews and studies show the highest proportion of high scores, while clinical texts display more variance.

	Single	Mixed
Document Type	8.4%	91.6%
Domain	44.7%	55.3%
Educational Score (1–5 scale)		
	Mean	**Median**
Intra-article std	0.68	0.70
Intra-article range	2.61	2.86

Table 1: Intra-document heterogeneity in 10K PMC articles (378K paragraphs). 91.6% of articles contain multiple document types. Educational scores vary by 2.6 points on average within articles, spanning more than half the 1–5 scale.

Label	Description
Document Type:	
Clinical Case	Report of symptoms, diagnosis, and treatment of individual patients
Study	Research with methods, results, and discussion
Review	Summary of current knowledge on a topic
Other	Editorials, commentaries, policy
Domain:	
Clinical	Patient care, clinical trials, case reports
Biomedical	Scientific aspects of medicine and biology
Other	Administrative, policy, general communications

Table 2: Annotation dimensions for document type and domain. Educational quality uses a 1-5 scale from basic (1) to outstanding pedagogical value (5).

biomedical corpora such as PubMed abstracts and PMC Open Access full-text articles to enhance domain knowledge in LLMs. BioMistral (Labrak et al., 2024), for instance, underwent continual pretraining on 3 billion tokens from the PMC Open Access Subset, while Meditron (Chen et al., 2023) fine-tuned Llama-2 on 46 billion tokens comprising PubMed abstracts, full papers, a general domain replay dataset, and clinical guidelines. Similarly, PMCLlama (Wu et al., 2024) processed 75 billion tokens from PMC Open Access and medical textbooks, achieving significant improvements on biomedical benchmarks.

However, this process is compute-intensive for a moderate increase in performance. Meditron-70B (Chen et al., 2023) required 128 A100 GPUs for 332 hours to achieve an average accuracy improvement of 1.8 percentage points on biomedical benchmarks, while BioMistral-7B (Labrak et al., 2024) used 32 A100 GPUs for 20 hours, resulting in a 0.9-point performance decrease initially, but a 2.9-point improvement when using ensembling with different merging techniques with the original model.

PMC Open Access contains significant diversity and heterogeneity. Researchers typically employ filtering and upsampling strategies to better control training data composition. For instance, BioMistral (Labrak et al., 2024) noted that 98.75% of PMC Open Access articles are in English, leading them to upsample non-English articles. Meditron (Chen et al., 2023) focused on high-quality, clinically relevant research by scoring articles (0–1) using MeSH tags, publication type, journal reputation, recency, and citation count. They filtered out low-scoring content while increasing the representation of higher-scoring articles. Notably, their ablation (Section 7.2) shows this document-level scheme

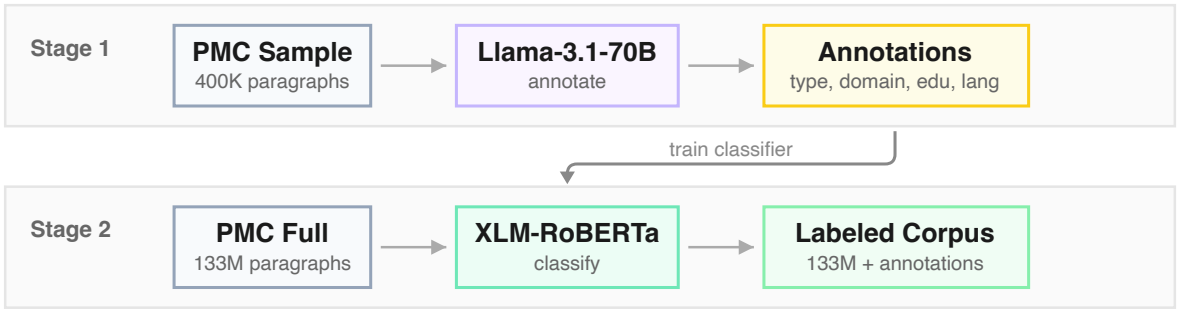

Figure 3: Two-stage annotation pipeline. Stage 1: Llama-3.1-70B annotates 0.33% of PMC paragraphs across four dimensions. Stage 2: XLM-RoBERTa distills these annotations to label the full corpus, enabling flexible filtering strategies.

underperforms unfiltered PMC, which we attribute to the intra-document heterogeneity reported in Table 1. Unlike web-scraped data where low-quality sources justify discarding entire documents, scientific articles are all potentially valuable. The real issue is variance within articles: as shown in Table 1, 91.6% of PMC articles mix multiple document types and educational scores vary by 2.6 points on average within articles (on a 1–5 scale). Some paragraphs contain excellent educational or clinical content, while others add little domain knowledge but remain difficult to predict, wasting compute during training. Unlike boilerplate sections that can be removed with simple heuristics, these low-value paragraphs require semantic understanding to identify.

