A Survey on Efficient LLM Training: From Data-centric Perspectives

Junyu Luo[1], **Bohan Wu**[1], **Xiao Luo**[2*], **Zhiping Xiao**[3*], **Yiqiao Jin**[4], **Rong-Cheng Tu**[5], **Nan Yin**[6], **Yifan Wang**[7], **Jingyang Yuan**[1], **Wei Ju**[1], **Ming Zhang**[1*]
[1] Peking University [2] University of California, Los Angeles
[3] University of Washington [4] Georgia Institute of Technology
[5] Nanyang Technological University [6] HKUST [7] UIBE
*Corresponding Author
https://github.com/luo-junyu/Awesome-Data-Efficient-LLM

Abstract

Post-training of Large Language Models (LLMs) is crucial for unlocking their task generalization potential and domain-specific capabilities. However, the current LLM post-training paradigm faces significant data challenges, including the high costs of manual annotation and diminishing marginal returns on data scales. Therefore, achieving data-efficient post-training has become a key research question. In this paper, we present the first systematic survey of data-efficient LLM post-training from a data-centric perspective. We propose a taxonomy of data-efficient LLM post-training methods, covering data selection, data quality enhancement, synthetic data generation, data distillation and compression, and self-evolving data ecosystems. We summarize representative approaches in each category and outline future research directions. By examining the challenges in data-efficient LLM post-training, we highlight open problems and propose potential research avenues. We hope our work inspires further exploration into maximizing the potential of data utilization in large-scale model training.

1 Introduction

Large Language Models (LLMs) post-training has emerged as a crucial stage for unlocking their domain adaptation capabilities and task generalization potential (Luo et al., 2025). This phase has effectively enhanced models' abilities in long-context reasoning (Zelikman et al., 2022), human alignment (Rafailov et al., 2024), instruction tuning (Zhang et al., 2023b), and domain-specific adaptation (Cheng et al., 2024).

During the LLM post-training phase, data is the essential driver of model evolution. However, the current paradigm faces a severe *data dilemma*: the cost of manually annotating high-quality data is rapidly growing, while simply scaling data volume yields diminishing returns. Moreover, static datasets inherently limit models from adapting to evolving real-world knowledge. The linear dependency between data volume and model performance fundamentally stems from the inefficient data usage in traditional post-training paradigms. Our work establishes the first systematic survey on data-efficient post-training, providing a unified, taxonomized framework to address the fragmented research landscape. Our survey reveals that breaking through efficiency bottlenecks requires establishing value extraction across the data lifecycle, rather than merely expanding data scale.

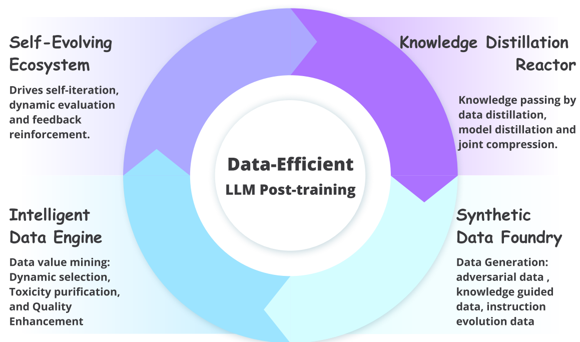

Figure 1: Illustration of the data flywheel in Data-Efficient LLM Post-Training, depicting the iterative cycle of knowledge distillation, synthetic data foundry, intelligent data engine, and self-evolving data ecosystems to maximize model performance with minimal data requirements.

Researchers have explored various approaches to fully exploit the data potential in LLM post-training (Jeong et al., 2024; Wang et al., 2024a; Luo et al., 2024b). While these methods have made notable progress in improving data efficiency, the field still lacks a comprehensive review. In this paper, we provide a comprehensive survey of data-efficient LLM post-training from a data-centric perspective. Specifically, we introduce the concept of a *data value flywheel* (as illustrated in Figure 1), which consists of four key components: knowledge distillation, synthetic data foundry, intelligent data engine, and self-evolving data ecosystems. Using this framework, we present a taxonomy of existing work, summarize key components, and identify promising research directions. We hope our work serves as both a useful roadmap for newcomers and a guide for future advancements in the field.

Differences from previous surveys. While several surveys have explored various aspects of LLMs post-training, including data selection (Wang et al., 2024b), synthetic data generation (Long et al., 2024; Tan et al., 2024), model self-feedback (Liang et al., 2024a; Pan et al., 2023), self-evolution (Tao et al., 2024), trustworthiness (Liu et al., 2023), and time-efficiency (Wan et al., 2023), these studies primarily focus on individual aspects rather than a holistic perspective. Our survey fills this gap by systematically examining these methods through the lens of data efficiency, offering critical insights into maximizing data value extraction.

2 Taxonomy

This section categorizes data-efficient post-training methods for LLMs into five core methodologies:

- **Data Selection**: *Filtering high-value subsets from raw data.* ❶ Static Filtering: Offline selection based on data properties; ❷ Dynamic Selection: Adjusting weights based on model uncertainty; ❸ Agent Strategy: Multi-model voting for reliable selection; ❹ Labeling Efficiency: Combining active learning and semi-supervised strategies for cost-effective sample coverage.
- **Data Quality Enhancement**: *Improving the utility of existing data.* ❶ Semantic Rewriting: Enhancing expression diversity through semantic-preserving transformations and generating variants while maintaining original meaning; ❷ Toxicity Control: Correcting harmful content; ❸ Distribution Stabilization: Adjusting data characteristics for robustness
- **Synthetic Data Generation**: *Creating new training data.* ❶ Instruction-Driven: Model-generated instruction-response pairs; ❷ Knowledge-Guided: Generation with structured knowledge; ❸ Adversarial Generation: Producing challenging samples.
- **Data Distillation and Compression**: *Extracting core knowledge for efficient training.* ❶ Model Distillation: Transferring large model output distributions to smaller models while preserving key knowledge; ❷ Data Distillation: Extracting high information density samples to construct compact datasets equivalent to full-scale data; ❸ Joint Compression: Combining model architecture compression with data selection strategies for end-to-end efficiency optimization
- **Self-Evolving Data Ecosystem**: *Building self-evolution mechanisms.* ❶ Self-Iterative Optimization: Using current model to generate data; ❷ Dynamic Evaluation Feedback: Real-time monitoring and adjustment; ❸ LLM-as-a-Judge: Feedback-Driven Data Optimization;

Category	Data Dependency	Compute Cost	Model Dependency	Data Value Mining
Data Selection	++	+	+	+++
Quality Enhance.	++	++	++	++
Synthetic Generation	+	+++	+++	+
Distill. & Compress.	+	+	+++	+++
Self-Evolving	+	+++	+++	+++

Table 1: Comparison of different data-efficient post-training methods across key dimensions.

Table 1 compares the five methodologies across key dimensions, where more '+' indicates higher requirements or better performance. Data selection shows high data efficiency but requires quality source data. Quality enhancement maintains balanced requirements across dimensions. Synthetic generation and self-evolving approaches demand more compute and model resources but reduce data dependency. Distillation methods excel in data efficiency while depending on model capabilities.

These five dimensions complement each other: selection filters quality data, enhancement improves utility, generation expands coverage, distillation concentrates knowledge, and self-evolution enables continuous improvement. Together, they pursue the goal of *less data, higher returns*.

3 Data Selection

Data selection is crucial for enhancing LLM post-training efficiency by identifying high-value data subsets. As shown in Figure 3, we divide existing approaches into four dimensions: (1) static filtering based on inherent data properties, (2) dynamic selection that adapts during training, (3) agent strategy using collaborative mechanisms, and (4) labeling efficiency through human-AI collaboration.

3.1 Static Filtering

Static filtering evaluates inherent data properties offline to identify samples with high information density and representativeness.

Quality-based Filtering. Alpagasus (Chen et al., 2023) achieves comparable performance using only

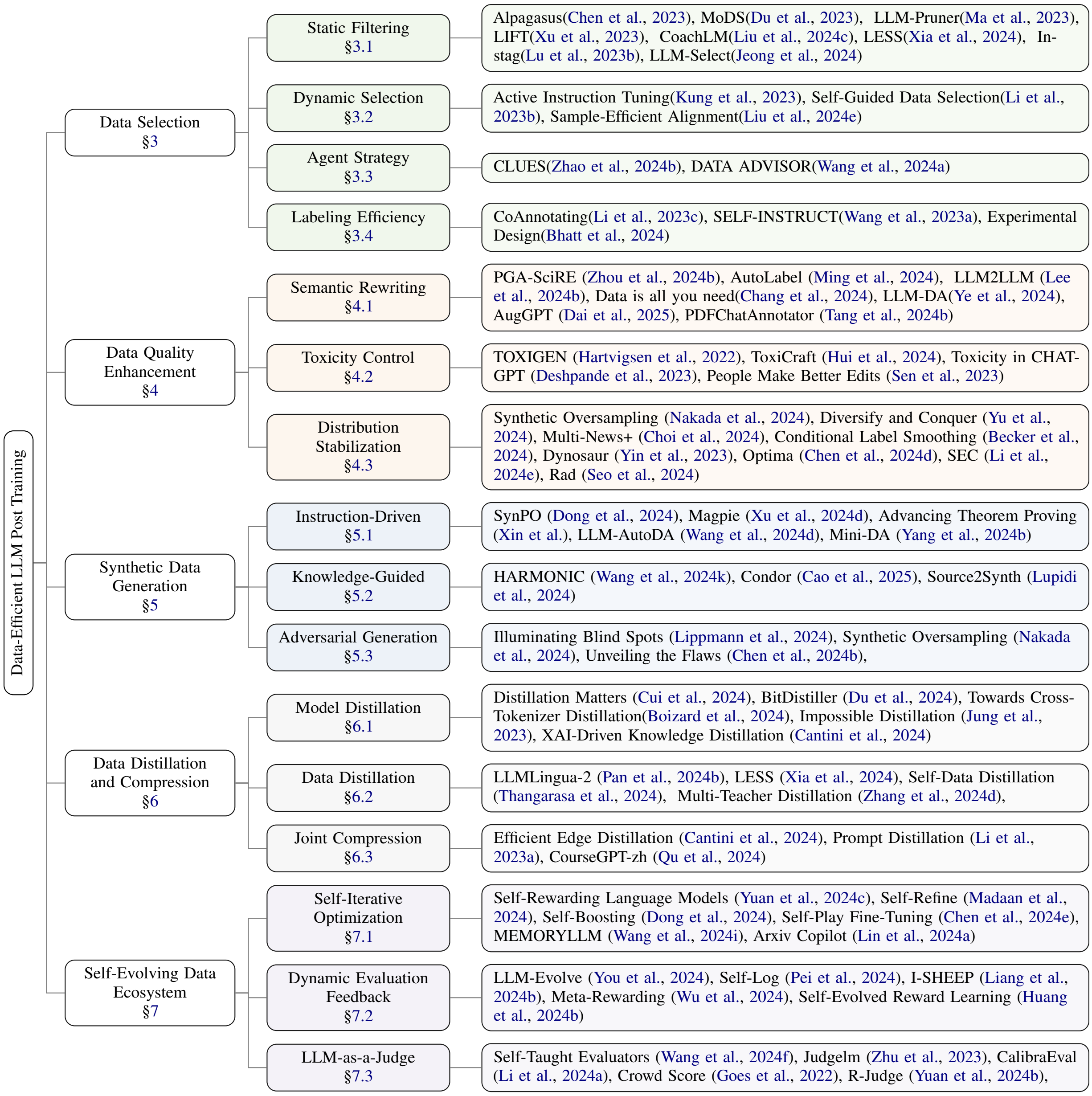

Figure 2: A taxonomy of Data-Efficient LLM Post Training.

17% of original data through complexity-based filtering (instruction length, diversity, and perplexity). MoDS (Du et al., 2023) employs multi-dimensional indicators and density peak clustering, while (Kang et al., 2024) uses KL-divergence-driven selection to align domain distributions.

Structure-aware Pruning. In code generation, (Tsai et al., 2024) combines static filtering with syntax tree analysis and execution verification. Works like LLM-Pruner (Ma et al., 2023) and (Zhang et al., 2024b; Kim and Baek, 2024; Yang et al., 2024c; Lu et al., 2024) leverage information entropy and multi-objective optimization. Additional studies (Azeemi et al., 2024; Guo et al., 2023; Dutta et al., 2024; Ling et al., 2024; Kim et al., 2024; Huang et al., 2023) explore grammar structure and contextual dependencies.

Semantic Enhancement. LIFT (Xu et al., 2023) and CoachLM (Liu et al., 2024c) enhance instruction quality through automatic revision. In recommendation systems, works like (Lin et al., 2024b; Lu et al., 2023b; Xia et al., 2024; Jeong et al., 2024) extend filtering methods using task-oriented scoring mechanisms and achieve better performance.

3.2 Dynamic Selection

Dynamic methods adapt data weights by evaluating sample importance based on model feedback.

Uncertainty-driven Selection. Active Instruction Tuning (Kung et al., 2023) prioritizes high-uncertainty tasks through prediction entropy. Self-

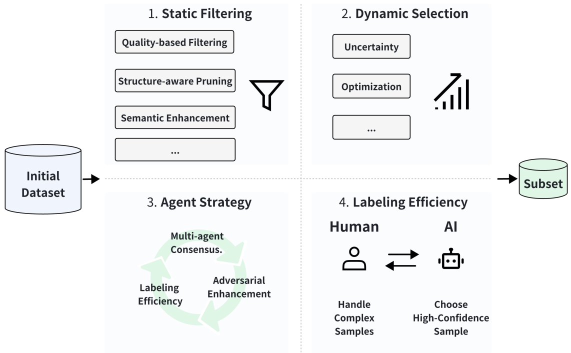

Figure 3: Overview of four major data selection approach categories: static filtering, dynamic selection, agent strategy, and labeling efficiency.

Guided Data Selection uses Instruction Following Difficulty (IFD) to measure loss variance and eliminate easily learned examples (Li et al., 2023b).

Optimization-based Selection. Compute-Constrained Data Selection (Yin and Rush, 2024) optimizes between data utility and computational cost. Sample-Efficient Alignment (Liu et al., 2024e) uses Thompson sampling to maximize contribution in preference alignment tasks.

3.3 Agent Strategy

Agent-based approaches leverage collaborative mechanisms for reliable selection.

Multi-agent Consensus. Multi-agent methods like CLUES (Zhao et al., 2024b) implement multi-model voting mechanisms based on training dynamics and gradient similarity metrics.

Adversarial Enhancement. Recent works like DATA ADVISOR (Wang et al., 2024a) uses red-team agents for safety filtering, while Automated Data Curation (Chen and Mueller, 2024) optimizes data through generator-discriminator frameworks.

3.4 Labeling Efficiency

These methods efficiently optimize annotation processes through iterative human-AI collaboration.

Human-AI Collaboration. Methods like LLMaAA (Zhang et al., 2023a) employ LLMs as annotators with uncertainty sampling. CoAnnotating (Li et al., 2023c) implements uncertainty-guided labor division between humans and AI.

Automated Generation. SELF-INSTRUCT (Wang et al., 2023a) enables autonomous self-generated instruction data, while (Li et al., 2023d) uses one-shot learning for rapid sample identification.

Workflow Optimization. Recent works establish scalabel efficient annotation workflows through adaptive experimental design (Bhatt et al., 2024) and systematic curation systems (Pang et al., 2024).

3.5 Discussion

Current data selection approaches face challenges in aligning static metrics with dynamic model requirements, managing computational complexity in optimization, and achieving cross-domain generalization (Xia et al., 2024; Yin and Rush, 2024; Zhao et al., 2024b). Future research points toward meta-learning-based selection frameworks, causal inference for sample analysis, and efficiency-aware optimization with hardware constraints, advancing data selection toward theoretical grounding.

4 Data Quality Enhancement

As illustrated in Figure 4, enhancing data quality is critical for maximizing the effectiveness of LLM post-training. Through semantic refinement, toxicity control, and distribution stabilization, researchers aim to improve the informativeness, safety, and robustness of training data. We categorize existing methods into three directions.

4.1 Semantic Rewriting

Semantic rewriting focuses on augmenting data diversity while preserving original meaning through controlled transformations. This can be achieved through several key approaches:

Instruction Refinement. CoachLM (Liu et al., 2024c) automatically revises complex instructions to reduce ambiguity, while (Li et al., 2024f) uses structured prompt chains for paraphrase generation, enhancing model generalization across tasks.

Domain-Specific Augmentation. Methods like (Jia et al., 2024) use curriculum learning for metaphor detection, while PGA-SciRE (Zhou et al., 2024b) injects structured knowledge for scientific relation extraction, adapting models to specialized tasks.

Automated Enhancement. AutoLabel (Ming et al., 2024) seamlessly integrates human feedback for quality rewriting, while LLM2LLM (Lee et al., 2024b) iteratively improves low-confidence samples. Recent studies extensively explore human-AI collaboration (Chung et al., 2023) and various data types: text (Dai et al., 2025), tabular (Banday et al., 2024), and multimodal (Tang et al., 2024b). Additional works (Zhou et al., 2024c; Chang et al., 2024; Ye et al., 2024; Zhang et al., 2025) survey generative paradigms across modalities.

