

# Knowledge Distillation and Dataset Distillation of Large Language Models: Emerging Trends, Challenges, and Future Directions \*

Luyang Fang<sup>1†</sup>, Xiaowei Yu<sup>2†</sup>, Jiazhang Cai<sup>1</sup>, Yongkai Chen<sup>3</sup>, Shushan Wu<sup>1</sup>, Zhengliang Liu<sup>4</sup>, Zhenyuan Yang<sup>4</sup>, Haoran Lu<sup>1</sup>, Xilin Gong<sup>1</sup>, Yufang Liu<sup>1</sup>, Terry Ma<sup>5</sup>, Wei Ruan<sup>4</sup>, Ali Abbasi<sup>6</sup>, Jing Zhang<sup>2</sup>, Tao Wang<sup>1</sup>, Ehsan Latif<sup>7</sup>, Weihang You<sup>4</sup>, Hanqi Jiang<sup>4</sup>, Wei Liu<sup>8</sup>, Wei Zhang<sup>9</sup>, Soheil Kolouri<sup>6</sup>, Xiaoming Zhai<sup>7</sup>, Dajiang Zhu<sup>2</sup>, Wenxuan Zhong<sup>1\*</sup>, Tianming Liu<sup>4\*</sup>, Ping Ma<sup>1\*</sup>

<sup>1</sup>\*Department of Statistics, University of Georgia, Athens, GA, USA.

<sup>2</sup>Department of Computer Science and Engineering, The University of Texas at Arlington, Arlington, TX, USA.

<sup>3</sup>Department of Statistics, Harvard University, Cambridge, MA, USA.

<sup>4</sup>\*School of Computing, University of Georgia, Athens, GA, USA.

<sup>5</sup>School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA.

<sup>6</sup>Department of Computer Science, Vanderbilt University, Nashville, TN, USA.

<sup>7</sup>AI4STEM Education Center, University of Georgia, Athens, GA, USA.

<sup>8</sup>Department of Radiation Oncology, Mayo Clinic Arizona, Phoenix, AZ, USA.

<sup>9</sup>School of Computer and Cyber Sciences, Augusta University, Augusta, GA, USA.

\*Corresponding author(s). E-mail(s): [wenxuan@uga.edu](mailto:wenxuan@uga.edu); [tliu@uga.edu](mailto:tliu@uga.edu); [pingma@uga.edu](mailto:pingma@uga.edu);

†These authors contributed equally to this work.

## Abstract

The exponential growth of Large Language Models (LLMs) continues to highlight the need for efficient strategies to meet ever-expanding computational and data demands. This survey provides a comprehensive analysis of two complementary paradigms: Knowledge Distillation (KD) and Dataset Distillation (DD), both aimed at compressing LLMs while preserving their advanced reasoning capabilities and linguistic diversity. We first examine key methodologies in KD, such as task-specific alignment, rationale-based training, and multi-teacher frameworks, alongside DD techniques that synthesize compact, high-impact datasets through optimization-based gradient matching, latent space regularization, and generative synthesis. Building on these foundations, we explore how integrating KD and DD can produce more effective and scalable compression strategies. Together, these approaches address persistent challenges in model scalability, architectural heterogeneity, and the preservation of emergent LLM abilities. We further highlight applications across domains such as healthcare and education, where distillation enables efficient deployment without sacrificing performance. Despite substantial progress, open challenges remain in preserving emergent reasoning and linguistic diversity, enabling efficient adaptation to continually evolving teacher models and datasets, and establishing comprehensive evaluation protocols. By synthesizing methodological innovations,

---

\*This version of the article has been accepted for publication after peer review but is not the Version of Record. The Version of Record is available at: <https://doi.org/10.1007/s10462-025-11423-3>.

theoretical foundations, and practical insights, our survey charts a path toward sustainable, resource-efficient LLMs through the tighter integration of KD and DD principles.

**Keywords:** Large Language Models, Knowledge Distillation, Dataset Distillation, Efficiency, Model Compression, Survey

## 1 Introduction

The emergence of Large Language Models (LLMs) like GPT-4 (Brown et al. 2020), DeepSeek (Guo et al. 2025), and LLaMA (Touvron et al. 2023) has transformed natural language processing, enabling unprecedented capabilities in tasks like translation, reasoning, and text generation. Despite these landmark achievements, these advancements come with significant challenges that hinder their practical deployment. First, LLMs demand immense computational resources, often requiring thousands of GPU hours for training and inference, which translates to high energy consumption and environmental costs. Second, their reliance on massive training datasets raises concerns about data efficiency, quality, and sustainability, as public corpora become overutilized and maintaining diverse, high-quality data becomes increasingly difficult (Hadi et al. 2023).

To surmount these challenges, distillation has emerged as a pivotal strategy, integrating Knowledge Distillation (KD) (Hinton et al. 2015) and Dataset Distillation (DD) (Wang et al. 2018), to tackle both model compression and data efficiency. Crucially, the success of KD in LLMs hinges on DD techniques, which enable the creation of compact, information-rich synthetic datasets that encapsulate the diverse and complex knowledge of the teacher LLMs.

KD transfers knowledge from a large, pre-trained *teacher* model to a smaller, more efficient *student* model by aligning outputs or intermediate representations. While effective for moderate-scale teacher models, traditional KD struggles with LLMs due to their vast scale, where knowledge is distributed across billions of parameters and intricate attention patterns. Moreover, the knowledge is not limited to output distributions or intermediate representations but also includes higher-order capabilities such as reasoning ability and complex problem-solving skills (Wilkins and Rodriguez 2024; Zhao et al. 2023; Latif et al. 2024). DD aims to condense large training datasets into compact synthetic datasets that retain the essential information required to train models efficiently. Recent work has shown that DD can significantly reduce the computational burden of LLM training while maintaining performance. For example, DD can distill millions of training samples into a few hundred synthetic examples that preserve task-specific knowledge (Cazenavette et al. 2022; Maekawa et al. 2024). When applied to LLMs, DD acts as a critical enabler for KD: it identifies high-impact training examples that reflect the teacher’s reasoning processes, thereby guiding the student to learn efficiently without overfitting to redundant data (Sorscher et al. 2022).

The scale of LLMs introduces dual challenges: reliance on unsustainable massive datasets (Hadi et al. 2023) and emergent abilities (e.g., chain-of-thought (CoT) reasoning (Wei et al. 2022)) requiring precise and sophisticated knowledge transfer techniques. These challenges necessitate a dual focus on KD and DD. While KD compresses LLMs by transferring knowledge to smaller models, traditional KD alone cannot address the data efficiency crisis: training newer LLMs on redundant or low-quality data yields diminishing returns (Albalak et al. 2024). DD complements KD by curating compact, high-fidelity datasets (e.g., rare reasoning patterns (Li et al. 2024)), as demonstrated in LIMA, where 1,000 examples achieved teacher-level performance (Zhou et al. 2023). This synergy leverages KD’s ability to transfer learned representations and DD’s capacity to generate task-specific synthetic data that mirrors the teacher’s decision boundaries. Together, they address privacy concerns, computational overhead, and data scarcity, enabling smaller models to retain both the efficiency of distillation and the critical capabilities of their larger counterparts.

