SciPedia: Unlocking the Value of Scientific Data for Pre-training

Yiwei Qin[1,2,4*] **Zhen Huang**[1,3,4*] **Tiantian Mi**[1,3,4*] **Weiye Si**[1,2,4]
Qipeng Guo[1] **Siyuan Feng**[1] **Pengfei Liu**[1,2,4†]

[1]Shanghai Innovation Institute [2]Shanghai Jiao Tong University
[3]Fudan University [4]GAIR
qinyiwei07@outlook.com, huangz25@m.fudan.edu.cn
tiantianmi2003@gmail.com, liupf@sjtu.edu.cn

Abstract

High-quality scientific data is critical for advancing LLMs, yet academic literature remains largely underutilized. This work addresses the fundamental question: *How can we systematically unlock scientific data's value for pre-training?* First, we construct a large-scale raw scientific corpus but identify a critical *Learnability Gap*, revealing that direct pre-training yields negligible gains. To bridge this, we develop a multi-stage pipeline featuring content cleaning and pedagogical augmentation, resulting in SciPedia, a 900B-token corpus. Finally, we establish a controlled verification framework: we develop SciPedia-Eval benchmark and conduct 600B tokens of continued pre-training (CPT) starting from transparent base models (3B/7B) trained from scratch. Compared to a CPT baseline trained with general-purpose data, our approach with SciPedia data boosts average performance by +2.12 (3B) and +2.95 (7B), reaching +5.60 and +8.40 on in-domain tasks. This setup further allows us to derive empirical guidelines for data composition and model configurations. [1]

1 Introduction

While Large Language Models (LLMs) have demonstrated impressive general capabilities, their progress toward rigorous scientific discovery remains constrained (Zhao et al., 2023; Hu et al., 2025). As web-scale data growth slows and high-quality signal becomes harder to obtain, it is increasingly critical to extract value from the highest-quality sources available (e.g., academic textbooks and research papers) (Luo et al., 2025).

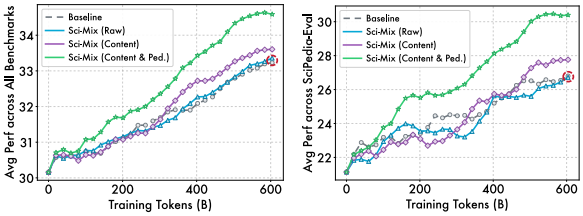


Figure 1: Comparison of training effectiveness across different data processing strategies.

However, unlike the readily consumable web, scientific literature presents unique systemic barriers, rendering it a largely untapped frontier in the current open pretraining landscape.

Consequently, existing efforts to utilize scientific data remain fragmented. Current public corpora are predominantly composed of web-filtered subsets (Penedo et al., 2023, 2024a; Soldaini et al., 2024; Paster et al., 2023) or restricted to narrow domains (Wang et al., 2024b; Zhou et al., 2025; Toshniwal et al., 2024), while their associated processing pipelines are typically optimized for web artifacts (Zhou et al., 2024; Su et al., 2024; Maini et al., 2024) or small-scope tasks (Du et al., 2025a), failing to address unique barriers in academic literature. Furthermore, even when such data is available (Tang et al., 2024), effective strategies for integrating it into pretraining remain largely underexplored. This necessitates a fundamental inquiry: ***How can we systematically unlock the value of scientific data for pretraining?***

To initiate this process, we first address the challenge of acquisition: ***How can we acquire scientific data at scale?*** Authentic academic books and research papers are predominantly locked in PDF documents, necessitating specialized extraction beyond standard text processing. To breach this barrier, we aggregate data directly from original sources and deploy a pipeline featuring OCR parsing and rule-based filtering to construct a large-scale raw scientific corpus. However, our diagnos-

[*] Equal contribution.
[†] Corresponding author.
[1]All artifacts are available at https://github.com/GAIR-NLP/Data-Darwinism, including the dataset, models, and benchmarks. The SciPedia dataset is also directly accessible at https://huggingface.co/datasets/GAIR/Darwin-Science.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 47112–47159
July 2-7, 2026

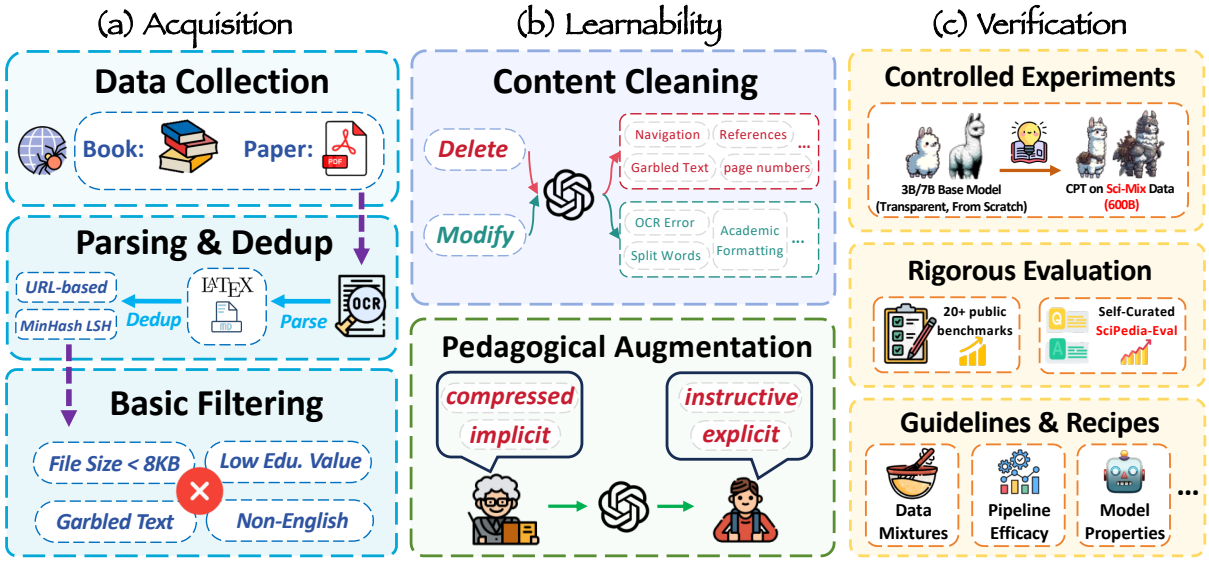


Figure 2: Overview of our work: **(a) Acquisition:** Constructing a large-scale raw corpus. **(b) Learnability:** Bridging the gap between compressed academic text and model uptake through content cleaning and pedagogical augmentation. **(c) Verification:** Deriving empirical guidelines through controlled training.

tic experiments yield a counter-intuitive finding: simply pre-training on this raw data provides negligible performance gains. This exposes a critical *Learnability Gap* stemming from the high information compression and implicit reasoning inherent in scientific literature, which renders raw content difficult for models to absorb directly. This observation aligns with recent findings from domain-specific pre-training: DeepSeekMath (Shao et al., 2024) reported that training on raw arXiv papers brought no notable improvements or even deterioration, on mathematical benchmarks, despite arXiv being widely used in math-focused corpora.

Naturally, we pivot to the challenge of learnability: ***How can we make scientific data learnable for pre-training?*** Our solution is a framework comprising two strategic phases: (1) *Content Cleaning*, which purifies the learning content by eliminating non-educational noise (e.g., metadata, OCR artifacts) and repairing structural fragmentation to isolate high-value academic content; and (2) *Pedagogical Augmentation*, which utilizes a teacher-assisted principle to let frontier LLMs bridge the gap between expert-level writing and model pre-training needs. By systematically expanding implicit logical leaps, contextualizing domain-specific terminology, and grounding abstract theories in intuitive analogies, this phase lowers the cognitive barrier for models to internalize complex scientific causality. We designate this rigorously processed corpus as SciPedia.

Finally, we confront the challenge of verification: ***How can we rigorously validate scientific data and determine optimal training recipes?*** Scientific pre-training currently lacks rigorous experimental control, often leaving it unclear whether performance gaps stem from data quality or model configurations. To disentangle these factors, we establish a controlled verification framework. We develop SciPedia-Eval to assess distribution-aligned domain knowledge and train transparent base models (3B and 7B) from scratch to ensure no prior scientific exposure. Starting from these checkpoints, we conduct 600B tokens of continued pre-training (CPT). The integration of SciPedia demonstrates robust efficacy: it outperforms a competitive baseline CPT by +2.12 (3B) and +2.95 (7B) points across 20+ diverse benchmarks, with gains amplifying to +5.60 and +8.40 on our specialized suite. Moreover, this controlled setting allows us to derive evidence-based guidelines regarding data mixtures, pipeline efficacy, and model properties (e.g., scale, context, training stage).

In summary, our contributions are three-fold:

- **The SciPedia Corpus:** We present a massive, high-quality scientific dataset comprising 900B tokens derived from academic textbooks and research papers, processed to overcome accessibility and learnability barriers.
- **The Processing Pipeline:** We open-source our comprehensive data processing pipeline,

transforming raw PDFs into high-value training data through robust extraction, content cleaning, and pedagogical augmentation.

- **The Verification Framework & Recipes:** We establish a controlled verification framework by training transparent base models from scratch and developing the SciPedia-Eval benchmark. This setup validates our approach's efficacy and provides actionable insights into data composition, pipeline design, and critical model factors.

2 Raw Data and Diagnostic

To address the challenge of *accessing high-quality scientific data*, we first detail the construction of our large-scale raw corpus (Fig. 2 (a)) and subsequently diagnose its learnability limitations.

2.1 Data Collection

Data Source Our original data corpus is constructed from two complementary sources: (1) PDF files of academic books and papers collected from publicly accessible online repositories, and (2) three scholarly paper collections from the open-source TxT360 dataset (Tang et al., 2024): PubMed Central, arXiv, and S2ORC full text. For scanned PDFs, we employ `olmOCR-7B-0225-preview` (Poznanski et al., 2025) to convert them into machine-readable text.

Data Deduplication To remove redundancy, we apply MinHash (Broder, 2000) with LSH from `datatrove` (Penedo et al., 2024b) using parameters $(n_b, n_h) = (14, 8)$ (112 hash features per document), removing 22% of documents.

Document Annotation and Filtering Following deduplication, we annotate all documents using `EAI-Distill-0.5B` (AI et al., 2025), a fine-tuned Qwen2.5-0.5B-Instruct model for 12-dimensional document classification covering aspects such as field of discipline classification (FDC), document type, and content quality. Based on these annotations, we apply a four-stage filtering pipeline: (1) **File Size**: we discard documents smaller than 8KB to remove spam and fragments; (2) **Category**: we remove non-educational content based on document type labels; (3) **Garbled Text**: we filter out documents with more than 50% garbled characters resulting from OCR errors; (4) **Language**: we retain only English documents using `fast-langdetect`. Detailed statistics and threshold justifications for each filtering stage are discussed in Appendix A.

Discipline Classification We further organize the retained documents into 9 major disciplines using FDC labels from `EAI-Distill-0.5B`: *computer science*, *medicine*, *biology*, *chemistry*, *mathematics*, *physics*, *human & social sciences*, *engineering*, and *other STEM fields*. The complete FDC mapping scheme is detailed in Appendix B.1.

Book-Paper Classification Finally, since books and papers exhibit different learnability characteristics that require different downstream processing, we classify all documents into *book* and *paper* categories. For data sources with explicit type metadata (e.g., arXiv papers, published books), we directly use the provided labels; for ambiguous cases, we employ `Qwen2.5-7B-Instruct` (Team, 2024) to determine whether each document is a book or a paper. The classification methodology and processing differences between the two categories are elaborated in Appendix B.2.

2.2 Learnability Challenge

After collecting our raw scientific corpus through OCR extraction and basic filtering, a natural question arises: can models effectively learn from this content directly? To investigate this, we compare the **Baseline** (general-purpose) with the **Sci-Mix (Raw)** (50% raw scientific content and 50% baseline mixture) trained for 600B tokens (details of evaluation and training in Sec. 4 and Sec. 5.1).

Fig. 1 reveals a surprising finding: Sci-Mix (Raw) performs comparably to the Baseline. This lack of improvement persists even on the distribution-aligned SciPedia-Eval, indicating that **raw scientific content provides minimal benefits**. This exposes a critical *Learnability Gap*: raw scientific text, despite its high conceptual density, remains largely unintelligible to the model. Consequently, simply increasing data volume is insufficient. These findings necessitate the deeper analysis of learning barriers and the systematic processing pipeline introduced in the following section.

3 Processing Pipeline

To bridge the *Learnability Gap* identified in Sec. 2.2 and address the challenge of *enhancing data learnability*, our qualitative analysis of the raw OCR corpus reveals two primary barriers: structural noise compromising data quality

and pedagogical gaps between expert-level content and model learnability. Consequently, we develop a two-stage pipeline (Fig. 2(b)): (1) **Content Cleaning** to preserve learnable signal, and (2) **Pedagogical Augmentation** to transform implicit expert logic into explicit pedagogical content. This decoupled design enables task specialization: Stage 1 removes structural noise, providing clean text for Stage 2's pedagogical synthesis.

Category	Samples (M)	Tokens (B)	Avg. Toks/Sample
Book	2.98	251.5	84 396
Content	2.98	251.5	84 396
Paper	47.81	655	13 700
Content	26.31	215	8172
Ped.	21.50	440	20 465
Total	50.79	906.5	17 848

Table 1: Dataset statistics across categories

3.1 Content Cleaning

This stage addresses the *data quality* barrier. While preliminary filtering in Sec. 2.1 ensures document-level relevance, scientific texts, particularly scanned ones, often contain non-educational noise, such as reference lists, malformed equations, and OCR-induced errors. These elements act as "distractions" during pretraining, disrupting the model's focus on core scientific logic and necessitating a process to purify the text for learning.

Approach Design To systematically address these quality issues, we develop an LLM-based cleaning approach guided by an empirical analysis (Appendix E.1). Our strategy centers on two core principles to minimize extraneous content while preserving high-value academic integrity, such as technical notation and educational materials (full prompt in Appendix E.2): **Deletion:** Removing minimal-value content such as structural elements (table of contents, references, headers/footers), non-academic artifacts (placeholders, URLs, advertisements), OCR errors (garbled text, encoding anomalies), and scanning duplications. **Modification:** Repairing formatting defects without altering semantics, such as merging fragmented text and restoring damaged formulas or tables.

Implementation We apply this cleaning pipeline to the entire OCR-processed corpus. Documents are segmented into 1,024 character chunks (≈256 token windows) and processed independently using `GPT-OSS-120B` (OpenAI, 2025), which is selected for its optimal balance of accuracy and throughput (see Appendix E.3). This granularity balances cleaning fidelity with context preservation: it is small enough to ensure strict adherence to cleaning principles, yet large enough to maintain textual coherence. The process results in a 20% reduction in corpus volume; additional implementation details and examples are provided in Appendices E.4 and E.5.

3.2 Pedagogical Augmentation

While the structural denoising described in Sec. 3.1 ensures data cleanliness, research corpus is typically written in an "Expert-to-Expert" paradigm, characterized by high information compression, implicit reasoning steps, and heavy reliance on assumed background knowledge. For a pre-training model, this creates a *understanding barrier*: the model encounters conclusions without witnessing the derivation process, leading to inefficient internalization of logic.

Principles of Pedagogical Rewriting To bridge this gap, we introduce a **Pedagogical Augmentation** strategy that harnesses rich knowledge from frontier models. We employ a pipeline designed to make implicit reasoning explicit. Specifically, the augmentation targets three key dimensions: (1) **Reasoning Reconstruction:** Expanding logical leaps (e.g., "it follows that") into step-by-step derivations, allowing the model to trace the causality between assumptions and conclusions. (2) **Terminological Explication:** Contextualizing domain-specific jargon and variable definitions within the narrative flow rather than assuming prior mastery. (3) **Pedagogical Bridging:** We ground abstract concepts in established knowledge through intuitive analogies. This involves introducing contextual bridges that link complex, isolated theoretical constructs to concrete physical examples, facilitating better concept association.

Implementation Given the high conceptual density of research papers and, we apply this augmentation exclusively to papers.[2] To ensure tractable processing while maintaining narrative consistency, we segment documents into 1,024 to-

[2]Pilot studies reveal that augmenting textbooks, which are already pedagogically structured, often produced redundant explanations with diminishing returns. Given the computational costs and these qualitative findings, we prioritize papers where learnability gains are more substantial.

ken windows (a larger window size compared to Sec. 3.1). The rewriting process is executed by `Qwen3-235B-A22B-Instruct` (Yang et al., 2025), guided by a structured prompt (see Appendix F.2) designed to strictly enforce the dimensions described above. The rewrite model selection is based on a preliminary study using an *LLM-as-a-Judge* framework (detailed in Appendix F.1).

Fidelity Evaluation Given the reliance on teacher models for content generation, we conduct systematic evaluation to ensure rewrite fidelity and control potential error propagation. We employ a two-tier validation approach: expert qualitative audits during the design phase confirmed that rewrites maintain mathematical rigor while enhancing pedagogical clarity, and systematic LLM-based evaluation on a representative sample (60 chunks across 20 papers) using a 10-point fidelity checklist across content preservation, hallucination control, and technical rigor dimensions. The evaluation yielded high average scores of 9.2/10 and 9.1/10 from two independent judge models, indicating minimal error propagation risk. Detailed evaluation methodology and results are provided in Appendix F.4.

3.3 Data Portrait

Benchmark Decontamination To ensure evaluation integrity, we decontaminate our dataset against benchmarks using exact 20-gram matching on concatenated problem-solution pairs, removing approximately 0.03% of contaminated documents.

Final Dataset Statistics The final SciPedia comprises 50M documents totaling approximately 900B tokens, with broad coverage across natural sciences, engineering, and social sciences. Detailed statistics are in Tab. 1 and Appendix B.3. We also construct SciPedia-Raw, containing 601B tokens (321B from books, 280B from papers) of original OCR-extracted text.

4 Evaluation Suite

To address the challenge of *rigorous scientific verification*, we establish a reliable testing ground via a dual strategy: a custom, distribution-aligned benchmark measuring in-domain mastery, and public benchmarks monitoring general capability. All reported average scores are computed as the simple average of accuracies across benchmarks unless otherwise specified.

4.1 SciPedia-Eval

Existing benchmarks target elementary science and lack the depth to capture the specialized knowledge enhanced by SciPedia. To address this, we introduce SciPedia-Eval, an academic benchmark designed to evaluate this specialized nature.

Construction Pipeline We generate seven-option multiple-choice questions via a four-stage pipeline. First, documents are segmented into 4096-token windows for semantic completeness. Remaining stages use `Qwen3-32B` (Team, 2025): we (1) identify knowledge points and generate questions grounded in source text rather than parametric knowledge; (2) filter questions requiring external info or referential expressions; and (3) validate that each answer has sufficient source support (prompts and samples in Appendix G.1).

Final Benchmark Following this pipeline, we construct SciPedia-Eval, comprising 140K questions from books and 10K questions from papers, all sourced from documents held out from the training data. To enable efficient evaluation during pretraining, we sample 1,500 questions from both book and paper to form two test sets: SciPedia-Eval-Book and SciPedia-Eval-Paper.