More sophisticated filtering approaches have emerged to enhance pre-training data quality. While basic heuristic filtering using rules and perplexity scores from small language models trained on Wikipedia showed improvements in language modeling, LLM-based semantic quality filtering has proven substantially more effective. FineWeb-Edu (Penedo et al., 2024) demonstrated the efficacy of model-based filtering by using Llama-3-70B-Instruct to annotate 500K documents from the FineWeb corpus based on educational value on a scale of 1 to 5. They then trained a smaller BERT-like model on these annotations and applied it to the entire FineWeb corpus, filtering out samples with scores below 3. Despite removing 92% of the initial dataset, this refined subset outperformed both the complete FineWeb corpus and other open web datasets on knowledge-intensive benchmarks like MMLU (Hendrycks et al., 2021), ARC (Clark et al., 2018), and OpenBookQA (Mihaylov et al., 2018). This document-level filtering is well-suited for web corpora where entire pages from low-quality sources can be discarded, but is less appropriate for scientific literature where articles should be kept while selectively filtering their paragraphs.

WebOrganizer (Wettig et al., 2025) takes a complementary approach by organizing web content into structured taxonomies based on both topic and format. Rather than focusing solely on quality metrics, it unpacks monolithic web corpora into well-defined categories by distilling annotations from large language models into efficient classifiers. This systematic organization enables more refined data mixing strategies that improve model performance on downstream tasks. Importantly, their work demonstrates that domain-based organization provides valuable complementary benefits to quality-based filtering methods, as the two approaches can be combined to further enhance performance.

3 Contributions

In our work, we develop a more refined approach for biomedical dataset curation through Biomed-Enriched, which applies LLM-driven annotation at the paragraph level rather than at the document level. Building on techniques from FineWeb-Edu (Penedo et al., 2024) and WebOrganizer (Wettig et al., 2025), we focus on biomedical content from PMC Open Access, creating rich metadata about paragraph type, domain, educational quality, and language. Since scientific articles mix high-value and low-value content within the same document, we filter at the paragraph level: upsampling high-value content (clinical cases, educational passages) while filtering out low-information segments that

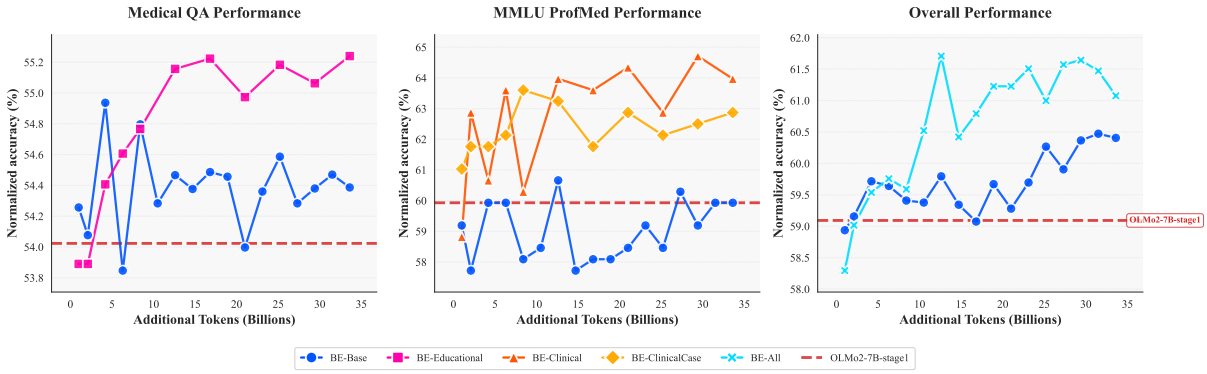

Figure 4: Performance comparison across dataset variants showing training progression. BE-All achieves target performance with approximately one-third of the training tokens required by BE-Base. The OLMo2-7B-stage1 horizontal line marks the starting checkpoint before continual pretraining; all BE variants receive 33.6B additional tokens with identical hyperparameters.

are difficult to detect with simple heuristics but add little domain knowledge during training. We use a two-stage annotation process based on a smaller follow-up model to efficiently process the entire corpus. The detailed annotation allows us to extract valuable subsets, particularly clinical case content, offering a public alternative to data typically restricted due to privacy concerns. Our experiments show that this targeted data curation substantially improves efficiency in biomedical pre-training, resulting in faster convergence and enhanced performance on domain-specific tasks.

Our contributions are:

- We release Biomed-Enriched: a paragraph-level annotated PMC corpus with 2M clinical case paragraphs extracted from published case reports, offering publicly accessible clinical case content without data use agreements. We also release the annotation classifier and our best encoder model.
- Our combined curation strategy (filtering + upsampling) matches standard continual pretraining using one-third of training tokens.
- We match BioClinical-ModernBERT on 11 clinical and biomedical tasks (77.3% vs 77.1% F1) using 2.5× fewer tokens and only public PubMed/PMC data.