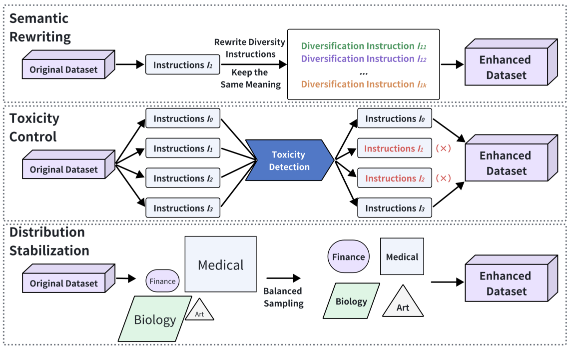

Figure 4: Three key approaches for data quality enhancement in LLM post-training: semantic rewriting for diversity, toxicity control for safety, and distribution stabilization for balanced representation.

4.2 Toxicity Control

Mitigating harmful content is crucial for data quality enhancement. Recent approaches focus on detection, benchmarking, and human collaboration:

Detection Frameworks. Methods like (Zhang et al., 2024a) effectively distill toxicity knowledge into compact detectors, while (Wang and Chang, 2022) strategically leverages generative prompts for zero-shot toxicity classification across diverse tasks.

Adversarial Benchmarking. Frameworks such as TOXIGEN (Hartvigsen et al., 2022) and ToxiCraft (Hui et al., 2024) generate adversarial datasets to stress-test models. Studies (Luong et al., 2024; Deshpande et al., 2023; Chetnani, 2023; Oh et al., 2024) examine the relationship between model size and toxicity generation, finding that smaller models often exhibit lower toxicity rates.

Human-AI Collaboration. Research demonstrates that human intervention significantly improves toxicity detection quality (Sen et al., 2023), particularly through counterfactual data augmentation. Additional work explores covert toxicity detection (Lee et al., 2024a), data contamination (Balloccu et al., 2024), and geometric interpretability (Balestriero et al.) to enhance model safety.

4.3 Distribution Stabilization

Stabilizing data distribution ensures that models generalize well across different tasks and domains. Several methods tackle issues like class imbalance, noise reduction, and domain adaptation:

Imbalance Mitigation. Approaches like Synthetic Oversampling (Nakada et al., 2024) and Diversify and Conquer (Yu et al., 2024) effectively address class imbalance through adaptive synthetic sample generation. Studies show significant improvements, with (Cai et al., 2023) demonstrating a 38% fairness boost in cross-disciplinary applications.

Noise Reduction. Multi-News+ (Choi et al., 2024) signficantly reduces annotation errors by 62% through automated label correction, while (Chen and Mueller, 2024) employs self-supervised filtering for robust fine-tuning data curation.

Domain Adaptation. ChatTS (Xie et al., 2024) uses Fourier transforms for time-series alignment, while (Becker et al., 2024) applies domain-specific label smoothing for clinical text. Advanced approaches like Dynosaur (Yin et al., 2023) and Optima (Chen et al., 2024d) leverage curriculum learning and multi-source coordination. Methods such as (Li et al., 2024e; Seo et al., 2024; Wang et al., 2024c) optimize data flows for multi-agent and edge deployment scenarios.

4.4 Discussion

The three key approaches—semantic rewriting, toxicity control, and distribution stabilization—form a comprehensive framework for data quality enhancement in LLM post-training. While each method addresses specific challenges, future research should focus on developing integrated solutions that combine these approaches efficiently, balancing quality improvements with compute costs.

5 Synthetic Data Generation

Generating synthetic training data is a powerful strategy to overcome data scarcity and enhance the robustness of LLM post-training. As illustrated in Figure 5, synthetic data generation methods can be categorized into three main approaches: *Instruction-Driven*, *Knowledge-Guided*, and *Adversarial Generation*, each serving distinct purposes in enhancing model capabilities.

5.1 Instruction-Driven Synthetic Data Generation

Instruction-driven methods harness LLMs' ability to produce new examples directly from task prompts. Recent works demonstrate diverse applications: SynPO (Dong et al., 2024) generates preference pairs for alignment (12% ROUGE-L improvement), Magpie (Xu et al., 2024d) enables template-free instruction generation (98% AlpacaEval accuracy), and Advancing Theorem Proving (Xin et al.) synthesizes Lean4 proof steps, boosting GPT-4's proving capabilities by 34%.

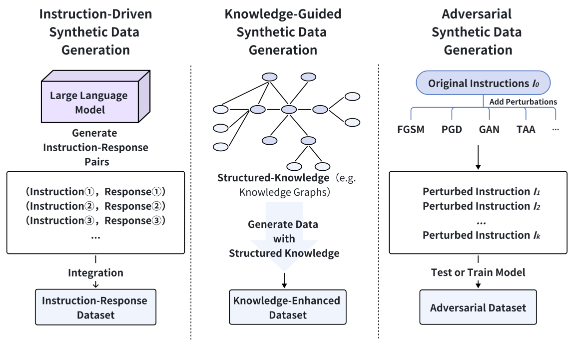

Figure 5: Three main approaches for data generation in LLM post-training: instruction-driven generation for creating instruction-response pairs, knowledge-guided generation using structured knowledge, and adversarial generation for testing model robustness.

5.2 Knowledge-Guided Synthetic Data Generation

Knowledge-guided approaches integrate external knowledge to steer data generation.
Theoretical Frameworks. Towards a Theoretical Understanding (Gan and Liu, 2024) rigorously establishes a reverse-bottleneck theory linking data diversity to enhanced model generalization.
Structured Data Synthesis. HARMONIC (Wang et al., 2024k) combines privacy-preserving tabular data generation with 0.92 F1-score on medical records. (Xu et al., 2024b) improves relational consistency through schema-aware fine-tuning.
Cost-Effective Strategies. (Chan et al., 2024) demonstrates hybrid generation methods reduce API costs by 70% while maintaining data utility. Source2Synth (Lupidi et al., 2024) improves factual accuracy through knowledge-graph alignment.

5.3 Adversarial Generation

Adversarial generation methods systematically probe model vulnerabilities to enhance robustness. Recent works demonstrate diverse approaches: Illuminating Blind Spots (Lippmann et al., 2024) uses agent-based simulations to generate edge cases, reducing errors by 19% on dialect variation; Unveiling Synthetic Data Flaws (Chen et al., 2024b) introduces contrastive unlearning to address data imperfections, yielding 32% quality improvements on GLUE; and ToxiCraft (Hui et al., 2024) generates subtle harmful content, revealing significant gaps in commercial safety filters.

5.4 Discussion

Each approach offers distinct trade-offs: instruction-driven methods enable rapid scaling but risk semantic drift; knowledge-guided approaches maintain fidelity through structured constraints; and adversarial generation strengthens robustness by exposing vulnerabilities. Future work should combine these strengths—for instance, merging privacy-preserving generation with adversarial testing. Key challenges persist in optimizing generation costs (Chan et al., 2024) and developing theoretical foundations (Gan and Liu, 2024).

6 Data Distillation and Compression

Data distillation and compression techniques enhance LLM post-training efficiency by reducing data complexity while preserving performance. As shown in Figure 6, this involves three complementary approaches: model distillation for knowledge transfer, data distillation for dataset compression, and joint compression for unified optimization.

6.1 Model Distillation

Model distillation transfers knowledge from large to smaller models while maintaining performance. Recent advances include Impossible Distillation (Jung et al., 2023), which creates high-quality models from low-quality teachers, and Performance-Guided Distillation (Di Palo et al., 2024), achieving 98% accuracy with 40% reduced costs. Cross-Tokenizer Distillation (Boizard et al., 2024) enables knowledge transfer between different architectures through universal logit distillation. For edge deployment, XAI-Driven Distillation (Cantini et al., 2024) produces interpretable medical models, while BitDistiller (Du et al., 2024) enables sub-4-bit precision with minimal accuracy loss. Multistage Collaborative Distillation (Zhao et al., 2024a) improves performance through multi-teacher coordination in low-resource settings.

6.2 Data Distillation

Data distillation focuses on selecting high-information-density samples to create compact yet representative datasets. Knowledge Distillation in Automated Annotation (Pangakis and Wolken, 2024) shows LLM-generated labels can effectively train classifiers comparable to human annotations. LESS (Xia et al., 2024) leverages influence functions for efficient instruction tuning, while LLMLingua-2 (Pan et al., 2024b) approaches prompt compression through token-level distillation. Domain-specific applications include Self-Data Distillation (Thangarasa et al.,

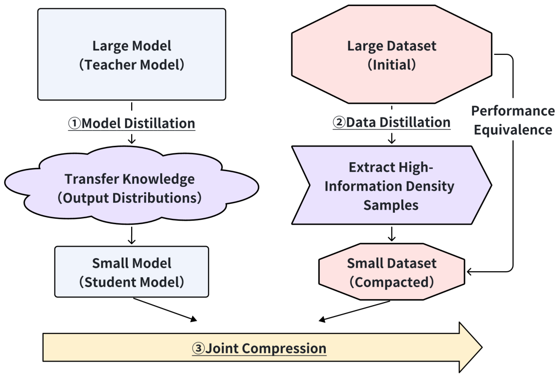

Figure 6: Data distillation and compression in LLM post-training: model distillation for knowledge transfer, data distillation for sample extraction, and joint compression for unified optimization.

2024) for model refinement, Multi-Teacher Distillation (Zhang et al., 2024d) for healthcare data integration, and techniques for reducing hallucination (McDonald et al., 2024).

6.3 Joint Compression

Joint compression combines model compression with data selection to optimize overall efficiency. Compact Language Models via Pruning and Distillation (Muralidharan et al., 2024) co-optimizes structural pruning and label smoothing, compressing LLaMA-7B to 2.8B parameters with minimal performance loss. Efficient Edge Distillation (Cantini et al., 2024) enables adaptive width scaling for edge devices through supernet training. In recommendation systems, Prompt Distillation (Li et al., 2023a) aligns ID-based and text-based representations, reducing inference time by 43%.

For multimodal applications, recent work demonstrates joint compression of graph and text encoders (Pan et al., 2024a) and curriculum-aligned prompt distillation for educational LLMs (Qu et al., 2024), achieving significant parameter reduction while maintaining performance.

6.4 Discussion

These three approaches offer complementary benefits for enhancing LLM efficiency: model distillation optimizes architecture, data distillation curates high-impact samples, and joint compression unifies model-data optimization. Future research should focus on integrating these methods, particularly for edge AI and low-resource applications.

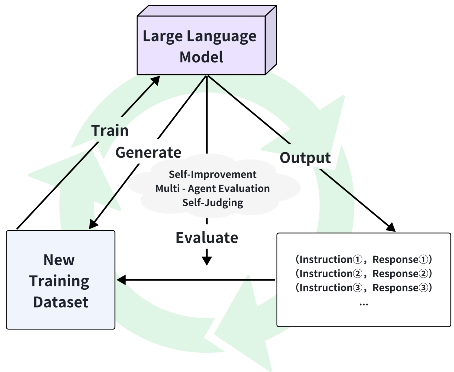

Figure 7: Self-Evolving Data Ecosystem: autonomous data generation, real-time feedback, and continuous learning.

7 Self-Evolving Data Ecosystem

The Self-Evolving Data Ecosystem strategically optimizes LLM post-training through autonomous data generation, real-time feedback, and continuous learning. As shown in Figure 7, this ecosystem forms a closed loop of generation, evaluation, and adaptive training. We discuss three key components: Self-Iterative Optimization, Dynamic Evaluation Feedback, and LLM-as-a-Judge.

7.1 Self-Iterative Optimization

Self-iterative optimization enables LLMs to use their own outputs to generate new training data, refining their capabilities autonomously. Several approaches illustrate this concept:

Self-Improvement Methods. Recent works like Self-Rewarding (Yuan et al., 2024c), Self-Refine (Madaan et al., 2024), and Self-Boosting (Dong et al., 2024) enable models to autonomously improve through iterative self-optimization. Self-Play Fine-Tuning (Chen et al., 2024e) extends this by leveraging competitive self-interaction, outperforming traditional methods like DPO (Rafailov et al., 2024).

Knowledge Retention. In the context of retaining knowledge while integrating new data, MemoryLLM (Wang et al., 2024i) enables continuous model updates while preserving existing knowledge. Automated Proof Generation (Chen and Mueller, 2024) and Arxiv Copilot (Lin et al., 2024a) demonstrate this capability in code verification and academic research tasks.

7.2 Dynamic Evaluation Feedback

Dynamic evaluation feedback systems allow models to make real-time adjustments based on their performance, optimizing their outputs on the fly.

Key contributions include:

Multi-Agent Evaluation. The Benchmark Self-Evolving Framework (Wang et al., 2024e) and LLM-Evolve (You et al., 2024) employ multi-agent systems to evaluate and adjust LLM performance dynamically. These frameworks enable the models to self-adjust in real-time across various benchmarks, ensuring continuous evolution.

Iterative Refinement Self-Refine (Madaan et al., 2024) and Self-Log (Pei et al., 2024) employ feedback loops for iterative refinement and log parsing, optimizing the model's output without requiring external retraining.I-SHEEP (Liang et al., 2024b) offers a resource-efficient paradigm that enhances performance through self-alignment, while Interactive Evolution: A Neural-Symbolic Self-Training Framework (Xu et al., 2024a) enables LLMs to autonomously train in neural-symbolic environments.

Improved Decision Making. For improving model alignment, Meta-Rewarding (Wu et al., 2024) and Self-Evolved Reward Learning (Huang et al., 2024b) leverage iterative feedback from their outputs to improve judgment skills, ensuring more accurate decision-making in complex tasks.

7.3 LLM-as-a-Judge

LLM-as-a-Judge systems utilize the model's own evaluations to guide real-time adjustments to its training data, ensuring that only the most relevant and accurate data is used. This category includes:

Self-Judging. Meta-rewarding (Wu et al., 2024), Self-Taught Evaluators (Wang et al., 2024f), and Judgelm (Zhu et al., 2023) demonstrate how LLMs can serve as judges to refine their own performance. These methods emphasize using LLM feedback to select, refine, and optimize their outputs, improving their ability to self-evaluate and adjust.

Bias Mitigation. CalibraEval (Li et al., 2024a), Crowd Score (Goes et al., 2022), and R-Judge (Yuan et al., 2024b) focus on bias mitigation, fairness, and evaluating model safety in real-time interactions. These tools enhance the judge-like capabilities of LLMs, ensuring that evaluations are more accurate and less prone to bias.

Adversarial Testing. For improving creativity and reducing bias in generative tasks, TOXIGEN (Hartvigsen et al., 2022), ToxiCraft (Hui et al., 2024), and Crowd Score (Goes et al., 2022) generate adversarial datasets to test model robustness under various scenarios, particularly in toxicity detection. These frameworks ensure models are more resistant to harmful content generation.

7.4 Discussion

The combination of Self-Iterative Optimization, Dynamic Evaluation Feedback, and LLM-as-a-Judge creates a robust framework for autonomous LLM improvement. While these approaches show promise in reducing manual intervention, future work should focus on unifying them into scalable frameworks that generalize across diverse tasks.

8 Challenges and Future Directions

Domain-Driven Data Synthesis and Refinement. While general-purpose models like GPT are commonly used for data generation (Di Palo et al., 2024), domain-specific models can better capture professional knowledge (Lightman et al., 2023). Future work should explore domain-specific pre-trained models for generating specialized data (Luo et al., 2023; Cheng et al., 2024), along with refinement techniques to optimize data quality while reducing annotation costs.

Scalability of Large-Scale Data Synthesis. As LLM pre-training demands increasingly larger and higher-quality datasets, efficient large-scale data generation becomes crucial. Current data synthesis and augmentation methods face scalability bottlenecks. Future work should focus on developing parallel, cost-effective, and efficient data generation frameworks that meet the demands of large-scale pre-training while maintaining a balance between data diversity and relevance (Karunya et al., 2023).

Reliable Quality Assessment Metrics. Current evaluation frameworks lack standardized metrics for assessing synthetic data quality (Zendel et al., 2024). Future research should develop metrics that evaluate semantic fluency, information accuracy, and potential biases (Chundawat et al., 2022; Gerstgrasser et al., 2024) to ensure robust selection.

9 Conclusion

In this paper, we presented a systematic review of LLM post-training research from a data efficiency perspective. We established the first taxonomic framework for data-efficient post-training, encompassing five core methodologies. Through detailed analysis of representative approaches within each category, we revealed that breaking through data efficiency bottlenecks requires establishing value extraction mechanisms across the entire data lifecycle. We aimed to highlight the current state and provide valuable insights for future work in this promising field of data-efficient LLM post-training.

Limitations

While our work presents the first comprehensive framework for analyzing data-efficient LLM post-training approaches, several limitations and opportunities for future research remain. First, given the explosive growth of this field, some emerging techniques may not be fully captured in our current taxonomic system, necessitating continuous updates to maintain comprehensiveness. Second, while data efficiency is crucial, the proposed methods may face additional challenges regarding trustworthiness and scalability that warrant further investigation. Furthermore, the synergistic effects and interaction mechanisms between different data efficiency enhancement techniques remain underexplored, calling for the development of cross-method optimization theories. We anticipate these open challenges will inspire deeper theoretical innovations and practical breakthroughs.

References

Swapnaja Achintalwar, Adriana Alvarado Garcia, Ateret Anaby-Tavor, Ioana Baldini, Sara E Berger, Bishwaranjan Bhattacharjee, Djallel Bouneffouf, Subhajit Chaudhury, Pin-Yu Chen, Lamogha Chiazor, et al. 2024. Detectors for safe and reliable llms: Implementations, uses, and limitations. *arXiv preprint arXiv:2403.06009*.