4.2 Public Benchmarks

Beyond our specialized SciPedia-Eval, we extend our evaluation to a comprehensive suite of established benchmarks to assess both broad generalization and vertical scientific competency. Since the evaluated models are base checkpoints without alignment, we employ both perplexity-based and generative strategies (see details in Appendix H).

General Capabilities To evaluate general reasoning and knowledge recall, we employ BBH (3-shot) (Suzgun et al., 2022), ARC-Easy and ARC-Challenge (0-shot) (Clark et al., 2018), MMLU (5-shot) (Hendrycks et al., 2020), MMLU-Pro (5-shot) (Wang et al., 2024a), DROP (5-shot) (Dua et al., 2019), OpenBookQA (5-shot) (Mihaylov et al., 2018), and PIQA (0-shot) (Bisk et al., 2020).

Scientific Capabilities We examine scientific domain performance using GSM-8K (8-shot) (Cobbe et al., 2021), MATH (4-shot) (Hendrycks et al., 2021b), GPQA-Main (5-shot) (Rein et al., 2024), SuperGPQA (5-shot) (Du et al., 2025b), MMLU-STEM (5-shot) (Hendrycks et al., 2020), MMLU-Pro-STEM (5-shot) (Wang

	S3B			S7B		
	Baseline	Sci-Mix	Delta	Baseline	Sci-Mix	Delta
General Tasks						
BBH	33.08	37.81	4.73	43.17	49.25	6.08
ARC-E	66.97	69.26	2.29	74.13	74.87	0.75
ARC-C	39.27	42.05	2.78	49.08	48.77	-0.31
MMLU	45.29	48.62	3.33	53.19	57.60	4.41
MMLU-P	13.42	16.94	3.52	22.66	27.36	4.70
DROP	29.61	31.44	1.82	35.70	37.57	1.87
OBQA	42.12	41.28	-0.84	45.00	46.28	1.28
PIQA	77.45	77.80	0.35	79.79	79.72	-0.08
Scientific Tasks						
GSM-8K	27.90	29.42	1.52	45.97	48.37	2.40
MATH	12.60	12.40	-0.20	20.68	20.44	-0.24
GPQA	25.80	26.07	0.27	28.66	27.28	-1.38
SupGPQA	12.11	13.90	1.79	15.09	17.34	2.25
M-STEM	40.22	39.89	-0.33	46.30	50.16	3.85
MP-STEM	13.41	15.67	2.26	20.72	25.12	4.40
SciBench	3.84	3.72	-0.12	6.51	7.35	0.84
Oly-MC	24.05	25.72	1.67	30.04	32.13	2.09
MedQA	31.11	33.61	2.50	38.76	45.78	7.03
MMCQA	33.42	34.53	1.11	37.20	41.42	4.22
PMQA	69.28	74.24	4.96	74.88	75.92	1.04
In-Domain Tasks						
SP-Book	30.77	36.31	5.53	47.44	53.60	6.16
SP-Paper	26.85	32.52	5.66	41.56	52.20	10.64
Average						
Avg-Gen	45.05	46.23	1.18	51.29	52.64	1.35
Avg-Sci	26.70	28.11	1.40	33.17	37.53	4.36
Avg-Dom	28.81	34.41	5.60	44.50	52.90	8.40
Avg-All	33.27	35.39	2.12	40.79	43.74	2.95

Table 2: Performance comparison between the Baseline and Sci-Mix configurations. The Delta column denotes the improvement achieved by Sci-Mix over the Baseline. Abbreviations for benchmarks are as follows: M(P)-STEM: MMLU(-Pro)-STEM; OBQA: OpenBookQA; Oly-MC: OlympicArena-MC; SP-Book/Paper: SciPedia-Eval-Book/Paper; MMC/PMQA: MedMC/PubMedQA.

et al., 2024a), SciBench (4-shot) (Wang et al., 2023), OlympicArena-MC (4-shot) (Huang et al., 2024), MedQA (0-shot) (Jin et al., 2021), MedMCQA (0-shot) (Pal et al., 2022), and PubMedQA (0-shot) (Jin et al., 2019).

5 Experiments

Leveraging the above testing ground, we implement a controlled setting and empirically validate our scientific corpus. Through comparative CPT, we quantify performance gains across varying model scales and benchmarks.

5.1 Experimental Setup

Models We train 3B and 7B parameter models from scratch, named S3B and S7B, to avoid opaque off-the-shelf models. By strictly excluding scientific content during the 930B pre-training tokens, we establish contamination-free base models possessing foundational capabilities yet unexposed to science (Appendix C). This setup allows us to precisely isolate performance gains attributable solely to our scientific data.

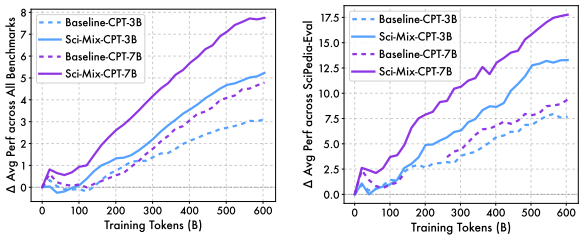


Figure 3: Performance gains of S3B and S7B models. In both plots, the y-axis denotes the relative improvement over the corresponding base models.

Data Configurations We compare two configurations to isolate the effect of scientific content:

- **Baseline:** 600B tokens of the pretraining mixture (80.2% CommonCrawl, 11.2% Code, 8.5% Math).
- **Sci-Mix:** 600B tokens with 50% from our scientific corpus (books:papers = 1:2) and 50% from the baseline pretraining mixture.[3]

Training Details All experiments utilize NVIDIA NeMo framework (NVIDIA, 2024) for 600B tokens of CPT with cosine decay ($3 \times 10^{-4} \rightarrow 3 \times 10^{-5}$), sequence length 4,096, and global batch size 4,096. To ensure robustness, we report the average of the final 5 checkpoints (520B–600B, saved every 1,200 steps) and smooth learning curves using a 5-point moving average.

5.2 Main Results

The quantitative results for S3B and S7B are summarized in Tab. 2 and Fig. 3. Overall, incorporating scientific data yields consistent performance improvements across the evaluation suite. We highlight three key findings:

Scientific Advantage Scales Scientific data yields robust gains, with S3B and S7B improving by **+2.12** and **+2.95** points on average. These improvements are more pronounced on scientific benchmarks (+1.40 vs. +1.18 for S3B; +4.36 vs. +1.35 for S7B) compared to general tasks. Critically, *the advantage over the baseline grows throughout the 600B token window with no sign of saturation* (Fig. 3). While both configurations

[3] Both the 50% scientific ratio and the 1:2 book-paper ratio are validated through experiments in Sec. 6.

continue to improve, the gap between them widens progressively, indicating that our enhanced corpus provides superior sustained learning value. This suggests that our two-stage processing pipeline produces content that remains highly effective even at extended training scales.

Model Capacity Amplifies Data Value A clear scaling pattern emerges: *larger models derive greater benefits from scientific data*. S7B gains +2.95 points from scientific data compared to +2.12 for S3B (Tab. 2), reflecting that larger models are better equipped to capture the complex reasoning and dense domain knowledge embedded in scientific texts. While smaller models do benefit, their capacity constraints limit the extent of their learning. This suggests that for high-complexity content, model scale becomes a critical determinant of data utilization, making capacity a key consideration for effective knowledge acquisition.

Aligned Eval Reveals Hidden Gains *Scientific data effectiveness depends heavily on the evaluation metric*. While standard benchmarks show gains of 1.76–2.38 points, aligned SciPedia-Eval yields 5.60–8.40 points—a more than threefold increase (Tab. 2). This stems from a distribution mismatch: standard benchmarks focus on standardized tests, whereas our training data comprises research-level content. Aligned benchmarks capture domain-specific gains that standard evaluations miss. Thus, relying solely on standard benchmarks can undervalue data sources, obscuring true gains without domain-matched evaluation.

Comparison with Existing Scientific Corpora Beyond validating our approach against general-purpose baselines, we further compare SciPedia with Proof-Pile-2 (Azerbayev et al., 2023), a widely-used mathematical and scientific corpus. Under identical CPT settings, SciPedia consistently outperforms Proof-Pile-2 across both model scales (detailed results in Appendix D), confirming both the overall effectiveness of our corpus and its competitive advantages over established scientific datasets.

6 Analysis

To move beyond validation to *optimal training recipes*, we investigate the mechanisms underlying this success. We systematically analyze both **Data-Centric** and **Model-Centric** factors through controlled ablations on S3B to isolate key drivers and establish evidence-based guidelines.

6.1 Data-Centric Analysis

We examine two fundamental dimensions of data preparation: the Composition Strategy to optimize data mixtures, and the Processing Strategy to maximize content learnability.

6.1.1 Composition Strategy

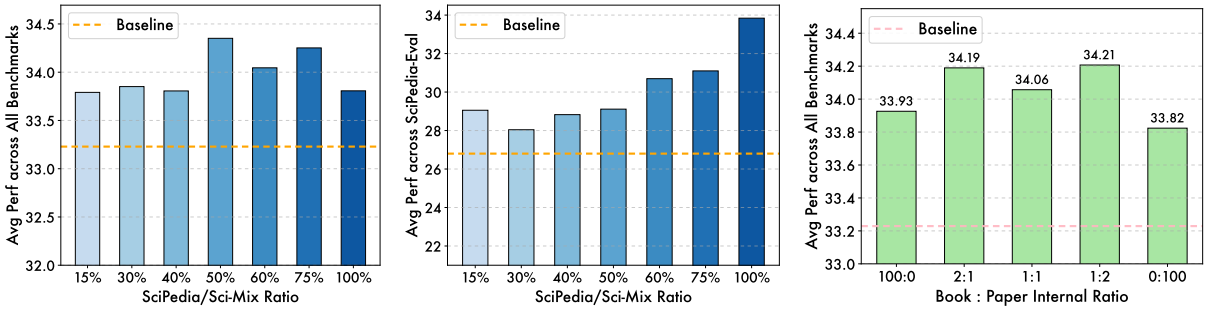
Scientific Content Ratio We evaluate scientific ratios from 15% to 100% (1:1 books-to-papers) and find that aggregated benchmarks follow an inverted-U pattern peaking at 50% (Fig. 4a). Pure scientific training lags behind this balanced mixture, suggesting that *general-purpose performance requires balancing domain focus with broad capabilities*. Specifically, ratios below 30% offer insufficient domain exposure, while excessive scientific data degrades general reasoning.

Conversely, aligned benchmark performance increases monotonically with scientific ratio (Fig. 4b). This divergence shows that *optimal composition is goal-dependent*: balanced mixes suit generalists, while specialized applications favor higher proportions. Thus, "saturation" observed on standard metrics may stem from target mismatch rather than purely from inherent data limits.

Book-Paper Balance Beyond the overall ratio, the internal composition between books and papers is also critical. Books provide systematic foundational knowledge with pedagogical structure, while papers present cutting-edge research with technical depth. Testing five book:paper ratios (100:0–0:100) at a fixed 50% scientific content shows stable performance across mixtures but degrades at extremes (Fig. 4c). This suggests that *books and papers provide complementary value*, and the model's relative insensitivity to precise proportions allows flexible adjustments based on data availability. Accordingly, we adopt a 1:2 ratio to match the natural distribution.

6.1.2 Processing Strategy

Processing Pipeline Effectiveness To validate our pipeline, we compare four configurations on the full corpus: Baseline, Raw, Content-Cleaned only, and Full (Content + Ped.). As shown in Fig. 1, raw data yields negligible benefit, confirming that unprocessed content is largely opaque to learning. While cleaning provides modest

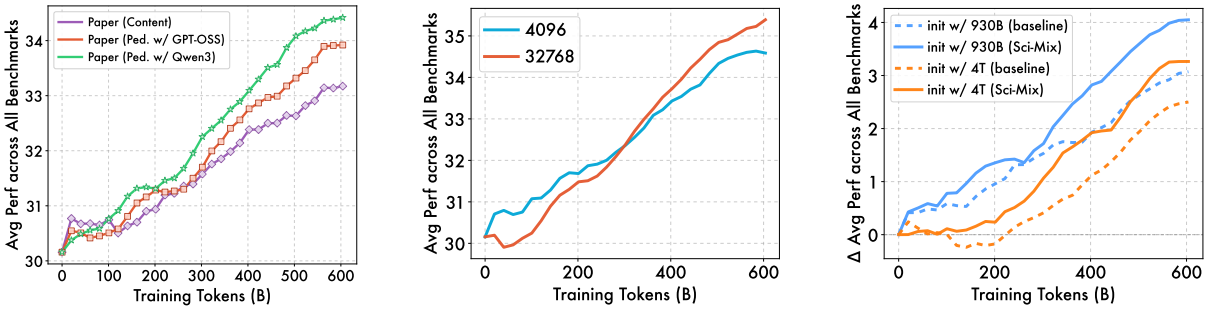


(a) Effect of different overall scientific-content ratios on all benchmarks.

(b) Effect of different overall scientific-content ratios on SciPedia-Eval.

(c) Effect of book-to-paper ratios within the scientific content on all benchmarks.

Figure 4: Data-centric analysis of data mixture ratios.



(a) Processing Strategy

(b) Context Length

(c) Training Checkpoints

Figure 5: (a) Dissection of processing strategies on scientific papers, contrasting content cleaning versus pedagogical augmentation. (b) Comparison of models trained with different context lengths. (c) Performance gains of Sci-Mix over the baseline starting from different training base checkpoints.

gains (+0.38 points), the full pipeline achieves substantial improvements (+1.36 points), empirically demonstrating that *systematic processing is essential to realize scientific data's full potential*.

Pedagogical Augmentation Value To isolate pedagogical augmentation's impact, we compare content-cleaned papers against their augmented counterparts on identical subsets. Additionally, we employ `GPT-OSS-120B` and `Qwen3-235B` as teacher models to assess the impact of generator capability (Fig. 5a). Both augmented variants surpass the cleaned baseline (`OSS-120B`: +0.75; `Qwen3-235B`: +1.27), confirming that *pedagogical augmentation adds distinct value beyond cleaning*. Furthermore, the +0.52 gap between teachers demonstrates that *teacher model quality is a critical determinant of augmentation effectiveness*.

6.2 Model-Centric Analysis

Data learnability is not intrinsic; it is also determined by the learner. Beyond model scale (discussed in Sec. 5.2), we investigate two additional properties that affect learning:

Context Length Requirements Since scientific reasoning involves long-range dependencies, we compare a standard 4K context window (RoPE base = 10,000) against an extended 32K (RoPE base = 1,000,000), finding that the 32K ultimately leads by **+0.80** points (Fig. 5b). Learning dynamics reveal an initial adaptation phase: while the 4K model leads early, the 32K version progressively pulls ahead, implying that *extended context yields superior long-term performance but requires adaptation time*. Practitioners should thus evaluate over sufficient training durations, as long-context advantages emerge gradually.

Training Stage Consistency We investigate whether the benefits of scientific data depend on model maturity by comparing early-stage (930B tokens) vs. late-stage (4T tokens) checkpoints, both continuing training for 600B tokens with Baseline and Sci-Mix configurations (Fig. 5c).

Both stages exhibit robust improvements over their respective baselines (Early: +0.98, Late: +0.76). This consistency yields two insights. First, the persistent gain at the late stage confirms that

scientific data remains effective even for mature models. Second, the comparable magnitude of these gains implies that *early checkpoints serve as reliable proxies for data evaluation*, enabling corpus assessment at a fraction of the compute cost.

7 Related Work

Domain-Specialized Scientific Corpora While general web corpora (Raffel et al., 2020; Penedo et al., 2023; Soldaini et al., 2024; Penedo et al., 2024a) enable scalable pre-training and data-mix studies (Li et al., 2024a), they lack deep domain specialization. Success in mathematics demonstrated by foundational corpora (Paster et al., 2023; Wang et al., 2024b; Han et al., 2024; Zhou et al., 2025; Lu et al., 2024) and instruction-tuning datasets (Toshniwal et al., 2024; Zeng et al., 2024) highlights the value of targeted data. Although efforts like MegaScience (Fan et al., 2025) explore science reasoning via benchmarks (Rein et al., 2024; Hendrycks et al., 2021a), the community remains without a comprehensive, multi-discipline corpus of **high-density, research-grade scientific texts** (e.g., physics, chemistry, biology) suitable for foundational pre-training.

Data Processing and Transformation Pre-training pipelines have evolved from standard filtering and deduplication (Penedo et al., 2024a; Soldaini et al., 2024; Broder, 1997) to sophisticated LLM-driven refinement. Recent methods utilize example-level editing (Zhou et al., 2024; Bi et al., 2025), synthetic rewriting (Su et al., 2024; Maini et al., 2024; Jiang et al., 2025), and automated cleaning workflows (Li et al., 2024b; Zhang et al., 2025). However, existing pipelines are optimized for web text rather than the symbol-rich content of **scientific books and papers**. We bridge this gap by introducing a hierarchy specifically designed to filter and pedagogically rewrite complex scientific literature for improved model training.

8 Conclusion

This work presents a systematic framework for unlocking the potential of scientific literature, addressing the full lifecycle of acquisition, learnability, and verification. Our investigation demonstrates that unleashing the value of complex domains requires not just data accumulation, but deep cultivation—prioritizing enhanced data learnability alongside rigorous training strategies. We release our corpus, pipeline, and verification suite to support the community in exploring the next frontier of scientific LLMs.

Limitations

Despite the robust performance of our framework, we acknowledge several limitations that point to directions for future research:

Pedagogical Restructuring and Distillation. While our ablations demonstrate substantial gains from pedagogical augmentation, fully disentangling the contribution of pedagogical formatting from knowledge distillation effects remains challenging, as both mechanisms are intertwined in our teacher-model-based pipeline. More fine-grained ablations would be needed to precisely quantify their relative contributions.

Dependency on Teacher Model Capabilities. Beyond the distillation concern, the quality and diversity of our synthesized data are inherently bounded by the teacher model's knowledge and instruction-following abilities. Any hallucinations or biases present in the teacher model could potentially propagate to our training corpus, despite our quality control measures.

Scale verification. Due to computational constraints, our controlled verification was conducted on models at the 3B and 7B parameter scales. While we observe consistent gains across these sizes, the efficacy of our method on significantly larger models (e.g., 70B+), where different scaling dynamics may emerge, remains to be empirically validated.

Benchmark Correctness. As a synthetically constructed benchmark, SciPedia-Eval faces inherent correctness challenges. We employ rigorous quality control measures, including grounding both questions and answers directly in source text (reducing dependency on model capabilities through a rephrasing paradigm) and automated correctness validation. Nevertheless, manual sampling reveals approximately 3% of questions contain errors.

