Novelty Assessment Report

Paper: How to train data-efficient LLMs
PDF URL: https://openreview.net/pdf?id=yKUbw7q1IA
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-29

Abstract

The training of large language models (LLMs) is expensive. In this paper, we study data-efficient approaches for pre-training LLMs, \ie, techniques that aim to optimize the Pareto frontier of model quality and training resource/data consumption. We seek to understand the tradeoffs associated with data selection routines based on (i) expensive-to-compute data-quality estimates, and (ii) maximization of coverage and diversity-based measures in the feature space. Our first technique, AskLLM, leverages the zero-shot reasoning capabilities of instruction-tuned LLMs to directly assess the quality of a training example. To target coverage, we propose density sampling, which models the data distribution to select a diverse sample. Testing the effect of 22 different data curation techniques on the pre-training of T5-style of models, involving hundreds of pre-training runs and post fine-tuning evaluation tasks, we find that AskLLM and density are the best methods in their respective categories. While coverage sampling techniques often recover the performance of training on the entire dataset, training on data curated via AskLLM consistently outperforms full-data training---even when we sample only 10\% of the original dataset, while converging up to 70\% faster.

Disclaimer

This report is **AI-GENERATED** using Large Language Models and WisPaper (a scholar search engine). It analyzes academic papers' tasks and contributions against retrieved prior work. While this system identifies **POTENTIAL** overlaps and novel directions, **ITS COVERAGE IS NOT EXHAUSTIVE AND JUDGMENTS ARE APPROXIMATE**. These results are intended to assist human reviewers and **SHOULD NOT** be relied upon as a definitive verdict on novelty.

Note that some papers exist in multiple, slightly different versions (e.g., with different titles or URLs). The system may retrieve several versions of the same underlying work. The current automated pipeline does not reliably align or distinguish these cases, so human reviewers will need to disambiguate them manually.

If you have any questions, please contact: mingzhang23@m.fudan.edu.cn

Core Task Landscape

This paper addresses: **Data-Efficient Pre-Training of Large Language Models**
A total of **50 papers** were analyzed and organized into a taxonomy with **28 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Data Selection and Curation Methods**
- **Data Synthesis and Augmentation**
- **Continual and Domain-Adaptive Pre-Training**
- **Low-Resource and Cross-Lingual Adaptation**
- **Model Compression and Efficient Architectures**
- **Training Optimization and Efficiency Techniques**
- **Post-Training Alignment and Fine-Tuning Efficiency**
- **Multimodal and Cross-Domain Adaptation**
- **Specialized Pre-Training Paradigms and Benchmarks**
- **Data-Centric Perspectives and Surveys**

Complete Taxonomy Tree

- Data-Efficient Pre-Training of Large Language Models Survey Taxonomy
- Data Selection and Curation Methods
  - Quality-Based Data Selection ★ (5 papers)
  - [0] How to train data-efficient LLMs (Anon et al., 2026) View paper
  - [2] Dataman: Data manager for pre-training large language models (Peng Ru, 2025) View paper
  - [5] Data-efficient pretraining with group-level data influence modeling (Zichun Yu, 2025) View paper
  - [10] Bertin: Efficient pre-training of a spanish language model using perplexity sampling (de la Rosa, 2022) View paper
  - [40] Ultra-fineweb: Efficient data filtering and verification for high-quality llm training data (Wang Yu-dong, 2025) View paper
  - Diversity and Coverage Sampling (3 papers)
  - [11] Efficient pretraining data selection for language models via multi-actor collaboration (Tianyi Bai, 2025) View paper
  - [19] Harnessing diversity for important data selection in pretraining large language models (Zhang Chi, 2024) View paper
  - [28] Sample Efficient Demonstration Selection for In-Context Learning (Purohit, 2025) View paper
  - Data Influence and Utility Modeling (2 papers)
  - [41] Data Efficacy for Language Model Training (Dai, 2025) View paper
  - [44] INGENIOUS: using informative data subsets for efficient pre-training of language models (H S V N S Kowndinya Renduchintala, 2023) View paper
- Data Synthesis and Augmentation
  - Synthetic Data Generation (2 papers)
  - [25] Rephrasing the web: A recipe for compute and data-efficient language modeling (Maini, 2024) View paper
  - [37] Demystifying synthetic data in llm pre-training: A systematic study of scaling laws, benefits, and pitfalls (Kang, 2025) View paper
  - Data Recycling and Reuse (1 papers)
  - [26] Recycling the Web: A Method to Enhance Pre-training Data Quality and Quantity for Language Models (Nguyen Thao, 2025) View paper