This survey comprehensively examines KD and DD techniques for LLMs, followed by a discussion of their integration. Traditional KD transfers knowledge from large teacher models to compact students, but modern LLMs’ unprecedented scale introduces challenges like capturing emergent capabilities and preserving embedded knowledge. DD addresses these challenges by synthesizing smaller, high-impact datasets that retain linguistic, semantic, and reasoning diversity for effective training. Our analysis prioritizes standalone advancements in KD and DD while exploring their

combined potential to enhance model compression, training efficiency, and resource-aware deployment. This survey underscores their collective role in overcoming scalability, data scarcity, and computational barriers.

While some prior surveys on KD and DD are available, our survey distinguishes itself from them in several significant respects. This survey, to the best of our knowledge, is the first to place KD and DD under a unifying framework, demonstrating how to jointly utilize them to compact LLMs without losing the capacity for reasoning. Earlier surveys, such as Xu et al. (2024) and Yu et al. (2023), treat the two subjects as separate, thus overlooking their interplays. We also cover newly developed techniques, such as rationale-based KD, uncertainty-aware KD, and generative model-based DD, which are particularly valuable for the field of LLMs. Moreover, we provide a summary of current theoretical guarantees, converting empirical success to explicit conditions for informing practice and future work. To enrich the DD landscape, we discuss modern data-selection methods that predate LLMs but offer useful ideas for distilled datasets. Additionally, we carefully review evaluation protocols that cover reasoning retention, calibration, robustness, memory usage, and compression level, which provide a useful toolkit for benchmarking distilled LLMs and reveal the compression-performance trade-offs in real-world settings. By integrating KD and DD, highlighting advanced methods and theories, incorporating data selection insights, and presenting a comprehensive evaluation toolkit, our survey bridges model- and data-centric views, offering a clear blueprint for building reliable and efficient LLMs.

The subsequent sections explore the following key aspects:

- Fundamentals of KD and DD (Section 2), distinguishing their roles in compressing LLMs and optimizing training efficiency.
- Methodologies for KD in LLMs (Section 3), including rationale-based distillation, uncertainty-aware approaches, multi-teacher frameworks, dynamic/adaptive strategies, and task-specific distillation. Additionally, we review theoretical studies that offer deeper insights into the underlying principles of KD.
- Methodologies for DD in LLMs (Section 4), covering optimization-based distillation, synthetic data generation, and complementary data selection strategies for compact training data.
- Integration of KD and DD (Section 5), presenting unified frameworks that combine KD and DD strategies for enhancing LLMs.
- Evaluation metrics (Section 6) for assessing the effectiveness of distillation in LLMs, focusing on performance retention, computational efficiency, and robustness.
- Applications across multiple domains (Section 7), including medical and health, education, and bioinformatics, demonstrating the practical benefits of distillation in real-world scenarios.
- Challenges and future directions (Section 8), identifying key areas for improvement.

The taxonomy of this survey is illustrated in Figure 1.

## 2 Fundamentals of Distillation

This section introduces the definition and core concepts of Knowledge Distillation (KD) and Dataset Distillation (DD). Table 1 presents a comparative summary of KD and DD. Knowledge distillation has consistently demonstrated high effectiveness across a wide range of benchmarks such as GLUE, SuperGLUE, and MMLU, where student models often retain over 95% of the teacher model’s performance while offering significant efficiency gains. These results establish KD as a reliable approach for compressing large models without substantial loss in accuracy. In contrast, dataset distillation is a more recent technique, with performance evaluations on datasets like ImageNet, AlpacaEval, and GSM8K showing promising results—often achieving 80–90% of the performance obtained using full real datasets. However, DD’s scalability to large-scale language models remains under active investigation. While KD is widely adopted in production for its stability and maturity, DD is gaining traction for scenarios requiring data efficiency, privacy preservation, or decentralized training environments. The following subsections discuss the significance of distillation in LLMs compared to traditional distillation methods.



**Fig. 1:** Taxonomy of Distillation of Large Language Models.

## 2.1 Knowledge Distillation

### 2.1.1 Definition and Core Concepts

KD is a model compression paradigm that transfers knowledge from a computationally intensive teacher model  $f_T$  to a compact student model  $f_S$ . Formally, KD trains  $f_S$  to approximate both the output behavior and intermediate representations of  $f_T$ . The foundational work of Hinton et al. (2015) introduced the concept of soft labels: instead of training on hard labels  $y$ , the student learns from the teacher's class probability distribution  $\mathbf{p}_T = \sigma(\mathbf{z}_T/\tau)$ , where  $\mathbf{z}_T$  are logits from  $f_T$ ,  $\sigma$  is the softmax function, and  $\tau$  is a temperature parameter that controls distribution smoothness. The student's objective combines a cross-entropy loss  $\mathcal{L}_{CE}$  (for hard labels) and a distillation loss  $\mathcal{L}_{KL}$ :

$$\mathcal{L}_{KD} = \alpha \cdot \mathcal{L}_{CE}(\sigma(\mathbf{z}_S(\mathbf{x})), y) + (1 - \alpha) \cdot \tau^2 \cdot \mathcal{L}_{KL}(\sigma(\mathbf{z}_T(\mathbf{x})/\tau), \sigma(\mathbf{z}_S(\mathbf{x})/\tau)), \quad (1)$$

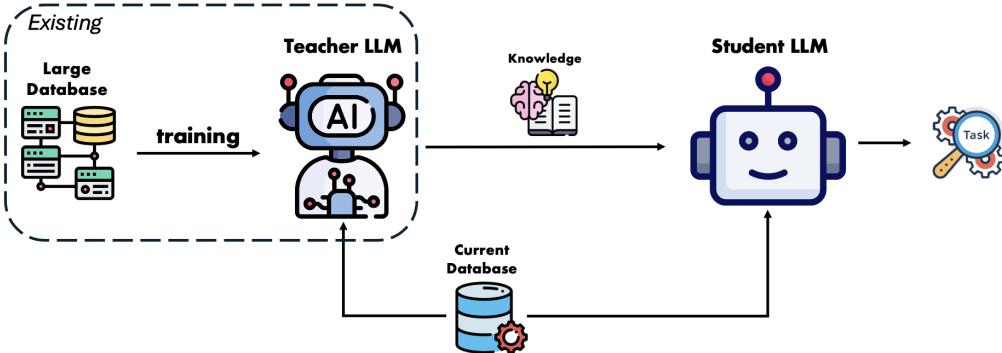
where  $\mathcal{L}_{KL}$  is the Kullback-Leibler (KL) divergence between student and teacher softened outputs, and  $\alpha$  balances the two terms. Beyond logits, later works generalized KD to transfer hidden state activations (Romero et al. 2014), intermediate layers (Sun et al. 2019), attention matrices (Jiao et al. 2019), or relational knowledge (Park et al. 2019), formalized as minimizing distance metrics (e.g.,  $\|\mathbf{h}_T - \mathbf{h}_S\|^2$ ) between teacher and student representations. This framework enables the student to inherit not only task-specific accuracy but also the teacher's generalization patterns, making KD a cornerstone for efficient model deployment.