Ethical Considerations

Data Provenance and Copyright Compliance We strictly adhere to intellectual property rights and legal regulations through a differentiated release strategy for our artifacts. For the **SciPedia dataset**, we exclusively release the subset verified

to be under permissive licenses (e.g., Open Access, CC-BY, Public Domain), with all publicly distributed data rigorously vetted for copyright compliance. The released **base models** (3B/7B) were trained from scratch strictly on publicly available, permissively licensed general corpora, ensuring they have not been exposed to any copyrighted scientific literature during training and serve as a legally compliant and contamination-free starting point for the community. For the **SciPedia-Eval benchmark**, we construct it entirely from documents with permissive licenses to ensure long-term accessibility and reproducibility, facilitating broad research adoption.

Potential Risks While scientific literature is generally objective, we acknowledge that it may reflect historical biases. Users should be aware of these limitations when deploying systems trained on this data.

References

Essential AI, :, Andrew Hojel, Michael Pust, Tim Romanski, Yash Vanjani, Ritvik Kapila, Mohit Parmar, Adarsh Chaluvaraju, Alok Tripathy, Anil Thomas, Ashish Tanwer, Darsh J Shah, Ishaan Shah, Karl Stratos, Khoi Nguyen, Kurt Smith, Michael Callahan, Peter Rushton, and 6 others. 2025. Essential-web v1.0: 24t tokens of organized web data. *Preprint*, arXiv:2506.14111.

Syeda Nahida Akter, Shrimai Prabhumoye, John Kamalu, Sanjeev Satheesh, Eric Nyberg, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. 2025. MIND: Math informed synthetic dialogues for pretraining LLMs. In *The Thirteenth International Conference on Learning Representations*.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, and 1 others. 2025. Smollm2: When smol goes big–data-centric training of a small language model. *arXiv preprint arXiv:2502.02737*.

Anthropic. 2025. System card: Claude opus 4 & claude sonnet 4. Technical report, Anthropic. Model string: claude-opus-4-20250514.

Zhangir Azerbayev, Hailey Schoelkopf, Keiran Paster, Marco Dos Santos, Stephen McAleer, Albert Q. Jiang, Jia Deng, Stella Biderman, and Sean Welleck. 2023. Llemma: An open language model for mathematics. *Preprint*, arXiv:2310.10631.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Baolong Bi, Shenghua Liu, Xingzhang Ren, Dayiheng Liu, Junyang Lin, Yiwei Wang, Lingrui Mei, Junfeng Fang, Jiafeng Guo, and Xueqi Cheng. 2025. Refinex: Learning to refine pre-training data at scale from expert-guided programs. *arXiv preprint arXiv:2507.03253*.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, and 1 others. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7432–7439.

Andrei Z Broder. 1997. On the resemblance and containment of documents. In *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No. 97TB100171)*, pages 21–29. IEEE.

Andrei Z. Broder. 2000. Identifying and filtering near-duplicate documents. In *Combinatorial Pattern Matching*, pages 1–10, Berlin, Heidelberg. Springer Berlin Heidelberg.

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*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Wei Du, Shubham Toshniwal, Branislav Kisacanin, Sadegh Mahdavi, Ivan Moshkov, George Armstrong, Stephen Ge, Edgar Minasyan, Feng Chen, and Igor Gitman. 2025a. Nemotron-math: Efficient long-context distillation of mathematical reasoning from multi-mode supervision. *arXiv preprint arXiv:2512.15489*.

Xinrun Du, Yifan Yao, Kaijing Ma, Bingli Wang, Tianyu Zheng, King Zhu, Minghao Liu, Yiming Liang, Xiaolong Jin, Zhenlin Wei, and 1 others. 2025b. Supergpqa: Scaling llm evaluation across 285 graduate disciplines. *arXiv preprint arXiv:2502.14739*.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. *arXiv preprint arXiv:1903.00161*.

Run-Ze Fan, Zengzhi Wang, and Pengfei Liu. 2025. Megascience: Pushing the frontiers of post-training datasets for science reasoning. *arXiv preprint arXiv:2507.16812*.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2024. The language model evaluation harness.

Xiaotian Han, Yiren Jian, Xuefeng Hu, Haogeng Liu, Yiqi Wang, Qihang Fan, Yuang Ai, Huaibo Huang, Ran He, Zhenheng Yang, and 1 others. 2024. Infimm-webmath-40b: Advancing multimodal pre-training for enhanced mathematical reasoning. *arXiv preprint arXiv:2409.12568*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

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

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021b. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Ming Hu, Chenglong Ma, Wei Li, Wanghan Xu, Jiamin Wu, Jucheng Hu, Tianbin Li, Guohang Zhuang, Jiaqi Liu, Yingzhou Lu, and 1 others. 2025. A survey of scientific large language models: From data foundations to agent frontiers. *arXiv preprint arXiv:2508.21148*.

Zhen Huang, Zengzhi Wang, Shijie Xia, Xuefeng Li, Haoyang Zou, Ruijie Xu, Run-Ze Fan, Lyumanshan Ye, Ethan Chern, Yixin Ye, and 1 others. 2024. Olympicarena: Benchmarking multi-discipline cognitive reasoning for superintelligent ai. *Advances in Neural Information Processing Systems*, 37:19209–19253.

Minqi Jiang, JoÃĢo GM AraÃšjo, Will Ellsworth, Sian Gooding, and Edward Grefenstette. 2025. Generative data refinement: Just ask for better data. *arXiv preprint arXiv:2509.08653*.

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. *arXiv preprint arXiv:1909.06146*.

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Yitzhak Gadre, Hritik Bansal, Etash Guha, Sedrick Scott Keh, Kushal Arora, and 1 others. 2024a. Datacomp-lm: In search of the next generation of training sets for language models. *Advances in Neural Information Processing Systems*, 37:14200–14282.

Lan Li, Liri Fang, Bertram Ludäscher, and Vetle I Torvik. 2024b. Autodcworkflow: Llm-based data cleaning workflow auto-generation and benchmark. *arXiv preprint arXiv:2412.06724*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*.

Zimu Lu, Aojun Zhou, Ke Wang, Houxing Ren, Weikang Shi, Junting Pan, Mingjie Zhan, and Hongsheng Li. 2024. Mathcoder2: Better math reasoning from continued pretraining on model-translated mathematical code. *arXiv preprint arXiv:2410.08196*.

Junyu Luo, Bohan Wu, Xiao Luo, Zhiping Xiao, Yiqiao Jin, Rong-Cheng Tu, Nan Yin, Yifan Wang, Jingyang Yuan, Wei Ju, and 1 others. 2025. A survey on efficient large language model training: From data-centric perspectives. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30904–30920.

Pratyush Maini, Skyler Seto, He Bai, David Grangier, Yizhe Zhang, and Navdeep Jaitly. 2024. Rephrasing the web: A recipe for compute and data-efficient language modeling. *arXiv preprint arXiv:2401.16380*.

Meta AI. 2024. Llama 3.3 model card. `https://github.com/meta-llama/llama-models/blob/main/models/llama3_3/MODEL_CARD.md`. 70B parameter model.

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 *EMNLP*.

NVIDIA, :, Aarti Basant, Abhijit Khairnar, Abhijit Paithankar, Abhinav Khattar, Adithya Renduchintala, Aditya Malte, Akhiad Bercovich, Akshay Hazare, Alejandra Rico, Aleksander Ficek, Alex Kondratenko, Alex Shaposhnikov, Alexander Bukharin, Ali Taghibakhshi, Amelia Barton, Ameya Sunil Mahabaleshwarkar, Amy Shen, and 198 others. 2025. Nvidia nemotron nano 2: An accurate and efficient hybrid mamba-transformer reasoning model. *Preprint*, arXiv:2508.14444.

NVIDIA. 2024. Nemo: A toolkit for conversational ai and large language models. Version 2.0.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, and 1 others. 2024. 2 olmo 2 furious. *arXiv preprint arXiv:2501.00656*.

OpenAI. 2025. gpt-oss-120b gpt-oss-20b model card. *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 *Conference on health, inference, and learning*, pages 248–260. PMLR.

Keiran Paster, Marco Dos Santos, Zhangir Azerbayev, and Jimmy Ba. 2023. Openwebmath: An open dataset of high-quality mathematical web text. *arXiv preprint arXiv:2310.06786*.

Guilherme Penedo, Hynek Kydlícek, Loubna Ben Allal, and Thomas Wolf. 2024a. Fineweb: decanting the web for the finest text data at scale. *HuggingFace. Accessed: Jul*, 12.

Guilherme Penedo, Hynek Kydlíek, Alessandro Cappelli, Mario Sasko, and Thomas Wolf. 2024b. Datatrove: large scale data processing.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

Jake Poznanski, Aman Rangapur, Jon Borchardt, Jason Dunkelberger, Regan Huff, Daniel Lin, Christopher Wilhelm, Kyle Lo, and Luca Soldaini. 2025. olmocr: Unlocking trillions of tokens in pdfs with vision language models. *arXiv preprint arXiv:2502.18443*.

Shrimai Prabhumoye Mostofa Patwary Mohammad Shoeybi Bryan Catanzaro Rabeeh Karimi Mahabadi, Sanjeev Satheesh. 2025. Nemotron-cc-math: A 133 billion-token-scale high quality math pretraining dataset.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2024. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

Salvatore Sanfilippo. 2009. Redis: In-memory data structure store. https://redis.io.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, and 1 others. 2024. Dolma: An open corpus of three trillion tokens for language model pretraining research. *arXiv preprint arXiv:2402.00159*.

Dan Su, Kezhi Kong, Ying Lin, Joseph Jennings, Brandon Norick, Markus Kliegl, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. 2024. Nemotron-cc: Transforming common crawl into a refined long-horizon pretraining dataset. *arXiv preprint arXiv:2412.02595*.

Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. 2021. Roformer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, and 1 others. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*.

Liping Tang, Nikhil Ranjan, Omkar Pangarkar, Xuezhi Liang, Zhen Wang, Li An, Bhaskar Rao, Linghao Jin, Huijuan Wang, Zhoujun Cheng, Suqi Sun, Cun Mu, Victor Miller, Xuezhe Ma, Yue Peng, Zhengzhong Liu, and Eric P. Xing. 2024. Txt360: A top-quality llm pre-training dataset requires the perfect blend. *Preprint*, arXiv:2411.05452.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Shubham Toshniwal, Wei Du, Ivan Moshkov, Branislav Kisacanin, Alexan Ayrapetyan, and Igor Gitman. 2024. Openmathinstruct-2: Accelerating ai for math with massive open-source instruction data. *arXiv preprint arXiv:2410.01560*.

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. 2023. Scibench: Evaluating college-level scientific problem-solving abilities of large language models. *arXiv preprint arXiv:2307.10635*.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, and 1 others. 2024a. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290.

Zengzhi Wang, Xuefeng Li, Rui Xia, and Pengfei Liu. 2024b. Mathpile: A billion-token-scale pretraining corpus for math. *Advances in Neural Information Processing Systems*, 37:25426–25468.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Liang Zeng, Liangjun Zhong, Liang Zhao, Tianwen Wei, Liu Yang, Jujie He, Cheng Cheng, Rui Hu, Yang Liu, Shuicheng Yan, and 1 others. 2024. Skywork-math: Data scaling laws for mathematical reasoning in large language models–the story goes on. *arXiv preprint arXiv:2407.08348*.

Shuo Zhang, Zezhou Huang, and Eugene Wu. 2025. Data cleaning using large language models. In *2025 IEEE 41st International Conference on Data Engineering Workshops (ICDEW)*, pages 28–32. IEEE.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, and 1 others. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2).

Fan Zhou, Zengzhi Wang, Qian Liu, Junlong Li, and Pengfei Liu. 2024. Programming every example: Lifting pre-training data quality like experts at scale. *arXiv preprint arXiv:2409.17115*.

Fan Zhou, Zengzhi Wang, Nikhil Ranjan, Zhoujun Cheng, Liping Tang, Guowei He, Zhengzhong Liu, and Eric P. Xing. 2025. Megamath: Pushing the limits of open math corpora. *arXiv preprint arXiv:2504.02807*. Preprint.

A Filtering Details

Based on the annotation results, we apply a four-stage filtering process to all documents (Table 3):

- **File Size Filtering**: We initially analyzed the distribution of text sizes and found that a portion of documents exhibited significantly shorter lengths. Subsequent sampling analysis indicated that these short texts were primarily spam, document fragments, or advertisements with limited educational value. To determine an appropriate threshold, we randomly sampled 100 documents and sorted them by file size in descending order.By examining document quality at different thresholds $\{20\text{KB}, 15\text{KB}, 12\text{KB}, 10\text{KB}, 8\text{KB}, 5\text{KB}\}$, we ultimately established 8KB as the filtering threshold and discard all documents smaller than 8KB. This step filtered about 13% of documents.

- **Category Filtering**: Among the 12-dimensional labels generated by EAI-Distill-0.5B, Document Type v1 and Document Type v2 identify the content type of documents within the classification system, such as News, Personal Blog, etc. We filter out documents with no educational value based on these labels, such as Advertisement. In this phase, we removed approximately 7% of documents.

- **Garbled Text Detection Filtering**: Due to insufficient PDF clarity or uncommon fonts, OCR model may fail to recognize text accurately, producing substantial garbled output characterized by repeated characters, abnormal symbol accumulation, etc. We consider such garbled text harmful to model training. Therefore, we detect the proportion of garbled characters in the text and discard documents where this proportion exceeds 50%. In this phase, we removed approximately 0.1% of documents.

- **Language Filtering**: Through sampling statistics, we observed significant language distribution variations across different data sources. Some sources contain only 54% English documents with substantial amounts of Russian and other languages, while others have English proportions as high as 91%.

To avoid bias in training results caused by uneven language distribution, we retain only English documents at this stage and do not consider multilingual scenarios. We employ the fast-langdetect tool for language detection, filtering out 24% of documents in this phase.

Stage	Criterion	Filtered	Retained
Initial	-	-	100%
File Size	<8KB	13%	87%
Category	Ads, etc.	7%	80.9%
Garbled	>50%	0.1%	80.8%
Language	Non-EN	24%	61.4%

Table 3: Document Filtering Statistics

B Classification

B.1 Discipline Classification Mapping Rules

The Dewey Decimal Classification (DDC) is a widely adopted library classification system that systematically organizes knowledge through decimal numerical codes. It employs a hierarchical structure where the hundreds digit represents the main class, the tens and ones digits subdivide into subclasses, and digits after the decimal point provide finer granularity, theoretically supporting hierarchical subdivision to arbitrary depth. While this natural hierarchical structure facilitates discipline classification, the classification granularity is overly fine-grained, and portions of the system originate from historical periods that do not adequately reflect contemporary disciplinary development and evolution. Therefore, we merged and remapped the numerical codes from FDC labels to align them with disciplines suitable for current research needs.You can see the mapping rules in Table 9

B.2 Book-Paper classification

Given the significant differences in knowledge density between books and papers, we first need to distinguish between these two types of documents to implement targeted processing strategies. We employ the Qwen2.5-7B-Instruct for this classification task, with the prompt design as follows Figure 6 .

B.3 SciPedia Domain Distribution

As mentioned earlier, we categorize SciPedia by discipline. The detailed domain distribution is provided in Table 4 and Table 5.

Domain	Tokens (B)	Percentage
Computer Science	10.52	4.18%
Engineering	22.19	8.83%
Human & Social	148.43	59.02%
Medicine	27.79	11.05%
Biology	8.44	3.36%
Chemistry	7.14	2.84%
Mathematics	11.18	4.44%
Physics	4.69	1.86%
STEM Others	11.12	4.42%
Total	**251.49**	**100.00%**

Table 4: Token Distribution by Domain in Book

Domain	Tokens (B)	Percentage
Computer Science	49.90	7.63%
Engineering	38.03	5.82%
Human & Social	45.35	6.93%
Medicine	255.05	38.87%
Biology	58.28	8.91%
Chemistry	42.85	6.55%
Mathematics	77.29	11.81%
Physics	57.49	8.79%
STEM Others	30.71	4.69%
Total	**655**	**100.00%**

Table 5: Token Distribution by Domain in Paper

C Foundation Model Training

C.1 Pretraining Dataset

Our foundation model training dataset consists of three parts: CC, Math, and Code, totaling 5.37T. The specific composition can be seen in Table 6

CC. Massive web data accounts for a significant portion of pretraining. We selected the non-synthetic subset of the Nemotron-CC (Su et al., 2024). Nemotron-CC is a 6.3T token dataset based on Common Crawl, consisting of 4.4T globally deduplicated original tokens and 1.9T synthetically generated tokens. To avoid introducing additional confounding factors, we only used the real data portion of Nemotron-CC, i.e., 4.4T tokens.

To facilitate subsequent control over the proportion of different academic domains, we also employed the eai-distill-0.5b model for annotation and performed domain classification based on FDC labels, dividing the data into *computer science*, *medicine*, *biology*, *chemistry*, *mathematics*, *physics*, *humansocial*, *engineering*, and *stem-others* (as shown in Appendix B.1). After accounting for losses during processing, our final CC dataset contains approximately 4.28T tokens.

Math. To enhance the model's scientific reasoning capabilities, we specifically collected two high-quality mathematical pretraining datasets: *MegaMath* (Zhou et al., 2025) and *Nemotron-CC-Math-v1* (Rabeeh Karimi Mahabadi, 2025).

MegaMath is currently the largest open-source English mathematics corpus. We selected three subsets:

- MegaMath-Web (264B tokens): The complete web dataset of MegaMath extracted from Common Crawl, employing optimized HTML parsing techniques to preserve mathematical formulas and symbols.
- MegaMath-Web-Pro (15B tokens): A high-quality subset obtained through language model scoring and LLM refinement, surpassing all existing open-source mathematical corpora in quality.
- MegaMath-Synth-Code (7B tokens): A synthetic dataset translating high-quality mathematical code from other programming languages into Python.

Our other mathematical data source is *Nemotron-CC-Math-v1*, a high-quality mathematical pretraining dataset extracted from Common Crawl. This dataset is constructed using an innovative Lynx + LLM processing pipeline and is divided into two subsets based on quality scores: Nemotron-CC-Math-v1-3 (score=3) and Nemotron-CC-Math-v1-4+ (scores 4-5). Additionally, based on the high-quality subset Nemotron-CC-Math-v1-4+, a structured mathematical dialogue dataset, Nemotron-CC-Math-v1-4+-MIND (74B tokens), was generated using the same synthetic method as Nemotron-MIND (Akter et al., 2025) (converting content into various conversational scenarios via LLM). Therefore, we utilized three datasets in total: Nemotron-CC-Math-v1-3 (81B tokens), Nemotron-CC-Math-v1-4+ (52B tokens), and Nemotron-CC-Math-v1-4+-MIND (74B tokens).

Code. Our code dataset is derived from three sources: self-crawled GitHub repositories, Nemotron-Pretraining-Code-v1 (NVIDIA et al., 2025), and txt360-stack-exchange.

For the self-crawled GitHub data, we first collected the metadata of all available GitHub repositories. We then filtered out repositories with fewer than ten stars to ensure basic quality and maintenance activity. Next, we organized all the source files from the remaining repositories at the file level and applied the OpenCoder filtering method to remove low-quality or non-informative code files. Through this process, we obtained approximately 187B tokens of high-quality code data. Finally, we categorized these code files according to their programming languages, resulting in a structured dataset organized by language and file-level code granularity.