- Continual and Domain-Adaptive Pre-Training
  - Domain-Specific Continual Pre-Training (4 papers)
  - [1] Towards effective and efficient continual pre-training of large language models (Chen Jie, 2025) View paper
  - [6] Efficient continual pre-training for building domain specific large language models (Xie Yong, 2024) View paper
  - [16] TCM-GPT: efficient pre-training of large language models for domain adaptation in traditional Chinese medicine (Yang GuoXing, 2024) View paper
  - [47] Continual Pre-training of Language Models (Ke, 2023) View paper
  - Continual Learning Strategies and Warm-Up (1 papers)
  - [20] Continual pre-training of large language models: How to (re) warm your model? (Gupta, 2023) View paper
  - Anti-Forgetting via Replay Mechanisms (1 papers)
  - [43] GeRe: Towards Efficient Anti-Forgetting in Continual Learning of LLM via General Samples Replay (Zhang Yunan, 2025) View paper
- Low-Resource and Cross-Lingual Adaptation
  - Low-Resource Language Pre-Training (3 papers)
  - [17] Adapting multilingual LLMs to low-resource languages using continued pre-training and synthetic corpus: A case study for Hindi LLMs (R Joshi, 2025) View paper
  - [33] Towards low-resource languages machine translation: A language-specific fine-tuning with LoRA for specialized large language models (Xiao LIANG, 2025) View paper
  - [38] Low-Resource Language Expansion and Translation Capacity Enhancement for LLM: A Study on the Uyghur (K Lu, 2025) View paper
  - Cross-Lingual Transfer and Data Efficiency (1 papers)
  - [31] Exploring the data efficiency of cross-lingual post-training in pretrained language models (Chanhee Lee, 2021) View paper
  - Prompt-Based Low-Resource Adaptation (2 papers)
  - [42] LLM as prompter: Low-resource inductive reasoning on arbitrary knowledge graphs (Wang Kai, 2024) View paper
  - [45] Democratizing LLMs for low-resource languages by leveraging their English dominant abilities with linguistically-diverse prompts (Nguyen, 2024) View paper
- Model Compression and Efficient Architectures
  - Quantization for Efficient Pre-Training (2 papers)
  - [4] Towards efficient pre-training: Exploring fp4 precision in large language models (Zhou Jie-cheng, 2025) View paper
  - [9] Exploring quantization for efficient pre-training of transformer language models (Fournier, 2024) View paper
  - Low-Rank Parameterization and Optimization (2 papers)
  - [3] Lost: Low-rank and sparse pre-training for large language models (Li Jiaxi, 2025) View paper
  - [49] I3s: Importance sampling subspace selection for low-rank optimization in llm pretraining (Haochen Zhang, 2025) View paper
  - Pruning and Structured Sparsity (1 papers)
  - [15] EfficientLLM: Scalable Pruning-Aware Pretraining for Architecture-Agnostic Edge Language Models (Xing, 2025) View paper
  - Flexible and Adaptive Architectures (1 papers)
  - [14] Flextron: Many-in-one flexible large language model (Cai, 2024) View paper
  - Comprehensive Compression Surveys (1 papers)
  - [18] Efficient compressing and tuning methods for large language models: A systematic literature review (Gun Il Kim, 2025) View paper
- Training Optimization and Efficiency Techniques
  - Parameter-Efficient Pre-Training (1 papers)
  - [35] STEP: Staged Parameter-Efficient Pre-training for Large Language Models (Kazuki Yano, 2024) View paper
  - Context Length Extension and Scaling (2 papers)
  - [13] Longrecipe: Recipe for efficient long context generalization in large language models (Hu Zhiyuan, 2025) View paper
  - [50] Efficient Pretraining Length Scaling (Wu Bohong, 2025) View paper
  - Model Reuse and Transfer Learning (1 papers)
  - [32] bert2bert: Towards reusable pretrained language models (Cheng Chen, 2022) View paper
- Post-Training Alignment and Fine-Tuning Efficiency
  - Data-Efficient Alignment and Preference Learning (3 papers)
  - [12] Efficient alignment of large language models via data sampling (Ghosh, 2024) View paper
  - [21] Sample-efficient LLM Optimization with Reset Replay (Liu Zi-chuan, 2025) View paper
  - [29] Leveraging sparsity for sample-efficient preference learning: A theoretical perspective (Yao, 2025) View paper
  - Sample-Efficient Instruction Tuning (1 papers)
  - [46] Federated Data-Efficient Instruction Tuning for Large Language Models (Qin Zhen, 2024) View paper
  - Reasoning Capability Enhancement (2 papers)
  - [34] InfiAlign: A Scalable and Sample-Efficient Framework for Aligning LLMs to Enhance Reasoning Capabilities (Cai Shuo, 2025) View paper
  - [48] From data-centric to sample-centric: Enhancing llm reasoning via progressive optimization (Chen Xin-jie, 2025) View paper
- Multimodal and Cross-Domain Adaptation
  - Vision-Language Model Adaptation (1 papers)
  - [7] Visualgpt: Data-efficient adaptation of pretrained language models for image captioning (Jun Chen, 2022) View paper
  - Sample-Efficient Modality Integration (1 papers)
  - [8] Sample-efficient Integration of New Modalities into Large Language Models (Martins, 2025) View paper
- Specialized Pre-Training Paradigms and Benchmarks
  - Alternative Pre-Training Objectives (1 papers)
  - [24] Should We Still Pretrain Encoders with Masked Language Modeling? (Boizard, 2025) View paper
  - Sample-Efficient Pre-Training Benchmarks (3 papers)
  - [22] [Call for Papers] The 2nd BabyLM Challenge: Sample-efficient pretraining on a developmentally plausible corpus (Choshen, 2024) View paper
  - [23] Call for Papers - The BabyLM Challenge: Sample-efficient pretraining on a developmentally plausible corpus (Warstadt, 2023) View paper