4 Method

We present Biomed-Enriched, a biomedical text dataset for enhanced biomedical training constructed through paragraph-level annotation and filtering. Our approach addresses the limitations of existing article-level filtering strategies by enabling more granular selection of high-value content. This is particularly relevant for clinical text which is traditionally difficult to access due to privacy constraints.

4.1 Data Collection and Preprocessing

We extracted text from the PubMed Central (PMC) Open Access Subset (National Library of Medicine, 2024), containing approximately 4.5 million full-text scientific articles. This corpus, while valuable, presents challenges including heterogeneous quality, predominance of English content ($\approx$ 98%), uneven representation of clinical cases, and variable educational value. Using a custom pipeline, we processed the raw XML files to extract article content, segment articles into 133M individual paragraphs, filter out non-textual elements, and retain only paragraphs containing a minimum of 64 tokens.

4.2 Two-Stage Annotation Framework

We follow a two-stage annotation process on a single 8×A100 node: 160 GPU-hours for the first stage and 80 GPU-hours for the second.

Small scale annotation with a LLM First, we used Llama-3.1-70B-Instruct (Touvron et al., 2024) to annotate a diverse subset of 400,000 paragraphs from the PMC Open Access ($\approx$ 0.33%) corpus across multiple dimensions: type classification, domain categorization, and educational quality (1-5 scale). Language was identified separately using langdetect. The annotation dimensions are described in Table 2.

Dimension	Agreement
Educational score	$r = 0.876$, MAE $= 0.326$
Domain	F1 $= 0.954$
Document type	F1 $= 0.925$, $\kappa = 0.821$

Table 3: Distillation quality on held-out paragraphs.

Large scale annotation with a fine-tuned SLM To scale annotation to the full corpus, we distilled the LLM annotations into a smaller XLM-RoBERTa-base model (Conneau et al., 2020) trained to jointly predict all annotation dimensions. On 39,800 held-out paragraphs, the student closely matches the Llama-3.1-70B labels (Table 3). It is important to note that these annotations are not gold-standard labels verified by medical experts, but neural heuristics for data curation. Similar to FineWeb-Edu (Penedo et al., 2024), they guide which content to prioritize during pretraining.

4.3 Dataset Construction and Filtering

Using the distilled model, we annotated the full corpus and constructed several dataset variants through strategic filtering and upsampling. Upsampling refers to duplicating articles in the training corpus, proportionally increasing their sampling probability.

BE-Base: The complete unmodified PMC Open Access Subset serving as baseline.

BE-Educational: Preserves all articles but removes paragraphs with educational quality scores below 3.

BE-Clinical: Upsamples 10x articles with predominantly clinical domain content.

BE-ClinicalCase: Upsamples 10x articles containing at least one clinical case paragraph.

BE-Prefix: Prefixes each paragraph with its predicted annotations to allow modeling of metadata-content relationships.

BE-French: Upsamples 10x articles containing French text.

BE-All: Combines quality filtering ($score \geq 3$), upsampling of clinical content, French text, and clinical cases, plus metadata prefixing.

For all variants, we preserved the original structure of the article. To maintain the contextual relationships between paragraphs within scientific articles, we employed an 8K context window during pre-training. This approach ensures that models can process complete scientific articles, allowing them to capture dependencies where information presented in earlier paragraphs is essential for properly understanding later content.

5 Data Analysis

Most PMC paragraphs receive an educational score of 4, with a mean of 3.48 and median of 4.00. This distribution varies by content type: reviews and studies contain the highest proportion of educational content (86.9% and 78.7% scoring 4, respectively), while clinical cases show more variance (57.0% rated 4). Domain-wise, biomedical paragraphs score higher (75.3% at score 4) than clinical text (44.0%), and content labeled "other" rarely reaches high scores (2.1%). These patterns support the score threshold (≥ 3) used in BE-Educational and BE-All, and explain why combining domain and quality filters yields consistent gains across tasks.

5.1 Continual Pre-training

Continual pre-training served as a method to evaluate the relevance and utility of our annotations. Our evaluation focuses on isolating the effects of data curation rather than pursuing state-of-the-art scores on benchmarks. A more powerful foundation model would likely yield higher absolute scores, but would probably obscure the precise impact of our dataset.

We selected OLMo2-7B-stage1 (OLMo et al., 2025) as our foundation model for continual pre-training, strategically choosing this intermediate checkpoint to more clearly attribute performance changes to our data curation. Although stage 1 has already developed strong language modeling capabilities, it precedes the knowledge-intensive tuning of stage 2, providing an ideal balance of baseline capabilities without the risk of catastrophic forgetting of instruction-following abilities during domain adaptation. Notably, the data mix used in stage 1 includes DCLM (Li et al., 2024), which is a dataset obtained by filtering web-data using a classifier trained on instruct-data. Hence, OLMo2-7B already has relatively strong question-answering capabilities after stage 1.