In addition to our self-crawled GitHub data, we incorporated Nemotron-Pretraining-Code-v1 as a supplement. This is a large-scale curated GitHub source code dataset processed through a multi-stage pipeline including deduplication and quality filtering, which provides original metadata for this portion of the data. We crawled additional original code based on the provided metadata, and then deduplicated it against our own crawled data, and ultimately obtained 220B tokens. Furthermore, this dataset also includes large-scale natural language-code paired data constructed via LLM across 11 programming languages, namely the Synthetic-Code subset. We also utilized all synthetic data from this dataset (171B tokens).

Additionally, to enrich code-related question-answer data, we incorporated the txt360-stack-exchange subset. This dataset aggregates question-answer data from the Stack Exchange platform, a network of Q&A websites covering various domains including programming, science, mathematics, and more, representing one of the largest publicly available Q&A data sources. Given its substantial collection of high-quality programming-related questions and answers, we included it as an important supplement to our code data sources, totaling approximately 20B tokens.

C.2 Pretraining Configuration

Model Architecture. Our main experiments utilize a 3B parameter base model following the Qwen2.5 architecture (Team, 2024). The model employs the Qwens tokenizer (Bai et al., 2023) with a vocabulary size of 151,643 tokens, a context length of 4,096 tokens, and Rotary Position Embeddings (RoPE, Su et al. 2021) with a base frequency of 10,000.

Optimization Setup. We train all models using the AdamW optimizer (Loshchilov and Hutter, 2019) with $\beta_1 = 0.9$, $\beta_2 = 0.95$, and $\epsilon = 1\text{e}{-8}$.

Dataset	Tokens
Common Crawl	**4.28T**
Nemotron-CC (actual)	4.28T
Math	**493B**
MegaMath	286B
MegaMath-Web	264B
MegaMath-Web-Pro	15B
MegaMath-Synth-Code	7B
Nemotron-CC-Math-v1	207B
Nemotron-CC-Math-v1-3	81B
Nemotron-CC-Math-v1-4+	52B
Nemotron-CC-Math-v1-4+-MIND	74B
Code	**598B**
Self-crawled GitHub (star>5)	187B
Nemotron-Pretraining-Code-v1	391B
Original	220B
Synthetic-Code	171B
txt360-stack-exchange	20B
Total	**5.37T**

Table 6: Foundation Model Pre-training Dataset Composition

The learning rate schedule incorporates a 2,000-step linear warmup phase followed by a constant peak learning rate of $3e{-}4$ throughout the remaining pretraining phase. All models use a micro-batch size of 4.

Training Schedule. We employ a progressive global batch size (GBS) scaling strategy:

- **Stage 1**: GBS=1,024 for 70,000 steps ($\sim$293.6B tokens)
- **Stage 2**: GBS=2,048 for 40,000 steps ($\sim$335.5B tokens)
- **Stage 3**: GBS=4,096 for the remaining steps

The final stage varies by model configuration to achieve target token counts. This results in two 3B model variants: S3B (18,000 steps in Stage 3, $\sim$302B tokens, 930B total), and S3B_{4T} (200,000 steps, $\sim$3.36T tokens, 4T total). The 7B model follows identical training recipes at the 930B token scale, denoted as S7B. All experiments are conducted using the NVIDIA NeMo framework (NVIDIA, 2024).

Data Mixture. Following the data composition strategy of Allal et al. (2025); OLMo et al. (2024), where Common Crawl dominates the pretraining mixture, we adopt a sampling ratio of 80.2% CC, 11.2% Code, and 8.5% Math.

C.3 Results

Table 7 presents the evaluation results of our pretrained models at different training stages. We report performance for S3B at 930B tokens, S3B_{4T} at 4T tokens, and S7B at 930B tokens across all benchmark categories. These pretrained models serve as capable starting points for our subsequent experiments, providing basic checkpoints with established general reasoning and scientific capabilities for investigating the impact of scientific data integration.

D Comparison with Existing Scientific Corpora

To contextualize SciPedia within the landscape of existing scientific corpora, we conduct a comparative experiment against Proof-Pile-2 (Azerbayev et al., 2023), a widely-used mathematical and formal reasoning dataset.

We perform continued pre-training on the S3B and S7B model using identical configurations to Section 5.1. We construct two data mixtures: (1) Sci-Mix, which combines 50% SciPedia with 50% baseline mixture (as described in Section 5.1), and (2) Proof-Pile-2 Mix, which replaces SciPedia with Proof-Pile-2 while maintaining the same 50% scientific data ratio and 50% baseline mixture.

Due to the size constraints of Proof-Pile-2, we limit the training duration to 160B tokens to ensure fair comparison without excessive data repetition. Both configurations are trained for the same number of tokens and evaluated on the full benchmark suite described in Section 4.

Table 8 presents the performance comparison between the two scientific data sources. At the 160B token checkpoint, models trained with SciPedia achieve average scores of 31.16 (S3B) and 37.93 (S7B), outperforming the Proof-Pile-2 configuration by +0.78 and +1.08 points respectively across all benchmarks. These results indicate that SciPedia provides a more effective learning signal than established scientific datasets, supporting our claim that the gains observed in Section 5 stem from our systematic processing pipeline rather than merely from the inclusion of scientific content.

The performance gap can be attributed to two

	Scientific Tasks											In-Domain Tasks	
	G-8K	MATH	GPQA	SupG	M-S	MP-S	SciB	Oly-MC	MQA	MMCQA	PMQA	SP-B	SP-P
S3B	20.02	11.00	23.88	8.44	34.92	10.09	3.34	19.18	30.95	32.92	66.00	23.27	19.00
$S3B_{4T}$	27.29	12.60	27.68	11.31	39.23	12.96	3.34	27.07	30.16	34.26	72.20	32.60	26.33
S7B	35.41	17.20	24.33	10.81	41.33	15.62	3.92	26.27	34.64	33.16	73.20	37.33	32.93
	General Tasks								Average				
	BBH	ARCE	ARCC	MMLU	MP	DROP	OBQA	PIQA	AVG-G	AVG-S	Avg-D	Avg-A	
S3B	32.31	65.49	36.52	40.48	11.20	27.04	38.80	78.45	41.29	23.70	21.13	30.16	
$S3B_{4T}$	33.47	68.64	41.38	45.89	13.70	29.64	40.80	77.86	43.92	21.85	29.46	33.73	
S7B	36.78	70.88	42.49	48.90	18.10	30.50	43.80	78.29	46.22	28.72	35.13	35.99	

Table 7: Evaluation results of pretrained models at different training stages and scales. Abbs: G-8K (GSM-8K), SupG (SuperGPQA), M-S (MMLU-STEM), MP-S (MMLU-Pro-STEM), SciB (SciBench), Oly-MC (OlympicArena-MC), MQA (MedQA), MMCQA (MedMCQA), PMQA (PubMedQA), SP-B (SciPedia-Eval-Book), SP-P (SciPedia-Eval-Paper), ARCE (ARC-Easy), ARCC (ARC-Challenge), MP (MMLU-Pro), OBQA (OpenBookQA), AVG-G/S/D/All (Average General/Science/In-Domain/All).

key factors. First, while Proof-Pile-2 is highly specialized in mathematics and formal reasoning, SciPedia covers a broader spectrum of scientific disciplines (computer science, physics, biology, chemistry, medicine, etc.), providing more diverse domain knowledge. Second, our two-stage processing pipeline, combining content cleaning with pedagogical augmentation, explicitly addresses the learnability gap identified in Section 2.2, transforming raw scientific text into more accessible training material. This comparison, combined with the ablation study in Section 6.1.2 (Figure 1), provides converging evidence that our processing methodology is the primary driver of performance improvements.

E Details of Content Cleaning

This appendix provides comprehensive supplementary materials for content cleaningg, including the empirical analysis that informed our cleaning protocol design, complete prompt specifications, representative cleaning examples, and evaluation protocols.

E.1 Quality Analysis Report Synthesis

To systematically identify common quality issues in our scientific corpus, we randomly sampled 20 documents and generated 40 detailed quality assessment reports using `Gemini-2.5-Pro` and `Claude-Sonnet-4.0`. Figure 7 presents the synthesized findings, organizing quality issues into two primary operation categoriesdeletion (removing unwanted content) and modification (repairing and standardizing existing content)with subcategories by content type and processing priority. These structured insights directly informed the design of our content cleaning rules and protection guidelines.

E.2 Content Cleaning Processing Prompt

Figure 8 provides the complete content cleaning prompt used in production, including detailed deletion rules, modification operations, and content protection guidelines. The prompt specifies explicit criteria for identifying and removing unwanted content while preserving all academically valuable material.

E.3 Evaluation and Model Selection

To ensure effective content cleaning and support iterative prompt refinement, we developed a comprehensive evaluation framework. Given that content cleaning operates through explicit deletion and correction rules, evaluation must assess both rule execution accuracy (whether rules are correctly applied) and rule completeness (whether the rule set covers all necessary cases). We adopted a hybrid strategy combining human inspection for identifying improvement opportunities with LLM-based evaluation for systematic quality assessment and quantitative comparison across different prompts and cleaning models.

Evaluation Dataset We randomly sampled 20 documents from our corpus to serve as representative evaluation cases, ensuring diversity across scientific domains and document types (books vs. papers).

Human Evaluation Human evaluators reviewed entire documents, with particular attention to high-risk sectionsdocument beginnings and endings where table of contents, reference lists,

	Scientific Tasks										In-Domain Tasks	
	G-8K	MATH	GPQA	SupG	M-S	MP-S	SciB	Oly-MC	MQA	MMCQA	PMQA	SP-B	SP-P
S3B-proof-pile	19.52	9.88	25.58	10.88	35.37	11.70	3.08	21.03	28.30	28.66	66.08	24.65	22.56
S3B-ScienPedia	18.64	9.12	26.34	11.11	35.72	11.74	2.94	22.10	29.30	31.94	70.88	24.97	23.33
S3B-Delta	-0.88	-0.76	0.76	0.22	0.35	0.04	-0.14	1.07	1.00	3.28	4.80	0.32	0.77
S7B-proof-pile	37.03	17.04	25.54	13.84	42.92	17.40	5.41	27.16	33.86	30.95	73.12	38.52	35.91
S7B-ScienPedia	36.13	14.96	26.79	14.23	42.54	17.74	4.97	27.64	38.15	34.12	75.56	42.89	40.53
S7B-Delta	-0.89	-2.08	1.25	0.39	-0.38	0.34	-0.44	0.48	4.29	3.17	2.44	4.37	4.62

	General Tasks								Average			
	BBH	ARCE	ARCC	MMLU	MP	DROP	OBQA	PIQA	AVG-G	Avg-S	Avg-D	Avg-A
S3B-proof-pile	33.87	65.08	37.27	38.76	12.04	27.93	39.24	76.56	41.34	23.64	23.61	30.38
S3B-ScienPedia	33.25	64.98	38.45	42.88	12.04	28.41	40.08	76.12	42.03	24.53	24.15	31.16
S3B-Delta	-0.62	-0.10	1.18	4.13	0.00	0.48	0.84	-0.45	0.68	0.88	0.55	0.78
S7B-proof-pile	40.39	70.66	43.99	49.34	19.26	32.89	40.36	78.14	46.88	29.48	37.21	36.84
S7B-ScienPedia	39.22	70.87	44.32	50.94	20.18	33.23	43.52	77.93	47.53	30.26	41.71	37.93
S7B-Delta	-1.17	0.21	0.32	1.61	0.92	0.35	3.16	-0.20	0.65	0.78	4.50	1.08

Table 8: Performance comparison between SciPedia and Proof-Pile-2 at 160B tokens. Delta denotes the advantage of SciPedia over Proof-Pile-2.

and structural artifacts typically appear. Evaluators identified execution failures and uncovered problematic content types not addressed by current rules, providing qualitative feedback that directly informed prompt iterations.

LLM-based Evaluation For systematic quantitative comparison across different prompts and cleaning models, we employed LLM-as-judge evaluation using `Claude-Sonnet-4.0` (Anthropic, 2025) and `Gemini-2.5-Pro` (Comanici et al., 2025). From each of the 20 documents, we sampled three groups of three consecutive chunks to ensure local coherence and representativeness. For each evaluation unit, both the original and cleaned versions were presented to the judge models.

The evaluation prompt (Figure 9) instructed judges to provide rule-by-rule analysis, concrete examples of execution failures, suggestions for missing or ambiguous rules, and an overall quality score with prioritized recommendations. This structured output enabled quantitative comparison of alternative approaches while simultaneously gathering actionable insights for rule refinement.

Model Selection We compared multiple language models for Content Cleaning cleaning using identical prompts and evaluation protocols: Qwen2.5 series (7B, 32B, 72B-Instruct) (Team, 2024), `Llama3.3-70B-Instruct` (Meta AI, 2024), Qwen3 series (8B, 14B, 32B, 235B, both thinking and non-thinking variants) (Team, 2025), and `GPT-OSS-120B` (OpenAI, 2025).

Our evaluation results show that Qwen3 series substantially outperformed Qwen2.5 and Llama3.3. Within Qwen3, thinking mode achieved higher accuracy but reduced throughput several-fold. Among Qwen3 models, 8B underperformed while 14B, 32B, and 235B showed comparable quality. `GPT-OSS-120B` demonstrated competitive cleaning accuracy while offering superior processing efficiency, making it our choice for production deployment.

E.4 Implementation Details

E.4.1 Quality Control

Not all model outputs are correct. Common failure modes include malformed output formats that prevent proper extraction of cleaned text, infinite repetition until reaching output length limits, and other processing errors. When such failures occur, we retain the original chunk unchanged. A document is considered successfully processed if at least 95% of its chunks are correctly cleaned; otherwise, it is marked as failed and queued for reprocessing.

E.4.2 Distributed Processing System

Processing pretraining-scale data requires flexible, scalable, and robust infrastructure. We adopt a producer-consumer architecture where a Redis (Sanfilippo, 2009) server acts as the task queue and GPU servers function as workers running vLLM servers that continuously fetch and process tasks.

This design addresses several critical challenges:

- **Dynamic resource allocation**: The availability of GPU nodes in our cluster varies over time. Our design allows seamless addition or

removal of worker nodes without interrupting the overall pipeline.

- **Orphan task management**: GPU servers may crash or be shut down unexpectedly, leaving tasks incomplete. We implement a heartbeat mechanism to monitor worker health, periodically detecting dead workers and reclaiming their orphaned tasks for reassignment.

- **Automatic recovery**: vLLM servers running on GPU workers may crash. Our system automatically detects failures and restarts crashed servers to maintain processing continuity.

- **Task retry mechanism**: Tasks that fail due to quality control issues or other errors are automatically re-queued for processing. Tasks exceeding a maximum retry threshold are marked as permanently failed.

- **Priority queuing**: The system supports priority-based task scheduling, allowing high-priority tasks to bypass the standard queue when necessary.

This architecture enables efficient processing of our large-scale scientific corpus while maintaining robustness against the inevitable failures that occur in distributed systems operating over extended periods.

E.5 Content Cleaning Processing Examples

This section presents representative before-and-after examples demonstrating content cleaning effects on real scientific documents across varying quality levels. Through side-by-side comparisons, we illustrate how Content Cleaning processing successfully removes front matter and structural artifacts, standardizes mathematical notation, corrects formatting inconsistencies, recovers text from severe OCR corruption, and preserves all academically valuable content. The examples span from well-formatted thesis documents to heavily damaged scanned texts, showcasing Content Cleaning's capability to handle diverse quality scenarios common in scientific corpora. Each example highlights specific aspects of the cleaning pipeline, showing the practical impactand limitationsof our deletion and modification operations on document quality under different degradation conditions.

Example 1: PhD Thesis Front Matter Cleaning

This example demonstrates Content Cleaning processing on a mathematics PhD thesis, showcasing the removal of typical academic front matter (title page, acknowledgments, table of contents) while preserving research content (abstracts, keywords) and standardizing mathematical notation.Figure 10, Figure 11, and Figure 23 present the original text, cleaned text, and expert commentary of Example 1.

Example 2: Severe OCR Corruption Recovery

This example illustrates Content Cleaning processing on heavily damaged scanned text from a mathematical paper, demonstrating successful recovery of fragmented formulas, removal of OCR artifacts, deletion of reference lists, and reconstruction of readable mathematical content from severely corrupted input.Figure 13, Figure 14, and Figure 26 present the original text, cleaned text, and expert commentary of Example 2.

E.6 Content Cleaning Evaluation Prompt

Figure 9 provides the complete evaluation prompt used to assess content cleaning quality through LLM-as-judge methodology. The prompt guides judge models to analyze rule execution accuracy, identify coverage gaps, provide concrete examples of failures, and generate prioritized recommendations for improvement.

F Details of Pedagogical Augmentation

To perform content-level rewriting under the **Pedagogical Augmentation** stage, we design a structured, topic-agnostic prompt to guide large language models in transforming dense, expert-level scientific text into pedagogically enriched material. The prompt operates at the level of text chunks (average size 1,024 tokens) and explicitly balances two goals: (i) ensuring absolute fidelity to the original content, and (ii) enhancing the clarity, narrative coherence, and educational depth of the rewritten output.

F.1 Pair-wise Evaluation Prompt for Pedagogical Augmentation

To ensure that our evaluation of pedagogical rewriting quality is both consistent and scalable, we employ an LLM-based pairwise comparison

framework. As shown in Figure 16 defines the precise evaluation setting used to compare different model and prompt configurations under the pedagogical augmentation processing stage. It formalizes the perspective, evaluation dimensions, and output format to guarantee reproducibility across multiple runs and domains.

F.2 Pedagogical Augmentation Prompt

Figure 17 provides the final version of the pedagogical augmentation prompt used in production. The prompt is meticulously designed to balance two competing goals: maintaining absolute fidelity to the original scientific content while transforming it into clear, pedagogically rich material. It explicitly instructs the model to reconstruct implicit reasoning, provide intuitive explanations, and enhance narrative flow without deviating from factual accuracy.

F.3 Pedagogical Augmentation Cases

To concretely illustrate the effects of Pedagogical Augmentation, we present side-by-side comparisons between raw academic texts and their pedagogically rewritten counterparts. These examples demonstrate how the augmentation process enhances conceptual clarity, narrative flow, and instructional value while maintaining rigorous fidelity to the original content.

We provide two representative cases. For Case 1, the original text is shown in Figure 21, the augmented version in Figure 22, and the detailed analysis in Figure 23. Similarly, Case 2 is presented in Figures 24, 25, and 26, respectively.

F.4 Rewrite Fidelity and Hallucination Control

Given the reliance on teacher models for pedagogical augmentation, we evaluate rewrite fidelity through two complementary approaches: expert qualitative audits and systematic LLM-based assessment.

Expert Qualitative Audit. During the initial design of the pedagogical augmentation pipeline, subject-matter experts (Mathematics PhDs) audited augmented versions of their own research papers. The experts confirmed that rewrites remained mathematically faithful to the original arguments without introducing distortions or factual errors. Specifically, they noted that while the restructuring significantly improved explanatory flow by explicitly articulating reasoning steps that were previously implicit, it maintained full technical rigor. This observation indicates that the pedagogical transformation does not compromise correctness for the sake of clarity, but rather makes existing rigorous logic more accessible through explicit articulation. A representative case study with detailed expert commentary is provided in Figure 20.