### 2.1.2 KD in the Era of LLMs vs. Traditional Models

The emergence of LLMs, exemplified by models like GPT-3 (175B parameters (Brown et al. 2020)), has necessitated rethinking traditional KD paradigms. While classical KD usually focuses on compressing task-specific models (e.g., ResNet-50 to MobileNet) with homogeneous architectures (Gou et al. 2021), LLM-driven distillation confronts four fundamental shifts:

**Table 1:** Comparison of Knowledge Distillation and Dataset Distillation.

Aspect	Knowledge Distillation (KD)	Dataset Distillation (DD)
Goal	Transfer knowledge from a large teacher model to a smaller student	Synthesize a small dataset that approximates the training signal
Output	A compact student model	A small, synthetic dataset
Efficiency Gains	Faster inference, reduced deployment cost	Faster training, reduced data storage or communication costs
Data Requirements	Often requires access to original or augmented data with teacher outputs	Can work with minimal or no access to original data
Scalability	Scales well to various downstream tasks and modalities	Scalability to large LLMs still limited and under investigation
Typical Benchmarks	GLUE, SQuAD, MMLU, SuperGLUE, GPQA	ImageNet, AlpacaEval, MT-Bench, MeetingBank, ZeroScrolls, GSM8K
Performance Trends	Distilled models can retain 95% of teacher performance	Synthesized data can match ~80–90% of full data performance on small tasks
Use Cases	Model compression, efficient inference, domain adaptation	Data-efficient training, privacy-aware learning, federated learning
Challenges	Maintaining generalization and robustness of student	Quality of synthetic data, difficulty scaling to high-capacity models



**Fig. 2:** Overview of Knowledge Distillation in LLMs. Knowledge is distilled from a teacher LLM, which has been trained on a large existing database. This knowledge, potentially enriched with current, task-specific data, is transferred to a smaller student LLM. By learning from both the teacher’s guidance and the current data, the student LLM becomes more efficient and effective at performing downstream tasks.

- **Scale-Driven Shifts:** Traditional KD operates on static output distributions (e.g., class probabilities), but autoregressive LLMs generate sequential token distributions over vocabularies of ~50k tokens. This demands novel divergence measures for sequence-level knowledge transfer (Shridhar et al. 2023), such as token-level Kullback-Leibler minimization or dynamic temperature scaling.
- **Architectural Heterogeneity:** Traditional KD often assumed matched or closely related teacher-student topologies (e.g., both CNNs). LLM distillation often bridges architecturally distinct models (e.g., sparse Mixture-of-Experts teachers to dense students (Fedus et al. 2022)). This requires layer remapping strategies (Jiao et al. 2019) and representation alignment (e.g., attention head distillation (Michel et al. 2019)) to bridge topological gaps while preserving generative coherence.
- **Knowledge Localization:** LLMs encode knowledge across deep layer stacks and multi-head attention mechanisms, necessitating distillation strategies that address:
  - *Structural patterns:* Attention head significance (Michel et al. 2019) and layer-specific functional roles (e.g., syntax vs. semantics).
  - *Reasoning trajectories:* Explicit rationales like CoT and implicit latent state progressions.

Unlike traditional model distillation, which often focuses on replicating localized features, LLM distillation must preserve cross-layer dependencies that encode linguistic coherence and logical inference (Sun et al. 2019).

- **Dynamic Adaptation:** LLM distillation increasingly employs iterative protocols where teachers evolve via reinforcement learning from human feedback (RLHF) (Ouyang et al. 2022) or synthetic data augmentation (Taori et al. 2023b), diverging from static teacher assumptions in classical KD.

## 2.2 Dataset Distillation

### 2.2.1 Overview of Dataset Distillation

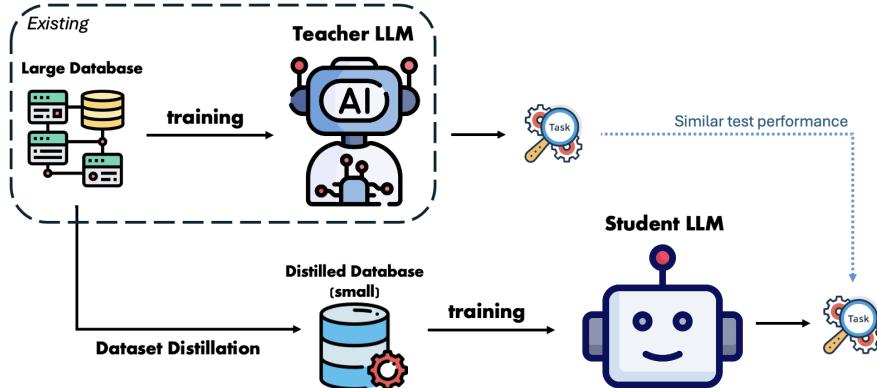
Dataset distillation (Wang et al. 2018) is a technique designed to condense knowledge from large datasets into significantly smaller, synthetic datasets while retaining the ability to train models effectively. Unlike data selection methods (e.g., data pruning or coresnet selection (Dasgupta et al. 2009)), which focus on choosing representative real samples, dataset distillation actively synthesizes new, compact samples that encapsulate the essential learning signal. The distilled dataset is often orders of magnitude smaller yet enables models to achieve comparable or even improved performance.

Formally, let  $\mathcal{D} \triangleq \{(\mathbf{x}_i, y_i)\}_{i=1}^{|\mathcal{D}|}$  be a large dataset, and  $\mathcal{D}_{\text{syn}} \triangleq \{(\tilde{\mathbf{x}}_i, \tilde{y}_i)\}_{i=1}^n$  be the distilled dataset with  $n \ll |\mathcal{D}|$ . For a learning model  $\Phi$ , let  $\theta^{\mathcal{D}}$  and  $\theta^{\mathcal{D}_{\text{syn}}}$  be the parameters learned from training on  $\mathcal{D}$  and  $\mathcal{D}_{\text{syn}}$ , respectively. Dataset distillation aims to make  $\theta^{\mathcal{D}}$  and  $\theta^{\mathcal{D}_{\text{syn}}}$  produce similar outcomes:

$$\arg \min_{\mathcal{D}_{\text{syn}}, n} \left( \sup_{\mathbf{x} \sim \mathcal{X}, y \sim \mathcal{Y}} \{ |l(\Phi_{\theta^{\mathcal{D}}}(\mathbf{x}), y) - l(\Phi_{\theta^{\mathcal{D}_{\text{syn}}}}(\mathbf{x}), y)| \} \right). \quad (2)$$

An  $\epsilon$ -approximate data summary satisfies:

$$\sup_{\mathbf{x} \sim \mathcal{X}, y \sim \mathcal{Y}} \{ |l(\Phi_{\theta^{\mathcal{D}}}(\mathbf{x}), y) - l(\Phi_{\theta^{\mathcal{D}_{\text{syn}}}}(\mathbf{x}), y)| \} \leq \epsilon. \quad (3)$$



**Fig. 3:** Overview of Dataset Distillation in LLMs. A teacher LLM is trained on a massive original database. Through dataset distillation, a compact, high-quality subset (Distilled Database) is synthesized to preserve essential knowledge. This smaller dataset is then used to train a student LLM, aiming to achieve similar performance as the teacher while requiring significantly fewer data.

With the increasing scale of modern deep learning models, dataset distillation has gained attention for accelerating training (Ding et al. 2024), enabling continual learning (Deng and Russakovsky 2022), and improving data efficiency in low-resource settings (Song et al. 2023). However, challenges remain, such as preserving sufficient diversity and robustness in the distilled data and avoiding overfitting (Cazenavette et al. 2023). In LLMs, dataset distillation is crucial for reducing

computational overhead while maintaining the rich semantic diversity needed for effective language modeling. Table 2 outlines some commonly used datasets for data distillation.