- [27] Findings of the Second BabyLM Challenge: Sample-Efficient Pretraining on Developmentally Plausible Corpora (Hu, 2024) View paper
- Domain-Specific Efficient Pre-Training (2 papers)
  - [30] Yulan-mini: An open data-efficient language model (Hu Yiwen, 2024) View paper
  - [36] LFDe: A Lighter, Faster and More Data-Efficient Pre-training Framework for Event Extraction (Zhigang Kan, 2024) View paper
- Data-Centric Perspectives and Surveys (1 papers)
  - [39] A Survey on Efficient Large Language Model Training: From Data-centric Perspectives (JunYu Luo, 2025) View paper

Narrative

Core task: data-efficient pre-training of large language models. The field has organized itself around several complementary strategies for reducing the computational and data costs of training large language models. Data Selection and Curation Methods focus on identifying high-quality subsets of training corpora through filtering, deduplication, and influence-based techniques, aiming to maximize model performance with fewer tokens. Data Synthesis and Augmentation explore generating or transforming training examples to enrich limited datasets, while Continual and Domain-Adaptive Pre-Training address how to efficiently update or specialize models for new domains without full retraining. Low-Resource and Cross-Lingual Adaptation tackles the challenge of extending models to languages and settings with scarce data, and Model Compression and Efficient Architectures pursue smaller, faster models through quantization, pruning, and architectural innovations. Training Optimization and Efficiency Techniques improve the training process itself via better optimizers, curriculum learning, and hardware utilization, whereas Post-Training Alignment and Fine-Tuning Efficiency streamline instruction tuning and preference learning. Multimodal and Cross-Domain Adaptation extends these principles beyond text, and Specialized Pre-Training Paradigms and Benchmarks provide controlled settings like the BabyLM Challenge to study data efficiency at small scale.

Within Data Selection and Curation Methods, a particularly active line of work examines quality-based filtering and influence estimation to prioritize informative training examples. Data-efficient LLMs[0] situates itself in this quality-focused branch, emphasizing principled data selection to reduce pre-training costs. Nearby efforts such as Dataman[2] and Group-level Data Influence[5] explore complementary angles on measuring and leveraging data quality, with Dataman[2] offering scalable curation pipelines and Group-level Data Influence[5] providing finer-grained attribution of training subsets to model behavior. Ultra-fineweb[40] represents another closely related effort, curating a high-quality web corpus through aggressive filtering. The central tension across these works lies in balancing the computational overhead of quality assessment against the downstream gains from cleaner data, and in determining whether coarse heuristics or fine-grained influence methods yield better trade-offs. Data-efficient LLMs[0] contributes to this landscape by synthesizing quality-based selection strategies, offering a perspective on how careful data curation can substantially reduce the scale requirements for effective pre-training.

Related Works in Same Category

The following **4 sibling papers** share the same taxonomy leaf node with the original paper:

1. Dataman: Data manager for pre-training large language models

Authors: Peng Ru, Yang Ke-xin, Ru Peng, Zeng Ya-wen, Kexin Yang, et al. (15 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

The performance emergence of large language models (LLMs) driven by data scaling laws makes the selection of pre-training data increasingly important. However, existing methods rely on limited heuristics and human intuition, lacking comprehensive and clear guidelines. To address this, we are inspired by ``reverse thinking'' -- prompting LLMs to self-identify which criteria benefit its performance. As its pre-training capabilities are related to perplexity (PPL), we derive 14 quality criteria fro...