Systematic LLM-based Evaluation. To provide broader quantitative evidence beyond individual expert assessment, we conducted systematic evaluation on a randomly sampled set of 20 papers, selecting 3 chunks from each (60 chunks total) to ensure coverage across different scientific domains and content types. We employed a 10-point Binary Fidelity Checklist organized across three evaluation pillars: (1) Content Fidelity, which assesses preservation of all original claims, entities, relationships, and logical structures without omission or modification; (2) Hallucination Control, which enforces zero tolerance for fabrication of data points, technical terms, citations, or unsupported assertions; and (3) Technical Rigor, which verifies maintenance of logical soundness, mathematical correctness, and argument validity. The complete evaluation prompt with detailed scoring criteria is provided in Figure 27.

We utilized `Claude-4.5-Sonnet` and `Gemini-2.5-Pro` as independent judges. Each model evaluated all 60 segment pairs (original versus augmented text), generating 120 evaluation reports in total. Judges compared augmented text against the original source, identified any instance of content addition, omission, or distortion, flagged potential hallucinations or unsupported claims, and assigned a fidelity score with detailed justification. The evaluation yielded high average fidelity scores of 9.2/10 (`Claude-4.5-Sonnet`) and 9.1/10 (`Gemini-2.5-Pro`). These scores indicate that the pedagogical pipeline preserves the technical integrity of source material with minimal error propagation.

G Benchmark Construction

G.1 Q&A generation

This prompt guides the Qwen3-32B model to generate high-quality seven-option multiple-choice questions from text segments. The prompt requires the model to first determine whether the text

is suitable for question generation, and if so, identify core knowledge points and generate questions accordingly. A key constraint is that the question statement, option content, and correct answer must all be directly derived from the original text, with the model only performing text refinement and reorganization. The complete prompt is shown in Figure 28.

G.2 Filter Stage 1: Completeness Filter

This prompt is used for first-stage completeness validation, assessing question independence and self-containment. The model determines whether the question relies on external information (such as figures, tables, or specific studies) and whether it contains referential expressions pointing to external content (such as "in the paper," "as described above," etc.). The model receives only the question itself as input and outputs an independence judgment with explanations. The complete prompt is shown in Figure 29.

G.3 Filter Stage 2: Correctness Filter

This prompt is used for second-stage correctness validation, with the core objective of verifying whether the labeled answer has sufficient support from the original text. The model receives the original text, question, and answer as input, and determines whether the answer can be verified from the original text. Only questions that pass verification are retained in the final benchmark. Figure 30.

G.4 MCQ Example

Figure 31 and Figure 32 is an example including the original text segment and the generated seven-option question, demonstrating how the model identifies key knowledge points from the source material and constructs evaluation questions directly grounded in the original text.

H Evaluation Details

Since the evaluated models are *base checkpoints*i.e., models not aligned or fine-tuned through post-trainingwe adopted both *few-shot prompting* and *perplexity-based* evaluation strategies to better reflect intrinsic model capability. Concretely, we used perplexity-based evaluation for **ARC-Easy**, **ARC-Challenge**, **MMLU**, **OpenBookQA**, **PIQA**, **GPQA-Main**, and **MMLU-STEM**, while generative evaluation was applied to the remaining benchmarks, particularly those requiring complex reasoning chains or CoT (Chain-of-Thought) generation. All evaluations were implemented using a slightly modified version of the `lm-evaluation-harness` (Gao et al., 2024) framework, with inference conducted under *greedy decoding* settings for consistency across experiments.

Higher Level	Code Range	Category
computer science	000-009	computer_science
engineer	355-359	military_science
	600-610, 620-621, 626, 629	engineering
	622	mining
	623	maritime
	624	civil
	625	railway
	627	water
	628	environment
	630-631, 632-635, 636-639	agriculture
	660-669	chemical
	670-689	manufacturing
	690-699	construction
mathematics	500-519	mathematics
physics	530-539	physics
chemistry	540-549	chemistry
biology	570-579	biology
medicine	610-619	medicine
stem-others	520-529	astronomy
	550-559	earth
	560-569	paleontology
	580-589	botany
	590-599	zoology
	910-919	geography
humansocial	010-099, 350-354, 640-649, 650-659	management
	100-129, 140-149, 160-199	philosophy
	130-139, 150-159	psychology
	200-299	religion
	300-319, 360-369, 380-399	sociology
	320-329	political_science
	330-339	economics
	340-349	law
	370-379	education
	400-499	linguistics
	700-709, 750-769	art_fine_arts
	710-729	architecture
	730-739	artifacts
	740-749	design
	770-779	photography
	780-789	music
	790-799	sports
	800-899	literature
	900-909, 920-999	history

Table 9: Mapping from FDC Code Ranges to Categories

Book Paper Split Prompt

```
Determine if this document is a scientific academic paper.

Note: The following is a sampled portion of a larger document.
Look for:
- Scientific research content with technical depth
- Formal academic writing style
- Dense technical terminology and concepts
- Complex analytical content
Exclude:
- News articles, interviews
- Blog posts, web content
- Documentation, manuals
- Simple explanatory content
Text sample from document:
text_sample
Please strictly return the result in the following JSON format,
```

```
do not add any other content:

"analysis": "analysis of why this is or isn't an academic
paper with sufficient complexity",
"is_article": true/false
```

Figure 6: Book Paper Split Prompt

Content Cleaning Quality Analysis Report Synthesis

```
## Overview

Our empirical analysis of 40 quality assessment reports identified two primary categories of
cleaning operations required for scientific texts: **Deletion** operations that remove
unwanted or low-value content, and **Modification** operations that repair and standardize
existing text without altering semantic meaning.

## Deletion Operations

Deletion operations identify and remove redundant, erroneous, low-information, or unnecessary
content from documents.

### Document Structural Deletion
* **Headers, footers, and page numbers**: Remove these non-content elements
* **Table of contents and navigation structures**: Remove page numbers and dotted leader lines
after chapter listings
* **Front and back matter**: Remove prefaces, acknowledgments, references, indexes, copyright
statements, and other standard book structural elements
* **Publication and metadata information**: Remove ISBN, publisher information, JSON format
metadata, and OCR internal markers

### Academic Content Deletion
* **Citation systems**: Remove in-text citation markers, complete reference lists, and
footnote systems
* **Cross-references**: Remove internal chapter references and editorial markers
* **Figure and table artifacts**: Remove image placeholders, invalid URLs, and figure
reference markers

### Invalid and Redundant Content Deletion
* **Placeholder text**: Remove incomplete content markers and page continuation notices
* **Decorative elements**: Remove ornamental symbols, pure separator lines, and excessive
formatting markers
* **Duplicate information**: Remove multiple occurrences of identical content
* **Non-target language content**: Remove text segments not in the target language

## Modification Operations

Modification operations adjust, standardize, correct, or transform existing text content to
make it more regular, readable, and suitable for model processing.

### Mathematical and Symbolic Standardization
* **LaTeX formula format unification**: Convert various LaTeX formats to standard Markdown
math notation
* **Escape character cleanup**: Clean redundant backslashes and LaTeX command residue
* **Special symbol handling**: Standardize representation of mathematical symbols, Greek
letters, and discipline-specific notation

### Text Continuity Repair
* **Word break and line break repair**: Merge words and sentences incorrectly split by hyphens
or line breaks
* **Paragraph structure reconstruction**: Restore paragraph logic and text flow damaged by OCR
* **Whitespace normalization**: Compress consecutive blank lines and standardize spacing

### Format Structure Unification
```

```
* **Heading level standardization**: Convert uniformly to Markdown heading format and
standardize hierarchy
* **Academic format standardization**: Unify format markers for theorems, definitions, proofs,
and other academic concepts
* **List and table repair**: Fix Markdown syntax errors and handle structured content

### Character and Encoding Correction
* **OCR recognition error correction**: Fix character recognition errors and spelling issues
* **Punctuation standardization**: Standardize usage of quotation marks, dashes, and other
punctuation
* **Character encoding handling**: Fix special Unicode characters and encoding display issues

### Semantic and Structural Preservation
* **Logical relationship maintenance**: Maintain semantic coherence and logical structure
during cleaning
* **Context integrity**: Ensure modification operations don't disrupt overall text meaning
* **Complex layout reconstruction**: Handle multi-column layouts, text-image mixed
arrangements, and other complex formatting

## Processing Principles

Two overarching principles guide Content Cleaning processing:

* **Deletion priority**: Systematically delete irrelevant content first, then perform format
modifications and text repairs
* **Semantic protection**: Prioritize maintaining original text meaning and logical
relationships throughout all operations
```

Figure 7: Content Cleaning Quality Analysis Report Synthesis

Content Cleaning Prompt

```
You are an expert document cleaner specialized in identifying and removing unwanted content
and correcting OCR errors from various document (mainly academic) chunks.

## Objective:
Clean and standardize OCR text by identifying and removing redundant, erroneous, or unwanted
content and correcting obvious OCR errors according to the rules below. Your task is to
identify and delete unnecessary content completely, fix technical errors, while preserving all
academic value.

## Deletion and Correction Rules:
### Document Structural Deletion
* Remove **table of contents and navigation structures**: Multiple consecutive chapter/section
titles listed together without accompanying text content
  - **Preserve content section headings in main text**: such as chapter headings, section
  titles followed by explanatory text or academic material
* Remove **reference lists completely**: numbered entries with author names, publication
titles, and years (e.g., "1. Smith, J. (2020). Title. Journal, 15(3), 123-145.") **[Delete
entire list regardless of format]**
* Remove **front matter and back matter**: such as prefaces, acknowledgments, copyright
statements, indexes, and other standard book structural elements
  - **Preserve sections with academic value**: such as abstracts, introductions, conclusions
  that present research background or methodology
* Remove **publication and metadata information**: such as ISBN, publisher information,
revision history, version numbers, institutional affiliations, author affiliations, addresses,
contact information
* Remove **page headers, page footers, and page numbers**

### Academic Content Deletion
* Remove **pure indexing appendices**: such as glossaries, symbol tables, abbreviation lists,
indexes, notations and other purely referential lookup content (entries that only provide
definitions without explanations, e.g., "a - alpha coefficient")
  - **Preserve**: appendices with learning value (e.g. mathematical derivations, proofs,
  technical explanations)
```

```
  - **Preserve**: explanatory content that directly supports main text elements (e.g.
  abbreviation/parameter explanations after tables/formulas/diagrams)
* Remove **image files and placeholders**: such as `<img>` tags, image file paths, image URLs,
markdown syntax and image placeholders (e.g. `[Image]`, `[Picture not available]`)
  - **Preserve**: figure/table titles, descriptive text (including content within markdown
  image formats: ![description](path)  description)
  - **Preserve**: in-text references (e.g., "as shown in Figure 1")

### Invalid and Redundant Content Deletion
* Remove **OCR processing artifacts**: such as garbled text, encoding artifacts, duplicate
characters, malformed special characters, OCR messages (`[OCR error]`), file paths,
timestamps, version numbers, revision history
* Remove **garbage content**: such as junk information, advertising content, placeholders (e.g.
[Insert citation here])
* Remove **duplicate content**: identical paragraphs or sections mainly caused by OCR errors
  - **Exception**: Carefully apply to technical formulas, equations, or specialized notation
  that may contain subtle but meaningful differences
  - **Exception**: Apply contextual analysis - preserve identical content that serves different
  semantic purposes or artistic purposes (e.g., poetic refrains, literary repetition)
* Remove **content and navigation markers**: [content missing], [page break], (Continued), and
similar placeholder markers
* Remove **URLs and links**: all web addresses, hyperlinks, and link information

### OCR Error Correction
* **Fix text fragmentation**: repair split words, broken sentences, erroneous line breaks and
paragraph divisions, missing spaces and punctuation
* **Fix fragmented structured content**: Repair OCR-damaged structured content (e.g. tables,
diagrams, formulas) appearing as consecutive lines of isolated words, single characters, or
short phrases
  - **Pattern**: Consecutive lines (5+) with 1-3 words/characters each
  - **Action**: Preserve content while indicating structural damage; delete if unrepairable
* **Standardize whitespace and formatting**: clean excessive whitespace, compress blank lines,
standardize spacing and indentation
* **Fix character and encoding errors**: correct obvious character errors, spelling issues,
and Unicode anomalies
* **Standardize punctuation**: unify quotation marks, dashes, hyphens, and other punctuation
* **Complete truncated words**: only fix obviously incomplete words from clear OCR errors,
avoid modifying content at chunk edges
* **Standardize academic formatting**: remove excessive LaTeX commands and unify notation
format

## Content Protection Rules:
### Always Preserve Academic and Educational Content
* Preserve **Technical and specialized content**: such as formulas, equations, proofs, symbols,
chemical structures, biological sequences and their original format
  - **Preserve exact content**: do not alter variables, coefficients, structures, sequences,
  or any technical details
* Preserve **In-text references and citations**: such as (Smith, 2020), [15], "see Chapter 2",
equation (5), "Figure 2.5", (pp. 3-7)
* Preserve **Table structures**: preserve academic table content, formatting and structural
markers (e.g. "|", HTML tags)
  - **Exception**: Does not apply to navigation tables (table of contents, indexes,
  glossaries) which should be removed
* Preserve **Code blocks and programming examples**: preserve code block markers (```language,
```, etc.) and internal code syntax and structure
* Preserve **Educational content**: such as exercises, questions, answers, solutions, case
studies, instructions, user guides
* Preserve **Explanatory content**: such as NOTE boxes, WARNING boxes, tips, author comments,
supplementary information, academic footnotes
* Preserve **Chunk boundary content**: incomplete sentences and words at chunk edges due to
text segmentation
* Preserve **Literary and humanities content**: including poetry, fiction, drama, creative
writing, literary analysis, philosophical texts, and other humanities scholarship with
educational value

Instructions:
- Carefully identify all content matching the deletion rules
- Remove completely any content that should be deleted
```
```

```
- Preserve all valuable academic content by applying protection rules and retaining content
that doesn't match deletion rules
- Apply OCR error corrections to fix obvious technical problems
- Ensure text flows naturally after corrections and deletions
- If the entire chunk should be deleted, leave the output tags completely empty
- **Important**: The content inside the <CLEANED_TEXT> tags must be exactly the text after
deletion, with no explanations, comments, or additional text inside the tags

## Input:
OCR document chunk:
[CHUNK]

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

Figure 8: Content Cleaning Prompt

Content Cleaning Evaluation Prompt

````
# Data Cleaning Quality Evaluation Prompt

You are an expert in data preprocessing and text cleaning quality assessment. Your task is to
evaluate text data cleaning quality by analyzing rule execution accuracy and rule completeness.
Focus specifically on the deletion and OCR error correction phases of cleaning - identifying
what was done incorrectly and what rules need improvement.

## Evaluation Focus
1. **Rule Execution Accuracy**: Check if cleaning rules were correctly applied to identify and
remove unwanted content, and if OCR correction rules were properly applied to fix technical
errors
2. **Rule Completeness**: Assess if the cleaning rules cover all necessary cases and are
clearly defined

Note: This evaluation focuses on deletion and OCR error correction phases. Advanced text
modification (restructuring, rewriting, semantic improvements) happens in a separate step and
should not be included in the scoring, but suggestions can be provided.

## Input:
Cleaning rules:
[CLEAN_RULES]

Text samples before and after cleaning:
[EVALUATION_INPUT]

## Output Format
```markdown
# Data Cleaning Quality Evaluation Report

## 1. Rule Execution Accuracy Analysis (By Rule)

*Evaluate each cleaning rule individually. Every rule must be assessed, even if it was
executed perfectly.*

### Rule 1: [Rule Name/Description]
**Execution Quality:** [Excellent/Good/Fair/Poor]
**Missed Deletions/Corrections:** [Number] instances (0 if none)
**Incorrect Deletions/Corrections:** [Number] instances (0 if none)

**Examples (if any issues):**
```
Chunk ID: [specify the chunk ID from the evaluation input]

Before cleaning: [copy exactly from the original text provided]
````

```
After cleaning: [copy exactly from the cleaned text provided - must be ACTUAL result, not what
you think it should be]

Problem: [highlight specific issues]
Explanation: [why this is problematic according to the rule]
```

If no issues: "No issues found - this rule was executed correctly throughout the text."

Rule 2: [Rule Name/Description]
Execution Quality: [Excellent/Good/Fair/Poor]
Missed Deletions/Corrections: [Number] instances (0 if none)
Incorrect Deletions/Corrections: [Number] instances (0 if none)

[Continue for EVERY cleaning rule provided - do not skip any rules]

Overall Execution Summary
Rules with Most Issues:
1. [Rule name] - [X missed deletions/corrections, X incorrect deletions/corrections]
2. [Rule name] - [X missed deletions/corrections, X incorrect deletions/corrections]
3. [Rule name] - [X missed deletions/corrections, X incorrect deletions/corrections]

2. Rule Completeness Analysis

2.1 Missing Rules (New cleaning needs discovered)
Content type found: [Describe unwanted content or OCR errors that current rules don't
address]
Suggested new rule: [Specific rule to handle this content/error]

Example:
```
Chunk ID: [specify the chunk ID from the evaluation input]

Problematic content found: [show the unwanted content or OCR error in text]

How it should be cleaned: [show desired result]
```

2.2 Existing Rules Needing Improvement
Rule name: [Specific rule that needs changes]
Problem: [What's wrong - ambiguity, inaccuracy, or other issues]
Suggested improvement: [How to fix the rule - modification or clarification]

Example:
```
Chunk ID: [specify the chunk ID from the evaluation input]

Current rule causes: [problematic cleaning result or inconsistent application]

After improvement should be: [improved result]
```

3. Additional Observations

Advanced Modification Suggestions (Not scored)
*Note: These are suggestions for advanced modification phase (restructuring, rewriting,
semantic improvements) and do not affect the current cleaning quality score.*

[Any suggestions for advanced text modification/restructuring improvements that should be
handled in the next phase]

4. Evaluation Summary

Overall Cleaning Quality
[Excellent/Good/Fair/Poor] - [Brief explanation of assessment reasoning based on deletion and
OCR correction accuracy]

Issues and Recommendations by Priority

```
**High Priority:** [Problems that significantly impact deletion or OCR correction accuracy]
- Issues: [specific problems]
- Recommendations: [concrete solutions]

**Medium Priority:** [Problems that affect cleaning consistency but not core quality]
- Issues: [specific problems]
- Recommendations: [concrete solutions]

**Low Priority:** [Minor cleaning optimization opportunities]
- Issues: [specific problems]
- Recommendations: [concrete solutions]
```

```
## Important Note
Provide honest assessment based on actual observations. Focus on whether content that should
be deleted was correctly identified and removed, and whether OCR errors were properly
corrected. If the cleaning quality is excellent with minimal issues, report that truthfully.
Don't artificially identify problems - accurate evaluation is more valuable than finding
issues where none exist.