Alon Albalak, Yanai Elazar, Sang Michael Xie, Shayne Longpre, Nathan Lambert, Xinyi Wang, Niklas Muennighoff, Bairu Hou, Liangming Pan, Haewon Jeong, et al. 2024. A survey on data selection for language models. *arXiv preprint arXiv:2402.16827*.

Abdul Hameed Azeemi, Ihsan Ayyub Qazi, and Agha Ali Raza. 2024. Language model-driven data pruning enables efficient active learning. *arXiv preprint arXiv:2410.04275*.

Randall Balestriero, Romain Cosentino, and Sarath Shekkizhar. Characterizing large language model geometry helps solve toxicity detection and generation. In *Forty-first International Conference on Machine Learning*.

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondřej Dušek. 2024. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source llms. *arXiv preprint arXiv:2402.03927*.

Banooqa Banday, Kowshik Thopalli, Tanzima Z Islam, and Jayaraman J Thiagarajan. 2024. On the role of prompt construction in enhancing efficacy and efficiency of llm-based tabular data generation. *arXiv preprint arXiv:2409.03946*.

Prosanta Barai, Gondy Leroy, Prakash Bisht, Joshua M Rothman, Sumi Lee, Jennifer Andrews, Sydney A Rice, and Arif Ahmed. 2024. Crowdsourcing with enhanced data quality assurance: An efficient approach to mitigate resource scarcity challenges in training large language models for healthcare. *AMIA Summits on Translational Science Proceedings*, 2024:75.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, et al. 2024. Llms instead of human judges? a large scale empirical study across 20 nlp evaluation tasks. *arXiv preprint arXiv:2406.18403*.

Luca Becker, Philip Pracht, Peter Sertdal, Jil Uboreck, Alexander Bendel, and Rainer Martin. 2024. Conditional label smoothing for llm-based data augmentation in medical text classification. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 833–840. IEEE.

Gantavya Bhatt, Yifang Chen, Arnav M Das, Jifan Zhang, Sang T Truong, Stephen Mussmann, Yinglun Zhu, Jeffrey Bilmes, Simon S Du, Kevin Jamieson, et al. 2024. An experimental design framework for label-efficient supervised finetuning of large language models. *arXiv preprint arXiv:2401.06692*.

Arezo Bodaghi, Benjamin CM Fung, and Ketra A. Schmitt. 2024. Augmentoxic: Leveraging reinforcement learning to optimize llm instruction finetuning for data augmentation to enhance toxicity detection. *ACM Transactions on the Web*.

Nicolas Boizard, Kevin El Haddad, Céline Hudelot, and Pierre Colombo. 2024. Towards cross-tokenizer distillation: the universal logit distillation loss for llms. *arXiv preprint arXiv:2402.12030*.

Xunxin Cai, Meng Xiao, Zhiyuan Ning, and Yuanchun Zhou. 2023. Resolving the imbalance issue in hierarchical disciplinary topic inference via llm-based data augmentation. In *2023 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 1424–1429. IEEE.

Riccardo Cantini, Alessio Orsino, and Domenico Talia. 2024. Xai-driven knowledge distillation of large language models for efficient deployment on low-resource devices. *Journal of Big Data*, 11(1):63.

Maosong Cao, Taolin Zhang, Mo Li, Chuyu Zhang, Yunxin Liu, Haodong Duan, Songyang Zhang, and Kai Chen. 2025. Condor: Enhance llm alignment with knowledge-driven data synthesis and refinement. *arXiv preprint arXiv:2501.12273*.

Razieh Chalehchaleh, Reza Farahbakhsh, and Noel Crespi. 2024. Enhancing multilingual fake news detection through llm-based data augmentation. In *The 13th International Conference on Complex Networks and their Applications*.

Yung-Chieh Chan, George Pu, Apaar Shanker, Parth Suresh, Penn Jenks, John Heyer, and Sam Denton.

2024. Balancing cost and effectiveness of synthetic data generation strategies for llms. *arXiv preprint arXiv:2409.19759*.

Kaiyan Chang, Kun Wang, Nan Yang, Ying Wang, Dantong Jin, Wenlong Zhu, Zhirong Chen, Cangyuan Li, Hao Yan, Yunhao Zhou, et al. 2024. Data is all you need: Finetuning llms for chip design via an automated design-data augmentation framework. In *Proceedings of the 61st ACM/IEEE Design Automation Conference*, pages 1–6.

Ting-Yun Chang and Robin Jia. 2022. Data curation alone can stabilize in-context learning. *arXiv preprint arXiv:2212.10378*.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yinuo Liu, Yaochen Wang, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. 2024a. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. *arXiv preprint arXiv:2402.04788*.

Jie Chen, Yupeng Zhang, Bingning Wang, Wayne Xin Zhao, Ji-Rong Wen, and Weipeng Chen. 2024b. Unveiling the flaws: Exploring imperfections in synthetic data and mitigation strategies for large language models. *arXiv preprint arXiv:2406.12397*.

Jiuhai Chen and Jonas Mueller. 2024. Automated data curation for robust language model fine-tuning. *arXiv preprint arXiv:2403.12776*.

Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, et al. 2023. Alpagasus: Training a better alpaca with fewer data. *arXiv preprint arXiv:2307.08701*.

Shengyuan Chen, Qinggang Zhang, Junnan Dong, Wen Hua, Qing Li, and Xiao Huang. 2024c. Entity alignment with noisy annotations from large language models. *arXiv preprint arXiv:2405.16806*.

Weize Chen, Jiarui Yuan, Chen Qian, Cheng Yang, Zhiyuan Liu, and Maosong Sun. 2024d. Optima: Optimizing effectiveness and efficiency for llm-based multi-agent system. *arXiv preprint arXiv:2410.08115*.

Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. 2024e. Self-play fine-tuning converts weak language models to strong language models. *arXiv preprint arXiv:2401.01335*.

Daixuan Cheng, Shaohan Huang, and Furu Wei. 2024. Adapting large language models via reading comprehension. In *The Twelfth International Conference on Learning Representations*.

Yash Prakash Chetnani. 2023. Evaluating the impact of model size on toxicity and stereotyping in generative llm. Master's thesis, State University of New York at Buffalo.

Juhwan Choi, Jungmin Yun, Kyohoon Jin, and YoungBin Kim. 2024. Multi-news+: Cost-efficient dataset cleansing via llm-based data annotation. *arXiv preprint arXiv:2404.09682*.

Vikram S Chundawat, Ayush K Tarun, Murari Mandal, Mukund Lahoti, and Pratik Narang. 2022. A universal metric for robust evaluation of synthetic tabular data. *IEEE Transactions on Artificial Intelligence*, 5(1):300–309.

John Joon Young Chung, Ece Kamar, and Saleema Amershi. 2023. Increasing diversity while maintaining accuracy: Text data generation with large language models and human interventions. *arXiv preprint arXiv:2306.04140*.

Simone Corbo, Luca Bancale, Valeria De Gennaro, Livia Lestingi, Vincenzo Scotti, and Matteo Camilli. 2025. How toxic can you get? search-based toxicity testing for large language models. *arXiv preprint arXiv:2501.01741*.

Yu Cui, Feng Liu, Pengbo Wang, Bohao Wang, Heng Tang, Yi Wan, Jun Wang, and Jiawei Chen. 2024. Distillation matters: empowering sequential recommenders to match the performance of large language models. In *Proceedings of the 18th ACM Conference on Recommender Systems*, pages 507–517.

Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Yihan Cao, Zihao Wu, Lin Zhao, Shaochen Xu, Fang Zeng, Wei Liu, et al. 2025. Auggpt: Leveraging chatgpt for text data augmentation. *IEEE Transactions on Big Data*.

Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. 2023. Toxicity in chatgpt: Analyzing persona-assigned language models. *arXiv preprint arXiv:2304.05335*.

Priyanshu Dhingra, Satyam Agrawal, Chandra Sekar Veerappan, Eng Siong Chng, and Rong Tong. 2024. Enhancing speech de-identification with llm-based data augmentation. In *2024 11th International Conference on Advanced Informatics: Concept, Theory and Application (ICAICTA)*, pages 1–5. IEEE.

G Dhruva, Ishani Bhat, Sanika M Rangayyan, and P Preethi. 2024. Synthetic data augmentation using large language models (llm): A case-study of the kamyr digester. In *2024 Third International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT)*, pages 1–7. IEEE.

Flavio Di Palo, Prateek Singhi, and Bilal Fadlallah. 2024. Performance-guided llm knowledge distillation for efficient text classification at scale. *arXiv preprint arXiv:2411.05045*.

Bosheng Ding, Chengwei Qin, Ruochen Zhao, Tianze Luo, Xinze Li, Guizhen Chen, Wenhan Xia, Junjie Hu, Luu Anh Tuan, and Shafiq Joty. 2024. Data augmentation using llms: Data perspectives, learning paradigms and challenges. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 1679–1705.

Qingxiu Dong, Li Dong, Xingxing Zhang, Zhifang Sui, and Furu Wei. 2024. Self-boosting large language models with synthetic preference data. *arXiv preprint arXiv:2410.06961*.

Vitor Gaboardi dos Santos, Guto Leoni Santos, Theo Lynn, and Boualem Benatallah. 2024. Identifying citizen-related issues from social media using llm-based data augmentation. In *International Conference on Advanced Information Systems Engineering*, pages 531–546. Springer.

Dayou Du, Yijia Zhang, Shijie Cao, Jiaqi Guo, Ting Cao, Xiaowen Chu, and Ningyi Xu. 2024. Bitdistiller: Unleashing the potential of sub-4-bit llms via self-distillation. *arXiv preprint arXiv:2402.10631*.

Qianlong Du, Chengqing Zong, and Jiajun Zhang. 2023. Mods: Model-oriented data selection for instruction tuning. *arXiv preprint arXiv:2311.15653*.

Oshin Dutta, Ritvik Gupta, and Sumeet Agarwal. 2024. Efficient llm pruning with global token-dependency awareness and hardware-adapted inference. In *Workshop on Efficient Systems for Foundation Models II@ ICML2024*.

Fahim Faisal, Md Mushfiqur Rahman, and Antonios Anastasopoulos. 2024. Dialectal toxicity detection: Evaluating llm-as-a-judge consistency across language varieties. *arXiv preprint arXiv:2411.10954*.

Chrisantha Fernando, Dylan Banarse, Henryk Michalewski, Simon Osindero, and Tim Rocktäschel. 2023. Promptbreeder: Self-referential self-improvement via prompt evolution. *arXiv preprint arXiv:2309.16797*.

Zeyu Gan and Yong Liu. 2024. Towards a theoretical understanding of synthetic data in llm post-training: A reverse-bottleneck perspective. *arXiv preprint arXiv:2410.01720*.

Gerald Gartlehner, Leila Kahwati, Rainer Hilscher, Ian Thomas, Shannon Kugley, Karen Crotty, Meera Viswanathan, Barbara Nussbaumer-Streit, Graham Booth, Nathaniel Erskine, et al. 2024. Data extraction for evidence synthesis using a large language model: A proof-of-concept study. *Research Synthesis Methods*.

Matthias Gerstgrasser, Rylan Schaeffer, Apratim Dey, Rafael Rafailov, Henry Sleight, John Hughes, Tomasz Korbak, Rajashree Agrawal, Dhruv Pai, Andrey Gromov, et al. 2024. Is model collapse inevitable? breaking the curse of recursion by accumulating real and synthetic data. *arXiv preprint arXiv:2404.01413*.

Akshay Goel, Almog Gueta, Omry Gilon, Chang Liu, Sofia Erell, Lan Huong Nguyen, Xiaohong Hao, Bolous Jaber, Shashir Reddy, Rupesh Kartha, et al. 2023. Llms accelerate annotation for medical information extraction. In *Machine Learning for Health (ML4H)*, pages 82–100. PMLR.

Fabricio Goes, Zisen Zhou, Piotr Sawicki, Marek Grzes, and Daniel G Brown. 2022. Crowd score: A method for the evaluation of jokes using large language model ai voters as judges. *arXiv preprint arXiv:2212.11214*.

Bastián González-Bustamante. 2024. Benchmarking llms in political content text-annotation: Proof-of-concept with toxicity and incivility data. *arXiv preprint arXiv:2409.09741*.

Saroj Gopali, Faranak Abri, Akbar Siami Namin, and Keith S Jones. 2024. The applicability of llms in generating textual samples for analysis of imbalanced datasets. *IEEE Access*.

Xinyu Guan, Li Lyna Zhang, Yifei Liu, Ning Shang, Youran Sun, Yi Zhu, Fan Yang, and Mao Yang. 2025. rstar-math: Small llms can master math reasoning with self-evolved deep thinking. *arXiv preprint arXiv:2501.04519*.

Song Guo, Jiahang Xu, Li Lyna Zhang, and Mao Yang. 2023. Compresso: Structured pruning with collaborative prompting learns compact large language models. *arXiv preprint arXiv:2310.05015*.

Xu Guoa and Yiqiang Chenb. 2023. Generative llms for synthetic data generation: Methods, challenges and the future. *International Journal of Information Technology*, 29(1).

Pengrui Han, Rafal Kocielnik, Adhithya Saravanan, Roy Jiang, Or Sharir, and Anima Anandkumar. 2024. Chatgpt based data augmentation for improved parameter-efficient debiasing of llms. *arXiv preprint arXiv:2402.11764*.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. *arXiv preprint arXiv:2203.09509*.

Xinmeng Hou. 2024. Mitigating biases to embrace diversity: A comprehensive annotation benchmark for toxic language. *arXiv preprint arXiv:2410.13313*.

Chen Huang, Yang Deng, Wenqiang Lei, Jiancheng Lv, and Ido Dagan. 2024a. Selective annotation via data allocation: These data should be triaged to experts for annotation rather than the model. *arXiv preprint arXiv:2405.12081*.

Chenghua Huang, Zhizhen Fan, Lu Wang, Fangkai Yang, Pu Zhao, Zeqi Lin, Qingwei Lin, Dongmei Zhang, Saravan Rajmohan, and Qi Zhang. 2024b. Self-evolved reward learning for llms. *arXiv preprint arXiv:2411.00418*.

Haoyu Huang, Chong Chen, Conghui He, Yang Li, Jiawei Jiang, and Wentao Zhang. 2024c. Can llms be good graph judger for knowledge graph construction? *arXiv preprint arXiv:2411.17388*.

Hui Huang, Yingqi Qu, Jing Liu, Muyun Yang, and Tiejun Zhao. 2024d. An empirical study of llm-as-a-judge for llm evaluation: Fine-tuned judge models are task-specific classifiers. *arXiv preprint arXiv:2403.02839*.

Jiangping Huang, Bochen Yi, Weisong Sun, Bangrui Wan, Yang Xu, Yebo Feng, Wenguang Ye, and Qinjun Qin. Enhancing review classification via llm-based data annotation and multi-perspective feature representation learning. *Available at SSRN 5002351*.

Jin Huang, Xinyu Li, Liang Gao, Qihao Liu, and Yue Teng. 2024e. Automatic programming via large language models with population self-evolution for dynamic job shop scheduling problem. *arXiv preprint arXiv:2410.22657*.

Xijie Huang, Li Lyna Zhang, Kwang-Ting Cheng, M Yang, and Mao Yang. 2023. Fewer is more: Boosting llm reasoning with reinforced context pruning. *arXiv preprint arXiv:2312.08901*.

Zheng Hui, Zhaoxiao Guo, Hang Zhao, Juanyong Duan, and Congrui Huang. 2024. Toxicraft: A novel framework for synthetic generation of harmful information. *arXiv preprint arXiv:2409.14740*.

Daniel P Jeong, Zachary C Lipton, and Pradeep Ravikumar. 2024. Llm-select: Feature selection with large language models. *arXiv preprint arXiv:2407.02694*.

Kaidi Jia, Yanxia Wu, and Rongsheng Li. 2024. Curriculum-style data augmentation for llm-based metaphor detection. *arXiv preprint arXiv:2412.02956*.

Xun Jiang, Feng Li, Han Zhao, Jiaying Wang, Jun Shao, Shihao Xu, Shu Zhang, Weiling Chen, Xavier Tang, Yize Chen, et al. 2024. Long term memory: The foundation of ai self-evolution. *arXiv preprint arXiv:2410.15665*.

Jaehun Jung, Peter West, Liwei Jiang, Faeze Brahman, Ximing Lu, Jillian Fisher, Taylor Sorensen, and Yejin Choi. 2023. Impossible distillation: from low-quality model to high-quality dataset & model for summarization and paraphrasing. *arXiv preprint arXiv:2305.16635*.

Feiyang Kang, Hoang Anh Just, Yifan Sun, Himanshu Jahagirdar, Yuanzhi Zhang, Rongxing Du, Anit Kumar Sahu, and Ruoxi Jia. 2024. Get more for less: Principled data selection for warming up fine-tuning in llms. *arXiv preprint arXiv:2405.02774*.

S Karunya, M Jalakandeshwaran, Thanuja Babu, and R Uma. 2023. Ai-powered real-time speech-to-speech translation for virtual meetings using machine learning models. In *2023 Intelligent Computing and Control for Engineering and Business Systems (ICCEBS)*, pages 1–6. IEEE.

Bo-Kyeong Kim, Geonmin Kim, Tae-Ho Kim, Thibault Castells, Shinkook Choi, Junho Shin, and Hyoung-Kyu Song. 2024. Shortened llama: A simple depth pruning for large language models. *arXiv preprint arXiv:2402.02834*, 11.