Each Biomed-Enriched variant was trained with the same amount of tokens, namely exactly 33.6 billion tokens, using identical hyperparameters (as shown in Table 6). We follow the annealing strategy of OLMo2 (OLMo et al., 2025) used in the mid-training phase. By maintaining strict parameter parity across experiments, we created a controlled

		Medical QA			MMLU Medical						
	Tokens	**MedQA**	**MedMCQA**	**PubMedQA**	**Anat**	**Clin**	**Bio**	**Med**	**Gen**	**Prof**	**Avg**
SOTA open-source models (for reference)											
Llama-3-8B	∼15T	59.70	57.47	74.80	68.89	74.72	78.47	61.85	83.00	70.22	69.90
Meditron-70B	2T + 48B	57.10	46.80	76.60	53.30	66.70	76.30	63.00	69.00	71.60	64.49
Benchmark Results by Dataset Variant											
OLMo2-7B-stage1	∼4T	45.33	41.14	75.60	54.81	63.40	69.44	53.18	69.00	59.93	59.09
+ BE-Base	4T + 33.6B	44.85	41.91	76.40	57.04	64.15	**70.83**	**59.54**	69.00	59.93	60.41
+ BE-Clinical	4T + 33.6B	41.95	39.35	76.60	53.33	63.40	65.28	58.38	66.00	**63.97**	58.70
+ BE-ClinicalCase	4T + 33.6B	42.11	39.52	76.60	57.04	64.91	66.67	**59.54**	69.00	62.87	59.81
+ BE-Prefix	4T + 33.6B	45.72	41.76	**77.80**	57.04	64.53	68.75	57.23	66.00	61.76	60.07
+ BE-Educational	4T + 33.6B	45.64	**43.08**	77.00	57.04	65.28	68.06	56.65	**71.00**	58.82	60.29
+ BE-All	4T + 33.6B	**47.21**	42.79	76.60	60.00	**65.66**	68.06	58.96	69.00	61.40	61.08
+ BE-All	4T + 12.6B	**47.21**	42.94	76.4	**64.44**	64.53	68.75	57.23	**71.00**	62.87	**61.71**

Note: MMLU abbreviations: Anat=Anatomy, Clin=Clinical Knowledge, Bio=College Biology, Med=College Medicine, Gen=Medical Genetics, Prof=Professional Medicine.

Table 4: Comprehensive performance results across medical QA benchmarks for different dataset enrichment strategies.

Parameter	**Value**
Peak learning rate	5e-4
Minimal LR	5e-7 (α_f=0.001)
LR Decay	one_minus_sqrt
Batch size	576
Weight decay	1e-5
Context length	8,192
Hardware	4× H100 (1 node)
Training tokens	70B

Table 5: Encoder hyperparameters for continual pretraining.

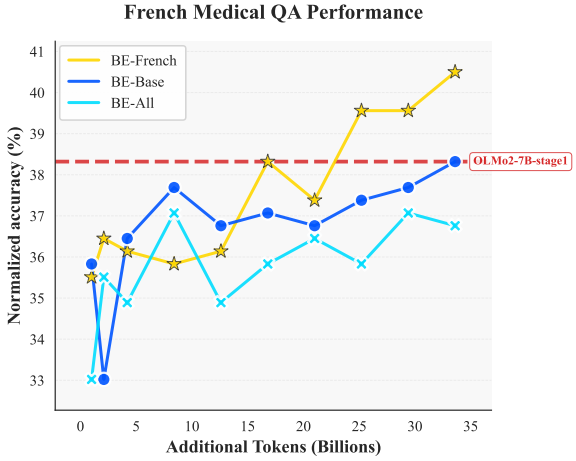

Figure 5: Performance on FrenchMedMCQA, with BE-French outperforming other variants, demonstrating effective language-specific improvement.

environment focused solely on measuring the effectiveness of our different data curation strategies.

We also conducted continual-pretraining experiments on encoder models. Starting from ModernBERT-base (Warner et al., 2024) after its context extension phase, we followed exactly the same hyperparameters as BioClinical-ModernBERT phase 1 (Table 5). Our baseline reproduces their standard continual pretraining setup on public PubMed/PMC. Our clinical-upsample recipe additionally applies educational filtering ($score \geq 3$) and upsamples clinical cases 100× and clinical domain content 10×.

5.2 Evaluation Framework

We evaluate how our annotation-guided corpus refinement affects model performance during continual pre-training. We measure performance at regular intervals throughout training on several biomedical benchmarks to understand how effectively models acquire domain knowledge from differently curated datasets.

Our evaluation consists of zero-shot testing on

Parameter	**Value**
Peak learning rate	6.15e-5
Minimal LR	6.15e-6
LR Decay	Linear
Batch size	1024
Weight decay	0.1
Context length	8,192 tokens
Hardware	128 MI250X GPUs
Training time (hours / GPU-hours)	68 / 8700

Table 6: Hyperparameters used for continual pretraining.