Note that not every section in the report needs to be filled. If there are no issues in a
particular category, you can leave that section empty or state "No issues found in this
category."
```

Figure 9: Content Cleaning Evaluation Prompt

Example 1 (Text Before Content Cleaning)

Towards an (,2)-category of homotopy coherent monads in an -cosmos

THÈSE Nř 7748 (2017)
PRÉSENTÉE LE 22 SEPTEMBRE 2017
À LA FACULTÉ DES SCIENCES DE LA VIE
LABORATOIRE POUR LA TOPOLOGIE ET LES NEUROSCIENCES
PROGRAMME DOCTORAL EN MATHÉMATIQUES

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Dimitri ZAGANIDIS

acceptée sur proposition du jury:
Prof. M. Troyanov, président du jury
Prof. K. Hess Bellwald, directrice de thèse
Prof. D. Verity, rapporteur
Prof. E. Riehl, rapporteuse
Prof. Z. Patakfalvi, rapporteur
I would like to warmly thank my advisor Prof. Kathryn Hess Bellwald for her constant support and encouragement. Thank you for your interest, confidence, optimism and enthusiasm! Thank you for welcoming me in your research group, which provided me with the best work environment I could dream of. You have been a source of inspiration from the beginning.

I would like to thank all the past and present members of Kathryns group, Varvara Karpova, Marc Stephan, Kay Werndli, Martina Rovelli, Rachel Jeitziner, Justin Young, Gavin Seal, Magdalena Kedziorek, Gard Spreemann, Martina Scolamiero, Jean Verette and Lyne Moser for their friendliness, help and great discussions.

A special thank to Jérôme Scherer, who has given me the opportunity to teach geometry to the high potential teenagers of the Euler program. It has been an interesting and challenging experience!

I want to warmly thank Prof. Emily Riehl and Prof. Dominic Verity. First of all, I am indebted to them mathematically speaking. The motivation for this thesis came from two different sources. Firstly, an article of Ross Street [50] that we studied in the Kan extension seminar, which was organized by Emily Riehl in 2014. Secondly, from the series of articles

by Emily Riehl and Dominic Verity [43, 45, 44]. Dominic Verity is the father of weak complicial sets, the idea to use some of them as models of $(\infty, 2)$-categories came from a lecture by Emily Riehl at the Higher Structure conference at MATRIX, Australia, in June 2016. This wonderful conference was organized by Marcy Robertson and Philip Hackney, and I take the opportunity to thank them for organizing such a great event. Finally, I am grateful that Emily Riehl and Dominic Verity accepted to be members of my thesis jury.

I also want to thank the other jury members Prof. Zsolt Patakfalvi and Prof. Marc Troyanov for their precious time.

I dedicate this thesis to my family, who has always been supportive and encouraging, and in particular to my wife Cynthia, which provides me with so much love and happiness.
Abstract

This thesis is part of a program initiated by Riehl and Verity to study the category theory of $(\infty, 1)$-categories in a model-independent way. They showed that most models of $(\infty, 1)$-categories form an ∞-cosmos $\mathcal{K}$, which is essentially a category enriched in quasi-categories with some additional structure reminiscent of a category of fibrant objects. Riehl and Verity showed that it is possible to formulate the category theory of $(\infty, 1)$-categories directly with ∞-cosmos axioms. This should also help organize the category theory of $(\infty, 1)$-categories with structure.

Given an ∞-cosmos $\mathcal{K}$, we build via a nerve construction a stratified simplicial set $N_{\mathrm{Mnd}}(\mathcal{K})$ whose objects are homotopy coherent monads in $\mathcal{K}$. If two ∞-cosmoi are weakly equivalent, their respective stratified simplicial sets of homotopy coherent monads are also equivalent. This generalizes a construction of Street for 2-categories. We also provide an $(\infty, 2)$-category $\mathrm{Adj}_r(\mathcal{K})$ whose objects are homotopy coherent adjunctions in $\mathcal{K}$, that we use to classify the 1-simplices of $N_{\mathrm{Mnd}}(\mathcal{K})$ up to homotopy.

Key words: higher category, ∞-cosmos, $(\infty, 2)$-category, $(\infty, 1)$-category, homotopy coherent monad, model category
Résumé

Cette thèse sinscrit dans un programme initié par Riehl et Verity pour étudier la théorie des (, 1)-catégories dune façon qui ne dépend pas du modèle choisi. Ils ont montré que la plupart des modèles de (, 1)-catégories forme un -cosmos, cest-à-dire essentiellement une catégorie enrichie sur les quasi-catégories, munie de plus dune structure rappelant celle dune catégorie dobjets fibrants. Riehl et Verity ont montré quil est possible de formuler la théorie des catégories satisfaite par les (, 1)-catégories directement à partir des axiomes d-cosmos. Ceci devrait également aider à organiser la théorie des (, 1)-catégories munies dune structure.

Étant donné un -cosmos $\mathcal{K}$, nous construisons, grâce à une construction de nerf, un ensemble simplicial stratifié $\mathcal{N}_{\mathrm{Mnd}}(\mathcal{K})$ dont les objets sont les monades homotopiquement cohérentes dans $\mathcal{K}$. Si deux -cosmoi sont faiblement équivalents, leurs ensembles simpliciaux stratifiés des monades homotopiquement cohérentes respectifs sont également équivalents. Ceci généralise une construction de Street pour les 2-catégories. Nous fournissons également une (, 2)-catégorie $\mathrm{Adj}_r(\mathcal{K})$ dont les objets sont les adjonctions homotopiquement cohérentes dans $\mathcal{K}$ et que nous utilisons pour classifier les 1-simplexes de $\mathcal{N}_{\mathrm{Mnd}}(\mathcal{K})$ à homotopie près.

Mots clefs: catégorie dordre supérieur, -cosmos, (, 2)-catégorie, (, 1)-catégorie, monades homotopiquement cohérentes, catégorie de modèles
Contents

1 Introduction 1
1.1 Historical motivations 1
1.1.1 Monads and adjunctions in classical category theory
.... 1
1.1.2 Higher category theory 2
1.1.3 ∞-Cosmoi 5
1.2 Main contributions and organization of the thesis
..... 6

Notations and Terminology 9

2 Background Material 11
2.1 Enriched categories 11
2.1.1 Simplicial categories 11
2.1.2 2-Categories 16
2.1.3 Weighted limits and enriched right Kan extensions 22
2.2 Higher categories 26
2.2.1 Quasi-categories 26
2.2.2 Weak complicial sets 31
2.3 The universal 2-category containing an adjunction 36
2.3.1 The 2-categorical model 36
2.3.2 The simplicial model 38
2.3.3 The isomorphism $\mathcal{A}\lceil| \cong \mathcal{A}\lceil|$

..... 41
2.3.4 The Eilenberg-Moore object of algebras as a weighted limit 46
2.4 ∞-Cosmoi and their homotopy 2-category 47
2.4.1 Homotopy coherent monads and adjunctions 51
2.4.2 Absolute left liftings and left exact transformations 52
2.4.3 Monadicity theorem 54
2.5 The homotopy coherent nerve 58
3 The 2-category $\mathrm{Adj}^S_{\mathrm{hc}}[n]$ 65

3.1 The 2-category $\mathrm{Adj}[n]$ 67
3.1.1 The 2-categorical model 67
3.1.2 The simplicial model 70

3.2 Description of $\mathrm{Adj}^S_{\mathrm{hc}}[n]$ 73
3.2.1 The simplicial and 2-categorical models 73
3.2.2 Non-degenerate morphisms of $\mathrm{Adj}^S_{\mathrm{hc}}[n]$ 77
3.2.3 Atomic morphisms of $\mathrm{Adj}^S_{\mathrm{hc}}[n]$ 79
3.2.4 Convenient 2-subcategories of $\mathcal{C}\Delta[n]^{\mathrm{co}}$ 87

4 The lifting theorem 91

4.1 Right fillability and parent-child relation 92
4.1.1 Right fillable morphism and its distinguished face 92
4.1.2 Parent-child relation 95
4.1.3 Right parental relative subcomputads 100

4.2 Decomposition as an L-cell complex 102
4.2.1 A single pushout against a morphism in L 102
4.2.2 Transfinite composition 106

4.3 Proof of Theorem A 109
4.3.1 Lifting against a morphism of L_2 109
4.3.2 The proof 113

5 Homotopy coherent diagrams 115

5.1 Universal property of $\mathrm{Adj}^S_{\mathrm{hc}}[n]$ and of $\mathrm{Mnd}^S_{\mathrm{hc}}[n]$ 116

5.2 Minimal data defining a ho. coh. diag. of adjunctions 118
5.3 Induced ho. coh. diag. of free-forgetful adjunctions 123
5.3.1 A projective cell-complex 124
5.3.2 Identifying the domain of the right adjoints 129

6 Towards an $(\infty, 2)$-category of ho. coh. monads 133

6.1 Several stratified nerve constructions 134
6.2 Discussion about Conjecture B 137
6.3 Proof of Theorem C 146
6.4 Classification results 149
6.4.1 Description of $h(\mathrm{Adj}_r(K)|_1)$ 150
6.4.2 A reflective subcategory 153

A Proofs of background results 167

B Presentation by computads 175

Figure 10: Example 1 (Text Before Content Cleaning)

Example 1 (Text After Content Cleaning)

Abstract

This thesis is part of a program initiated by Riehl and Verity to study the category theory of $(\infty, 1)$-categories in a model-independent way. They showed that most models of $(\infty, 1)$-categories form an ∞-cosmos $\mathcal{K}$, which is essentially a category enriched in quasi-categories with some additional structure reminiscent of a category of fibrant objects. Riehl

and Verity showed that it is possible to formulate the category theory of $(\infty, 1)$-categories directly with ∞-cosmos axioms. This should also help organize the category theory of $(\infty, 1)$-categories with structure.

Given an ∞-cosmos $\mathcal{K}$, we build via a nerve construction a stratified simplicial set $N_{\text{Mnd}}(\mathcal{K})$ whose objects are homotopy coherent monads in $\mathcal{K}$. If two ∞-cosmoi are weakly equivalent, their respective stratified simplicial sets of homotopy coherent monads are also equivalent. This generalizes a construction of Street for 2-categories. We also provide an $(\infty, 2)$-category $\text{Adj}_r(\mathcal{K})$ whose objects are homotopy coherent adjunctions in $\mathcal{K}$, that we use to classify the 1-simplices of $N_{\text{Mnd}}(\mathcal{K})$ up to homotopy.

Key words: higher category, ∞-cosmos, $(\infty, 2)$-category, $(\infty, 1)$-category, homotopy coherent monad, model category

Résumé

Cette thèse sinscrit dans un programme initié par Riehl et Verity pour étudier la théorie des $(\infty, 1)$-catégories dune façon qui ne dépend pas du modèle choisi. Ils ont montré que la plupart des modèles de $(\infty, 1)$-catégories forment un ∞-cosmos, cestàdire essentiellement une catégorie enrichie sur les quasicatégories, munie de plus dune structure rappelant celle dune catégorie dobjets fibrants. Riehl et Verity ont montré quil est possible de formuler la théorie des catégories satisfaite par les $(\infty, 1)$-catégories directement à partir des axiomes d∞-cosmos. Ceci devrait également aider à organiser la théorie des $(\infty, 1)$-catégories munies dune structure.

Étant donné un ∞-cosmos $\mathcal{K}$, nous construisons, grâce à une construction de nerf, un ensemble simplicial stratifié $\mathcal{N}_{\text{Mnd}}(\mathcal{K})$ dont les objets sont les monades homotopiquement cohérentes dans $\mathcal{K}$. Si deux ∞-cosmoi sont faiblement équivalents, leurs ensembles simpliciaux stratifiés des monades homotopiquement cohérentes respectifs sont également équivalents. Ceci généralise une construction de Street pour les 2catégories. Nous fournissons également une $(\infty, 2)$-catégorie $\text{Adj}_r(\mathcal{K})$ dont les objets sont les adjonctions homotopiquement cohérentes dans $\mathcal{K}$ et que nous utilisons pour classifier les 1simplexes de $\mathcal{N}_{\text{Mnd}}(\mathcal{K})$ à homotopie près.

Figure 11: Example 1 (Text After Content Cleaning)

Commentary for Example 1

This example illustrates effective application of Content Cleaning deletion and modification operations:

Structural Deletion: The processing correctly identified and removed all standard thesis front matter elementsdegree information, institutional affiliations, jury composition, acknowledgments, and table of contentswhile preserving both English and French abstracts that contain essential research summaries.

Mathematical Notation Standardization: Several improvements to mathematical formatting enhance readability and consistency: (1) LaTeX expressions are uniformly formatted (e.g., $(\infty, 1)$-categories maintains consistent spacing); (2) special characters in the French abstract (e.g., "c'estàdire") are properly rendered with Unicode hyphens rather than simple dashes; (3) mathematical symbols within bilingual text preserve their LaTeX notation across both languages, ensuring technical precision.

Content Protection: All academically valuable elements were preserved intactresearch abstracts in both languages, keyword lists, mathematical definitions, and technical terminologydemonstrating the effectiveness of our content protection guidelines in distinguishing structural artifacts from substantive academic content.

Overall: This example validates Content Cleaning's ability to clean academic front matter comprehensively while maintaining the integrity of research content and improving the standardization of mathematical notation across multilingual documents.

Figure 12: Commentary for Example 1

Example 2 (Text Before Content Cleaning)

Consider SDE

$$d\xi^x(t) = \exp_{\xi^x(t)} \left(\left(C_{\xi^x(t)}(t)dt + C_{\xi^x(t)}(t)dw(t) \right) + \int_{\mathbb{E}} \gamma(\xi^x(t), \alpha)\gamma(\xi^x(t), \alpha) \right)$$

where $\mathbb{E}$ is a typical layer of the bundle π, $\gamma(\xi^x(t), \alpha)$ be a Poisson random measure on $\mathbb{E} \times \mathbb{E}$ with a mean value $\mathbb{E}\gamma(\xi^x(t), \alpha) = \gamma(t, \alpha)$ and $\gamma(t, \alpha)$ be a bounded measure on $\mathbb{E}$,

$$C_{\xi^x} = (\alpha_x, B_x h - \Gamma^x_{ij}(\alpha_x, h)),$$

$$\gamma(\alpha) = (0, f_x(\alpha)h),$$

527

z C
B*

xe
6- [(E j

y

J

h

e

B

,

*

L

)

W , g))

In a local trivialization the equation {Q)
form
deit)
: c>H-lcH s
fit)
(n$^r tt$
n
$S.t$;
$>$
2 −

has the

I

1 (r"

f-L

č tY)

5.11 j

Y

/rt

,

,

It)) V -It

’ c

Theorem 2. Let

I H d w - i T* ^
£if-J
Z

-

A-

i* 9lt)+

f^J

(

,

f"

<f) <fK
S’"
is)

ii'
fit, ’ fit)

’ 6 , fc) (? fi), d h’if)) i

theorem 1 conditions be valid and
C 1 - smooth

> /x, f x 5 e

bounded fields on ty

Then there exists a uni que Markov

process

fying

%’s.) ^{z}.

)

and the condition

) satis-

In a local trivialization the process

has the

representation
Csrt), $'^{-4}$
– (S l i . s) ’
where

is a random evolutionary family of
ma^{s}of

y

, Sf-O$^{/}Oaresolutioaacf$

correspondingly. The relation

-

defines a multiplicative operator functional
6

process

(5), IG)

acting from

of the

Ji'Vs'**)

to

3 Consider parabolic equations both with respect to
scalar functions
x.) and sections of vector
bundles
$^{i}s, x)$

- Vv

+

v

-

o

ń

+

528

* J- Tr f V
t- 5

£"foe,

VT' *

c$^{H}lcl\}$

r O

g

?

£8)

with V j N?
being covariant derivatives corresponding to r "j r" ana
Theorem 3. Let the conditions of theorems 1 and

2 are valid. Then there exists a unique classical solution of the
equation (7) such that
t
and a unique classical solution of 18satisfying;
being C smooth bounded functions. Those solutions
may be represented in the form
ut*.*
) x
. -f (s.i+0,
H }> 9
(
s
.
*
.
if
i
-x
C
R e f e r e n c e s
1. Belopolskaya Ya. I., Dalecky Yu. L. Ito equations
and differential geometry.- Usp. mat. nauk, N 3ż 1982,
p. 95 - 14-2.
2. Belopolskaya Ya. I., Dalecky Yu. L. Diffusion
processes in Banach spaces and manifolds. Tr. MMO,
v. 37, M.t Nauka, 1978, p.107 - 141.
3. Belopolskaya Ya.I. On stochastic equations with
unbounded coefficients for jump processes. Lecture notes in
Control, Springer, 1980, p.245 - 254.
4. Belopolskaya Ya.I. Markov processes with jumps
and integrodifferential systems. Tr. of intern, symp.
on differential equations. Vilnus, 1978.

The purpose of this lecture will be to report on the developments of the last five years on the above topic. The subject may be phrased in non-probabilistic terms as the study of local solutions of certain partial differential equations on Riemannian manifolds.

In 1976 Debiard, Gaveau and Mazet [DGM] discovered comparison theorems for the transition function and exit time of the Brownian motion from a geodesic sphere of a Riemannian manifold, and expressed the results in terms of sectional curvature of the metric. These theorems, which correspond to the Rauch comparison theorems of non-stochastic differential geometry, do not give sharp results when applied to a small geodesic ball. Meanwhile Gray and VanHecke in 1979 [GV] made a (non-probabilistic) study of the volume of small geodesic balls in a Riemannian manifold, in an effort to use the volume to characterize the metric, at least for a class of model spaces. This effort succeeded only in dimension less than 4, where certain unpleasant examples were found. This led us to attempt to use the mean exit time of Brownian motion as a stochastic substitute for the volume. By refining the Debiard-Gaveau-Mazet methods to obtain an asymptotic expansion of the mean exit time [GP], we obtain many new candidates for "domain functionals", in order to characterize the metric by a global geometric quantity. These issues, which may be categorized under the heading "Can you feel the shape of a manifold by Brownian motion", are discussed in a survey paper of the same title [Pi].

Limit theorems for Brownian motion in a small ball may be discussed by analogy with classical limit theorems of probability theory. The exit time from a small ball obeys a sort of "central limit theorem", 529
where the limit law is that of the exit time from the unit ball of $\mathbb{R}^n$, irrespective of the Riemannian metric (the law of large numbers is trivial here, since the mean zero condition is assured by the absence of a drift term in the generator of the diffusion (we are considering pure Brownian motion, defined solely by the metric of the Riemannian manifold). In order to refine the central limit theorem, we seek an asymptotic expansion, analogous to results in classical probability theory. The coefficients in the expansion are geometric invariants.

Figure 13: Example 2 (Text Before Content Cleaning)

Example 2 (Text After Content Cleaning)

Consider SDE

$$d\xi^x(t) = \exp_{\xi^x(t)} \left(\left(C_{\xi^x(t)}(t)\, dt + C_{\xi^x(t)}(t)\, dw(t) \right) + \int_{\mathbb{E}} \gamma(\xi^x(t), \alpha)\, \gamma(\xi^x(t), \alpha) \right)$$

where $\mathbb{E}$ is a typical layer of the bundle π, $\gamma(\xi^x(t), \alpha)$ is a Poisson random measure on $\mathbb{E} \times \mathbb{E}$ with a mean value $\mathbb{E}\gamma(\xi^x(t), \alpha) = \gamma(t, \alpha)$ and $\gamma(t, \alpha)$ is a bounded measure on $\mathbb{E}$,

$$C_{\xi^x(t)} = (\alpha_x, B_x h - \Gamma^x_{ij}(\alpha_x, h)),$$

$$\gamma(\alpha) = (0, f_x(\alpha)h).$$

In a local trivialization the equation has the form and has the required smoothness and boundedness conditions.

Theorem 2. Let the conditions of Theorem 1 be valid and the fields be smooth and bounded. Then there exists a unique Markov process satisfying the required condition.

In a local trivialization the process has the representation

$$\text{process} = \text{random evolutionary family of operators},$$

where the family is a random evolutionary family of solutions of the corresponding equations, and the solutions are the solutions of the corresponding equations. The relation defines a multiplicative operator functional acting from the appropriate space to the target space.

Consider parabolic equations both with respect to scalar functions and sections of vector bundles.

Theorem 3. Let the conditions of Theorems 1 and 2 be valid. Then there exists a unique classical solution of equation (7) and a unique classical solution satisfying the condition $c_i O - c_i^*$ for R_1, M_{ji} being smooth bounded functions. Those solutions may be represented in the form

The purpose of this lecture will be to report on the developments of the last five years on the above topic. The subject may be phrased in nonprobabilistic terms as the study of local solutions of certain partial differential equations on Riemannian manifolds.

In 1976 Debiard, Gaveau and Mazet [DGM] discovered comparison theorems for the transition function and exit time of Brownian motion from a geodesic sphere of a Riemannian manifold, and expressed the results in terms of sectional curvature of the metric. These theorems, which correspond to the Rauch comparison theorems of nonstochastic differential geometry, do not give sharp results when applied to a small geodesic ball. Meanwhile Gray and VanHecke in 1979 [GV] made a nonprobabilistic study of the volume of small geodesic balls in a Riemannian manifold, in an effort to use the volume to characterize the metric, at least for a class of model spaces. This effort succeeded only in dimension less than 4, where certain unpleasant examples were found. This led us to attempt to use the mean exit time of Brownian motion as a stochastic substitute for the volume. By refining the DebiardGaveauMazet methods to obtain an asymptotic expansion of the mean exit time [GP], we obtain many new candidates for domain functionals in order to characterize the metric by a global geometric quantity. These issues, which may be categorized under the heading Can you feel the shape of a manifold by Brownian motion, are discussed in a survey paper of the same title [Pi].

Limit theorems for Brownian motion in a small ball may be discussed by analogy with classical limit theorems of probability theory. The exit time from a small ball obeys a sort of "central limit theorem", where the limit law is that of the exit time from the unit ball of $\mathbb{R}^n$, irrespective of the Riemannian metric (the law of large numbers is trivial here, since the mean zero condition is assured by the absence of a drift term in the generator of the diffusion (we are considering pure Brownian motion, defined solely by the metric of the Riemannian manifold). In order to refine the central limit theorem, we seek an asymptotic expansion, analogous to results in classical probability theory. The coefficients in the expansion are geometric invariants.

Figure 14: Example 2 (Text After Content Cleaning)

Commentary for Example 2

This example showcases Content Cleaning's capability to handle severe OCR damagea common challenge in digitized mathematical literaturewhile successfully applying both deletion and modification operations:

OCR Corruption Recovery: The original text exhibited extreme OCR damage with extensive character-level corruption: isolated single characters scattered across lines (e.g., "z C", "B*", "xe",

"6- [(E j", "y", "J", "h"), fragmented mathematical expressions, and garbled text throughout. Content Cleaning processing successfully identified these as OCR artifacts rather than meaningful content, removing the unrepairable fragments while attempting to preserve salvageable portions. The mathematical equations at the beginning were partially reconstructed, maintaining their LaTeX structure and essential notation.

Structural Deletion: The reference section at the document's end was correctly identified and completely removed, including the "References" heading and all four bibliographic entries (Belopolskaya and Dalecky citations). Additionally, the page number "527" and "529" were removed as page footer artifacts.

Text Continuity Restoration: Severely fragmented theorem statements were partially recovered. For instance, "Theorem 2" and "Theorem 3" sections, while heavily damaged in the original, were reconstructed into coherent (though simplified) statements. The cleaning preserved the logical structuretheorem numbering, conditions, and conclusionseven when full mathematical precision could not be recovered from corrupted input.

Mathematical Notation Preservation: Despite extreme corruption, critical mathematical elements were protected: LaTeX equation environments remained intact, variable names and operators in recoverable formulas were preserved (e.g., $\xi^x(t)$, $\mathbb{E}$, γ), and the scholarly narrative in the less-damaged final paragraphs (discussing Brownian motion and Riemannian manifolds) was fully retained with proper citation markers [DGM], [GV], [GP], [Pi].

Limitations and Trade-offs: This example also illustrates the boundaries of Content Cleaning processing: when OCR damage is catastrophic (as in the middle section with pure gibberish), the model cannot reconstruct missing mathematical content and instead produces simplified placeholder text. This represents a conservative approachpreserving what can be verified rather than hallucinating mathematical statementswhich is appropriate for maintaining corpus integrity even when complete recovery is impossible.

Overall: This extreme case validates Content Cleaning's robustness in handling severely corrupted scientific documents. While perfect reconstruction was impossible, the processing successfully removed artifacts, preserved salvageable content, maintained document structure, and produced output substantially more usable than the original corrupted textdemonstrating practical value even in worst-case OCR scenarios common in digitized legacy scientific literature.

Figure 15: Commentary for Example 2

Pair-wise Evaluation Prompt for Pedagogical Augmentation