Relationship Analysis

Both papers belong to the Quality-Based Data Selection category, using model-based assessments to identify high-value training examples for LLM pre-training. They overlap in leveraging LLM capabilities for data quality evaluation: the original paper's ASK-LLM directly prompts instruction-tuned LLMs to assess training example quality, while the candidate paper's DataMan uses a fine-tuned LLM to annotate 14 quality criteria derived from perplexity analysis. The key difference is that ASK-LLM performs direct zero-shot quality assessment via prompting for binary decisions, whereas DataMan employs a trained model for multi-dimensional pointwise ratings across comprehensive quality criteria and domain types, offering more granular and structured quality annotations.

---

2. Data-efficient pretraining with group-level data influence modeling

Authors: Zichun Yu, Fei Peng, Jie Lei, Arnold Overwijk, Wen-tau Yih, et al. (6 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Data-efficient pretraining has shown tremendous potential to elevate scaling laws. This paper argues that effective pretraining data should be curated at the group level, treating a set of data points as a whole rather than as independent contributors. To achieve that, we propose Group-Level Data Influence Modeling (Group-MATES), a novel data-efficient pretraining method that captures and optimizes group-level data utility. Specifically, Group-MATES collects oracle group-level influences by loca...

Relationship Analysis

Both papers belong to the Quality-Based Data Selection category, using model-based assessments to identify high-value training examples for data-efficient LLM pre-training. They overlap in their focus on leveraging LLM capabilities for quality scoring: the original paper's ASK-LLM directly prompts instruction-tuned LLMs to assess training example quality, while the candidate paper's Group-MATES uses a relational data influence model that captures group-level interactions among training data. The key difference is that the original paper evaluates individual example quality through zero-shot LLM reasoning, whereas the candidate paper models complex data relationships and group-level influences through parameterized influence functions trained on sampled trajectories.

---

3. Bertin: Efficient pre-training of a spanish language model using perplexity sampling

Authors: de la Rosa, Javier, Javier de la Rosa, Ponferrada, Eduardo G., et al. (20 authors total) | **Year/Venue**: 2022 | **URL**: View paper

Abstract

The pre-training of large language models usually requires massive amounts of resources, both in terms of computation and data. Frequently used web sources such as Common Crawl might contain enough noise to make this pre-training sub-optimal. In this work, we experiment with different sampling methods from the Spanish version of mC4, and present a novel data-centric technique which we name $\textit{perplexity sampling}$ that enables the pre-training of language models in roughly half the amount ...

Relationship Analysis

Both papers belong to the Quality-Based Data Selection category, using quality metrics to select high-value training examples for data-efficient LLM pre-training. They overlap in exploring perplexity-based filtering as a data quality measure, with the original paper (ASK-LLM) comparing perplexity filtering against LLM-based quality assessment across multiple sampling rates and model sizes. The key difference is that the candidate paper (BERTIN) focuses specifically on perplexity sampling using KenLM models to train Spanish

RoBERTa models efficiently, while the original paper proposes ASK-LLM as a novel alternative that uses instruction-tuned LLMs for contextualized quality scoring, demonstrating superior performance over perplexity filtering.

4. Ultra-fineweb: Efficient data filtering and verification for high-quality llm training data

Authors: Wang Yu-dong, FU Zixuan, Yudong Wang, Cai Jie, Zixuan Fu, et al. (27 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Data quality has become a key factor in enhancing model performance with the rapid development of large language models (LLMs). Model-driven data filtering has increasingly become a primary approach for acquiring high-quality data. However, it still faces two main challenges: (1) the lack of an efficient data verification strategy makes it difficult to provide timely feedback on data quality; and (2) the selection of seed data for training classifiers lacks clear criteria and relies heavily on h...

Relationship Analysis

Both papers belong to the Quality-Based Data Selection category, employing model-based assessments to identify high-value training examples for LLM pre-training. They overlap in using LLM-based quality scoring (the original paper's ASK-LLM and the candidate's LLM annotation approach) and both validate their methods through extensive pre-training experiments. The key difference is that the original paper focuses on comparing quality versus coverage sampling strategies and proposes ASK-LLM as a direct prompting approach, while the candidate paper emphasizes an efficient verification strategy for rapid data quality assessment and uses a lightweight fastText classifier for scalable filtering, ultimately producing the Ultra-FineWeb dataset.

Contributions Analysis

Overall novelty summary. The paper proposes two data selection techniques—AskLLM, which uses instruction-tuned LLMs to assess training example quality, and density sampling, which models data distributions for diverse subset selection—and benchmarks 22 curation methods across hundreds of pre-training runs. It resides in the Quality-Based Data Selection leaf, which contains five papers including the original work. This leaf sits within the broader Data Selection and Curation Methods branch, indicating a moderately populated research direction focused on identifying high-value training subsets through quality metrics and model-based assessments.