Minsang Kim and Seungjun Baek. 2024. Measuring sample importance in data pruning for training llms from a data compression perspective. *arXiv preprint arXiv:2406.14124*.

Hyukhun Koh, Dohyung Kim, Minwoo Lee, and Kyomin Jung. 2024. Can llms recognize toxicity? a structured investigation framework and toxicity metric. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6092–6114.

Achintya Kundu, Fabian Lim, Aaron Chew, Laura Wynter, Penny Chong, and Rhui Dih Lee. 2024. Efficiently distilling llms for edge applications. *arXiv preprint arXiv:2404.01353*.

Po-Nien Kung, Fan Yin, Di Wu, Kai-Wei Chang, and Nanyun Peng. 2023. Active instruction tuning: Improving cross-task generalization by training on prompt sensitive tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1813–1829.

Dong-Ho Lee, Hyundong Cho, Woojeong Jin, Jihyung Moon, Sungjoon Park, Paul Röttger, Jay Pujara, and Roy Ka-Wei Lee. 2024a. Improving covert toxicity detection by retrieving and generating references. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, pages 266–274.

Nicholas Lee, Thanakul Wattanawong, Sehoon Kim, Karttikeya Mangalam, Sheng Shen, Gopala Anumanchipalli, Michael W Mahoney, Kurt Keutzer, and Amir Gholami. 2024b. Llm2llm: Boosting llms with novel iterative data enhancement. *arXiv preprint arXiv:2403.15042*.

Haitao Li, Junjie Chen, Qingyao Ai, Zhumin Chu, Yujia Zhou, Qian Dong, and Yiqun Liu. 2024a. Calibraeval: Calibrating prediction distribution to mitigate selection bias in llms-as-judges. *arXiv preprint arXiv:2410.15393*.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024b. Llms-as-judges: a comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*.

Jiawei Li, Chong Feng, and Yang Gao. 2024c. Meteor: Evolutionary journey of large language models from guidance to self-growth. *arXiv preprint arXiv:2411.11933*.

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

Lei Li, Yongfeng Zhang, and Li Chen. 2023a. Prompt distillation for efficient llm-based recommendation. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 1348–1357.

Ming Li, Yong Zhang, Zhitao Li, Jiuhai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. 2023b. From quantity to quality: Boosting llm performance with self-guided data selection for instruction tuning. *arXiv preprint arXiv:2308.12032*.

Minzhi Li, Taiwei Shi, Caleb Ziems, Min-Yen Kan, Nancy F Chen, Zhengyuan Liu, and Diyi Yang. 2023c. Coannotating: Uncertainty-guided work allocation between human and large language models for data annotation. *arXiv preprint arXiv:2310.15638*.

Xinjin Li, Yu Ma, Yangchen Huang, Xingqi Wang, Yuzhen Lin, and Chenxi Zhang. 2024e. Synergized data efficiency and compression (sec) optimization for large language models. In *2024 4th International Conference on Electronic Information Engineering and Computer Science (EIECS)*, pages 586–591. IEEE.

Yichuan Li, Kaize Ding, Jianling Wang, and Kyumin Lee. 2024f. Empowering large language models for textual data augmentation. *arXiv preprint arXiv:2404.17642*.

Yunshui Li, Binyuan Hui, Xiaobo Xia, Jiaxi Yang, Min Yang, Lei Zhang, Shuzheng Si, Junhao Liu, Tongliang Liu, Fei Huang, et al. 2023d. One shot learning as instruction data prospector for large language models. *arXiv preprint arXiv:2312.10302*.

Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. 2023e. Synthetic data generation with large language models for text classification: Potential and limitations. *arXiv preprint arXiv:2310.07849*.

Xun Liang, Shichao Song, Zifan Zheng, Hanyu Wang, Qingchen Yu, Xunkai Li, Rong-Hua Li, Yi Wang, Zhonghao Wang, Feiyu Xiong, et al. 2024a. Internal consistency and self-feedback in large language models: A survey. *arXiv preprint arXiv:2407.14507*.

Yiming Liang, Ge Zhang, Xingwei Qu, Tianyu Zheng, Jiawei Guo, Xinrun Du, Zhenzhu Yang, Jiaheng Liu, Chenghua Lin, Lei Ma, et al. 2024b. I-sheep: Self-alignment of llm from scratch through an iterative self-enhancement paradigm. *arXiv preprint arXiv:2408.08072*.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. *arXiv preprint arXiv:2305.20050*.

Guanyu Lin, Tao Feng, Pengrui Han, Ge Liu, and Jiaxuan You. 2024a. Arxiv copilot: A self-evolving and efficient llm system for personalized academic assistance. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 122–130.

Xinyu Lin, Wenjie Wang, Yongqi Li, Shuo Yang, Fuli Feng, Yinwei Wei, and Tat-Seng Chua. 2024b. Data-efficient fine-tuning for llm-based recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 365–374.

Zi Lin, Zihan Wang, Yongqi Tong, Yangkun Wang, Yuxin Guo, Yujia Wang, and Jingbo Shang. 2023. Toxicchat: Unveiling hidden challenges of toxicity detection in real-world user-ai conversation. *arXiv preprint arXiv:2310.17389*.

Gui Ling, Ziyang Wang, Yuliang Yan, and Qingwen Liu. 2024. Slimgpt: Layer-wise structured pruning for large language models. *arXiv preprint arXiv:2412.18110*.

Philip Lippmann, Matthijs TJ Spaan, and Jie Yang. 2024. Illuminating blind spots of language models with targeted agent-in-the-loop synthetic data. *arXiv preprint arXiv:2403.17860*.

Deli Liu, Xiaoping Zhou, and Yu Li. 2025. Balancing performance and cost of llms in a multi-agent framework for bim data retrieval. *Architectural Engineering and Design Management*, pages 1–18.

Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, et al. 2024a. Best practices and lessons learned on synthetic data for language models. *arXiv preprint arXiv:2404.07503*.

Xiaoqun Liu, Jiacheng Liang, Luoxi Tang, Chenyu You, Muchao Ye, and Zhaohan Xi. 2024b. Buckle up: Robustifying llms at every customization stage via data curation. *arXiv preprint arXiv:2410.02220*.

Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. 2023. Trustworthy llms: A survey and guideline for evaluating large language models' alignment. *arXiv preprint arXiv:2308.05374*.

Yilun Liu, Shimin Tao, Xiaofeng Zhao, Ming Zhu, Wenbing Ma, Junhao Zhu, Chang Su, Yutai Hou, Miao Zhang, Min Zhang, et al. 2024c. Coachlm: Automatic instruction revisions improve the data quality in llm instruction tuning. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*, pages 5184–5197. IEEE.

Yizhu Liu, Ran Tao, Shengyu Guo, and Yifan Yang. 2024d. Improving topic relevance model by mix-structured summarization and llm-based data augmentation. *arXiv preprint arXiv:2404.02616*.

Zichen Liu, Changyu Chen, Chao Du, Wee Sun Lee, and Min Lin. 2024e. Sample-efficient alignment for llms. *arXiv preprint arXiv:2411.01493*.

Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. 2024. On llms-driven synthetic data generation, curation, and evaluation: A survey. *arXiv preprint arXiv:2406.15126*.

Jianqiao Lu, Wanjun Zhong, Wenyong Huang, Yufei Wang, Fei Mi, Baojun Wang, Weichao Wang, Lifeng Shang, and Qun Liu. 2023a. Self: Language-driven self-evolution for large language model. *arXiv preprint arXiv:2310.00533*.

Keming Lu, Hongyi Yuan, Zheng Yuan, Runji Lin, Junyang Lin, Chuanqi Tan, Chang Zhou, and Jingren Zhou. 2023b. # instag: Instruction tagging for analyzing supervised fine-tuning of large language models. In *The Twelfth International Conference on Learning Representations*.

Lei Lu, Zhepeng Wang, Ruexue Bao, Mengbing Wang, Fangyi Li, Yawen Wu, Weiwen Jiang, Jie Xu, Yanzhi Wang, and Shangqian Gao. 2024. All-in-one tuning and structural pruning for domain-specific llms. *arXiv preprint arXiv:2412.14426*.

Junyu Luo, Xiao Luo, Xiusi Chen, Zhiping Xiao, Wei Ju, and Ming Zhang. 2024a. Semievol: Semi-supervised fine-tuning for llm adaptation. *arXiv preprint arXiv:2410.14745*.

Junyu Luo, Xiao Luo, Kaize Ding, Jingyang Yuan, Zhiping Xiao, and Ming Zhang. 2024b. Robustft: Robust supervised fine-tuning for large language models under noisy response. *Preprint*, arXiv:2412.14922.

Junyu Luo, Weizhi Zhang, Ye Yuan, Yusheng Zhao, Junwei Yang, Yiyang Gu, Bohan Wu, Binqi Chen, Ziyue Qiao, Qingqing Long, et al. 2025. Large language model agent: A survey on methodology, applications and challenges. *arXiv preprint arXiv:2503.21460*.

Ziyang Luo, Can Xu, Pu Zhao, Qingfeng Sun, Xiubo Geng, Wenxiang Hu, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. 2023. Wizardcoder: Empowering code large language models with evol-instruct. *arXiv preprint arXiv:2306.08568*.

Tinh Son Luong, Thanh-Thien Le, Linh Ngo Van, and Thien Huu Nguyen. 2024. Realistic evaluation of toxicity in large language models. *arXiv preprint arXiv:2405.10659*.

Alisia Lupidi, Carlos Gemmell, Nicola Cancedda, Jane Dwivedi-Yu, Jason Weston, Jakob Foerster, Roberta Raileanu, and Maria Lomeli. 2024. Source2synth: Synthetic data generation and curation grounded in real data sources. *arXiv preprint arXiv:2409.08239*.

Xinyin Ma, Gongfan Fang, and Xinchao Wang. 2023. Llm-pruner: On the structural pruning of large language models. *Advances in neural information processing systems*, 36:21702–21720.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2024. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36.

Daniel McDonald, Rachael Papadopoulos, and Leslie Benningfield. 2024. Reducing llm hallucination using knowledge distillation: A case study with mistral large and mmlu benchmark. *Authorea Preprints*.

Xuran Ming, Shoubin Li, Mingyang Li, Lvlong He, and Qing Wang. 2024. Autolabel: Automated textual data annotation method based on active learning and large language model. In *International Conference on Knowledge Science, Engineering and Management*, pages 400–411. Springer.

Clint Morris, Michael Jurado, and Jason Zutty. 2024. Llm guided evolution-the automation of models advancing models. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 377–384.

Saurav Muralidharan, Sharath Turuvekere Sreenivas, Raviraj Bhuminand Joshi, Marcin Chochowski, Mostofa Patwary, Mohammad Shoeybi, Bryan Catanzaro, Jan Kautz, and Pavlo Molchanov. 2024. Compact language models via pruning and knowledge distillation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Ryumei Nakada, Yichen Xu, Lexin Li, and Linjun Zhang. 2024. Synthetic oversampling: Theory and a practical approach using llms to address data imbalance. *arXiv preprint arXiv:2406.03628*.

Sejoon Oh, Yiqiao Jin, Megha Sharma, Donghyun Kim, Eric Ma, Gaurav Verma, and Srijan Kumar. 2024. Uniguard: Towards universal safety guardrails for jailbreak attacks on multimodal large language models. *arXiv:2411.01703*.

Bo Pan, Zheng Zhang, Yifei Zhang, Yuntong Hu, and Liang Zhao. 2024a. Distilling large language models for text-attributed graph learning. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 1836–1845.

Liangming Pan, Michael Saxon, Wenda Xu, Deepak Nathani, Xinyi Wang, and William Yang Wang. 2023. Automatically correcting large language models: Surveying the landscape of diverse self-correction strategies. *arXiv preprint arXiv:2308.03188*.

Zhuoshi Pan, Qianhui Wu, Huiqiang Jiang, Menglin Xia, Xufang Luo, Jue Zhang, Qingwei Lin, Victor Rühle, Yuqing Yang, Chin-Yew Lin, et al. 2024b. Llmlingua-2: Data distillation for efficient and faithful task-agnostic prompt compression. *arXiv preprint arXiv:2403.12968*.

Jinlong Pang, Jiaheng Wei, Ankit Parag Shah, Zhaowei Zhu, Yaxuan Wang, Chen Qian, Yang Liu, Yujia Bao, and Wei Wei. 2024. Improving data efficiency via curating llm-driven rating systems. *arXiv preprint arXiv:2410.10877*.

Nicholas Pangakis and Samuel Wolken. 2024. Knowledge distillation in automated annotation: Supervised text classification with llm-generated training labels. *arXiv preprint arXiv:2406.17633*.

Changhua Pei, Zihan Liu, Jianhui Li, Erhan Zhang, Le Zhang, Haiming Zhang, Wei Chen, Dan Pei, and Gaogang Xie. 2024. Self-evolutionary group-wise log parsing based on large language model. In *2024 IEEE 35th International Symposium on Software Reliability Engineering (ISSRE)*, pages 49–60. IEEE.

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*.

Zheyan Qu, Lu Yin, Zitong Yu, Wenbo Wang, et al. 2024. Coursegpt-zh: an educational large language model based on knowledge distillation incorporating prompt optimization. *arXiv preprint arXiv:2405.04781*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Yi Ren, Shangmin Guo, Linlu Qiu, Bailin Wang, and Danica J Sutherland. Bias amplification in language model evolution: An iterated learning perspective. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Gaurav Sahu and Issam H Laradji. 2024. Mixsumm: Topic-based data augmentation using llms for low-resource extractive text summarization. *arXiv preprint arXiv:2407.07341*.

Gayathri Saranathan, Mahammad Parwez Alam, James Lim, Suparna Bhattacharya, Soon Yee Wong, Martin Foltin, and Cong Xu. Dele: Data efficient llm evaluation. In *ICLR 2024 Workshop on Navigating and Addressing Data Problems for Foundation Models*.

Shouvon Sarker, Xishuang Dong, Xiangfang Li, and Lijun Qian. 2024. Enhancing llm fine-tuning for text-to-sqls by sql quality measurement. *arXiv preprint arXiv:2410.01869*.

Maximilian Schmidhuber and Udo Kruschwitz. 2024. Llm-based synthetic datasets: Applications and limitations in toxicity detection. *LREC-COLING 2024*, page 37.

Nabeel Seedat, Nicolas Huynh, Boris van Breugel, and Mihaela van der Schaar. 2023. Curated llm: Synergy of llms and data curation for tabular augmentation in ultra low-data regimes. *arXiv preprint arXiv:2312.12112*.

Indira Sen, Dennis Assenmacher, Mattia Samory, Isabelle Augenstein, Wil van der Aalst, and Claudia Wagner. 2023. People make better edits: Measuring the efficacy of llm-generated counterfactually augmented data for harmful language detection. *arXiv preprint arXiv:2311.01270*.

Minju Seo, Jinheon Baek, James Thorne, and Sung Ju Hwang. 2024. Retrieval-augmented data augmentation for low-resource domain tasks. *arXiv preprint arXiv:2402.13482*.

Yiping Song, Juhua Zhang, Zhiliang Tian, Yuxin Yang, Minlie Huang, and Dongsheng Li. 2024a. Llm-based privacy data augmentation guided by knowledge distillation with a distribution tutor for medical text classification. *arXiv preprint arXiv:2402.16515*.

Yuncheng Song, Liang Ding, Changtong Zan, and Shujian Huang. 2024b. Self-evolution knowledge distillation for llm-based machine translation. *arXiv preprint arXiv:2412.15303*.

Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara. 2024. Large language models are inconsistent and biased evaluators. *arXiv preprint arXiv:2405.01724*.

Sowmya S Sundaram, Benjamin Solomon, Avani Khatri, Anisha Laumas, Purvesh Khatri, and Mark A Musen. 2024. Use of a structured knowledge base enhances metadata curation by large language models. *arXiv preprint arXiv:2404.05893*.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation and synthesis: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 930–957.

Shuo Tang, Xianghe Pang, Zexi Liu, Bohan Tang, Rui Ye, Xiaowen Dong, Yanfeng Wang, and Siheng Chen. 2024a. Synthesizing post-training data for llms through multi-agent simulation. *arXiv preprint arXiv:2410.14251*.

Yi Tang, Chia-Ming Chang, and Xi Yang. 2024b. Pdfchatannotator: A human-llm collaborative multi-modal data annotation tool for pdf-format catalogs. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, pages 419–430.

Zhengwei Tao, Ting-En Lin, Xiancai Chen, Hangyu Li, Yuchuan Wu, Yongbin Li, Zhi Jin, Fei Huang, Dacheng Tao, and Jingren Zhou. 2024. A survey on self-evolution of large language models. *arXiv preprint arXiv:2404.14387*.

Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. 2024. Judging the judges: Evaluating alignment and vulnerabilities in llms-as-judges. *arXiv preprint arXiv:2406.12624*.

Vithursan Thangarasa, Ganesh Venkatesh, Mike Lasby, Nish Sinnadurai, and Sean Lie. 2024. Self-data distillation for recovering quality in pruned large language models. *arXiv preprint arXiv:2410.09982*.

Yun-Da Tsai, Mingjie Liu, and Haoxing Ren. 2024. Code less, align more: Efficient llm fine-tuning for code generation with data pruning. *arXiv preprint arXiv:2407.05040*.