**Table 2:** Common Datasets for Data Distillation.

Dataset	Size	Category	Related Works
SVHN	60K	Image	HaBa ( <a href="#">Liu et al. 2022</a> ), FreD ( <a href="#">Shin et al. 2023</a> )
CIFAR-10	60K	Image	FreD ( <a href="#">Shin et al. 2023</a> ), IDC ( <a href="#">Kim et al. 2022</a> )
CIFAR-100	60K	Image	IDM ( <a href="#">Zhao et al. 2023</a> ), DD ( <a href="#">Wang et al. 2018</a> )
TinyImageNet	100K	Image	CMI ( <a href="#">Zhong et al. 2024</a> ), DD ( <a href="#">Wang et al. 2018</a> )
ImageNet-1K	14000K	Image	Teddy ( <a href="#">Yu et al. 2024</a> ), TESLA ( <a href="#">Cui et al. 2023</a> )
TDBench	23 datasets	Table	TDCoLER ( <a href="#">Kang et al. 2025</a> )
Speech Commands	8K	Audio	IDC ( <a href="#">Kim et al. 2022</a> )
AMC AIME STaR	4K	Text	Numinamath ( <a href="#">Li et al. 2024</a> )
R1-Distill-SFT	17K	Text	QWTHN ( <a href="#">Kong et al. 2025</a> )
OpenThoughts-114k	114K	Text	FreeEvalLM ( <a href="#">Zhao et al. 2025</a> )

### 2.2.2 Methods and Approaches

Dataset distillation has evolved through various methodological advancements, each aiming to compress large datasets into small yet highly informative subsets ([Sachdeva and McAuley 2023](#); [Yin and Shen 2024](#); [Yu et al. 2023](#)). From a high-level perspective, two main categories can be distinguished: *optimization-based dataset distillation* and *synthetic data generation for dataset distillation*.

#### *Optimization-Based Methods.*

These methods distill datasets by aligning training dynamics (e.g., gradients, trajectories) or final model performance between synthetic and original data.

- **Meta-Learning Optimization:** A bi-level optimization approach that trains a model on a synthetic dataset and evaluates its performance on the original dataset. The synthetic samples are iteratively refined to match the model's performance when trained on the full dataset ([Wang et al. 2018](#)).
- **Gradient Matching:** Proposed by [Zhao et al. \(2020\)](#), it aligns the gradients from one-step updates on real vs. synthetic datasets. The objective is to make gradients consistent so that training on synthetic data closely resembles training on real data over short time spans.
- **Trajectory Matching:** To address the limitation of single-step gradient matching, multi-step parameter matching (a.k.a. trajectory matching) was introduced by [Cazenavette et al. \(2022\)](#). It aligns the endpoints of training trajectories over multiple steps, making the distilled dataset more robust over longer training horizons.

#### *Synthetic Data Generation.*

These methods directly create artificial data that approximates the distribution of the original dataset. The representative technique here is *Distribution Matching* (DM) ([Zhao and Bilen 2023](#)), aiming to generate synthetic data whose empirical distribution aligns with the real dataset using metrics like Maximum Mean Discrepancy (MMD).

### 2.2.3 Applications and Use Cases

Dataset distillation has wide-ranging applications where data redundancy must be reduced without sacrificing performance:

- **LLM Fine-Tuning and Adaptation:** By creating a distilled subset of domain-specific data, researchers can rapidly adapt large-scale LLMs to specialized tasks without incurring the full computational cost.
- **Low-Resource Learning:** In federated learning and edge AI scenarios ([Wu et al. 2024](#); [Qu et al. 2025](#)), a distilled dataset reduces communication and computational overhead.

- **Neural Architecture Search and Hyperparameter Tuning:** Distilled datasets provide a proxy for evaluating model variants quickly, cutting down on expensive full-dataset training ([Prabhakar et al. 2022](#)).
- **Privacy and Security:** Distilled datasets can serve as privacy-preserving proxies for sensitive data (e.g., medical or financial records), reducing exposure of individual-level information ([Yu et al. 2023](#)).
- **Continual Learning:** By summarizing past tasks, distilled datasets help mitigate catastrophic forgetting when models learn incrementally over time ([Binici et al. 2022](#)).

### 3 Methodologies and Techniques for Knowledge Distillation in LLMs

In this section, we review methodologies for KD in LLMs, including rationale-based approaches, uncertainty-aware techniques, multi-teacher frameworks, dynamic/adaptive strategies, and task-specific distillation. We also explore theoretical studies that uncover the foundational mechanisms driving KD’s success in LLMs.

#### 3.1 Rationale-Based KD

Rationale-Based Knowledge Distillation (RBKD) improves traditional knowledge distillation by allowing the student model to learn not only the teacher’s final predictions but also the reasoning process behind them, like CoT reasoning. This makes the student model more interpretable and reduces the need for large amounts of labeled data ([Hsieh et al. 2023](#)). Instead of merely imitating the teacher’s outputs, the student develops a deeper understanding of problem-solving, leading to better generalization and adaptability to new tasks.

Formally, given a dataset  $\mathcal{D} = \{(x_i, y_i, r_i)\}_{i=1}^n$ , where  $x_i$  is the input,  $y_i$  is the label, and  $r_i$  is the rationale generated by the teacher, the student is trained to jointly predict both the rationale and the final answer. This objective can be formulated as:

$$\mathcal{L}_{\text{RBKD}} = -\frac{1}{n} \sum_{i=1}^n \log P_\theta(r_i, y_i | x_i), \quad (4)$$

where  $P_\theta$  denotes the student model parameterized by  $\theta$ . This formulation encourages the student to internalize the reasoning path rather than shortcircuiting to the answer.

By incorporating reasoning steps, RBKD enhances transparency, making models more reliable in fields like healthcare and law, where understanding decisions is crucial. It also improves efficiency, as smaller models can achieve strong performance without requiring extensive computational resources. This makes RBKD a practical approach that balances accuracy, interpretability, and resource efficiency ([Chu et al. 2023](#)).

One promising direction is Keypoint-based Progressive CoT Distillation (KPOD), which addresses both token significance and the order of learning ([Feng et al. 2024](#)). In KPOD, the difficulty of each reasoning step is quantified using a weighted token generation loss:

$$d_k^{(i)} = - \sum_{j=p_k}^{q_k} \widehat{w}_j^{(i)} \log P(r_j^{(i)} | r_{<j}^{(i)}, x^{(i)}; \theta_s), \quad (5)$$

where  $d_k^{(i)}$  is the difficulty score of the  $k$ -th step in the  $i$ -th rationale,  $p_k$  and  $q_k$  denote the start and end positions of that step, and  $\widehat{w}_j^{(i)}$  represents the normalized importance weight for token  $j$ . This difficulty-aware design allows the student model to acquire reasoning skills progressively, from simpler to more complex steps, resulting in stronger generalization and interoperability.