The taxonomy reveals neighboring leaves addressing diversity-focused sampling (three papers) and data influence modeling (two papers), suggesting the field has organized quality-based, diversity-based, and influence-based selection into distinct but complementary categories. The paper's dual focus on quality (AskLLM) and coverage (density sampling) bridges these categories. Sibling papers in the same leaf include Dataman and Group-level Data Influence, which explore scalable curation pipelines and fine-grained attribution respectively. The taxonomy's scope notes clarify that quality-based selection excludes diversity sampling and domain-specific filtering, positioning this work at the intersection of quality assessment and distribution coverage.

Among 30 candidates examined, none clearly refute the three main contributions: AskLLM sampling (10 candidates, 0 refutable), density sampling (10 candidates, 0 refutable), and the large-scale empirical benchmark (10 candidates, 0 refutable). This suggests that within the limited search scope, the specific combination of LLM-based quality assessment, density-based diversity sampling, and comprehensive benchmarking of 22 techniques appears relatively novel. The absence of refutable candidates across all contributions indicates that the paper's integrated approach and empirical scale may distinguish it from prior work, though the search examined only top-30 semantic matches rather than an exhaustive literature review.

Based on the limited search scope of 30 candidates, the work appears to occupy a moderately explored area with distinct methodological contributions. The taxonomy structure shows active research in quality-based selection (five papers in the leaf), but the specific techniques and large-scale benchmarking approach may offer new empirical insights. The analysis does not cover potential overlaps beyond the top-30 semantic matches or recent concurrent work, so the novelty assessment remains provisional pending broader literature examination.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: ASK-LLM sampling technique

Description: The authors propose ASK-LLM, a data selection method that leverages instruction-tuned LLMs to directly assess training example quality through zero-shot reasoning. This technique consistently outperforms other data curation routines and enables training models that exceed full-dataset performance while using only a fraction of the data.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Learning to generate instruction tuning datasets for zero-shot task adaptation

URL: View paper

Brief Assessment

Generate Instruction Datasets[53] focuses on generating instruction tuning datasets from unannotated text for zero-shot task adaptation, not on using instruction-tuned LLMs to assess training example quality for data selection during pre-training.

2. A zero-shot and few-shot study of instruction-finetuned large language models applied to clinical and biomedical tasks

URL: View paper

Brief Assessment

Clinical Biomedical Instructions[54] focuses on evaluating instruction-tuned LLMs on clinical/biomedical NLP tasks (NER, QA, RE, NLI), not on data selection methods for pre-training. The candidate does not address training data quality assessment or curation techniques.

3. EchoQA: A Large Collection of Instruction Tuning Data for Echocardiogram Reports

URL: View paper

Brief Assessment

EchoQA[59] focuses on creating a question-answering dataset for echocardiogram reports and fine-tuning models for cardiac diagnosis, not on data selection methods for pre-training. The candidate does not address training data curation or quality assessment techniques for LLM pre-training.

4. Instructretro: Instruction tuning post retrieval-augmented pretraining

URL: View paper

Brief Assessment

Instructretro[60] focuses on retrieval-augmented pretraining and instruction tuning for LLMs, not on using instruction-tuned LLMs for zero-shot quality assessment of training data selection during pretraining.

5. Unsupervised text representation learning via instruction-tuning for zero-shot dense retrieval

URL: View paper

Brief Assessment

Instruction-tuning Retrieval[58] focuses on unsupervised text representation learning for dense retrieval tasks, not on data selection for LLM pre-training. The candidate paper uses instruction-tuned models to generate synthetic queries for corpus representation augmentation in retrieval systems, which is a fundamentally different application domain than ASK-LLM's use of instruction-tuned LLMs to assess training data quality for pre-training.

---

6. Enhancing zero-shot facial expression recognition by llm knowledge transfer

URL: View paper

Brief Assessment

Facial Expression LLM[55] focuses on zero-shot facial expression recognition using LLMs for knowledge transfer to enhance vision-language models, not on training data selection for LLM pre-training. The candidate addresses a different task domain (facial expression analysis) rather than data curation for language model training.

---

7. Instructblip: Towards general-purpose vision-language models with instruction tuning

URL: View paper

Brief Assessment

Instructblip[51] focuses on vision-language instruction tuning for multimodal models, not on data selection methods for LLM pre-training. The candidate does not address zero-shot quality assessment for training data curation.

---

8. Evaluating instruction-tuned large language models on code comprehension and generation

URL: View paper

Brief Assessment

Code Instruction Tuning[52] evaluates instruction-tuned LLMs on code tasks but does not propose a data selection method using LLM-based quality assessment for training data curation. The candidate focuses on evaluating existing instructed models rather than developing sampling techniques.

---

9. Llms as zero-shot graph learners: Alignment of gnn representations with llm token embeddings

URL: View paper

Brief Assessment

Zero-shot Graph Learners[56] focuses on aligning GNN representations with LLM token embeddings for graph machine learning tasks, not on using instruction-tuned LLMs to assess training data quality for dataset curation in language model pre-training.