MMLU medical subcategories, MedQA (Jin et al., 2021), MedMCQA (Pal et al., 2022), and PubMedQA (Jin et al., 2019). For the French adaptation assessment, we use 5-shot evaluation on FrenchMedMCQA (Labrak et al., 2022). We compare our models against three baselines: OLMo2-7B-stage1, Llama-3-8B, and Meditron-70B.

For the encoder models, we evaluate on 11 tasks spanning clinical and biomedical domains. Clinical tasks include ChemProt (Krallinger et al., 2017), Phenotype (Moseley et al., 2020), COS (Klassen et al., 2014), SocialHistory (Yetisgen and Vanderwende, 2017), and DEID (Neamatullah et al., 2008). Biomedical tasks include AnatEM (Pyysalo and Ananiadou, 2014), BC5CDR (Li et al., 2016), JNLPBA (Collier and Kim, 2004), NCBI Disease (Doğan et al., 2014), GAD (Bravo et al., 2015), and Hallmarks of Cancer (Baker et al., 2016). Following the evaluation recipe of BioClinical-ModernBERT, we finetune for 10 epochs on clinical tasks and 20 epochs on biomedical tasks, with early stopping patience of 3 epochs for all. We compare against BioClinical ModernBERT (Sounack et al., 2025), which was trained on 169B tokens from PubMed, PMC, and 20 clinical datasets including MIMIC-III (Johnson et al., 2016), MIMIC-IV (Johnson et al., 2023), CheXpert Plus radiology reports (Chambon et al., 2024), and various clinical notes from multiple institutions, most requiring data use agreements.

6 Results

Table 4 presents decoder results on medical QA benchmarks. The combined strategy BE-All achieves the best average performance (61.08%), with the strongest gains on MedQA (+2.4 pts) and MMLU Clinical Knowledge (+1.5 pts). Different enrichment strategies show complementary strengths: clinical upsampling yields a +4 point improvement on MMLU Professional Medicine, while educational filtering improves Medical Genetics by +2 points and MedMCQA by +1.2 points. As shown in Figure 4, BE-All reaches target performance using roughly one-third of the training tokens compared to BE-Base, with stable improvements visible from early checkpoints. BE-French also demonstrates successful adaptation to French medical QA (40.5% vs 38.3% baseline, Figure 5), showing that language-specific upsampling generalizes beyond English.

Table 7 presents encoder results across 11 clinical and biomedical benchmarks. Our BE-Clinical recipe achieves 77.08% average F1, matching BioClinical-ModernBERT (77.05%) while using 2.5× fewer tokens and only public PubMed/PMC data. Adding a decay phase with high-quality educational content brings slight additional gains: the Biomedical decay variant (edu≥4) achieves 77.32%. Different decay targets show task-specific strengths: Study decay yields the best results on AnatEM (79.7%) and NCBI Disease (81.2%), Review decay excels on GAD (80.2%), and Clinical Case decay achieves the highest SocialHistory score (57.0%). Figure 6 shows that clinical-upsample reaches higher performance faster on clinical tasks with more stable training. These results suggest that public PMC data has more potential for clinical NLP than commonly exploited, and that targeted curation can close the gap with models trained on restricted datasets.

7 Discussion

The core advantage of paragraph-level annotation lies in its ability to identify valuable content that would be missed by document-level approaches. Clinical case descriptions, for instance, often appear in isolated sections of broader scientific articles. By operating at the paragraph level, we extracted 2 million clinical case paragraphs from PMC Open Access. This content is otherwise difficult to obtain at scale due to privacy restrictions on hospital records.

For decoders, our results reveal clear task-specific benefits from different enrichment strategies. Educational filtering improves knowledge-intensive QA tasks, clinical upsampling enhances clinical reasoning benchmarks, and the combination of both yields the best overall performance with faster convergence. The successful French adaptation through targeted upsampling further demonstrates that this framework generalizes beyond English. However, aggressive clinical upsampling can reduce performance on general biomedical tasks like College Biology, suggesting a trade-off between specialization and breadth. Our decoder experiments used 7B parameter models, and larger models may respond differently to data curation choices.

For encoders, the decay phase experiments reveal that different paragraph types benefit different downstream tasks: Study paragraphs improve NER performance, Clinical Case paragraphs improve so-