#### *Challenges and Future Directions.*

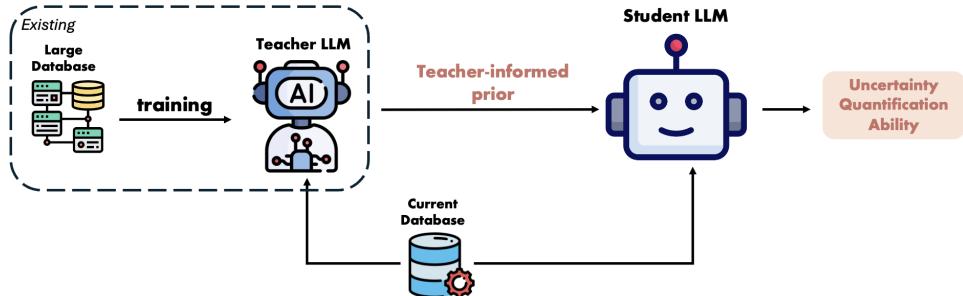
The growing focus on distilling richer knowledge from teacher LLMs, such as reasoning and other higher-order capabilities, raises important questions about which knowledge to extract and how to extract it effectively. Since many teacher LLMs are closed-source, capturing their advanced capabilities is more challenging than merely collecting hard-label predictions. This highlights several specific research gaps that warrant investigation. First, how can we develop standardized metrics

to evaluate the quality of extracted reasoning traces, particularly when ground truth reasoning paths are unavailable? Second, what novel prompting strategies can reliably elicit intermediate reasoning steps from black-box teacher LLMs without access to their internal representations? Third, existing methods primarily focus on single-step reasoning extraction. It is an interesting direction to explore how we can effectively capture and transfer sequential reasoning chains and complex problem decomposition strategies.

### 3.2 Uncertainty-Aware KD

Traditional KD methods mainly focus on matching the predictions or latent representations of the teacher model to improve the generalization of the student model. While effective, this approach overlooks the inherent uncertainty in the teacher’s predictions, which can provide critical insights into noisy samples. To address this limitation, the study of uncertainty in KD has become an active field driven by two primary goals: (1) distilling the uncertainty from probabilistic teachers and (2) quantifying the uncertainty of the distilled models.

In the first scenario, the teacher model generates a predictive distribution, such as the conditional probability measure  $p_T(y|\mathbf{x})$ . When the teacher model is trained as a Bayesian neural network (BNN) (MacKay 1992) or a Monte Carlo ensembles of models,  $p_T(y|\mathbf{x})$  usually represents the empirical distribution derived from Monte Carlo samples. Compared to the poor approximation of variational inference in general neural networks or the large computational complexity of the Monte Carlo approximation, Knowledge distillation provides a general and simple way to train a small model  $p_S(y|\mathbf{x})$  that provides direct estimation of conditional probability  $p_T(y|\mathbf{x})$ . This approach usually minimizes the KL divergence between  $p_S(y|\mathbf{x})$  and  $p_T(y|\mathbf{x})$ . Koratikara Balan et al. (2015) introduces Bayesian Dark Knowledge (BDK) by parameterizing  $p_S(y|\mathbf{x})$  as  $[p_S(y=1|\mathbf{x}), \dots, p_S(y=K|\mathbf{x})]$  for  $K$ -class classification, and as  $\mathcal{N}(y|\mu(\mathbf{x}), \sigma^2(\mathbf{x}))$  for regression, where  $\mathcal{N}$  is the probability density function of the normal distribution with mean  $\mu(\mathbf{x})$  and variance  $\sigma^2(\mathbf{x})$ . The effectiveness is demonstrated in distilling ensembles of teacher models trained using stochastic Langevin Monte Carlo. Subsequent extensions include Vadera et al. (2020), which generalizes BDK by distilling posterior expectations, and Malinin et al. (2019), which proposes Ensemble Distribution Distillation to capture both the ensemble mean and diversity within a single model.



**Fig. 4:** Bayesian Knowledge Distillation. Knowledge distilled from the teacher LLM is incorporated as a teacher-informed prior over the student LLM’s parameters. This leads to a posterior distribution that equips the student LLM with uncertainty quantification capabilities.

In the second scenario, the researchers aim to quantify the uncertainty of the distilled models, avoiding the propagation of incorrect knowledge to the student and reducing the risk of making overconfident predictions. A consensus strategy is to interpret KD within a Bayesian inference framework. Specifically, Fang et al. (2024) proposes a systematic statistical Bayesian framework to interpret the distillation process, with the workflow available in Figure 4. Bayesian Knowledge Distillation (BKD) is introduced to establish the equivalence between KD and a Bayesian model. It defines a proper prior distribution for the student model’s parameter based on the prediction outcomes of the teacher models, referred to as the teacher-informed prior (TIP). For example, in the  $K$ -class classification task, given the predicted class probabilities  $\{\mathbf{p}_i\}_{i=1}^N$ , where

$\mathbf{p}_i = (p_{i1}, \dots, p_{iK})^T$ , by the teacher model  $M_t$ , the prior distribution  $\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}; \{\mathbf{p}_i\}_{i=1}^N)$  on the student model's weights  $\boldsymbol{\theta}$  is defined by,

$$\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}; \{\mathbf{p}_i\}_{i=1}^N) \propto \prod_{i=1}^N \pi_{\mathbf{q}}(\mathbf{q}_i; \mathbf{p}_i), \quad (6)$$

where  $\propto$  denotes proportionality, and  $\pi_{\mathbf{q}}(\mathbf{q}; \mathbf{p})$  is a probability density function defined for a  $K$ -dimensional probability vector  $\mathbf{q} = (q_1, \dots, q_K)^T$  with the parameter  $\mathbf{p}$ . Following the idea of KD, the prior distribution  $\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}; \{\mathbf{p}_i\}_{i=1}^N)$  should assign a larger weight to the parameter  $\boldsymbol{\theta}$  that results in the probability function  $\pi_{\mathbf{q}}(\mathbf{q}_i; \mathbf{p}_i)$  having a high probability around  $\mathbf{p}_i$ . It is shown by taking  $\pi_{\mathbf{q}}(\mathbf{q}; \mathbf{p}_i)$  as the density function of a Dirichlet distribution  $Dir(\mathbf{1}_K + \lambda \mathbf{p}_i)$ , where  $\lambda$  is a tuning parameter controlling the confidence in the predicted probabilities of the teacher model  $M_t$ , the optimization of maximizing the posterior distribution of  $\boldsymbol{\theta}$  is equivalent to solving the objective function of KD.

Another advantage of BKD lies in its straightforward uncertainty quantification, which is achieved via the posterior distribution of the student model's weights. Within this framework, a suite of Bayesian inference tools enables posterior sampling for uncertainty quantification in the student model. For example, the stochastic Gradient Langevin Dynamics (SGLD) ([Welling and Teh 2011](#)) can efficiently sample the student model's weights from the derived posterior distribution. Using these posterior samples, we can compute the expected value of predictive evaluation metrics, including variance in regression or deviance in classification, by computing their posterior mean. This could provide a robust estimate of uncertainty in the population.

#### *Challenges and Future Directions.*

Uncertainty quantification typically incurs significantly higher computational costs than standard prediction tasks, particularly when relying on large-scale sampling or Monte Carlo methods. In contrast, variational inference has demonstrated strong performance as an efficient alternative to sampling-based methods in various settings ([Blei et al. 2017](#)). A promising yet underexplored direction is the systematic integration of variational inference into knowledge distillation frameworks to balance accuracy and computational efficiency. Meanwhile, the Bayesian framework for knowledge distillation opens up rich possibilities for advancing uncertainty quantification in this domain. Key future directions include (1) designing more flexible teacher-informed priors that incorporate statistical justification and (2) developing scenario-specific adaptations that account for different data distributions or task requirements.

These research directions would help solidify the theoretical foundations of uncertainty-aware knowledge distillation while enhancing its practical applicability.