---

10. Mods: Model-oriented data selection for instruction tuning

URL: View paper

Brief Assessment

Mods[57] focuses on instruction tuning data selection using quality evaluation models and k-center greedy algorithms, not on pre-training data selection using zero-shot LLM reasoning as in ASK-LLM.

Contribution 2: DENSITY sampling technique

Description: The authors introduce DENSITY, a coverage-maximizing sampler that estimates local density in the embedding space using kernel sums. This method aims to maximize topic coverage by downsampling redundant high-density regions and boosting under-represented portions of the input domain.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Entropic Analysis of Time Series through Kernel Density Estimation

URL: View paper

Brief Assessment

Entropic KDE[74] focuses on time series analysis using kernel density estimation for entropy metrics and change point detection in signal processing domains. The original paper's DENSITY sampler addresses data selection for LLM pre-training by estimating local density in embedding space to maximize topic coverage—a fundamentally different application domain and methodology.

---

2. Density estimation and modeling on symmetric spaces

URL: View paper

Brief Assessment

Symmetric Spaces Density[76] focuses on kernel density estimation on symmetric spaces (geometric manifolds), not on diverse data sampling in embedding spaces for LLM pre-training. The mathematical context and application domain are fundamentally different.

---

3. Learnable Kernel Density Estimation for Graphs

URL: View paper

Brief Assessment

Learnable KDE Graphs[73] focuses on learning kernel density estimation for graph-structured data using graph neural networks, not on sampling diverse data from embedding spaces for LLM pre-training. The technical approaches and application domains are fundamentally different.

---

4. Nonparametric Estimation with Kernel Mean Embeddings

URL: View paper

Brief Assessment

Kernel Mean Embeddings[71] focuses on kernel-based conditional density estimation methods in reproducing kernel Hilbert spaces, which is a different technical approach from the DENSITY sampler's use of kernel density estimation for coverage-maximizing data sampling in LLM pre-training.

---

5. Variational Kernel Density Estimation Recommendation Algorithm for Users with Diverse Activity Levels

URL: View paper

Brief Assessment

Variational KDE Recommendation[69] focuses on recommendation systems for users with diverse activity levels, not on data-efficient LLM pre-training or coverage-maximizing sampling for training data selection.

---

6. Layer-constrained variational autoencoding kernel density estimation model for anomaly detection

URL: View paper

Brief Assessment

Layer-constrained VAE KDE[77] focuses on anomaly detection using kernel density estimation in latent space for identifying outliers, not on diverse data sampling for training data-efficient LLMs. The methods serve fundamentally different purposes.

---

7. Kernel conditional density operators

URL: View paper

Brief Assessment

Kernel Conditional Density[78] focuses on reconstructing probability densities from kernel mean embeddings for conditional density estimation, not on coverage-maximizing sampling for LLM pre-training data selection.

---

8. Kernel based method for distributed derived feature tracking in high dimensions

URL: View paper

Brief Assessment

Kernel Distributed Tracking[75] focuses on distributed sensor networks for target tracking using kernel density estimation, not on data sampling for LLM pre-training. The application domains and technical objectives are fundamentally different.

---

9. Kernel density estimation in metric spaces

URL: View paper

Brief Assessment

KDE Metric Spaces[72] focuses on probability density estimation in metric spaces for statistical analysis of non-Euclidean data (e.g., hippocampal surfaces). DENSITY is a coverage-maximizing sampler for LLM pre-training that uses kernel density estimation in embedding space to diversify data selection. These are fundamentally different applications of kernel density estimation.

---

10. Exploiting probability density function of deep convolutional autoencoders' latent space for reliable COVID-19 detection on CT scans

URL: View paper

Brief Assessment

COVID Detection PDF[70] uses kernel density estimation (KDE) for COVID-19 classification in medical imaging, not for diverse data sampling in LLM pre-training. The technical application domains are fundamentally different.

Contribution 3: Large-scale empirical benchmark of data curation techniques

Description: The authors conduct an extensive comparative study testing 22 data curation techniques across hundreds of pre-training runs and over a thousand fine-tuning evaluations. This exhaustive benchmark provides new insights into the roles of coverage, quality, and sampling cost in LLM pre-training.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Data selection for language models via importance resampling

URL: View paper

Brief Assessment

Importance Resampling[68] focuses on a specific data selection method (DSIR) rather than providing a comprehensive benchmark. While it compares DSIR against 7-8 other methods, this is substantially smaller in scope than the original paper's claim of testing 22 techniques across hundreds of pre-training runs and over a thousand fine-tuning evaluations.