Praneeth Vadlapati. 2024. Autopuredata: Automated filtering of web data for llm fine-tuning. *arXiv preprint arXiv:2406.19271*.

Veniamin Veselovsky, Manoel Horta Ribeiro, Akhil Arora, Martin Josifoski, Ashton Anderson, and Robert West. 2023. Generating faithful synthetic data with large language models: A case study in computational social science. *arXiv preprint arXiv:2305.15041*.

Tu Vu, Kalpesh Krishna, Salaheddin Alzubi, Chris Tar, Manaal Faruqui, and Yun-Hsuan Sung. 2024. Foundational autoraters: Taming large language models for better automatic evaluation. *arXiv preprint arXiv:2407.10817*.

Zhongwei Wan, Xin Wang, Che Liu, Samiul Alam, Yu Zheng, Jiachen Liu, Zhongnan Qu, Shen Yan, Yi Zhu, Quanlu Zhang, et al. 2023. Efficient large language models: A survey. *arXiv preprint arXiv:2312.03863*.

Fei Wang, Ninareh Mehrabi, Palash Goyal, Rahul Gupta, Kai-Wei Chang, and Aram Galstyan. 2024a. Data advisor: Dynamic data curation for safety alignment of large language models. *arXiv preprint arXiv:2410.05269*.

Jiahao Wang, Bolin Zhang, Qianlong Du, Jiajun Zhang, and Dianhui Chu. 2024b. A survey on data selection for llm instruction tuning. *arXiv preprint arXiv:2402.05123*.

Jianwei Wang, Kai Wang, Ying Zhang, Wenjie Zhang, Xiwei Xu, and Xuemin Lin. 2025. On llm-enhanced mixed-type data imputation with high-order message passing. *arXiv preprint arXiv:2501.02191*.

Peidong Wang, Ming Wang, Zhiming Ma, Xiaocui Yang, Shi Feng, Daling Wang, and Yifei Zhang. 2024c. Language models as continuous self-evolving data engineers. *arXiv preprint arXiv:2412.15151*.

Pengkun Wang, Zhe Zhao, HaiBin Wen, Fanfu Wang, Binwu Wang, Qingfu Zhang, and Yang Wang. 2024d. Llm-autoda: Large language model-driven automatic data augmentation for long-tailed problems. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Siyuan Wang, Zhuohan Long, Zhihao Fan, Zhongyu Wei, and Xuanjing Huang. 2024e. Benchmark self-evolving: A multi-agent framework for dynamic llm evaluation. *arXiv preprint arXiv:2402.11443*.

Tianlu Wang, Ilia Kulikov, Olga Golovneva, Ping Yu, Weizhe Yuan, Jane Dwivedi-Yu, Richard Yuanzhe Pang, Maryam Fazel-Zarandi, Jason Weston, and Xian Li. 2024f. Self-taught evaluators. *arXiv preprint arXiv:2408.02666*.

Tianlu Wang, Ilia Kulikov, Olga Golovneva, Ping Yu, Weizhe Yuan, Jane Dwivedi-Yu, Richard Yuanzhe Pang, Maryam Fazel-Zarandi, Jason Weston, and Xian Li. 2024g. Self-taught evaluators. *arXiv preprint arXiv:2408.02666*.

Xinru Wang, Hannah Kim, Sajjadur Rahman, Kushan Mitra, and Zhengjie Miao. 2024h. Human-llm collaborative annotation through effective verification of llm labels. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–21.

Yau-Shian Wang and Yingshan Chang. 2022. Toxicity detection with generative prompt-based inference. *arXiv preprint arXiv:2205.12390*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023a. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 13484–13508.

Yu Wang, Yifan Gao, Xiusi Chen, Haoming Jiang, Shiyang Li, Jingfeng Yang, Qingyu Yin, Zheng Li, Xian Li, Bing Yin, et al. 2024i. Memoryllm: Towards self-updatable large language models. *arXiv preprint arXiv:2402.04624*.

Yuxia Wang, Haonan Li, Xudong Han, Preslav Nakov, and Timothy Baldwin. 2023b. Do-not-answer: A dataset for evaluating safeguards in llms. *arXiv preprint arXiv:2308.13387*.

Yuxia Wang, Haonan Li, Xudong Han, Preslav Nakov, and Timothy Baldwin. 2024j. Do-not-answer: Evaluating safeguards in llms. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 896–911.

Yuxin Wang, Duanyu Feng, Yongfu Dai, Zhengyu Chen, Jimin Huang, Sophia Ananiadou, Qianqian Xie, and Hao Wang. 2024k. Harmonic: Harnessing llms for tabular data synthesis and privacy protection. *arXiv preprint arXiv:2408.02927*.

Zhenhua Wang, Guang Xu, and Ming Ren. 2024l. Llm-generated natural language meets scaling laws: New explorations and data augmentation methods. *arXiv preprint arXiv:2407.00322*.

Martin Weyssow, Aton Kamanda, Xin Zhou, and Houari Sahraoui. 2024. Codeultrafeedback: An llm-as-a-judge dataset for aligning large language models to coding preferences. *arXiv preprint arXiv:2403.09032*.

Chenxi Whitehouse, Monojit Choudhury, and Alham Fikri Aji. 2023. Llm-powered data augmentation for enhanced cross-lingual performance. *arXiv preprint arXiv:2305.14288*.

Minghao Wu and Alham Fikri Aji. 2023. Style over substance: Evaluation biases for large language models. *arXiv preprint arXiv:2307.03025*.

Tianhao Wu, Weizhe Yuan, Olga Golovneva, Jing Xu, Yuandong Tian, Jiantao Jiao, Jason Weston, and Sainbayar Sukhbaatar. 2024. Meta-rewarding language models: Self-improving alignment with llm-as-a-meta-judge. *arXiv preprint arXiv:2407.19594*.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. 2024. Less: Selecting influential data for targeted instruction tuning. *arXiv preprint arXiv:2402.04333*.

Zhe Xie, Zeyan Li, Xiao He, Longlong Xu, Xidao Wen, Tieying Zhang, Jianjun Chen, Rui Shi, and Dan Pei. 2024. Chatts: Aligning time series with llms via synthetic data for enhanced understanding and reasoning. *arXiv preprint arXiv:2412.03104*.

Huajian Xin, Daya Guo, Zhihong Shao, ZZ Ren, Qihao Zhu, Bo Liu, Chong Ruan, Wenda Li, and Xiaodan Liang. Advancing theorem proving in llms through large-scale synthetic data. In *The 4th Workshop on Mathematical Reasoning and AI at NeurIPS'24*.

Fangzhi Xu, Qiushi Sun, Kanzhi Cheng, Jun Liu, Yu Qiao, and Zhiyong Wu. 2024a. Interactive evolution: A neural-symbolic self-training framework for large language models. *arXiv preprint arXiv:2406.11736*.

Shengzhe Xu, Cho-Ting Lee, Mandar Sharma, Raquib Bin Yousuf, Nikhil Muralidhar, and Naren Ramakrishnan. 2024b. Are llms naturally good at synthetic tabular data generation? *arXiv preprint arXiv:2406.14541*.

Xiaohan Xu, Ming Li, Chongyang Tao, Tao Shen, Reynold Cheng, Jinyang Li, Can Xu, Dacheng Tao, and Tianyi Zhou. 2024c. A survey on knowledge distillation of large language models. *arXiv preprint arXiv:2402.13116*.

Yang Xu, Yongqiang Yao, Yufan Huang, Mengnan Qi, Maoquan Wang, Bin Gu, and Neel Sundaresan. 2023. Rethinking the instruction quality: Lift is what you need. *CoRR*.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. 2024d. Magpie: Alignment data synthesis from scratch by prompting aligned llms with nothing. *arXiv preprint arXiv:2406.08464*.

Chuanpeng Yang, Yao Zhu, Wang Lu, Yidong Wang, Qian Chen, Chenlong Gao, Bingjie Yan, and Yiqiang Chen. 2024a. Survey on knowledge distillation for large language models: methods, evaluation, and application. *ACM Transactions on Intelligent Systems and Technology*.

Shuangtao Yang, Xiaoyi Liu, Xiaozheng Dong, and Bo Fu. 2024b. Mini-da: Improving your model performance through minimal data augmentation using llm. In *Proceedings of the Fifth Workshop on Data Science with Human-in-the-Loop (DaSH 2024)*, pages 25–30.

Yingxuan Yang, Huayi Wang, Muning Wen, Xiaoyun Mo, Qiuying Peng, Jun Wang, and Weinan Zhang. 2024c. P3: A policy-driven, pace-adaptive, and diversity-promoted framework for data pruning in llm training. *arXiv preprint arXiv:2408.05541*.

Junjie Ye, Nuo Xu, Yikun Wang, Jie Zhou, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024. Llm-da: Data augmentation via large language models for few-shot named entity recognition. *arXiv preprint arXiv:2402.14568*.

Da Yin, Xiao Liu, Fan Yin, Ming Zhong, Hritik Bansal, Jiawei Han, and Kai-Wei Chang. 2023. Dynosaur: A dynamic growth paradigm for instruction-tuning data curation. *arXiv preprint arXiv:2305.14327*.

Junjie Oscar Yin and Alexander M Rush. 2024. Compute-constrained data selection. *arXiv preprint arXiv:2410.16208*.

Jiaxuan You, Mingjie Liu, Shrimai Prabhumoye, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. 2024. Llm-evolve: Evaluation for llm's evolving capability on benchmarks. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16937–16942.

Simon Yu, Liangyu Chen, Sara Ahmadian, and Marzieh Fadaee. 2024. Diversify and conquer: Diversity-centric data selection with iterative refinement. *arXiv preprint arXiv:2409.11378*.

Jiayi Yuan, Ruixiang Tang, Xiaoqian Jiang, and Xia Hu. 2023. Llm for patient-trial matching: Privacy-aware data augmentation towards better performance and generalizability. In *American Medical Informatics Association (AMIA) Annual Symposium*.

Jiayi Yuan, Ruixiang Tang, Xiaoqian Jiang, and Xia Hu. 2024a. Large language models for healthcare data augmentation: An example on patient-trial matching. In *AMIA Annual Symposium Proceedings*, volume 2023, page 1324.

Tongxin Yuan, Zhiwei He, Lingzhong Dong, Yiming Wang, Ruijie Zhao, Tian Xia, Lizhen Xu, Binglin Zhou, Fangqi Li, Zhuosheng Zhang, et al. 2024b. R-judge: Benchmarking safety risk awareness for llm agents. *arXiv preprint arXiv:2401.10019*.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2024c. Self-rewarding language models. *arXiv preprint arXiv:2401.10020*.

Yuanhao Yue, Chengyu Wang, Jun Huang, and Peng Wang. 2024. Building a family of data augmentation models for low-cost llm fine-tuning on the cloud. *arXiv preprint arXiv:2412.04871*.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. 2022. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488.

Oleg Zendel, J Shane Culpepper, Falk Scholer, and Paul Thomas. 2024. Enhancing human annotation: Leveraging large language models and efficient batch processing. In *Proceedings of the 2024 Conference on Human Information Interaction and Retrieval*, pages 340–345.

Min Zeng, Caiquan Liu, Shiqi Zhang, Li Xie, Chen Sang, and Xiaoxin Chen. 2024. Data quality enhancement on the basis of diversity with large language models for text classification: Uncovered, difficult, and noisy. *arXiv preprint arXiv:2412.06575*.

Jiang Zhang, Qiong Wu, Yiming Xu, Cheng Cao, Zheng Du, and Konstantinos Psounis. 2024a. Efficient toxic content detection by bootstrapping and distilling large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 21779–21787.

Nan Zhang, Yanchi Liu, Xujiang Zhao, Wei Cheng, Runxue Bao, Rui Zhang, Prasenjit Mitra, and Haifeng Chen. 2024b. Pruning as a domain-specific llm extractor. *arXiv preprint arXiv:2405.06275*.

Ruoyu Zhang, Yanzeng Li, Yongliang Ma, Ming Zhou, and Lei Zou. 2023a. Llmaaa: Making large language models as active annotators. *arXiv preprint arXiv:2310.19596*.

Shenglin Zhang, Pengtian Zhu, Minghua Ma, Jiagang Wang, Yongqian Sun, Dongwen Li, Jingyu Wang, Qianying Guo, Xiaolei Hua, Lin Zhu, et al. 2024c. Enhanced fine-tuning of lightweight domain-specific q&a model based on large language models. In *2024 IEEE 35th International Symposium on Software Reliability Engineering Workshops (ISSREW)*, pages 61–66. IEEE.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. 2023b. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*.

Xinghua Zhang, Bowen Yu, Haiyang Yu, Yangyu Lv, Tingwen Liu, Fei Huang, Hongbo Xu, and Yongbin Li. 2023c. Wider and deeper llm networks are fairer llm evaluators. *arXiv preprint arXiv:2308.01862*.

Yuzhe Zhang, Huan Liu, Yang Xiao, Mohammed Amoon, Dalin Zhang, Di Wang, Shusen Yang, and Chai Quek. 2024d. Llm-enhanced multi-teacher knowledge distillation for modality-incomplete emotion recognition in daily healthcare. *IEEE Journal of Biomedical and Health Informatics*.

Ziyan Zhang, Yang Hou, Chen Gong, and Zhenghua Li. 2025. Data augmentation for cross-domain parsing via lightweight llm generation and tree hybridization. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 11235–11247.

Jiachen Zhao, Wenlong Zhao, Andrew Drozdov, Benjamin Rozonoyer, Md Arafat Sultan, Jay Yoon Lee, Mohit Iyyer, and Andrew McCallum. 2024a. Multistage collaborative knowledge distillation from a large language model for semi-supervised sequence generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14201–14214.

Wanru Zhao, Hongxiang Fan, Shell Xu Hu, Wangchunshu Zhou, and Nicholas Donald Lane. 2024b. CLUES: Collaborative private-domain high-quality data selection for llms via training dynamics. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems (NeurIPS)*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

Xuanhe Zhou, Xinyang Zhao, and Guoliang Li. 2024a. Llm-enhanced data management. *arXiv preprint arXiv:2402.02643*.

Yang Zhou, Shimin Shan, Hongkui Wei, Zhehuan Zhao, and Wenshuo Feng. 2024b. Pga-scire: Harnessing llm on data augmentation for enhancing scientific relation extraction. *arXiv preprint arXiv:2405.20787*.

Yue Zhou, Chenlu Guo, Xu Wang, Yi Chang, and Yuan Wu. 2024c. A survey on data augmentation in large model era. *arXiv preprint arXiv:2401.15422*.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. 2023. Judgelm: Fine-tuned large language models are scalable judges. *arXiv preprint arXiv:2310.17631*.

A Statistics

To demonstrate the research momentum in data-efficient LLM post-training, we conducted a statistical analysis of the surveyed papers. As shown in Figure 8, there has been a remarkable growth trajectory in this field: from merely 3 publications in 2022 to 31 papers in 2023, followed by a substantial surge to 158 papers in 2024, with 23 additional publications already recorded by February 2025. This trend clearly indicates the academic community's growing interest in this research direction, with the momentum continuing to accelerate. The rapid growth also underscores the critical importance of data-efficient post-training approaches in the LLM domain.

Furthermore, we performed a word frequency analysis on the titles of all surveyed papers and generated a word cloud visualization (Figure 9). The word cloud prominently features terms like

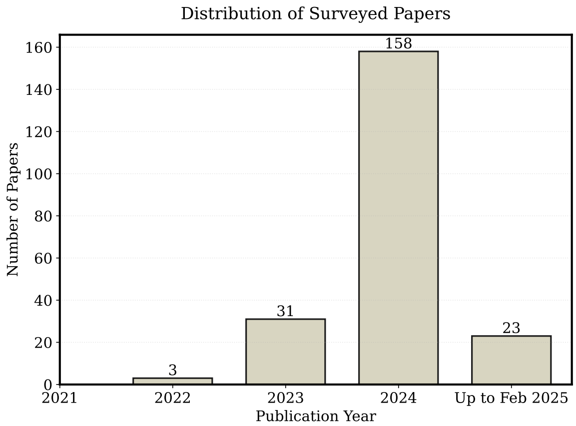

Figure 8: Distribution on publication year of surveyed papers.

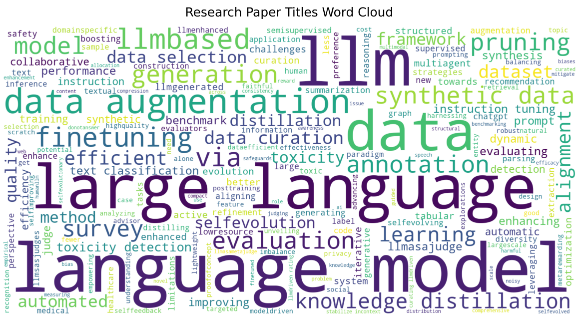

Figure 9: Word cloud of research paper titles.

language, *model*, and *large*, reflecting the centrality of large language models in this research area. The significant presence of keywords such as *data*, *efficient*, *finetuning*, and *distillation* highlights the emphasis on data efficiency and model optimization in current research endeavors. These visualizations strongly support our paper's focus on the data-centric perspective and its timeliness in the field.

The analysis demonstrates that data-efficient approaches to LLM post-training represent not only an emerging trend but also a fundamental research direction with significant implications for the advancement of language models.