### 3.3 Multi-Teacher KD

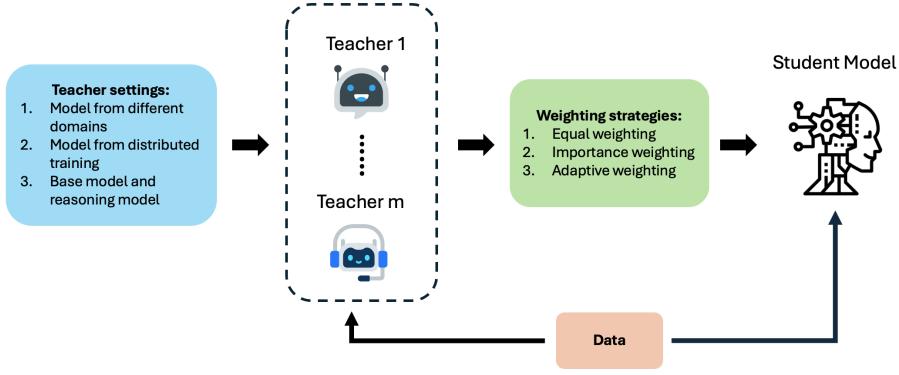
Multi-teacher knowledge distillation offers a powerful extension to the single-teacher paradigm by consolidating heterogeneous expertise from multiple teacher models into a single student. Conceptually, it aims to utilize the diversity of teacher networks to yield richer supervision and improved generalization. The key challenge is to fuse potentially conflicting signals from distinct teachers in a principled manner. Formally, suppose we have  $K$  trained teacher models  $T_k$  (where  $k \in \{1, \dots, K\}$ ). For a given input  $\mathbf{x}$ , each teacher produces a predictive distribution  $\mathbf{p}_k(\mathbf{x})$ . A simple ensemble strategy averages over these distributions:

$$\widehat{\mathbf{p}}(\mathbf{x}) = \frac{1}{K} \sum_{k=1}^K \mathbf{p}_k(\mathbf{x}). \quad (7)$$

Then, the student model  $S$  with parameters  $\theta$  can be updated by minimizing a distillation loss, for example KL divergence, between  $\widehat{\mathbf{p}}(\mathbf{x})$  and the student's output  $\mathbf{p}_S(\mathbf{x}; \theta)$ :

$$\mathcal{L} = D_{KL}(\widehat{\mathbf{p}}(\mathbf{x}) \parallel \mathbf{p}_S(\mathbf{x}; \theta)), \quad (8)$$

where  $D_{KL}(a \parallel b)$  denotes the KL divergence of  $a$  and  $b$ . While this uniform average provides an effective first step, more sophisticated methods address the variability in teacher quality and teacher



**Fig. 5:** Multi-Teacher Knowledge Distillation. The variants of Multi-Teacher KD include different weighting strategies and teacher settings.

conflicts. Some refined approaches take different weighting strategies for the teachers. For example, [Zhang et al. \(2022\)](#) assigns large weights to teacher predictions that closely match the ground truth, thus mitigating the adverse influence of unhelpful teachers. On the other hand, adaptive weighting has also been investigated to ensure a more dynamic synthesis of teacher signals. [Du et al. \(2020\)](#) explored an adaptive strategy that merges multi-teacher outputs in gradient space, making it possible to resolve conflicts without discarding the diversity they offer.

Further, instead of using a simple loss, such as the mean squared error, [You et al. \(2017\)](#) proposed aligning the student’s feature embeddings with those of the teachers via a distance metric, capturing complementary knowledge.

Recent research has focused on using different teacher settings in multi-teacher KD to enhance specific skills. [Tian et al. \(2024\)](#) proposed TinyLLM, a multi-teacher KD method that transfers reasoning skills from large LLMs to smaller ones by not only focusing on correct answers but also capturing the underlying rationales. TinyLLM uses an in-context example generator and a teacher-forcing CoT strategy to ensure reasoning accuracy and diversity of knowledge sources. The method demonstrated significant performance improvements across reasoning tasks by effectively synthesizing the diverse problem-solving strategies of different teachers. [Khanuja et al. \(2021\)](#) introduced MERGEDISTILL, which merges pre-trained multilingual and monolingual LLMs into a single task-agnostic student model. The framework employs offline teacher evaluation and vocabulary mapping to integrate knowledge from diverse models, enabling the student model to benefit from both multilinguality and the specialized knowledge of individual languages. MERGEDISTILL’s approach is particularly effective in scenarios with overlapping language sets, showing that leveraging diverse teacher models can create a robust and versatile student model. [Liu et al. \(2024\)](#) developed DIVERSEDISTILL, a framework that addresses the challenge of teacher-student heterogeneity by introducing a teaching committee comprising various teachers. The method dynamically weights teacher contributions based on the student’s understanding of each teacher’s expertise, which helps in scenarios with significant architectural or distributional differences among the teachers. The method is highly adaptable and has demonstrated improved performance in both vision and recommendation tasks. Lastly, other than focusing on the training process, [Wadhwa et al. \(2025\)](#) explored an interesting concept of ‘teacher footprints’ in distilled models, proposing methods to identify which teacher model was used to train a student LLM. Their work emphasizes the importance of understanding teacher influence, particularly when using proprietary models, and developed discriminative models based on syntactic patterns to trace teacher origins.

#### ***Challenges and Future Directions.***

Multi-teacher KD for LLMs introduces new complexities beyond the traditional single-teacher paradigm. On one hand, integrating expertise from multiple teachers can greatly enrich a student’s generalization and domain coverage, especially when those teachers specialize in different areas (e.g., biomedical vs. legal). However, practical constraints arise in handling teacher availability: many approaches assume full access to all teacher models, which is often unfeasible due to licensing, privacy, or asynchronous deployment. Moreover, distilling from multiple teachers incurs

high computational overhead, pairwise interactions among  $K$  teachers scale roughly as  $O(K^2)$ , making it hard to synchronize their distinct objectives and representations. Aligning knowledge across teachers with heterogeneous architectures or tokenization schemes also demands careful design to reconcile discrepancies. A straightforward research step to probe these issues is to evaluate multi-teacher students on domain-specific benchmarks, ensuring that each teacher’s expertise is retained. For instance, if one teacher excels in biomedical QA and another in legal reasoning, the student could be tested on a biomedical QA dataset and a legal reasoning dataset to verify it meets or exceeds each teacher’s performance in-domain. Such experiments would reveal whether multi-teacher distillation truly broadens the student’s skill set or whether some knowledge gets diluted.

Another key challenge is managing conflicts when teachers produce divergent or contradictory outputs. Without careful resolution, the student may either homogenize the teachers’ errors or fail to benefit from each teacher’s unique perspective. While various weighting and ensemble strategies attempt to reconcile teacher signals, they remain limited in open-ended LLM scenarios where disagreements can be subtle or domain-specific. An important open question is how a student can detect and resolve contradictory guidance from different teachers. One possible experimental setup to explore this would be to inject intentionally conflicting answers from teachers for certain training queries and then evaluate different conflict-resolution strategies (e.g., confidence-based weighting, context-dependent teacher selection) on held-out test cases to see which yields the most accurate and consistent student responses. In addition, tracing or quantifying each teacher’s individual contribution to the student is non-trivial, especially in collaborative, multi-domain settings where teachers’ knowledge areas overlap or their stylistic preferences blur together. Recent work by [Wadhwa et al. \(2025\)](#) takes an initial step toward this by identifying teacher footprints in the student, but we lack general methods to robustly attribute portions of the student’s knowledge to specific teachers. A concrete research question here is whether we can develop interpretability tools or influence-function analyses to trace knowledge transfer: for example, performing leave-one-teacher-out ablation studies (training students while withholding one teacher at a time) and measuring performance drops in that teacher’s specialty could quantify its influence. Developing such diagnostics would not only strengthen trust in distilled models but also guide the design of better fusion mechanisms. Ultimately, creating lightweight yet robust methods to fuse multi-teacher signals, resolve domain conflicts, and accommodate partial teacher availability stands as a central objective for advancing multi-teacher KD in LLMs. By formulating explicit evaluation protocols, such as cross-domain performance checks and contradiction-handling tests, future research can make multi-teacher distillation more principled and actionable.