---

2. Data selection via optimal control for language models

URL: View paper

Brief Assessment

Optimal Control Selection[66] focuses on a theoretical optimal control framework for data selection, not on conducting comparative benchmarks of multiple data curation techniques. The paper does not present extensive comparative testing of 22 different techniques as claimed in the original contribution.

---

3. Qurating: Selecting high-quality data for training language models

URL: View paper

Brief Assessment

Qurating[61] focuses on quality rating via LLM judgments for data selection, not on comparative benchmarking of multiple data curation techniques. The original paper tests 22 techniques across hundreds of runs, while Qurating[61] develops and evaluates a single quality-based selection method.

---

4. Datacomp-lm: In search of the next generation of training sets for language models

URL: View paper

Brief Assessment

[Final Audit Failure] The model insisted on a refutation claim but failed to provide verifiable evidence after multiple retries. Marked as cannot_refute for safety. Please manually verify the candidate text.

---

5. Dataman: Data manager for pre-training large language models

URL: View paper

Brief Assessment

Dataman[2] focuses on developing quality criteria and a data manager for pre-training data selection, not on conducting a comparative benchmark of existing data curation techniques across hundreds of pre-training runs.

---

6. Rule-based data selection for large language models

URL: View paper

Brief Assessment

Rule-based Selection[62] focuses on rule-based data selection using orthogonality metrics and DPP sampling for LLM fine-tuning, not on conducting a large-scale comparative benchmark of 22 different data curation techniques across hundreds of pre-training runs as described in the original paper.

---

7. Mates: Model-aware data selection for efficient pretraining with data influence models

URL: View paper

Brief Assessment

Mates[65] focuses on model-aware dynamic data selection using data influence models during pretraining, not on comparative benchmarking of multiple static data curation techniques across hundreds of runs.

---

8. Harnessing diversity for important data selection in pretraining large language models

URL: View paper

Brief Assessment

Diversity Harnessing[19] focuses on balancing quality and diversity in data selection using influence functions and multi-armed bandits, rather than conducting a comprehensive comparative benchmark of 22 different data curation techniques across hundreds of pre-training runs as described in the original paper.

---

9. Regmix: Data mixture as regression for language model pre-training

URL: View paper

Brief Assessment

Regmix[63] focuses on automated data mixture optimization through regression modeling rather than comparative benchmarking of data curation techniques. The candidate does not conduct an extensive comparative study of 22 different data curation techniques as described in the original contribution.

---

10. Generating datasets with pretrained language models

URL: View paper

Brief Assessment

Generating Datasets[67] focuses on automatically generating labeled sentence-pair datasets using pretrained language models for semantic similarity tasks, not on benchmarking data curation techniques for LLM pre-training.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] How to train data-efficient LLMs View paper
- [1] Towards effective and efficient continual pre-training of large language models View paper
- [2] Dataman: Data manager for pre-training large language models View paper
- [3] Lost: Low-rank and sparse pre-training for large language models View paper
- [4] Towards efficient pre-training: Exploring fp4 precision in large language models View paper
- [5] Data-efficient pretraining with group-level data influence modeling View paper
- [6] Efficient continual pre-training for building domain specific large language models View paper
- [7] Visualgpt: Data-efficient adaptation of pretrained language models for image captioning View paper
- [8] Sample-efficient Integration of New Modalities into Large Language Models View paper
- [9] Exploring quantization for efficient pre-training of transformer language models View paper
- [10] Bertin: Efficient pre-training of a spanish language model using perplexity sampling View paper
- [11] Efficient pretraining data selection for language models via multi-actor collaboration View paper
- [12] Efficient alignment of large language models via data sampling View paper
- [13] Longrecipe: Recipe for efficient long context generalization in large language models View paper
- [14] Flextron: Many-in-one flexible large language model View paper
- [15] EfficientLLM: Scalable Pruning-Aware Pretraining for Architecture-Agnostic Edge Language Models View paper
- [16] TCM-GPT: efficient pre-training of large language models for domain adaptation in traditional Chinese medicine View paper
- [17] Adapting multilingual LLMs to low-resource languages using continued pre-training and synthetic corpus: A case study for Hindi LLMs View paper
- [18] Efficient compressing and tuning methods for large language models: A systematic literature review View paper
- [19] Harnessing diversity for important data selection in pretraining large language models View paper
- [20] Continual pre-training of large language models: How to (re) warm your model? View paper
- [21] Sample-efficient LLM Optimization with Reset Replay View paper
- [22] [Call for Papers] The 2nd BabyLM Challenge: Sample-efficient pretraining on a developmentally plausible corpus View paper
- [23] Call for Papers - The BabyLM Challenge: Sample-efficient pretraining on a developmentally plausible corpus View paper
- [24] Should We Still Pretrain Encoders with Masked Language Modeling? View paper
- [25] Rephrasing the web: A recipe for compute and data-efficient language modeling View paper
- [26] Recycling the Web: A Method to Enhance Pre-training Data Quality and Quantity for Language Models View paper
- [27] Findings of the Second BabyLM Challenge: Sample-Efficient Pretraining on Developmentally Plausible Corpora View paper
- [28] Sample Efficient Demonstration Selection for In-Context Learning View paper
- [29] Leveraging sparsity for sample-efficient preference learning: A theoretical perspective View paper
- [30] Yulan-mini: An open data-efficient language model View paper
- [31] Exploring the data efficiency of cross-lingual post-training in pretrained language models View paper
- [32] bert2bert: Towards reusable pretrained language models View paper
- [33] Towards low-resource languages machine translation: A language-specific fine-tuning with LoRA for specialized large language models View paper
- [34] InfiAlign: A Scalable and Sample-Efficient Framework for Aligning LLMs to Enhance Reasoning Capabilities View paper
- [35] STEP: Staged Parameter-Efficient Pre-training for Large Language Models View paper
- [36] LFDe: A Lighter, Faster and More Data-Efficient Pre-training Framework for Event Extraction View paper