```
You are a PhD student who has just started your research journey. You often encounter complex
academic papers that are difficult to understand, and you greatly appreciate materials that
can explain concepts in a more accessible and educational way.

Now you need to compare two different text processing methods to see which one better
transforms academic content into something you can easily comprehend and learn from.

(Due to context length limitations, the text provided below is a fragment of a paper, not the
complete document.)

## Original Academic Text
{original_text}
```

```
## Version Processed by {prompt_A}
{Rewritten text by prompt_A}

## Version Processed by {prompt_B}
{Rewritten text by prompt_B}

## Evaluation Perspective

As a PhD student still building your research foundation, please evaluate these two versions
based on:

1. **Absolute Fidelity to Original Content**
   - CRITICAL: Zero tolerance for factual errors, hallucinations, or content that contradicts
   the original text
   - Complete preservation of the original hierarchical structure (section headers,
   subsections, numbered points, etc.)
   - All essential technical details, definitions, theorems, and mathematical relationships
   must remain intact

2. **Educational Accessibility and Pedagogical Value**
   - Does the text transform dense academic jargon into language that a beginning graduate
   student can understand?
   - Are complex concepts broken down with helpful explanations, intuitive descriptions, or
   motivating examples?
   - Does it provide the kind of step-by-step reasoning and context that helps bridge
   knowledge gaps?
   - Are abstract ideas made more concrete through analogies or clearer exposition?
   - **Bonus points**: Thoughtful knowledge supplementation that aids comprehension without
   distorting original meaning

3. **Textual Flow and Coherence**
   - Does the text read smoothly and naturally, especially at section transitions?
   - Are connections between ideas made explicit and easy to follow?
   - Is the logical progression of arguments clear and well-maintained?
   - Does the text avoid awkward phrasings or abrupt transitions that might result from
   processing methods?


Please provide your output in the following format:

## Analysis
<Detailed analysis of the {prompt_A} and {prompt_B} cleaning methods, including their
respective advantages and disadvantages>

## Winner
{prompt_A} OR {prompt_A}

Note: You must choose one winner based on the comprehensive evaluation of the above three
dimensions. If they are very close, choose the one that performs slightly better overall.
```

Figure 16: Pair-wise Evaluation Prompt for Pedagogical Augmentation

Pedagogical Augmentation Prompt

```
You are a master science communicator and pedagogical expert. Your mission is to transform the
following dense, expert-level text chunk into vibrant, crystal-clear educational material.
Imagine you are creating a definitive learning resource for a bright but novice audience. Your
goal is not merely to simplify, but to deeply elucidate, making the complex intuitive and the
implicit explicit.

Your transformation will be governed by two sets of principles: the Core Mandate (what you
must actively do) and the Unbreakable Rules (what you must never violate).

### The Unbreakable Rules: Fidelity and Integrity
This principle is of paramount importance and must be followed without exception to ensure the
output is valid.
```

```
* **(a) Scientific and Factual Correctness:** Maintain absolute rigor. All data, formulas,
definitions, theories, experimental results, and logical arguments must be preserved without
altering their meaning or context. Your additions must clarify, not contradict.

* **(b) Structural Integrity:** Preserve the original structure flawlessly. Keep ALL section
headers (`##`, `###`), figure/table labels, equation numbers, 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 Mandate: Deep Pedagogical Transformation
This is your primary objective. Be bold and proactive in adding educational value. Your goal
is to weave a rich tapestry of understanding.

* **(a) Deconstruct and Narrate the 'Why':** This is your primary mode of explanation.
Actively expand on logical leaps. When the text says "it follows that," "clearly," or
"trivially," you must step in and meticulously detail the intermediate steps. More importantly,
you must articulate the expert's internal monologue. When faced with an equation, a problem,
or a logical step, explain the strategy. Ask and answer questions like: "Okay, what's our goal
here?" "What's the first thing I should look for when I see an equation like this?" "We're
going to use technique X, and here's why it's the right tool for this specific job." Your
mission is to reveal the problem-solving journey, making every single connection transparent.

* **(b) From Jargon to Insight:** When you encounter a crucial technical term, or a
significant variable within a formula, you must deliberately pause the narrative to explain it.
Don't just provide a dry definition. Elucidate its importance: What role does this term or
variable play? Why does it matter? Crucially, you must then use simpler language, vivid
analogies, or concrete examples to build a strong and intuitive mental model for the reader
before you continue with the main explanation. This ensures no reader is left behind due to
unfamiliar notation or terminology.

* **(c) Invent Vivid Analogies and Concrete Examples:** Go beyond the text. Where a concept is
abstract, create a simple, concrete example to illustrate it. Invent memorable analogies that
connect the new information to a learner's existing knowledge (e.g., electron shells as floors
in a hotel).

* **(d) Create Contextual Bridges:** Weave a narrative thread by connecting the current idea
to the broader field of knowledge. Hint at future applications or link back to more
foundational concepts. For instance: "This principle of [X] is a cornerstone of the field and
will be essential for understanding [Y] later on."

* **(e) Think Like a Learner:** Proactively identify points of potential confusion. What
questions would a curious student ask here? Answer them before they are asked. A great teacher
warns students about common mistakes. Where applicable, insert brief, helpful asides that feel
like a mentor's margin notes.

* **(f) Prioritize Narrative Flow and Clean Formatting:** When you encounter messy or noisy
original LaTeX formatting, convert it to a clean and pristine style (especially for formulas
and tables). Above all, strive for a smooth, cohesive, and engaging narrative. Your writing
should feel like a continuous, guided tour through the material, not a collection of
disconnected facts and callouts. To that end, you must avoid the overuse of overly-structured,
point-by-point expressions. Let the main text flow logically and tell a story, adopting the
persona of an extremely patient and encouraging teacher.

---
Summary of Principles Above: To sum it up, you must strictly respect the accuracy and
structure of the original chunk while doing everything possible to make the rewritten text
easier to learn and to lower the reader's cognitive load. Consequently, the rewritten text
will typically be more detailed and thus longer than the original.
---