B Takeaway Insights

B.1 Key Findings

Recent advancements in data-efficient LLM post-training reveal fundamental principles governing data-model interactions:

(1) The ***data flywheel*** paradigm integrates selection, augmentation, and evolution into a closed-loop lifecycle. This self-reinforcing mechanism enables continuous quality improvement through iterative refinement, transcending traditional linear data consumption

(2) **Value-centric data curation** outperforms scale-driven approaches in low-resource scenarios. Techniques like adaptive importance weighting and uncertainty-aware sampling maximize information density per training instance

(3) **Model-data co-optimization** enables joint improvements in efficiency and performance through innovations like dynamic token pruning and parameter-efficient adaptation

B.2 Paradigm Shifts

The field is witnessing fundamental changes in data utilization:

(1) Evolution from static datasets to **dynamic value-flow ecosystems** where data continuously evolves through model feedback. This necessitates new frameworks for monitoring data quality and lineage across iterations

(2) Emergence of **human-AI collaborative frameworks** combining automated generation with expert oversight. These hybrid pipelines leverage LLMs for initial labeling while preserving human judgment for critical cases

(3) Development of **cross-modal distillation** techniques that maintain semantic fidelity while reducing architectural constraints through learned alignment spaces

B.3 Critical Limitations

Current approaches face several key challenges:

(1) **Limited domain expertise** in data synthesis and refinement, where general-purpose models may fail to capture specialized knowledge and nuances required for professional domains

(2) **Scalability bottlenecks** in large-scale data generation, particularly in balancing computational costs with the need for diverse, high-quality datasets for pre-training

(3) Absence of **standardized metrics** for assessing synthetic data quality, especially in evaluating semantic fluency, information accuracy, and potential biases

B.4 Future Directions

Addressing these limitations requires advances in:

(1) **Domain-specific** pre-trained models and refinement techniques that can better capture professional knowledge while optimizing data quality and reducing annotation costs

(2) **Parallel and cost-effective frameworks** for large-scale data generation that maintain an optimal balance between data diversity and relevance
(3) **Robust evaluation metrics** and frameworks that can reliably assess synthetic data quality across different domains and use cases

C Acknowledgment of AI Assistance in Writing and Revision

We acknowledge the use of LLMs for grammar checking and language enhancement. This usage complies with the ACL Policy on AI Writing Assistance. All content and technical contributions remain original to the authors.

D Literature Review Summary

To provide a comprehensive overview of the surveyed literature, we present a detailed summary table of all referenced papers. The table includes seven key fields for each paper: **Title** (the paper's full title), **Citation** (reference key), **TLDR** (a brief summary of the paper's main contributions), **Category** (the paper's primary research direction within data-efficient LLM post-training), **Year** (publication year), **Venue** (publication venue), and **Link** (direct link to the paper). This structured compilation offers readers quick access to the original papers, enables easy tracking of research evolution across different categories, and facilitates future research by providing a comprehensive reference database of the field's development.

Table 2: Summary of Referenced Papers

Title	Citation	TLDR	Category	Year	Venue	Link
Data-efficient Fine-tuning for LLM-based Recommendation	(Lin et al., 2024b)	Propose data pruning method for efficient LLM - based recommendation.	Data Selection	2024	ACM	link
CoachLM: Automatic Instruction Revisions Improve the Data Quality in LLM Instruction Tuning	(Liu et al., 2024c)	CoachLM automatically revises samples to enhance instruction dataset quality.	Data Selection, Data Quality Enhancement	2023	IEEE	link
Alpagasus: Training a Better Alpaca with Fewer Data	(Chen et al., 2023)	Propose data selection strategy, filter low - quality data for IFT, ALPAGASUS as example.	Data Selection	2024	NIPS / ICML / ICLR	link
From Quantity to Quality: Boosting LLM Performance with Self-Guided Data Selection for Instruction Tuning	(Li et al., 2023b)	Introduce self - guided method for LLMs to select samples, key innovation IFD metric.	Data Selection	2024	*ACL	link
Rethinking the Instruction Quality: LIFT is What You Need	(Xu et al., 2023)	LIFT elevates instruction quality by broadening data distribution.	Data Selection	2023	arxiv	link
Instag: Instruction tagging for analyzing supervised fine-tuning of large language models.pdf	(Lu et al., 2023b)	Propose INSTAG to tag instructions, find benefits for LLMs, and a data sampling procedure.	Data Selection	2024	NIPS / ICML / ICLR	link
MoDS: Model-oriented Data Selection for Instruction Tuning	(Du et al., 2023)	MoDS selects instruction data by quality, coverage and necessity.	Data Selection	2023	arxiv	link
SELF-INSTRUCT: Aligning Language Models with Self-Generated Instructions	(Wang et al., 2023a)	SELF - INSTRUCT bootstraps from LM for instruction - following, nearly annotation - free.	Data Selection	2023	*ACL	link
Active Instruction Tuning: Improving Cross-Task Generalization by Training on Prompt Sensitive Tasks	(Kung et al., 2023)	Propose active IT based on prompt uncertainty to select tasks for LLM tuning.	Data Selection	2023	*ACL	link
Automated Data Curation for Robust Language Model Fine-Tuning	(Chen and Mueller, 2024)	Introduce CLEAR for data curation in LLM fine - tuning without extra computations.	Data Selection	2024	*ACL	link
CLUES: Collaborative Private-domain High-quality Data Selection for LLMs via Training Dynamics	(Zhao et al., 2024b)	Propose data quality control via training dynamics for collaborative LLM training.	Data Selection	2024	NIPS / ICML / ICLR	link
Compute-Constrained Data Selection	(Yin and Rush, 2024)	Formalize data selection problem cost - aware, model trade - offs.	Data Selection	2025	NIPS / ICML / ICLR	link
DATA ADVISOR: Dynamic Data Curation for Safety Alignment of Large Language Models	(Wang et al., 2024a)	DATA ADVISOR for data generation to enhance LLM safety.	Data Selection	2024	*ACL	link
Data Curation Alone Can Stabilize In-context Learning	(Chang and Jia, 2022)	Two methods curate training data subsets to stabilize ICL without algorithm changes.	Data Selection	2023	*ACL	link
Get more for less: Principled Data Selection for Warming Up Fine-Tuning in LLMs	(Kang et al., 2024)	Select data to nudge pre - training dist. closer to target dist. for cost - effective fine - tuning.	Data Selection	2024	NIPS / ICML / ICLR	link
Improving Data Efficiency via Curating LLM-Driven Rating Systems	(Pang et al., 2024)	DS2 corrects LLM - based scores for data selection promoting diversity.	Data Selection	2025	NIPS / ICML / ICLR	link
LLM-Select: Feature Selection with Large Language Models	(Jeong et al., 2024)	LLMs can select predictive features without seeing training data.	Data Selection	2024	Journal	link
One-Shot Learning as Instruction Data Prospector for Large Language Models	(Li et al., 2023d)	NUGGETS uses one - shot learning to select high - quality instruction data.	Data Selection	2024	*ACL	link
SAMPLE-EFFICIENT ALIGNMENT FOR LLMS	(Liu et al., 2024e)	Introduce unified algorithm for LLM alignment based on Thompson sampling.	Data Selection	2024	arxiv	link
LESS: Selecting Influential Data for Targeted Instruction Tuning	(Xia et al., 2024)	Propose LESS to select data for targeted instruction tuning in LLMs.	Data Selection	2024	NIPS / ICML / ICLR	link
An Experimental Design Framework for Label-Efficient Supervised Finetuning of Large Language Models	(Bhatt et al., 2024)	Propose experimental design for SFT in LLMs to mitigate annotation cost.	Data Selection	2024	*ACL	link
DELE: Data Efficient LLM Evaluation	(Saranathan et al.)	Propose adaptive sampling for LLM evaluation to reduce cost without losing integrity.	Data Selection	2024	NIPS / ICML / ICLR	link

Continued on next page

Table 2 – Continued

Title	Citation	TLDR	Category	Year	Venue	Link
Towards a Theoretical Understanding of Synthetic Data in LLM Post-Training: A Reverse-Bottleneck Perspective	(Gan and Liu, 2024)	Model synthetic data gen process, relate generalization & info gain.	Data Synthesis	2024	arxiv	link
Advancing Theorem Proving in LLMs through Large-Scale Synthetic Data	(Xin et al.)	Generate Lean 4 proof data to enhance LLM theorem - proving, without experimental focus.	Data Synthesis	2024	NIPS / ICML / ICLR	link
Are LLMs Naturally Good at Synthetic Tabular Data Generation?	(Xu et al., 2024b)	LLMs as-is or fine - tuned are bad at tabular data generation; permutation - aware can help.	Data Synthesis	2024	arxiv	link
Balancing Cost and Effectiveness of Synthetic Data Generation Strategies for LLMs	(Chan et al., 2024)	Group synthetic data strategies, study LLM training, propose selection framework.	Data Synthesis	2024	NIPS / ICML / ICLR	link
Best Practices and Lessons Learned on Synthetic Data for Language Models	(Liu et al., 2024a)	The paper focuses on synthetic data for LMs, its use, challenges and responsible use.	Data Synthesis	2024	arxiv	link
ChatTS: Aligning Time Series with LLMs via Synthetic Data for Enhanced Understanding and Reasoning	(Xie et al., 2024)	ChatTS, a TS - MLLM, uses synthetic data for time series analysis.	Data Synthesis	2024	arxiv	link
Data extraction for evidence synthesis using a large language model: A proof-of-concept study	(Gartlehner et al., 2024)	The study assesses Claude 2's data extraction in evidence synthesis.	Data Synthesis	2024	Journal	link
Illuminating Blind Spots of Language Models with Targeted Agent-in-the-Loop Synthetic Data	(Lippmann et al., 2024)	Use intelligent agents as teachers to generate samples for blind spot mitigation.	Data Synthesis	2024	arxiv	link
Generating Faithful Synthetic Data with Large Language Models: A Case Study in Computational Social Science	(Veselovsky et al., 2023)	The paper studies strategies to increase synthetic data faithfulness.	Data Synthesis	2023	arxiv	link
Generative LLMs for Synthetic Data Generation: Methods, Challenges and the Future	(Guoa and Chenb, 2023)	The paper focuses on using LLMs for synthetic data generation & related aspects.	Data Synthesis	2023	Journal	link
HARMONIC: Harnessing LLMs for Tabular Data Synthesis and Privacy Protection	(Wang et al., 2024k)	Introduce HARMONIC for tabular data synth & privacy, use LLMs w/ fine - tuning.	Data Synthesis	2024	NIPS / ICML / ICLR	link
Magpie: Alignment Data Synthesis from Scratch by Prompting Aligned LLMs with Nothing	(Xu et al., 2024d)	MAGPIE self - synthesizes alignment data from aligned LLMs without human prompts.	Data Synthesis	2024	arxiv	link
Synthesizing Post-Training Data for LLMs through Multi-Agent Simulation	(Tang et al., 2024a)	MATRIX multi - agent simulator creates scenarios for data synthesis in LLM post - training.	Data Synthesis	2025	NIPS / ICML / ICLR	link
Synthetic Data Generation with Large Language Models for Text Classification: Potential and Limitations	(Li et al., 2023e)	Explore factors moderating LLM - generated data effectiveness in text classification.	Data Synthesis	2023	*ACL	link
Synthetic Oversampling: Theory and A Practical Approach Using LLMs to Address Data Imbalance	(Nakada et al., 2024)	Develop theoretical foundations for synthetic oversampling using LLMs.	Data Synthesis	2024	arxiv	link
Unveiling the Flaws: Exploring Imperfections in Synthetic Data and Mitigation Strategies for Large Language Models	(Chen et al., 2024b)	This paper explores synthetic data flaws in LLM & presents a mitigation method.	Data Synthesis	2024	*ACL	link
Condor: Enhance LLM Alignment with Knowledge-Driven Data Synthesis and Refinement	(Cao et al., 2025)	Condor generates high - quality SFT data with two - stage framework for LLMs.	Data Synthesis	2025	arxiv	link
Data Augmentation using LLMs: Data Perspectives, Learning Paradigms and Challenges	(Ding et al., 2024)	The paper explores LLM - based data augmentation, challenges & learning paradigms.	Data Augmentation	2024	*ACL	link
Data is all you need: Finetuning LLMs for Chip Design via an Automated design-data augmentation framework	(Chang et al., 2024)	Propose an automated design - data augmentation framework for LLMs in chip design.	Data Augmentation	2024	ACM	link

Continued on next page

Table 2 – Continued

Title	Citation	TLDR	Category	Year	Venue	Link
LLM-powered Data Augmentation for Enhanced Cross-lingual Performance	(Whitehouse et al., 2023)	Uses LLMs for data augmentation in limited multilingual datasets.	Data Augmentation, Survey	2023	*ACL	link
LLM-DA: Data Augmentation via Large Language Models for Few-Shot Named Entity Recognition	(Ye et al., 2024)	LLM - DA augments data at context/entity levels for few - shot NER.	Data Augmentation	2024	arxiv	link
LLM-Generated Natural Language Meets Scaling Laws: New Explorations and Data Augmentation Methods	(Wang et al., 2024l)	Calculates LLMNL and HNL by scaling laws, proposes ZGPTDA for data augmentation.	Data Augmentation	2024	arxiv	link
A Survey on Data Augmentation in Large Model Era	(Zhou et al., 2024c)	Paper reviews large - model - driven data aug. methods, applications & future challenges.	Data Augmentation	2024	arxiv	link
ChatGPT Based Data Augmentation for Improved Parameter-Efficient Debiasing of LLMs	(Han et al., 2024)	Use ChatGPT to generate data for LLM debiasing with two strategies.	Data Augmentation	2024	COLM	link
A Guide To Effectively Leveraging LLMs for Low-Resource Text Summarization: Data Augmentation and Semi-supervised Approaches	(Sahu and Laradji, 2024)	Two new methods for low - resource text summarization are proposed.	Data Augmentation	2025	*ACL	link
Empowering Large Language Models for Textual Data Augmentation	(Li et al., 2024f)	Propose a solution to auto - generate LLM augmentation instructions for quality data.	Data Augmentation	2024	*ACL	link
LLM-AutoDA: Large Language Model-Driven Automatic Data Augmentation for Long-tailed Problems	(Wang et al., 2024d)	Proposes LLM - AutoDA for long - tailed data augmentation by leveraging large - scale models.	Data Augmentation	2024	NIPS / ICML / ICLR	link
Building a Family of Data Augmentation Models for Low-cost LLM Fine-tuning on the Cloud	(Yue et al., 2024)	Present data augmentation models for low - cost LLM fine - tuning with key functionalities.	Data Augmentation	2025	*ACL	link
Mini-DA: Improving Your Model Performance through Minimal Data Augmentation using LLM	(Yang et al., 2024b)	Mini - DA selects challenging samples for augmentation, improving resource utilization.	Data Augmentation	2024	*ACL	link
Data Augmentation for Text-based Person Retrieval Using Large Language Models	(Lin et al., 2024b)	Propose LLM - DA for TPR, use TFF & BSS to augment data concisely & efficiently.	Data Augmentation	2024	*ACL	link
Data Augmentation for Cross-domain Parsing via Lightweight LLM Generation and Tree Hybridization	(Zhang et al., 2025)	Propose data augmentation via LLM & tree hybridization for cross - domain parsing.	Data Augmentation	2025	*ACL	link
AugGPT: Leveraging ChatGPT for Text Data Augmentation	(Dai et al., 2025)	Propose AugGPT for text data augmentation, rephrasing training samples.	Data Augmentation	2025	IEEE	link
PGA-SciRE: Harnessing LLM on Data Augmentation for Enhancing Scientific Relation Extraction	(Zhou et al., 2024b)	Propose PGA framework for RE in scientific domain, two data aug. ways.	Data Augmentation	2024	arxiv	link
Improving Topic Relevance Model by Mix-structured Summarization and LLM-based Data Augmentation	(Liu et al., 2024d)	Use query/doc summaries & LLM data augmentation for topic relevance modeling.	Data Augmentation	2024	arxiv	link
Retrieval-Augmented Data Augmentation for Low-Resource Domain Tasks	(Seo et al., 2024)	Propose RADA framework to augment data for low - resource domain tasks.	Data Augmentation	2024	arxiv	link
The Applicability of LLMs in Generating Textual Samples for Analysis of Imbalanced Datasets	(Gopali et al., 2024)	The paper compares approaches for handling text data class imbalance.	Data Augmentation	2024	IEEE	link
Self-Rewarding Language Models	(Yuan et al., 2024c)	Study self - rewarding LMs, use LLM - as - a - Judge for self - rewards during training.	Self Evolution	2024	NIPS / ICML / ICLR	link
Self-Play Fine-Tuning Converts Weak Language Models to Strong Language Models	(Chen et al., 2024e)	Propose SPIN method for LLM, self - play mechanism refines its own capabilities.	Self Evolution	2024	NIPS / ICML / ICLR	link