	Tokens	Clinical Tasks					Biomedical Tasks						Avg
		ChemPr Cls	Pheno Cls	COS NER	Social NER	DEID NER	AnatEM NER	BC5CDR NER	JNLPBA NER	NCBI NER	GAD Cls	HoC Cls	
Base Model													
ModernBERT-base (Warner et al., 2024)	2T	89.5	48.4	94.0	53.1	78.3	77.2	87.9	74.3	77.7	76.8	66.6	74.89
Reference Model													
BioClinical-ModernBERT	2T + 169B†	90.0	**60.7**	94.8	56.0	**81.8**	**79.2**	88.7	74.8	78.7	75.8	67.0	77.05
Our Models (Public Data Only)													
Baseline	2T + 70B	89.8	59.6	94.5	55.5	78.5	78.8	88.7	74.8	**81.3**	76.2	67.0	76.79
BE-Clinical	2T + 67B	89.9	59.9	94.8	56.8	79.7	78.5	88.1	74.6	79.4	77.7	68.5	77.08
With Decay Phase													
+ Review (edu≥3)	2T + 67B	**90.3**	59.7	94.6	55.4	79.6	78.5	88.5	74.7	79.6	**80.2**	68.8	77.26
+ Study (edu≥3)	2T + 67B	**90.3**	**60.3**	94.7	55.6	78.6	**79.7**	88.4	**74.8**	81.2	78.7	67.7	77.27
+ Clinical Case (edu≥3)	2T + 67B	89.5	59.4	94.5	**57.0**	79.3	78.7	88.3	74.7	80.0	77.9	68.1	77.04
+ Biomedical (edu≥4)	2T + 67B	90.1	59.9	**94.9**	56.4	79.6	79.0	**88.6**	74.7	79.9	78.4	**69.0**	**77.32**

† Trained on 169B tokens including 20 clinical datasets (MIMIC-III/IV, CheXpert, radiology reports, etc.). Our models use only public PubMed/PMC.
Abbreviations: ChemPr=ChemProt, Pheno=Phenotype, Social=SocialHistory, Cls=Classification, NER=Named Entity Recognition.
Significance (paired t-tests over 5 seeds vs. Baseline): Study decay improves AnatEM +0.8 (p=0.002), ChemProt +0.5 (p=0.018), GAD +2.4 (p=0.032); Review decay improves GAD +4.1 (p=0.023); Biomedical decay improves ChemProt +0.3 (p=0.047), HoC +2.0 (p=0.049); Clinical Case decay improves SocialHistory +1.5 (p=0.086).

Table 7: Encoder results across 11 clinical and biomedical benchmarks (mean over 5 seeds). Our best decay variant matches BioClinical-ModernBERT (77.3% vs. 77.1%) using 2.5× fewer tokens and only public PubMed/PMC data. Different decay targets show task-specific strengths: Study decay excels on AnatEM and NCBI, Review on GAD, Clinical Case on SocialHistory.

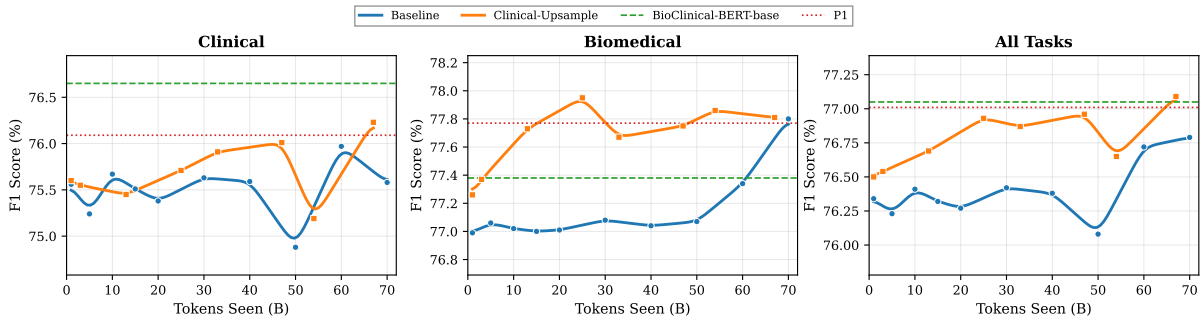

Figure 6: Encoder training curves on clinical, biomedical, and combined benchmarks. Our clinical-upsample recipe (orange) reaches higher performance faster and more stably on clinical tasks, and also outperforms baseline (blue) on biomedical tasks until 70B tokens. At 67B tokens, we match BioClinical-ModernBERT P1 (green dashed), their phase 1 checkpoint before decay on clinical data, which required 169B tokens and 20 clinical datasets.

cial history extraction, and Review paragraphs excel on relation extraction. The distinct task-specific strengths of each decay variant suggest a promising direction: model merging or checkpoint averaging could combine the benefits of specialized training runs without requiring additional compute. Rather than training general-purpose models on massive undifferentiated corpora, these findings support a more modular approach where data composition is tailored to downstream requirements. In practice, practitioners can reuse the released annotations and classifier to compose training mixes tailored to their needs, or apply the same pipeline to new corpora with their own annotation dimensions.

8 Conclusion

We labeled PMC paragraphs for type, domain, and educational quality using Llama-3.1-70B (400K paragraphs) and a fine-tuned XLM-RoBERTa (full corpus). From these annotations, we extracted 2M clinical case paragraphs and constructed dataset variants for different downstream needs.