**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.**

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

**Original text:**
{chunk}

**Refined text:**
```

Figure 17: Pedagogical Augmentation Prompt

Example provided for Human Expert (Text Before Pedagogical Augmentation Processing)

[Theorem 3.55, (**?**, Theorem 5.11)] Let T be a compact operator on a complex, infinite-dimensional Banach space. Then T is not hypercyclic.

Proof. We argue by contradiction. Suppose that T is a compact hypercyclic operator on a complex, infinite-dimensional Banach space X, and let $\lambda \in \sigma(T)$ with $\lambda \neq 0$.
First, $\mathrm{ran}(\lambda I - T)$ must be dense in X. Otherwise, its closure $\overline{\mathrm{ran}(\lambda I - T)}$ would be a proper closed subspace of X that is invariant under T. However, since T is hypercyclic, its only closed invariant subspaces are $\{0\}$ and X, which leads to a contradiction.
By a previous theorem, $\mathrm{ran}(\lambda I - T)$ is also closed. Hence it must be the whole space X, and therefore $\lambda I - T$ is surjective. This implies that λ is an eigenvalue of T.
The spectrum of a compact operator consists either of a sequence of eigenvalues converging to 0 together with 0, or of a finite set containing 0. In either case, there exists a connected component of $\sigma(T)$ that does not intersect the unit circle. This contradicts Kitais Theorem (**?**, Theorem 5.6), which states that if T is hypercyclic, then every connected component of $\sigma(T)$ must intersect the unit circle. □

Figure 18: Example provided for Human Expert (Text Before Pedagogical Augmentation Processing)

Example provided for Human Expert (Text After Pedagogical Augmentation Processing)

[Theorem 3.55, (**?**, Theorem 5.11)] Let T be a compact operator on a complex, infinite-dimensional Banach space. Then T cannot be hypercyclic.

Proof. We again argue by contradiction.
Assume that there exists a complex infinite-dimensional Banach space X and an operator $T\colon X \to X$ which is both compact and hypercyclic.
Recall that hypercyclicity means the existence of a vector $x \in X$ whose orbit

$$\{T^n x \mid n \geq 0\}$$

is dense in X. Intuitively, repeated application of T allows the vector x to approach arbitrarily closely any point in the space.
Now consider the spectrum $\sigma(T)$. Since T is compact and X is infinite-dimensional, Theorem 3.54 ensures that $0 \in \sigma(T)$. Besides 0, the spectrum may contain nonzero points. Choose $\lambda \in \sigma(T)$ with $\lambda \neq 0$.
We analyze the range of the operator $\lambda I - T$. First, we claim that $\mathrm{ran}(\lambda I - T)$ must be dense in X. Indeed, if it were not dense, then its closure

$$\overline{\mathrm{ran}(\lambda I - T)}$$

would be a proper closed subspace of X. Moreover, this subspace is invariant under T. However, a fundamental property of hypercyclic operators is that they admit no nontrivial closed invariant subspaces; the only such subspaces are $\{0\}$ and X. This contradiction shows that $\mathrm{ran}(\lambda I - T)$ is dense.

On the other hand, by standard results in the spectral theory of compact operators, when $\lambda \neq 0$ the range $\mathrm{ran}(\lambda I - T)$ is not only dense but also closed. A set that is both dense and closed must coincide with the whole space. Hence,

$$\mathrm{ran}(\lambda I - T) = X,$$

so $\lambda I - T$ is surjective.
Since T is compact and $\lambda \neq 0$, surjectivity of $\lambda I - T$ forces it to be invertible, which would imply $\lambda \notin \sigma(T)$. This contradicts our choice of λ. Therefore, our assumption that T is hypercyclic must be false.
Equivalently, every nonzero spectral value of a compact operator must be an eigenvalue, and the spectrum of T consists either of finitely many points (including 0) or of a sequence of eigenvalues converging to 0 together with 0. In particular, $\sigma(T)$ has a connected component that lies entirely away from the unit circle.
This contradicts Kitais Theorem, which states that if T is hypercyclic, then every connected component of $\sigma(T)$ must intersect the unit circle $\{z \in \mathbb{C} : |z| = 1\}$.
Hence, no compact operator on a complex infinite-dimensional Banach space can be hypercyclic. □

Figure 19: Example provided for Human Expert (Text After Pedagogical Augmentation Processing)

Expert Manual Verification for Example

Expert Comment:
The rewritten proof is mathematically faithful to the original argument while offering a noticeably clearer and more pedagogical presentation. All essential logical stepssuch as the use of invariant subspaces, the closedness and density of the range of $\lambda I - T$, and the spectral structure of compact operatorsare preserved without distortion or oversimplification. The final contradiction with Kitais theorem is reached in exactly the same way as in the original text.
What distinguishes the rewrite is its improved explanatory flow. Key concepts like hypercyclicity and the behavior of compact operators on infinite-dimensional spaces are briefly recalled at the points where they become relevant, which makes the argument easier to follow for readers who may not have all results immediately at hand. Subtle points that were implicit before, especially the role of surjectivity and its incompatibility with spectral membership, are now articulated more transparently.
Overall, the rewritten version maintains full mathematical rigor while enhancing readability and educational value. It would be particularly well suited for inclusion in a graduate-level thesis or lecture notes, where clarity and conceptual guidance are as important as correctness.

Figure 20: Expert Manual Verification for Example

Example 1 (Text Before Pedagogical Augmentation Processing)

Theorem 1. *Assume that $f_0 \in L^2(\mathbb{R}^N, \mathcal{M}(v)^{-1}\mathrm{d}v\mathrm{d}x)$, where $\mathcal{M}(v)$ is the unique normalized equilibrium distribution that satisfies . Then, up to a subsequence, the solution f^ε of converges weakly in $L^\infty(0,T; L^2(\mathbb{R}^{2d}, \mathcal{M}(v)^{-1}\mathrm{d}v\mathrm{d}x))$ to $\rho(t,x)\mathcal{M}(v)$ as $\varepsilon \to 0$, where $\rho(t,x)$ solves*

$$\begin{cases} \partial_t \rho + (-\Delta_x)^s \rho = 0\,, \\ \rho(0,x) = \rho_{in}(x) := \int_{\mathbb{R}^d} f_{in}(x,v)\mathrm{d}v\,. \end{cases}$$

In the classical case (i.e., s=1) when $\mathcal{M}$ is a fast decaying function such as Gaussian, one rescales t as $t \mapsto \varepsilon^2 t$ and the resulting macroscopic equation is the diffusion equation :

$$\partial_t \rho + \nabla_x \cdot (D \nabla_x \rho) = 0,$$

where D is the diffusion matrix

$$D = \int v \otimes v \mathcal{M} \mathrm{d}v\,.$$

Clearly the fat tail equilibrium renders the above integral unbounded and therefore invalids the classical diffusion limit. Conversely, the anomalous scaling is necessary. Similar scaling has also been investigated in the framework of linear Boltzmann equation, see for a reference.

Figure 21: Example 1 (Text Before Pedagogical Augmentation Processing)

Example 1 (Text After Pedagogical Augmentation Processing)

Theorem 1. *Assume that $f_0 \in L^2\big(\mathbb{R}^N, \mathcal{M}(v)^{-1}\,\mathrm{d}v\,\mathrm{d}x\big)$, where $\mathcal{M}(v)$ is the unique normalized equilibrium distribution satisfying eqn:equilibrium. Then, up to extraction of a subsequence, the solutions f^ε of eqn:111 converge weakly in*

$$L^\infty\big(0,T;\, L^2(\mathbb{R}^{2d}, \mathcal{M}(v)^{-1}\,\mathrm{d}v\,\mathrm{d}x)\big)$$

*to a limit of the form $\rho(t,x)\,\mathcal{M}(v)$ as $\varepsilon \to 0$. The macroscopic density $\rho(t,x)$ solves the **fractional diffusion equation***

$$\begin{cases} \partial_t \rho + (-\Delta_x)^{\,s}\,\rho = 0, \\ \rho(0,x) = \rho_{\mathrm{in}}(x) := \displaystyle\int_{\mathbb{R}^d} f_{\mathrm{in}}(x,v)\,\mathrm{d}v. \end{cases}$$

In other words, the kinetic model converges to a non-local heat equation in which the Laplacian is replaced by its fractional counterpart $(-\Delta_x)^s$. *
Here the weak convergence means that for any test function $\phi(t,x,v)$ smooth and compactly supported, the integrals

$$\int_0^T \int_{\mathbb{R}^{2d}} f^\varepsilon(t,x,v)\,\phi(t,x,v)\,\mathcal{M}(v)^{-1}\,\mathrm{d}v\,\mathrm{d}x\,\mathrm{d}t$$

converge to the corresponding integrals with $\rho(t,x)\mathcal{M}(v)$ in place of f^ε.
Thus, on the large-scale, long-time limit, the Lévy-FokkerPlanck dynamics are faithfully described by a fractional diffusion equation for the particle density.
In the classical case (i.e., $s = 1$) when the equilibrium distribution $\mathcal{M}$ is a fast-decaying function such as a Gaussian, one rescales the time variable as $t \mapsto \varepsilon^2 t$. Under this diffusive scaling the macroscopic limit is the standard diffusion equation

$$\partial_t \rho + \nabla_x \cdot \big(D\,\nabla_x \rho\big) = 0,$$

where the diffusion matrix D is given by the second-order moment of the equilibrium

$$D = \int v \otimes v\,\mathcal{M}\,\mathrm{d}v.$$

(Here $v \otimes v$ denotes the outer product of the velocity vector v with itself, producing a matrix whose (i,j) entry is $v_i v_j$.)
When $\mathcal{M}$ has a **fat tail** (i.e., decays only algebraically), the integral defining D diverges: the heavy-tailed equilibrium carries so much mass at large velocities that the second moment is infinite. Consequently the classical diffusion limit is no longer valid, and one must employ the **anomalous scaling** appropriate for fractional diffusion. Similar anomalous scalings have also been investigated for the linear Boltzmann equation; see, e.g., reference for a detailed discussion.

Figure 22: Example 1 (Text after Pedagogical Augmentation Processing)

Commentary for Example 1

From a mathematical standpoint, the processed version introduces several substantive pedagogical improvements that make the theorem far more accessible while maintaining full analytical rigor:

1. **Clarification of the analytical setting.** The rewritten text explicitly specifies the *type of convergence* (weak convergence in $L^\infty_t L^2_{x,v}$) and the precise functional framework that were only implicitly stated in the original. It also explains that the limiting quantity $\rho(t,x)\mathcal{M}(v)$ represents the macroscopic or averaged particle density, making the physical meaning of the limit transparent.

2. **Explicit linkage between microscopic and macroscopic equations.** The processed version draws a clear connection between the *kinetic equation* and the resulting *fractional diffusion equation*, explicitly stating that the kinetic model converges to a non-local heat equation. This pedagogical bridge helps readers understand how a kinetic transport process yields a macroscopic PDE in the asymptotic limit.

3. **Unpacking of key analytical concepts.** Several important notionssuch as *weak convergence*, *diffusive scaling*, and the *diffusion matrix*are explained in context. The text introduces the definition of weak convergence via test functions, clarifies the meaning of $v \otimes v$ in the diffusion

tensor D, and situates each concept within the logic of the proof. These elaborations convert terse symbolic statements into stepwise reasoning units suitable for learning.

4. **Intuitive explanation of anomalous diffusion.** The processed text goes beyond the formal statement to explain why the classical diffusion limit fails under heavy-tailed equilibria. It highlights that when $\mathcal{M}$ decays algebraically, the second moment diverges, invalidating the standard diffusion approximation and motivating the need for a fractional operator. This connects the analysis to the physical intuition of Lévy-type anomalous transport.

Overall: The processed version reconstructs the reasoning structure behind the theorem, linking functional analysis (weak convergence), PDE asymptotics (fractional diffusion), and probabilistic intuition (heavy-tailed transport). It transforms a compact specialist statement into a mathematically transparent and pedagogically rich exposition suitable for advanced teaching or graduate-level study.

Figure 23: Commentary for Example 1

Example 2 (Text Before Pedagogical Augmentation Processing)

The τ pair is produced back-to-back in the CM frame. As a result, the decay products of the two τ leptons can be separated from each other by dividing the event into two hemispheres the "signal" hemisphere and the "tag" hemisphere using the event thrust axis. The event thrust axis is calculated using all charged particles and all photon candidates in the entire event. We select events with one prompt track and a candidate reconstructed in the signal hemisphere, and exactly one oppositely charged prompt track in the tag hemisphere. A prompt track is defined to be a track with its point of closest approach to the beam spot being less than 1.5 in the plane transverse to the beam axis and less than 2.5 in the direction of the beam axis. Furthermore, if a pair of tracks is consistent with coming from a K^0_S or Λ decay, or from a conversion after a mass cut and a displaced vertex cut, neither track can be a prompt track. The components of momentum transverse to the beam axis for each of these two prompt tracks must be greater than 0.1 in the laboratory frame. The event is rejected if the prompt track in the signal hemisphere is identified to be coming from a charged kaon. A candidate is defined as a pair of oppositely charged pion candidates with invariant mass between 0.488 and 0.508; furthermore, the distance between the beam spot and the vertex must be at least three times its uncertainty (the will be referred to as the "candidate daughters"). To reduce backgrounds from nonτpair events, we require that the momentum of the charged particle in the tag hemisphere be less than 4 in the CM frame and be identified as an electron (e-tag) or a muon (μ-tag). To reduce backgrounds from Bhabha, $\mu^+\mu^-$, and $q\bar{q}$ events, we require the magnitude of the event thrust to be between 0.92 and 0.99.

Figure 24: Example 2 (Text Before Pedagogical Augmentation Processing)

Example 2 (Text After Pedagogical Augmentation Processing)

The τ-pair is produced back-to-back in the centre-of-mass (CM) frame. Because the two τ leptons fly in opposite directions, the particles that emerge from each τ can be cleanly separated by dividing the whole event into two opposite hemispheres. We call one hemisphere the signal hemisphere (where we look for the decay of interest) and the other the tag hemisphere (which we use to identify the partner τ). This division is made with respect to the *event thrust axis* a direction that maximises the sum of the longitudinal momenta of all particles, analogous to the axis of a pencil that best aligns with the flow of the event.

How the thrust axis is built. The thrust axis is calculated using **all** charged particles and all photon candidates in the event. In practice, every charged track and each neutral energy deposit in the electromagnetic calorimeter (EMC) contributes to determining this axis, ensuring that the hemispheres faithfully reflect the true geometry of τ-pair production.

Selection of tracks in the two hemispheres. We require exactly one *prompt* track in the signal hemisphere together with a reconstructed candidate, and exactly one oppositely charged prompt track in the tag hemisphere. A prompt track is defined as a track whose point of closest approach (PCA) to the beam spot satisfies: (1) transverse distance $< 1.5\,\mathrm{mm}$; and (2) longitudinal distance $< 2.5\,\mathrm{mm}$. These tight spatial cuts ensure that the track originates from the primary interaction point rather than from a secondary decay or detector artifact.

Removing tracks from secondary sources. If a pair of tracks is consistent with coming from a K^0_S or Λ decay, or from a photon conversion, we apply a mass cut and a displaced-vertex cut; when both are satisfied, neither of these tracks is considered prompt. This step prevents contamination from long-lived neutral particle decays and conversion backgrounds.

Transverse-momentum requirement. For each prompt track the transverse momentum must exceed $0.1\,\mathrm{GeV}/c$ in the laboratory frame, which suppresses very soft tracks that are difficult to reconstruct accurately.

Kaon veto in the signal hemisphere. If the prompt track in the signal hemisphere is identified as a charged kaon ($K^{\pm}$), the event is rejected. This veto efficiently removes strange-particle backgrounds that could mimic the signal topology.

Definition of the signal candidate. A candidate is reconstructed from two oppositely charged pion-like tracks ($\pi^+\pi^-$) with invariant mass between 0.488 and $0.508\,\mathrm{GeV}/c^2$. This narrow window centres on the ρ^0 resonance commonly produced in τ decays. The distance between the beam spot and the reconstructed vertex must be at least three times its uncertainty; the two pion tracks satisfying these criteria are referred to as the *candidate daughters*.

Tag-hemisphere momentum and lepton identification. To reduce non-τ backgrounds, the charged particle in the tag hemisphere must have momentum $< 4\,\mathrm{GeV}/c$ in the CM frame and must be identified as an electron (e-tag) or muon (μ-tag), consistent with typical leptonic τ decays.

Thrust-magnitude cut to reject Bhabha-type backgrounds. Events from Bhabha scattering ($e^+e^- \to e^+e^-$) or $\mu^+\mu^-$ production tend to have thrust ≈ 1. We therefore require the event thrust magnitude to lie between 0.92 and 0.99. This preserves the characteristic back-to-back topology of τ events while excluding overly collimated or spherical configurations.

Figure 25: Example 2 (Text After Pedagogical Augmentation Processing)

Commentary for Example 2

The L4-processed version transforms a dense detector-selection description into a clear, instructional narrative while retaining all experimental criteria. The main pedagogical advances are:

[label=**(8)**, leftmargin=1.5em, itemsep=4pt, topsep=4pt]

1. **Improved structural organization.** The rewritten text introduces subsections (How the thrust axis is built, Kaon veto, etc.), converting a monolithic paragraph into logically ordered experimental steps. This mirrors how an analysis procedure would be taught or reproduced in a lab manual.

2. **Physical and conceptual explanations.** Each selection criterion is accompanied by a brief rationalee.g., why the thrust axis isolates hemispheres, why small PCA cuts enforce promptness, and why the kaon veto suppresses strange backgrounds. These contextual notes transform a technical checklist into meaningful reasoning.

3. **Clarified numerical details and units.** Ambiguous quantities (e.g., distances 1.5 or 2.5) are expressed with explicit units and physical interpretation, ensuring dimensional clarity.

4. **Consistent particle-physics notation and readability.** The use of Greek letters, superscripts, and resonance names (ρ^0, K_S^0) follows standard conventions, improving precision and readability.

Overall: The Pedagogical Augmentation rewrite turns a procedural data-selection paragraph into a didactic exposition. It not only describes *what* cuts are applied but also *why* each exists, thereby serving both as documentation and as an educational explanation of event-selection logic in high-energy physics.

Figure 26: Commentary for Example 2

Rewrite Fidelity Evaluation Prompt

```
### Role:
You are a rigorous academic auditor with expertise across diverse scientific and technical
domains. Your task is to evaluate the fidelity of a "Pedagogically Restructured" version of a
scientific text against its original source.

### Inputs:
[EVALUATION_INPUT]

### Instructions:
```

```
Evaluate the Rewritten Text based on the following 10 distinct criteria. For each, answer ONLY
"Yes" or "No". "Yes" indicates the criteria is met (high fidelity/no hallucination).

#### Part 1: Content Fidelity (Preservation)
1. Core Claims: Does it preserve all primary conclusions and findings of the original text?
2. Key Entities: Are all essential technical terms, symbols, and domain-specific entities
retained correctly?
3. Logical Architecture: Does it maintain the original sequence of reasoning or the
fundamental causal chain?
4. Constraint Preservation: Does it respect the original boundary conditions, assumptions, or
limitations?

#### Part 2: Hallucination Control (No Inventions)
5. Zero Fabrication: Does it refrain from introducing non-existent data points, experimental
results, or citations?
6. Terminological Integrity: Are there NO "hallucinated" or made-up technical terms that do
not exist in the original context?
7. Attribution Accuracy: Does it avoid assigning claims to authors or sources that were not
mentioned in the original?

#### Part 3: Technical Rigor (No Distortion)
8. Logical Soundness: Is the step-by-step "decompressed" explanation free of logical fallacies
or non-sequiturs?
9. Semantic Consistency: Does the pedagogical rephrasing maintain the exact same meaning as
the original dense prose?
10. Outcome Alignment: Does the rewritten version lead to the same final conclusion without
over-simplification or distortion?

### Output Format:
[Criterion 1-10]: Yes/No
Total Fidelity Score: [X]/10
Summary of Discrepancies (if any): [Briefly list any 'No' findings]
```

Figure 27: Rewrite Fidelity Evaluation Prompt

Q&A Generation Prompt

```
First, evaluate whether the provided content contains sufficient professional knowledge to
create a challenging expert-level question. If the content is fragmented (like indexes/lists),
lacks substantial professional/technical content, or is unsuitable for professional knowledge
testing, directly return "No QA".

If suitable, create a multiple choice question based on the professional knowledge in the
provided content. The correct answer must be verifiable from the provided content. Use this
JSON format with 7 options:

"question": "the question",
"correct_option": "the accurate choice supported by the content",
"reference": "exact text excerpt that supports the correct answer",
"incorrect_option_1": "the first incorrect choice",
"incorrect_option_2": "the second incorrect choice",
"incorrect_option_3": "the third incorrect choice",
"incorrect_option_4": "the fourth incorrect choice",
"incorrect_option_5": "the fifth incorrect choice",
"incorrect_option_6": "the sixth incorrect choice"


Requirements:
- Design an expert-level challenging question that tests professional field knowledge
- Focus on professional information of the field rather than the methods or results
specifically designed in the provided content
- Create a standalone question with sufficient context - test takers will NOT see the original
provided content
- When multiple professional concepts are present, select the most theoretically important or
technically advanced one
- Include sophisticated incorrect options that require professional expertise to eliminate
```

```
- Ensure all options are factually distinguishable and avoid creating additional correct
answers

Key requirement: The correct answer must be verifiable from the provided content.

**CRITICAL**: Never reference the source with phrases like "in the text", "according to the
study", "as mentioned", "from the experimental results" etc.

**IMPORTANT**: Text suitability must be evaluated first before attempting question creation.

[Provided content]:
chunk_text
```

Figure 28: Q&A Generation Prompt

Completeness Filter Prompt

```
You are a QA validator. Check if this MCQ is suitable for standalone assessment of domain
experts.

**MCQ**
extracted_json_qa

**VALIDATION CRITERIA:**
**Question Independence:**
    - Question should not rely on external figures, tables, specific studies, experiments or
    references not included in the question
    - No phrases like: "in the paper", "as described above", "from the study", "referring to
    the table", etc.

**OUTPUT:**
```json

"is_valid": true/false,
"validation_result":
"question_independence": "PASS/FAIL - explanation"
,
"overall_assessment": "Brief explanation",
"specific_issues": ["Problems found, if any"]

Please validate this QA pair according to the criteria above:
```

Figure 29: Completeness Filter Prompt

Correctness Filter Prompt

```
You are a QA validator. Check if the correct answer can be verified from the original text.

**ORIGINAL TEXT:**
text

**GENERATED QA:**
extracted_json_qa

**VALIDATION CRITERIA:**

**Answer Verifiability (CRITICAL):**
- The correct answer must be explicitly stated or directly derivable from the original text
- If the original text contains questions without answers, and the MCQ uses those questions,
it's INVALID
- The answer cannot be generated/inferred by the model if not clearly supported by the text
- The correct answer must be factually accurate and directly supported by the original text
- The correct answer must be the ONLY correct option among all choices
```
```

OUTPUT:
```json

"is_valid": true/false,
"validation_result":
"answer_verifiability": "PASS/FAIL - explanation with specific text evidence"
,
"overall_assessment": "Brief explanation",
"specific_issues": ["Problems found, if any"]

Please validate this QA pair according to the criteria above:
```

Figure 30: Correctness Filter Prompt

Original Chunk For Q&A Generation

...
The star-sudifferential of φ, see e.g. is defined as $\partial^*\varphi(x) := \{g \in \mathbb{R}^n : \langle g, y-x\rangle < 0\, \forall y \in L_\varphi(x)\}$, where $L_\varphi(x) := \{y \in \mathbb{R}^n : \varphi(y) < \varphi(x)\}$ is the level set of φ at the level $\varphi(x)$. Clearly, if $\bar{L}_\varphi(x)$ is the closure of $L_\varphi(x)$, then $\partial^*\varphi(x) := \{g \in \mathbb{R}^n;\ \langle g, y-x\rangle \le 0\, \forall y \in \bar{L}_\varphi(x)\}$.
Hence $\partial^*\varphi(x) \equiv \mathbb{R}^n$ if x is a minimizer of φ over $\mathbb{R}^n$, and if φ is continuous on $\mathbb{R}^n$ then $\partial^*\varphi(x)$ is the normal cone of $\bar{L}_\varphi(x)$, that is $\partial^*\varphi(x) = N(\bar{L}_\varphi(x), x) := \{g \in \mathbb{R}^n : \langle g, y-x\rangle \le 0\, \forall y \in \bar{L}_\varphi(x)\}$.
Furthermore $\partial^*\varphi(x)$ contains nonzero vector. This subdifferential thus is also called normal-subdifferential.
Lemma 2. Assume that $\varphi : \mathbb{R}^n \to \mathbb{R}$ is continuous and quasiconvex. Then $\partial^*\varphi(x) \neq \emptyset \quad \forall x \in \mathbb{R}^n$, $0 \in \partial^*\varphi(x) \Leftrightarrow x \in \operatorname{argmin}\{\varphi(y) : y \in \mathbb{R}^n\}$.
...

Figure 31: MCQ Example:original chunk

Question & Answer

Q: In the context of quasiconvex optimization, which of the following statements accurately characterizes the star-subdifferential of a continuous quasiconvex function $\varphi : \mathbb{R}^n \to \mathbb{R}$?

Options:

A. The star-subdifferential $\partial^*\varphi(x)$ is defined only when φ is strongly convex in a neighborhood of x.

B. The star-subdifferential $\partial^*\varphi(x)$ coincides with the classical subdifferential for convex functions.

C. The star-subdifferential $\partial^*\varphi(x)$ is empty at local minima of φ.

D. The star-subdifferential $\partial^*\varphi(x)$ is always a singleton set for non-differentiable points.

E. The star-subdifferential $\partial^*\varphi(x)$ requires φ to be twice continuously differentiable.

F. The star-subdifferential $\partial^*\varphi(x)$ is guaranteed to be non-empty for every $x \in \mathbb{R}^n$.

G. The star-subdifferential $\partial^*\varphi(x)$ is equivalent to the Clarke subdifferential for all quasiconvex functions.

Answer: **F** — The star-subdifferential $\partial^* \varphi(x)$ is guaranteed to be non-empty for every $x \in \mathbb{R}^n$.

Figure 32: MCQ Example: Question&Answer