### 3.4 Dynamic and Adaptive KD

Previous approaches rely on a static, pre-trained teacher LLM to transfer knowledge to a student LLM. Dynamic and adaptive approaches introduce a paradigm shift by fostering a bidirectional collaboration between models. These frameworks emphasize two key strategies: (1) simultaneous co-evolutionary training, where teacher and student are both refined during joint optimization, and (2) self-distillation, which bypasses the need for a pre-trained teacher entirely by enabling a single model to generate and distill its own knowledge. By departing from fixed hierarchies, such methods enhance adaptability, mitigate the limitations of static teacher expertise, and improve generalization in evolving or resource-constrained scenarios.

#### 3.4.1 Simultaneous Training of Teacher and Student model

Simultaneous training of teacher and student models in knowledge distillation is an approach where both models are optimized together during the training process, rather than using a pre-trained teacher model to guide the student passively ([Sun et al. 2021; Chang et al. 2022](#)). This approach allows for dynamic interaction, where the student continuously refines its learning based on real-time feedback from the teacher, and the teacher can adapt its guidance based on the student’s progress.

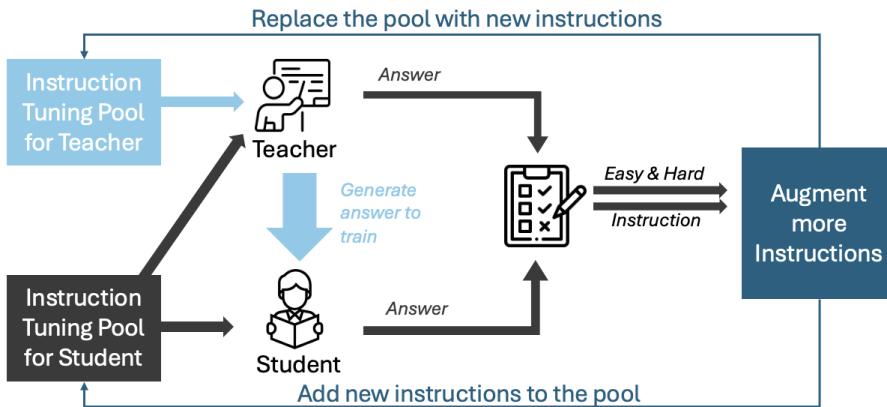
Some methods incorporate constraints in the loss function to facilitate joint learning of the student and teacher models. For instance, [Li et al. \(2024\)](#) proposed a new loss function named Bi-directional Logits Difference. Instead of directly aligning their logits, it computes pairwise differences between the top-k logits for both models. This results in two types of differences: the

teacher-led logits difference (t-LD) loss, which captures key knowledge from the teacher’s top-k logits, and the student-led logits difference (s-LD) loss, which forces the teacher to consider the student’s perspective. The BiLD loss is given by

$$L_{\text{BiLD}} = D_{\text{KL}}(\mathbf{p}_{t-LD} \parallel \mathbf{p}_{s-LD}) + D_{\text{KL}}(\mathbf{p}_{s-LD} \parallel \mathbf{p}_{t-LD}), \quad (9)$$

where  $D_{\text{KL}}(a \parallel b)$  denotes the KL divergence of  $a$  and  $b$ ,  $\mathbf{p}_a$  denotes the probability distribution of  $a$ .

Other methods employ specially designed training frameworks to enable joint learning. Li et al. (2023) introduced a Competitive Multi-modal Distillation framework that introduces a bidirectional feedback loop through multi-modal competitive distillation, the overall framework is shown in Figure 6. It consists of three key phases: instruction tuning, instruction evaluation, instruction augmentation. In the first phase, the teacher model is first tuned using its designated instructions before generating responses for questions from the instruction tuning pool. The student model is then trained to align its responses with the teacher’s outputs. In the evaluation phase, both the teacher and student models generate responses for the instructions, which are then evaluated by an assessor model. Finally, in the augmentation phase, new instructions are generated, then replace or add to the original instruction pool.



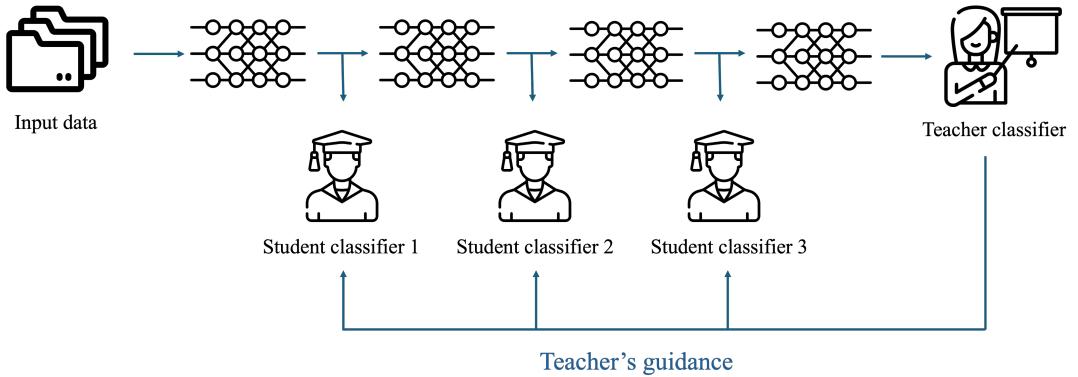
**Fig. 6:** Competitive Multi-modal Distillation Framework for Joint Training of Teacher and Student Models. It includes instruction tuning, evaluation, and augmentation phases, with a bidirectional feedback loop to iteratively improve both models. The assessor model guides instruction refinement based on performance comparison.

### 3.4.2 Self-distillation

In self-distillation, instead of using a separate teacher model to guide students, the model is trained so that the teacher and students learn simultaneously. The overall framework is shown in Figure 7. The architecture is divided into multiple blocks, with intermediate outputs extracted from earlier blocks. These outputs pass through an attention module and are fed into a shallow classifier, which makes predictions based on only the initial blocks. In contrast, the final classifier utilizes all blocks before making predictions. Here, the final classifier acts as the “teacher”, while the shallow classifiers serve as the “students” (Zhang et al. 2019).

The teacher’s guidance in self-distillation is enforced through a carefully designed loss function composed of three components. The first is the cross-entropy loss between the true labels and each classifier’s predictions. The second measures the KL divergence between the softmax output probabilities of each shallow (student) classifier and the final (teacher) classifier. The third imposes an L2-norm penalty on the difference between the features fed into each shallow classifier and those used by the final classifier.