Our experiments show consistent gains on both encoder and decoder architectures. For decoders, combining quality filtering with clinical upsampling matches standard continual pretraining using one-third of training tokens, with task-specific benefits: +4 points on MMLU Professional Medicine from clinical upsampling, +2 points on Medical Genetics from educational filtering. For encoders, we match BioClinical-ModernBERT on 11 clinical and biomedical tasks (77.3% vs 77.1%) using 2.5× fewer tokens and only public PubMed/PMC data. Adding a decay phase with paragraph-type-specific data reveals task-specific benefits: Study paragraphs improve NER tasks, Clinical Case paragraphs improve SocialHistory extraction. The French upsampling results suggest this approach generalizes to other languages.

Paragraph-level curation of public PMC data of-

fers a reproducible alternative to restricted clinical datasets, suggesting that strategic data selection matters as much as data access for domain adaptation.

Acknowledgments

This work was granted access to the HPC resources of IDRIS under the allocation 2025-AD011014393R2 made by GENCI.

9 Limitations

Our decoder experiments used 7B parameter models. Larger models may respond differently to data curation choices, and our findings may not directly transfer to other model scales. The two-stage annotation pipeline relies on a relatively small classifier (XLM-RoBERTa-base), which may have limited capacity compared to larger encoder models. Our clinical case paragraphs come from published case reports, which may differ stylistically from actual hospital records. The educational quality scores are neural heuristics for data curation rather than gold-standard pedagogical assessments. Finally, our evaluation focuses primarily on English benchmarks, with only preliminary results on French, leaving the generalization to other languages as future work.

References

Simon Baker, Ilona Silins, Yufan Guo, Imran Ali, Johan Högberg, Ulla Stenius, and Anna Korhonen. 2016. Automatic semantic classification of scientific literature according to the hallmarks of cancer. *Bioinformatics*, 32(3):432–440.

Àlex Bravo, Janet Piñero, Núria Queralt-Rosinach, Michael Rautschka, and Laura I Furlong. 2015. Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research. *BMC Bioinformatics*, 16:55.

Pierre Chambon, Jean-Benoit Delbrouck, Thomas Sounack, Shih-Cheng Huang, Zhihong Chen, Maya Varma, Steven QH Truong, Chu The Chuong, and Curtis P Langlotz. 2024. Chexpert plus: Augmenting a large chest x-ray dataset with text radiology reports, patient demographics and additional image formats. *arXiv preprint arXiv:2405.19538*.

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*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Nigel Collier and Jin-Dong Kim. 2004. Introduction to the bio-entity recognition task at JNLPBA. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP)*, pages 73–78. COLING.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. Ncbi disease corpus: A resource for disease name recognition and concept normalization. *Journal of Biomedical Informatics*, 47:1–10.

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 (ICLR)*.

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 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 (EMNLP-IJCNLP)*, pages 2567–2577. Association for Computational Linguistics.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-Wei H Lehman, Leo A Celi, and Roger G Mark. 2023. Mimic-iv, a freely accessible electronic health record dataset. *Scientific Data*, 10:1.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-Wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3:160035.

Prescott Klassen, Fei Xia, Lucy Vanderwende, and Meliha Yetisgen. 2014. Annotating clinical events

in text snippets for phenotype detection. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2753–2757, Reykjavik, Iceland. European Language Resources Association (ELRA).

Martin Krallinger, Obdulia Rabal, Analia Lourenco, and 1 others. 2017. Overview of the biocreative vi chemical-protein interaction track. In *Proceedings of the BioCreative VI Workshop*, pages 141–146.

Yanis Labrak, Adrien Bazoge, Richard Dufour, Beatrice Daille, Pierre-Antoine Gourraud, Emmanuel Morin, and Mickael Rouvier. 2022. FrenchMedMCQA: A French multiple-choice question answering dataset for medical domain. In *Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI)*, pages 41–46, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. BioMistral: A collection of open-source pretrained large language models for medical domains. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5848–5864, Bangkok, Thailand. Association for Computational Linguistics.

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Gadre, Hritik Bansal, Etash Guha, Sedrick Keh, Kushal Arora, and 1 others. 2024. Datacomp-lm: In search of the next generation of training sets for language models. In *Advances in Neural Information Processing Systems*, volume 37.

Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016:baw068.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.

Edward T. Moseley, Joy T. Wu, Jonathan Welt, John Foote, Patrick D. Tyler, David W. Grant, Eric T. Carlson, Sebastian Gehrmann, Franck Dernoncourt, and Leo Anthony Celi. 2020. A corpus for detecting high-context medical conditions in intensive care patient notes focusing on frequently readmitted patients. In *Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC 2020)*, pages 1362–1367. European Language Resources Association.

National Library of Medicine. 2024. PMC open access subset. `https://www.ncbi.nlm.nih.gov/pmc/tools/textmining/`. Accessed 2025-03-29.

Ishna Neamatullah, Margaret Douglass, Li-Wei Lehman, Andrew Reisner, Mauricio Villarroel, William J Long, Peter Szolovits, George B Moody, Roger G Mark, and Gari D Clifford. 2008. Automated de-identification of free-text medical records. *BMC Medical Informatics and Decision Making*, 8:32.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, and 21 others. 2025. 2 olmo 2 furious. *Preprint*, arXiv:2501.00656.