- [37] Demystifying synthetic data in llm pre-training: A systematic study of scaling laws, benefits, and pitfalls View paper
- [38] Low-Resource Language Expansion and Translation Capacity Enhancement for LLM: A Study on the Uyghur View paper
- [39] A Survey on Efficient Large Language Model Training: From Data-centric Perspectives View paper
- [40] Ultra-fineweb: Efficient data filtering and verification for high-quality llm training data View paper
- [41] Data Efficacy for Language Model Training View paper
- [42] LLM as prompter: Low-resource inductive reasoning on arbitrary knowledge graphs View paper
- [43] GeRe: Towards Efficient Anti-Forgetting in Continual Learning of LLM via General Samples Replay View paper
- [44] INGENIOUS: using informative data subsets for efficient pre-training of language models View paper
- [45] Democratizing LLMs for low-resource languages by leveraging their English dominant abilities with linguistically-diverse prompts View paper
- [46] Federated Data-Efficient Instruction Tuning for Large Language Models View paper
- [47] Continual Pre-training of Language Models View paper
- [48] From data-centric to sample-centric: Enhancing llm reasoning via progressive optimization View paper
- [49] I3s: Importance sampling subspace selection for low-rank optimization in llm pretraining View paper
- [50] Efficient Pretraining Length Scaling View paper
- [51] Instructblip: Towards general-purpose vision-language models with instruction tuning View paper
- [52] Evaluating instruction-tuned large language models on code comprehension and generation View paper
- [53] Learning to generate instruction tuning datasets for zero-shot task adaptation View paper
- [54] A zero-shot and few-shot study of instruction-finetuned large language models applied to clinical and biomedical tasks View paper
- [55] Enhancing zero-shot facial expression recognition by llm knowledge transfer View paper
- [56] Llms as zero-shot graph learners: Alignment of gnn representations with llm token embeddings View paper
- [57] Mods: Model-oriented data selection for instruction tuning View paper
- [58] Unsupervised text representation learning via instruction-tuning for zero-shot dense retrieval View paper
- [59] EchoQA: A Large Collection of Instruction Tuning Data for Echocardiogram Reports View paper
- [60] Instructretro: Instruction tuning post retrieval-augmented pretraining View paper
- [61] Qurating: Selecting high-quality data for training language models View paper
- [62] Rule-based data selection for large language models View paper
- [63] Regmix: Data mixture as regression for language model pre-training View paper
- [64] Datacomp-lm: In search of the next generation of training sets for language models View paper
- [65] Mates: Model-aware data selection for efficient pretraining with data influence models View paper
- [66] Data selection via optimal control for language models View paper
- [67] Generating datasets with pretrained language models View paper
- [68] Data selection for language models via importance resampling View paper
- [69] Variational Kernel Density Estimation Recommendation Algorithm for Users with Diverse Activity Levels View paper
- [70] Exploiting probability density function of deep convolutional autoencoders' latent space for reliable COVID-19 detection on CT scans View paper
- [71] Nonparametric Estimation with Kernel Mean Embeddings View paper
- [72] Kernel density estimation in metric spaces View paper
- [73] Learnable Kernel Density Estimation for Graphs View paper
- [74] Entropic Analysis of Time Series through Kernel Density Estimation View paper
- [75] Kernel based method for distributed derived feature tracking in high dimensions View paper
- [76] Density estimation and modeling on symmetric spaces View paper
- [77] Layer-constrained variational autoencoding kernel density estimation model for anomaly detection View paper
- [78] Kernel conditional density operators View paper