Continued on next page

Table 2 – Continued

Title	Citation	TLDR	Category	Year	Venue	Link
Self-Boosting Large Language Models with Synthetic Preference Data	(Dong et al., 2024)	SynPO self - boosts LLMs via synthetic preference data, eliminating large - scale annotation.	Self Evolution	2024	arxiv	link
MEMORYLLM: Towards Self-Updatable Large Language Models	(Wang et al., 2024i)	MEMORYLLM is self - updatable, can integrate new knowledge and retain long - term info.	Self Evolution	2024	NIPS / ICML / ICLR	link
Self-Refine: Iterative Refinement with Self-Feedback	(Madaan et al., 2024)	Self - Refine iteratively refines LLM outputs without extra training data or RL.	Self Evolution	2023	NIPS / ICML / ICLR	link
META-REWARDING LANGUAGE MODELS: Self-Improving Alignment with LLM-as-a-Meta-Judge	(Wu et al., 2024)	Introduce Meta - Rewarding step for self - improving LLMs' judgment skills.	Self Evolution	2024	arxiv	link
Automated Proof Generation for Rust Code via Self-Evolution	(Chen and Mueller, 2024)	SAFE framework enables Rust code proof generation via self - evolving cycle.	Self Evolution	2025	NIPS / ICML / ICLR	link
Arxiv Copilot: A Self-Evolving and Efficient LLM System for Personalized Academic Assistance	(Lin et al., 2024a)	Arxiv Copilot is a self - evolving LLM system for personalized academic assistance.	Self Evolution	2024	*ACL	link
Automatic programming via large language models with population self-evolution for dynamic job shop scheduling problem	(Huang et al., 2024e)	This paper proposes SeEvo method for HDRs design inspired by experts' strategies.	Self Evolution	2024	arxiv	link
Benchmark Self-Evolving: A Multi-Agent Framework for Dynamic LLM Evaluation	(Wang et al., 2024e)	A multi - agent framework for dynamic LLM evaluation through instance reframing.	Self Evolution	2025	*ACL	link
Bias Amplification in Language Model Evolution: An Iterated Learning Perspective	(Ren et al.)	Draws parallels between LLM behavior & human culture evolution via Iterated Learning.	Self Evolution	2024	NIPS / ICML / ICLR	link
Enhanced Fine-Tuning of Lightweight Domain-Specific Q&A Model Based on Large Language Models	(Zhang et al., 2024c)	Propose Self - Evolution framework for lightweight LLM fine - tuning.	Self Evolution	2024	IEEE	link
Interactive Evolution: A Neural-Symbolic Self-Training Framework For Large Language Models	(Xu et al., 2024a)	Propose ENVISIONS to self - train LLMs in neural - symbolic scenarios, overcoming two challenges.	Self Evolution	2024	arxiv	link
I-SHEEP: Self-Alignment of LLM from Scratch through an Iterative Self-Enhancement Paradigm	(Liang et al., 2024b)	I - SHEEP paradigm enables LLMs to self - improve iteratively in low - resource scenarios.	Self Evolution	2024	arxiv	link
Language Models as Continuous Self-Evolving Data Engineers	(Wang et al., 2024c)	Propose LANCE for LLMs to self - train by auto - data operations, reducing post - training cost.	Self Evolution	2024	arxiv	link
LLM Guided Evolution - The Automation of Models Advancing Models	(Morris et al., 2024)	GE uses LLMs to directly modify code for model evolution.	Self Evolution	2024	*ACL	link
LLM-Evolve: Evaluation for LLM's Evolving Capability on Benchmarks	(You et al., 2024)	Proposes LLM - Evolve framework to evaluate LLMs' evolving ability on benchmarks.	Self Evolution	2024	*ACL	link
Long Term Memory : The Foundation of AI Self-Evolution	(Jiang et al., 2024)	This paper explores AI self - evolution with LTM, not on experimental performance.	Self Evolution	2024	arxiv	link
METEOR: Evolutionary Journey of Large Language Models from Guidance to Self-Growth	(Li et al., 2024c)	Propose Meteor method for model evolution with 3 training phases to maximize domain capabilities.	Self Evolution, Distillation	2024	arxiv	link
Promptbreeder: Self-referential self-improvement via prompt evolution	(Fernando et al., 2023)	Promptbreeder self - improves prompts via self - referential evolution.	Self Evolution	2024	NIPS / ICML / ICLR	link
rStar-Math: Small LLMs Can Master Math Reasoning with Self-Evolved Deep Thinking	(Guan et al., 2025)	rStar - Math uses deep thinking via MCTS for SLMs to master math reasoning.	Self Evolution	2025	arxiv	link
Self: Language-driven self-evolution for large language model	(Lu et al., 2023a)	SELF enables LLMs to self - evolve without human intervention via language feedback.	Self Evolution	2024	NIPS / ICML / ICLR	link

Continued on next page

Table 2 – Continued

Title	Citation	TLDR	Category	Year	Venue	Link
Self-Evolution Fine-Tuning for Policy Optimization	(Chen et al., 2024e)	SEFT for policy optimization eliminates need for annotated samples.	Self Evolution	2024	*ACL	link
Self-Evolutionary Group-wise Log Parsing Based on Large Language Model	(Pei et al., 2024)	SelfLog self - evolves by LLM - extracted similar pairs and uses N - Gram - based methods.	Self Evolution	2024	IEEE	link
Self-Evolutionary Large Language Models through Uncertainty-Enhanced Preference Optimization	(Wang et al., 2024f)	UPO framework mitigates noisy pref data for LLM self - evolution via reliable feedback.	Self Evolution	2024	arxiv	link
Self-Evolved Reward Learning for LLMs	(Huang et al., 2024b)	Self - Evolved Reward Learning (SER) iteratively improves RM with self - generated data.	Self Evolution	2025	NIPS / ICML / ICLR	link
AugmenToxic: Leveraging Reinforcement Learning to Optimize LLM Instruction Fine-Tuning for Data Augmentation to Enhance Toxicity Detection	(Bodaghi et al., 2024)	Propose RL - based method for LLM fine - tuning to augment toxic language data.	Toxicity / Trust-worthy	2024	ACM	link
Benchmarking LLMs in Political Content Text-Annotation: Proof-of-Concept with Toxicity and Incivility Data	(González-Bustamante, 2024)	Benchmarked LLMs in political text -annotation, not focusing on exp. performance.	Toxicity / Trust-worthy	2024	arxiv	link
Can LLMs Recognize Toxicity? A Structured Investigation Framework and Toxicity Metric	(Koh et al., 2024)	Introduce LLM - based toxicity metric, analyze factors, evaluate its performance.	Toxicity / Trust-worthy	2024	*ACL	link
Characterizing Large Language Model Geometry Helps Solve Toxicity Detection and Generation	(Balestriero et al.)	The paper uses geometry to understand LLMs and solve toxicity - related issues.	Toxicity / Trust-worthy	2024	NIPS / ICML / ICLR	link
Detectors for Safe and Reliable LLMs: Implementations, Uses, and Limitations	(Achintalwar et al., 2024)	Paper presents detectors library for LLM harms, uses & challenges, not exp perf.	Toxicity / Trust-worthy	2024	arxiv	link
Do-Not-Answer: A Dataset for Evaluating Safeguards in LLMs	(Wang et al., 2023b)	This paper creates an open - source dataset to evaluate LLM safeguards.	Toxicity / Trust-worthy	2023	arxiv	link
Effcient Toxic Content Detection by Bootstrapping and Distilling Large Language Models	(Zhang et al., 2024a)	BD - LLM bootstraps & distills LLMs for toxic content detection via DToT.	Toxicity / Trust-worthy	2024	AAAI/IJCAL	link
Evaluating the Impact of Model Size on Toxicity and Stereotyping in Generative LLM	(Chetnani, 2023)	Explore LLM size's relation to toxicity & stereotyping, smallest model performs best.	Toxicity / Trust-worthy	2023	Journal	link
How Toxic Can You Get? Search-based Toxicity Testing for Large Language Models	(Corbo et al., 2025)	EvoTox tests LLM toxicity post - alignment via iterative evolution strategy.	Toxicity / Trust-worthy	2025	arxiv	link
Improving Covert Toxicity Detection by Retrieving and Generating References	(Lee et al., 2024a)	This paper explores refs' potential for covert toxicity detection.	Toxicity / Trust-worthy	2024	*ACL	link
Leak, Cheat, Repeat: Data Contamination and Evaluation Malpractices in Closed-Source LLMs	(Balloccu et al., 2024)	The paper analyzes data contamination & eval malpractices in closed - source LLMs.	Toxicity / Trust-worthy	2024	*ACL	link
LLM-Based Synthetic Datasets: Applications and Limitations in Toxicity Detection	(Schmidhuber and Kruschwitz, 2024)	The paper explores LLM - based synthetic data in toxicity detection, its potential and limits.	Toxicity / Trust-worthy	2024	*ACL	link
Mitigating Biases to Embrace Diversity: A Comprehensive Annotation Benchmark for Toxic Language	(Hou, 2024)	New annotation benchmark reduces bias, shows LLM annotation value.	Toxicity / Trust-worthy	2024	*ACL	link
People Make Better Edits: Measuring the Efficacy of LLM-Generated Counterfactually Augmented Data for Harmful Language Detection	(Sen et al., 2023)	Assess if CAD generation for harmful lang. detection can be automated using NLP models.	Toxicity / Trust-worthy	2023	*ACL	link
Realistic Evaluation of Toxicity in Large Language Models	(Luong et al., 2024)	New TET dataset helps rigorously evaluate toxicity in popular LLMs.	Toxicity / Trust-worthy	2024	*ACL	link

Continued on next page

Table 2 – Continued

Title	Citation	TLDR	Category	Year	Venue	Link
TOXICCHAT: Unveiling Hidden Challenges of Toxicity Detection in Real-World User-AI Conversation	(Lin et al., 2023)	This paper isn't about Efficient LLM Post Training, so can't provide relevant summary.	Toxicity / Trust-worthy	2023	*ACL	link
Toxicity Detection with Generative Prompt-based Inference	(Wang and Chang, 2022)	Explore generative zero - shot prompt - based toxicity detection.	Toxicity / Trust-worthy	2022	arxiv	link
Toxicity in CHATGPT: Analyzing Persona-assigned Language Models	(Deshpande et al., 2023)	The paper evaluates ChatGPT toxicity based on persona - assigned language models.	Toxicity / Trust-worthy	2023	*ACL	link
ToxiCraft:A Novel Framework for Synthetic Generation of Harmful Information	(Hui et al., 2024)	The paper proposes ToxiCraft to generate harmful info datasets, addressing two issues.	Toxicity / Trust-worthy	2024	*ACL	link
TOXIGEN: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection	(Hartvigsen et al., 2022)	Create TOXIGEN dataset, new method for generating text, human evaluation.	Toxicity / Trust-worthy	2022	arxiv	link
Dialectal Toxicity Detection: Evaluating LLM-as-a-Judge Consistency Across Language Varieties	(Faisal et al., 2024)	This paper focuses on dialectal toxicity detection in LLMs, not relevant to efficient post - training.	Toxicity / Trust-worthy, LLM-as-Judger	2024	arxiv	link
Do-Not-Answer: Evaluating Safeguards in LLMs	(Wang et al., 2024j)	The paper curates a dataset to evaluate LLM safeguards for safer deployment.	Toxicity / Trust-worthy	2024	*ACL	link
An Empirical Study of LLM-as-a-Judge for LLM Evaluation: Fine-tuned Judge Model is not a General Substitute for GPT-4	(Huang et al., 2024d)	Fine - tuned judge models have limitations, integrated method improves them.	LLM-as-Judger	2024	*ACL	link
CalibraEval: Calibrating Prediction Distribution to Mitigate Selection Bias in LLMs-as-Judges	(Li et al., 2024a)	CalibraEval mitigates LLM - as - Judges selection bias via NOA.	LLM-as-Judger	2024	arxiv	link
Can LLMs be Good Graph Judger for Knowledge Graph Construction?	(Huang et al., 2024c)	The paper proposes GraphJudger to address KG construction challenges.	LLM-as-Judger	2024	arxiv	link
CodeUltraFeedback: An LLM-as-a-Judge Dataset for Aligning Large Language Models to Coding Preferences	(Weyssow et al., 2024)	Propose LLM - as - a - Judge methodology for evaluating LLM coding preference alignment.	LLM-as-Judger	2024	arxiv	link
Crowd score: A method for the evaluation of jokes using large language model AI voters as judges	(Goes et al., 2022)	Crowd Score method assesses joke funniness via LLMs as AI judges.	LLM-as-Judger	2022	arxiv	link
Foundational Autoraters: Taming Large Language Models for Better Automatic Evaluation	(Vu et al., 2024)	Introduce FLAMe, trained on quality tasks, less biased than other LLM - as - a - Judge models.	LLM-as-Judger	2024	*ACL	link
Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena	(Zheng et al., 2024)	Use LLM - as - a - judge to evaluate chat assistants, verify with two benchmarks.	LLM-as-Judger	2023	NIPS / ICML / ICLR	link
Judgelm: Fine-tuned large language models are scalable judges	(Zhu et al., 2023)	Fine - tune LLMs as scalable judges, propose dataset & techniques.	LLM-as-Judger	2023	arxiv	link
Judging the Judges: Evaluating Alignment and Vulnerabilities in LLMs-as-Judges	(Thakur et al., 2024)	The paper studies LLM - as - judges, judges' performance and vulnerabilities.	LLM-as-Judger	2024	arxiv	link
Large Language Models are Inconsistent and Biased Evaluators	(Stureborg et al., 2024)	LLMs are inconsistent/biased evaluators; recipes to mitigate limitations are shared.	LLM-as-Judger	2024	arxiv	link
Llm-as-a-judge & reward model- What they can and cannot do	(Song et al., 2024a)	Analysis of automated evaluators: English eval & limitations.	LLM-as-Judger	2024	arxiv	link
LLMs instead of Human Judges? A Large Scale Empirical Study across 20 NLP Evaluation Tasks	(Bavaresco et al., 2024)	Evaluated 11 LLMs on 20 datasets; LLMs need human - validation before use as evaluators.	LLM-as-Judger	2024	arxiv	link
Meta-rewarding language models: Self-improving alignment with llm-as-a-meta-judge	(Wu et al., 2024)	Introduce Meta - Rewarding step to self - improve LLM's judgment skills.	LLM-as-Judger	2024	arxiv	link

Continued on next page

Table 2 – Continued

Title	Citation	TLDR	Category	Year	Venue	Link
MLLM-as-a-Judge: Assessing Multimodal LLM-as-a-Judge with Vision-Language Benchmark	(Chen et al., 2024a)	This paper introduces MLLM - as - a - Judge benchmark to assess MLLMs' judging ability.	LLM-as-Judger	2024	NIPS / ICML / ICLR	link
R-Judge: Benchmarking Safety Risk Awareness for LLM Agents	(Yuan et al., 2024b)	R - Judge benchmarks LLM agents' safety risk awareness in interactions.	LLM-as-Judger	2024	arxiv	link
Self-Taught Evaluators	(Wang et al., 2024g)	An approach improves evaluators using only synthetic training data.	LLM-as-Judger	2024	arxiv	link
Style Over Substance: Evaluation Biases for Large Language Models	(Wu and Aji, 2023)	Study shows evaluation bias for LLMs, proposes MERS to improve LLM - based evaluations.	LLM-as-Judger	2025	*ACL	link
Wider and Deeper LLM Networks are Fairer LLM Evaluators	(Zhang et al., 2023c)	The paper uses wider & deeper LLM networks for fairer LLM evaluation.	LLM-as-Judger	2023	arxiv	link
Internal Consistency and Self-Feedback in Large Language Models: A Survey	(Liang et al., 2024a)	This paper uses internal consistency perspective to explain LLM issues and introduce Self - Feedback.	Survey	2024	arxiv	link
A Survey on Self-Evolution of Large Language Models	(Tao et al., 2024)	The paper surveys self - evolution in LLMs, including its process and challenges.	Survey, Self Evolution	2024	arxiv	link
Automatically Correcting Large Language Models: Surveying the Landscape of Diverse Automated Correction Strategies	(Pan et al., 2023)	Reviews advances in auto - correcting LLMs via feedback, categorizes approaches.	Survey	2024	Journal	link
A Survey on Data Selection for LLM Instruction Tuning	(Wang et al., 2024b)	This paper surveys data selection for LLM instruction tuning.	Survey, Data Selection	2024	arxiv	link
Large Language Models for Data Annotation and Synthesis: A Survey	(Tan et al., 2024)	This paper focuses on LLM post - training from a data - centric view.	Survey, Data Synthesis	2024	*ACL	link
On LLMs-Driven Synthetic Data Generation, Curation, and Evaluation: A Survey	(Long et al., 2024)	The paper organizes LLMs - driven data gen. studies to show research gaps and future ways.	Survey	2024	*ACL	link
Trustworthy LLMs: A survey and guideline for evaluating large language models' alignment	(Liu et al., 2023)	The paper surveys LLM trustworthiness dimensions for alignment evaluation.	Survey, Toxicity / Trust-worthy	2024	NIPS / ICML / ICLR	link
A Survey on Data Selection for Language Models	(Albalak et al., 2024)	Comprehensive review of data selection for LMs to accelerate related research.	Survey, Data Selection	2024	Journal	link
LLMs-as-Judges: A Comprehensive Survey on LLM-based Evaluation Methods	(Li et al., 2024b)	I'm sorry, but the given data is about "LLMs - as - Judges" not "Efficient LLM Post Training: A Data - centric Perspective", so I can't provide a relevant summary.	Survey, LLM-as-Judger	2024	arxiv	link
A Survey on Data Synthesis and Augmentation for Large Language Models	(Wang et al., 2024b)	Reviews LLM data generation techniques, discusses constraints.	Survey, Data Synthesis, Data Augmentation	2024	arxiv	link
A Survey on Knowledge Distillation of Large Language Models	(Xu et al., 2024c)	Comprehensive survey on KD in LLMs: mechanisms, skills, verticalization & DA interplay.	Survey, Distillation	2024	arxiv	link
Survey on Knowledge Distillation for Large Language Models: Methods, Evaluation, and Application	(Yang et al., 2024a)	Survey on LLM knowledge distillation methods, evaluation & application, not exp perf.	Survey, Distillation	2024	ACM	link
Impossible Distillation: from Low-Quality Model to High-Quality Dataset & Model for Summarization and Paraphrasing	(Jung et al., 2023)	Impossible Distillation: distill high - quality from low - quality for summarization & paraphrasing.	Distillation	2023	arxiv	link
Prompt Distillation for Efficient LLM-based Recommendation	(Li et al., 2023a)	Propose prompt distillation to bridge IDs & words & reduce inference time.	Distillation	2023	ACM	link
Performance-Guided LLM Knowledge Distillation for Efficient Text Classification at Scale	(Di Palo et al., 2024)	PGKD for text classification, an LLM distillation method with versatile framework.	Distillation	2024	*ACL	link