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*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben Allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro von Werra, and Thomas Wolf. 2024. The fineweb datasets: Decanting the web for the finest text data at scale. *arXiv preprint arXiv:2406.17557*.

Sampo Pyysalo and Sophia Ananiadou. 2014. Anatomical entity mention recognition at literature scale. *Bioinformatics*, 30(6):868–875.

Thomas Sounack, Joshua Davis, Brigitte Durieux, Antoine Chaffin, Tom J. Pollard, Eric Lehman, Alistair E. W. Johnson, Matthew McDermott, Tristan Naumann, and Charlotta Lindvall. 2025. Bioclinical modernbert: A state-of-the-art long-context encoder for biomedical and clinical nlp. *Preprint*, arXiv:2506.10896.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *Preprint*, arXiv:2412.13663.

Alexander Wettig, Kyle Lo, Sewon Min, Hannaneh Hajishirzi, Danqi Chen, and Luca Soldaini. 2025. Organize the web: Constructing domains enhances pre-training data curation. *arXiv preprint arXiv:2502.10341*.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. 2024. PMC-LLaMA: toward building open-source language models for

medicine. *Journal of the American Medical Informatics Association*, 31(9):1833–1843.

Meliha Yetisgen and Lucy Vanderwende. 2017. Automatic identification of substance abuse from social history in clinical text. In *Artificial Intelligence in Medicine: 16th Conference on Artificial Intelligence in Medicine, AIME 2017*, pages 171–181. Springer.

Appendices

A Annotation Prompt

Below is the prompt used to instruct Llama-3.1-70B-Instruct for the first stage of our annotation process:

```
Below is an extract from a scientific article. Evaluate whether the extract has a high educational value
     and could be useful in an educational setting for teaching at the college level in biomedical sciences
     using the additive 5-point scoring system described below.

Points are accumulated based on the satisfaction of each criterion:
- Add 1 point if the extract provides some basic information relevant to biomedical topics, even if it
     includes some irrelevant or non-academic content like advertisements and promotional material.
- Add another point if the extract addresses certain elements pertinent to biomedical education but does
     not align closely with academic standards. It might mix educational content with non-educational
     material, offering a superficial overview of potentially useful topics, or presenting information in a
     disorganized manner and incoherent writing style.
- Award a third point if the extract is appropriate for educational use and introduces key concepts
     relevant to college-level biomedical curricula. It is coherent though it may not be comprehensive or
     could include some extraneous information. It may resemble an introductory section of a textbook or a
     basic tutorial that is suitable for learning but has notable limitations like treating concepts that
     are too complex for introductory students.
- Grant a fourth point if the extract is highly relevant and beneficial for educational purposes at the
     college level, exhibiting a clear and consistent writing style. It could be similar to a chapter from
     a textbook or a tutorial, offering substantial educational content, including exercises and solutions,
     with minimal irrelevant information, and the concepts are appropriate for college students. The
     content is coherent, focused, and valuable for structured learning.
- Bestow a fifth point if the extract is outstanding in its educational value, perfectly suited for
     teaching at the college level in biomedical sciences. It follows detailed reasoning, the writing style
     is easy to follow and offers profound and thorough insights into the subject matter, devoid of any
     non-educational or overly complex content.

Based on these factors, give a score from 1 to 5.

Also, classify the relevant domain as either "biomedical", "clinical", or "other" following the guidelines
     provided below:
- Clinical: Extract appears to be written in a clinical context by a healthcare professional. It should
     contain information directly related to patient care, such as details from clinical trials, case
     reports, or clinical guidelines.
- Biomedical: Extract contains substantive information on biomedical sciences. It could be from a research
     paper or textbook, focusing on the scientific aspects of medicine and biology.
- Other: Extract mentions biomedical or clinical topics but doesn't provide substantive content in these
     areas. This category includes:
  1. Administrative or funding information about biomedical research
  2. General news or public communications about medical topics
  3. Policy discussions related to healthcare
  4. Any content that talks about biomedical or clinical subjects without providing actual scientific or
       medical information

Additionally, identify the type of document, this category includes:
1. Clinical case: A detailed report of the symptoms, signs, diagnosis, treatment, follow-up, etc. of an
     individual patient.
2. Study: Research-based document that includes methods, results, and discussions about experiments or
     observations, often involving multiple subjects or data points.
3. Review: A document that summarizes or evaluates the current state of knowledge on a specific topic.
4. Other: Any other type of document not fitting the above categories.

After examining the extract:
- Briefly justify your quality classification, up to 100 words on one line using the format: "Explanation:
     <justification>"
- Conclude with the quality classification using the format: "Educational score: <classification>"
- Conclude with the domain classification using the format: "Domain: <classification>"
- Conclude with the document type classification using the format: "Document type: <classification>"
```