Continued on next page

Table 2 – Continued

Title	Citation	TLDR	Category	Year	Venue	Link
Knowledge Distillation in Automated Annotation: Supervised Text Classification with LLM-Generated Training Labels	(Pangakis and Wolken, 2024)	The paper tests LLM - generated labels for supervised text classification workflows.	Distillation	2024	*ACL	link
Multistage Collaborative Knowledge Distillation from a Large Language Model for Semi-Supervised Sequence Generation	(Zhao et al., 2024a)	Propose MCKD for semi - supervised seq. gen., iteratively improve pseudolabels.	Distillation	2024	*ACL	link
Self-Data Distillation for Recovering Quality in Pruned Large Language Models	(Thangarasa et al., 2024)	Self - data distillation fine - tuning mitigates quality loss from pruning and SFT.	Distillation	2024	NIPS / ICML / ICLR	link
Distillation Matters: Empowering Sequential Recommenders to Match the Performance of Large Language Models	(Cui et al., 2024)	Proposes DLLM2Rec for LLM-based rec. model distillation to sequential models.	Distillation	2024	ACM	link
Towards Cross-Tokenizer Distillation: the Universal Logit Distillation Loss for LLMs	(Boizard et al., 2024)	Introduce ULD loss for cross - tokenizer distillation in LLMs.	Distillation	2025	Journal	link
Self-Evolution Knowledge Distillation for LLM-based Machine Translation	(Song et al., 2024b)	Self - Evolution KD dynamically integrates prior knowledge for better knowledge transfer.	Distillation, Self Evolution	2025	*ACL	link
Efficiently Distilling LLMs for Edge Applications	(Kundu et al., 2024)	Propose MLFS for parameter - efficient supernet training of LLMs.	Distillation	2024	*ACL	link
Xai-driven knowledge distillation of large language models for efficient deployment on low-resource devices	(Cantini et al., 2024)	DiXtill uses XAI to distill LLM knowledge into a self - explainable student model.	Distillation	2024	Journal	link
Compact Language Models via Pruning and Knowledge Distillation	(Muralidharan et al., 2024)	Develop compression practices for LLMs via pruning and distillation.	Distillation	2024	NIPS / ICML / ICLR	link
LLM-Enhanced Multi-Teacher Knowledge Distillation for Modality-Incomplete Emotion Recognition in Daily Healthcare	(Zhang et al., 2024d)	Propose LLM - enhanced multi - teacher KD for emotion rec in modality - incomplete cases.	Distillation	2024	IEEE	link
BitDistiller: Unleashing the Potential of Sub-4-Bit LLMs via Self-Distillation	(Du et al., 2024)	BitDistiller combines QAT and KD for sub - 4 - bit LLMs with new techniques.	Distillation	2024	*ACL	link
Reducing LLM Hallucination Using Knowledge Distillation: A Case Study with Mistral Large and MMLU Benchmark	(McDonald et al., 2024)	Knowledge distillation reduces LLM hallucination via specific methods.	Distillation	2024	arxiv	link
Distilling Large Language Models for Text-Attributed Graph Learning	(Pan et al., 2024a)	Propose distilling LLMs into local graph model for TAG learning, novel training method.	Distillation	2024	ACM	link
CourseGPT-zh: an Educational Large Language Model Based on Knowledge Distillation Incorporating Prompt Optimization	(Qu et al., 2024)	CourseGPT - zh uses prompt optimization in a distillation framework for educational LLM.	Distillation	2024	arxiv	link
LLMLingua-2: Data Distillation for Efficient and Faithful Task-Agnostic Prompt Compression	(Pan et al., 2024b)	Propose data distillation for prompt compression, formulate as token classification.	Distillation	2024	*ACL	link
Fewer is More: Boosting LLM Reasoning with Reinforced Context Pruning	(Huang et al., 2023)	CoT - Influx maximizes concise CoT examples input to boost LLM math reasoning.	Data Pruning	2024	*ACL	link
LLM for Patient-Trial Matching: Privacy-Aware Data Augmentation Towards Better Performance and Generalizability	(Yuan et al., 2023)	Propose LLM - PTM for patient - trial match, ensure data privacy in methodology.	Application	2023	Others	link
LLM-Assisted Data Augmentation for Chinese Dialogue-Level Dependency Parsing	(Zhang et al., 2024d)	Present 3 LLM - based strategies for Chinese dialogue - level dependency parsing.	Application	2024	Others	link

Continued on next page

Table 2 – Continued

Title	Citation	TLDR	Category	Year	Venue	Link
Resolving the Imbalance Issue in Hierarchical Disciplinary Topic Inference via LLM-based Data Augmentation	(Cai et al., 2023)	Use Llama V1 to augment data for balancing disciplinary topic inference.	Application	2023	IEEE	link
LLM-based Privacy Data Augmentation Guided by Knowledge Distillation with a Distribution Tutor for Medical Text Classification	(Song et al., 2024a)	Propose a DP - based DA method for text classification in private domains.	Application	2024	Others	link
Large Language Models for Healthcare Data Augmentation: An Example on Patient-Trial Matching	(Yuan et al., 2024a)	An LLM - based patient - trial matching approach with privacy - aware data augmentation.	Application	2024	Others	link
Identifying Citizen-Related Issues from Social Media Using LLM-Based Data Augmentation	(dos Santos et al., 2024)	Propose LLM - based method for data augmentation to extract citizen - related data from tweets.	Application	2024	Others	link
Synthetic Data Augmentation Using Large Language Models (LLM): A Case-Study of the Kamyr Digester	(Dhruva et al., 2024)	Introduces LLM - based data augmentation technique for data scarcity.	Application	2024	IEEE	link
Conditional Label Smoothing For LLM-Based Data Augmentation in Medical Text Classification	(Becker et al., 2024)	Propose CLS for data augmentation in medical text classification.	Application	2024	IEEE	link
Curriculum-style Data Augmentation for LLM-based Metaphor Detection	(Jia et al., 2024)	Propose open - source LLM fine - tuning and CDA for metaphor detection.	Application	2024	arxiv	link
Enhancing Speech De-Identification with LLM-Based Data Augmentation	(Dhingra et al., 2024)	A novel data augmentation method for speech de - id using LLM and end - to - end model.	Application	2024	IEEE	link
Enhancing Multilingual Fake News Detection through LLM-Based Data Augmentation	(Chalehchaleh et al., 2024)	Use Llama 3 via LLM - based data augmentation to enrich fake news datasets.	Application	2024	Others	link
LLMs Accelerate Annotation for Medical Information Extraction	(Goel et al., 2023)	Propose LLM - human combo for medical text annotation, reducing human burden.	Application	2023	Others	link
Crowd sourcing with Enhanced Data Quality Assurance: An Efficient Approach to Mitigate Resource Scarcity Challenges in Training Large Language Models for Healthcare	(Barai et al., 2024)	Propose CS framework with quality control for LLM in healthcare, address resource scarcity.	Application	2024	Others	link
LLM2LLM: Boosting LLMs with Novel Iterative Data Enhancement	(Lee et al., 2024b)	LLM2LLM iteratively augments data for LLM fine - tuning in low - data scenarios.	Application	2024	*ACL	link
Data Quality Enhancement on the Basis of Diversity with Large Language Models for Text Classification: Uncovered, Difficult, and Noisy	(Zeng et al., 2024)	Propose DQE method for text classification with LLMs, select data by greedy algorithm.	Application	2025	*ACL	link
Multi-News+: Cost-efficient Dataset Cleansing via LLM-based Data Annotation	(Choi et al., 2024)	Use LLM for data cleansing in Multi - News dataset, no need for costly human annotators.	Application	2024	*ACL	link
LLM-Enhanced Data Management	(Zhou et al., 2024a)	LLMDB for data management: avoid hallucination, reduce cost, improve accuracy.	Application	2024	ACM	link
Enhancing LLM Fine-tuning for Text-to-SQLs by SQL Quality Measurement	(Sarker et al., 2024)	Propose using SQL Quality Measurement to enhance LLM - based Text - to - SQLs performance.	Application	2024	arxiv	link
On The Role of Prompt Construction In Enhancing Efficacy and Efficiency of LLM-Based Tabular Data Generation	(Banday et al., 2024)	Enriching prompts with domain insights improves LLM - based tabular data generation.	Application	2024	arxiv	link
On LLM-Enhanced Mixed-Type Data Imputation with High-Order Message Passing	(Wang et al., 2025)	Propose UnIMP with BiHMP and Xfusion for mixed - type data imputation.	Application	2025	arxiv	link

Continued on next page

Table 2 – Continued

Title	Citation	TLDR	Category	Year	Venue	Link
SEMIEVOL: Semi-supervised Fine-tuning for LLM Adaptation	(Luo et al., 2024a)	SEMIEVOL, a semi - supervised LLM fine - tuning framework, propagates and selects knowledge.	Application	2024	arxiv	link
Curated LLM: Synergy of LLMs and Data Curation for tabular augmentation in low-data regimes	(Seedat et al., 2023)	Introduce CLLM for tabular augmentation in low - data, with curation mechanism for data.	Application	2024	NIPS/ICML/ICLR	link
Data to Defense: The Role of Curation in Customizing LLMs Against Jailbreaking Attacks	(Liu et al., 2024b)	Propose data curation approach & mitigation framework to counter jailbreaking.	Application	2024	arxiv	link
Data Curation Alone Can Stabilize In-context Learning	(Chang and Jia, 2022)	Two methods curate data subsets to stabilize ICL without algorithm changes.	Application	2023	*ACL	link
The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data Only	(Penedo et al., 2023)	Show web data alone can lead to powerful models without curated data.	Application	2023	NIPS/ICML/ICLR	link
Use of a Structured Knowledge Base Enhances Metadata Curation by Large Language Models	(Sundaram et al., 2024)	LLMs can improve metadata curation with a structured knowledge base.	Application	2024	arxiv	link
Source2Synth: Synthetic Data Generation and Curation Grounded in Real Data Sources	(Lupidi et al., 2024)	Source2Synth generates synth data from real sources without human annotations.	Application	2024	arxiv	link
AutoDCWorkflow: LLM-based Data Cleaning Workflow Auto-Generation and Benchmark	(Li et al., 2024d)	Investigated LLM's data - cleaning workflow auto - gen, proposed a benchmark.	Application	2024	arxiv	link
Dynosaur: A Dynamic Growth Paradigm for Instruction-Tuning Data Curation	(Yin et al., 2023)	Dynosaur automatically builds instruction - tuning data, leveraging existing datasets to reduce costs.	Application	2023	*ACL	link
AutoPureData: Automated Filtering of Web Data for LLM Fine-tuning	(Vadlapati, 2024)	Proposes system to auto - filter web data for LLM training with trusted AI models.	Application	2024	arxiv	link
Automatic Dataset Construction (ADC): Sample Collection, Data Curation, and Beyond	(Huang et al., 2024e)	Propose ADC for efficient dataset construction, offer benchmarks.	Application	2024	arxiv	link
Diversify and Conquer: Diversity-Centric Data Selection with Iterative Refinement	(Yu et al., 2024)	Proposes k - means & iterative refinement for data selection to finetune LLMs.	Application	2025	NIPS/ICML/ICLR	link
Increasing Diversity While Maintaining Accuracy: Text Data Generation with Large Language Models and Human Interventions	(Chung et al., 2023)	Explore human - AI partnerships for high - quality LLM - based text data generation.	Application	2023	*ACL	link
Balancing performance and cost of LLMs in a multi-agent framework for BIM data retrieval	(Liu et al., 2025)	Propose MAS method to match queries with LLMs for balanced BIM data retrieval.	Application	2025	Others	link
Optima: Optimizing Effectiveness and Efficiency for LLM-Based Multi-Agent System	(Chen et al., 2024d)	Optima framework in LLM - based MAS improves communication and task effectiveness via LLM training.	Application	2025	NIPS/ICML/ICLR	link
Synergized Data Efficiency and Compression (SEC) Optimization for Large Language Models	(Li et al., 2024e)	Propose SEC for LLMs to enhance efficiency without sacrificing performance.	Application	2024	Others	link
LLMaAA: Making Large Language Models as Active Annotators	(Zhang et al., 2023a)	LLMaAA uses LLMs as annotators in active learning loop, optimizing annotation and training.	Application	2023	*ACL	link
Enhancing Review Classification Via Llm-Based Data Annotation and Multi-Perspective Feature Representation Learning	(Huang et al.)	Propose MJAR dataset and MPFR approach for review classification.	Application	2024	Others	link
AutoLabel: Automated Textual Data Annotation Method Based on Active Learning and Large Language Model	(Ming et al., 2024)	AutoLabel uses LLM and active learning to assist text data annotation.	Application	2024	Others	link

Continued on next page

Table 2 – Continued

Title	Citation	TLDR	Category	Year	Venue	Link
Human-LLM Collaborative Annotation Through Effective Verification of LLM Labels	(Wang et al., 2024h)	A multi - step human - LLM collaborative approach for accurate annotations.	Application	2024	ACM	link
PDFChatAnnotator: A Human-LLM Collaborative Multi-Modal Data Annotation Tool for PDF-Format Catalogs	(Tang et al., 2024b)	PDFChatAnnotator links data & extracts info, user can guide LLM annotations.	Application	2024	ACM	link
Selective Annotation via Data Allocation: These Data Should Be Triaged to Experts for Annotation Rather Than the Model	(Huang et al., 2024a)	Propose SANT for selective annotation, allocating data to expert and model effectively.	Application	2024	*ACL	link
Entity Alignment with Noisy Annotations from Large Language Models	(Chen et al., 2024c)	Propose LLM4EA framework for entity alignment with reduced annotation space and label refiner.	Active Annotation	2024	NIPS/ICML/ICLR	link
CoAnnotating: Uncertainty-Guided Work Allocation between Human and Large Language Models for Data Annotation	(Li et al., 2023c)	The paper proposes CoAnnotating for human - LLM co - annotation using uncertainty.	Active Annotation	2023	*ACL	link
Code Less, Align More: Efficient LLM Fine-tuning for Code Generation with Data Pruning	(Tsai et al., 2024)	Present techniques to enhance code LLM training efficiency with data pruning.	Data Pruning	2024	*ACL	link
LLM-Pruner: On the Structural Pruning of Large Language Models	(Ma et al., 2023)	LLM - Pruner compresses LLMs task - agnostically via structural pruning.	Data Pruning	2023	NIPS/ICML/ICLR	link
Pruning as a Domain-specific LLM Extractor	(Zhang et al., 2024b)	Introduce D - Pruner for domain - specific LLM compression by dual - pruning.	Data Pruning	2024	*ACL	link
Measuring Sample Importance in Data Pruning for Language Models based on Information Entropy	(Kim and Baek, 2024)	Rank training samples by informativeness via entropy for data - pruning of LLMs.	Data Pruning	2024	arxiv	link
P3: A Policy-Driven, Pace-Adaptive, and Diversity-Promoted Framework for data pruning in LLM Training	(Yang et al., 2024c)	P3 optimizes LLM fine - tuning via iterative data pruning with 3 key components.	Data Pruning	2024	NIPS/ICML/ICLR	link
All-in-One Tuning and Structural Pruning for Domain-Specific LLMs	(Lu et al., 2024)	ATP is a unified approach to pruning and fine-tuning LLMs via a trainable generator.	Data Pruning	2024	arxiv	link
Language Model-Driven Data Pruning Enables Efficient Active Learning	(Azeemi et al., 2024)	ActivePrune, a novel pruning strategy for AL, uses LMs to prune unlabeled data.	Data Pruning	2025	NIPS/ICML/ICLR	link
Compresso: Structured Pruning with Collaborative Prompting Learns Compact Large Language Models	(Guo et al., 2023)	Compresso: Structured Pruning via algo - LLM collaboration, uses LoRA and prompt.	Data Pruning	2024	NIPS/ICML/ICLR	link
Efficient LLM Pruning with Global Token-Dependency Awareness and Hardware-Adapted Inference	(Dutta et al., 2024)	Propose VIB - based pruning method, post - pruning for LLMs to compress and speed up.	Data Pruning	2024	Others	link
SlimGPT: Layer-wise Structured Pruning for Large Language Models	(Ling et al., 2024)	SlimGPT, a fast LLM pruning method, uses strategies for near - optimal results.	Data Pruning	2024	NIPS/ICML/ICLR	link
Shortened LLaMA: A Simple Depth Pruning for Large Language Models	(Kim et al., 2024)	Simple depth pruning can compete with width pruning in zero - shot LLM task.	Data Pruning	2024	NIPS/ICML/ICLR	link