In addition to this three-part loss function, various loss function designs capture the relationships between students and teachers. For instance, rather than computing KL divergence solely



**Fig. 7:** Framework for Self-Distillation. The model is partitioned into sequential sections, each followed by a shallow student classifier. The deepest classifier acts as the teacher, guiding earlier classifiers through knowledge distillation.

between the final classifier and each shallow classifier, it can be computed iteratively: first between the final classifier and its predecessor, then progressively between successive classifiers moving backward through the network. This approach is known as transitive teacher distillation. Alternatively, instead of measuring the KL divergence between each shallow classifier and the final classifier, it can be computed between each classifier’s output and the ensemble prediction of all classifiers, a technique called ensemble teacher distillation (Zhang et al. 2022). Furthermore, Allen-Zhu and Li (2020) provides a theoretical guarantee that self-distillation with an ensemble of independently trained neural networks can improve test accuracy when data exhibits a multi-view structure, reinforcing the understanding that the dark knowledge embedded in ensemble outputs contributes to model generalization.

In the development of AlphaFold, self-knowledge distillation played a pivotal role in enhancing the model’s accuracy. Initially, AlphaFold was trained using known protein structures from the Protein Data Bank. Subsequently, the model predicted structures for approximately 300,000 diverse protein sequences obtained from Uniclust. These predictions were then incorporated back into the training process, allowing the model to learn from its own predictions. This approach effectively utilized unlabeled sequence data, leading to significant improvements in the model’s performance (Jumper et al. 2021).

Self-distillation technique is particularly valuable as it enables the model to leverage vast amounts of unlabeled data, overcoming the limitations imposed by the availability of labeled training data. By refining its predictions through iterative learning from its own outputs, AlphaFold achieved unprecedented accuracy in protein structure prediction (Yang et al. 2023).

### 3.4.3 Challenges and Future Directions

While dynamic and adaptive KD enables flexible collaboration between models, practical challenges remain. First, simultaneous training of teacher and student models risks unstable convergence due to interdependent learning dynamics, requiring carefully designed loss functions (e.g., bidirectional feedback) to prevent conflicting objectives. Second, self-distillation avoids reliance on external teachers but can amplify inherent model biases or errors when its guidance lacks calibration. It is necessary to develop automatic bias detectors that flag flawed knowledge and enable self-correction through adversarial questioning or consistency checks across multiple inference paths. Another promising direction is to add minimal external supervision, such as human-annotated bias signals and uncertainty-aware loss functions, to recalibrate guidance without sacrificing efficiency. Third, real-time interactions between models significantly increase computational overhead, posing greater demands on computational resources. Future research may focus on adaptive training protocols that dynamically balance knowledge exchange, leveraging lightweight feedback or noise-robust distillation strategies. Iterative self-correction using both labeled and unlabeled data, as exemplified by AlphaFold, may further improve reliability. Enhancing computational efficiency through modular architectures or sparse interaction mechanisms could increase scalability. Finally,

developing interpretable metrics to quantify knowledge transfer would help build trust in dynamic KD systems.

Some existing work has proposed potential solutions. SCKD (Wang et al. 2024) and ISD-QA (Ramamurthy and Aakur 2022) use iterative self-correction with unlabeled data to enhance robustness. SparseBERT (Shi et al. 2021) improve scalability through modular or sparse designs. Proto2Proto (Keswani et al. 2022) and SSCBM (Hu et al. 2025) develop interpretable metrics to quantify and visualize knowledge transfer. While these methods offer valuable insights and techniques, they have yet to be extensively explored or applied to LLMs, highlighting an important direction for future research.

### 3.5 Vision KD for Autonomous Driving

Recent advancements in Large Vision-Language Models (VLMs) have demonstrated exceptional capabilities in semantic scene understanding, causal reasoning, and open-vocabulary object detection. In the domain of Autonomous Driving (AD), these capabilities are crucial for handling complex, long-tail scenarios that traditional geometric planners fail to address (e.g., interpreting hand gestures, navigating construction zones, or reasoning about occluded agents). However, the high computational latency and memory footprint of billion-parameter VLMs prohibit their direct deployment on resource-constrained vehicle edge devices. Consequently, a new paradigm of *Vision-Language Distillation* has emerged to transfer the reasoning capabilities of VLMs into lightweight, real-time driving policies.

Unlike traditional KD which typically focuses on logit matching, vision distillation in AD requires transferring spatial-temporal awareness and causal logic. We categorize these methodologies into three primary strategies: Rationale-Based Feature Alignment, Surrogate Task Distillation, and Contrastive Planning.

**Table 3:** Comparison of distillation methodologies transferring knowledge from VLMs to autonomous driving agents.

Method	Teacher Model	Student Model	Knowledge Type	Distillation Mechanism
VERDI (Feng et al. 2025)	VLM (e.g., Qwen-VL)	End-to-End Planner	Causal Rationales	Latent Space Feature Alignment
DiMA (Hegde et al. 2025)	Multi-modal LLM	Vision Planner	World Knowledge	Joint Training w/ Surrogate Tasks
VLP (Pan et al. 2024)	Large Language Model	BEV Backbone	Semantic Prototypes	Agent-Centric Contrastive Learning
Vi-LAD (Elnoor et al. 2025)	VLM	Navigation Policy	Social Attention	SSIM Loss on Attention Maps
DSDrive (Liu et al. 2025)	VLM (CoT)	Compact LLM	Reasoning + Waypoints	Dual-Head Coordination Loss
Lingo-1 (Wayve 2023)	Expert Driver + LLM	VLA Model	Explanations	Action-Commentary Generation

#### 3.5.1 Rationale-Based Feature Alignment

Driving decisions are often driven by causal logic (e.g., “I am stopping *because* the pedestrian is distracted”). Feng et al. (2025) introduces a framework, **VERDI** (VLM-Embedded Reasoning for Autonomous Driving), where a VLM teacher generates textual rationales for ground-truth trajectories. These rationales are encoded into a semantic latent space. The student model is trained to align its intermediate feature maps (from perception, prediction, and planning modules) with these linguistic embeddings via a learnable projector  $\mathcal{P}$ . The alignment loss is formulated as:

$$\mathcal{L}_{align} = \sum_{k \in \{per, pred, plan\}} (1 - \cos(\mathcal{P}_k(F_k), E_{text}(R))) \quad (10)$$

where  $F_k$  represents the student’s feature map at stage  $k$ , and  $E_{text}(R)$  is the embedding of the VLM-generated rationale  $R$ . This forces the student to organize its latent space semantically, improving robustness in zero-shot scenarios. Similarly, **DSDrive** (Liu et al. 2025) employs a waypoint-driven dual-head coordination module to synchronize the distilled reasoning with kinematic planning outputs.

#### 3.5.2 Surrogate Task Distillation

To enforce deeper scene understanding, the **DiMA** (Distilling Multi-modal LLMs) framework (Hegde et al. 2025) utilizes a joint training strategy where a shared Scene Encoder serves as a

tokenizer for an MLLM teacher. The framework introduces surrogate tasks such as *Masked Token Reconstruction* and *Future Prediction*. The gradients from the MLLM’s reasoning objectives back-propagate into the shared encoder, enriching the representations used by the lightweight planner. The reconstruction objective is defined as:

$$\mathcal{L}_{recon} = \|\hat{B}_{masked} - B_{gt}\|_2^2 \quad (11)$$

where  $\hat{B}_{masked}$  represents the reconstructed Bird’s-Eye-View (BEV) tokens. This enables the student planner to leverage the MLLM’s world knowledge during training while remaining independent at inference time.

### 3.5.3 Contrastive Vision-Language-Planning (VLP)

The **VLP** framework (Pan et al. 2024) addresses generalization by employing contrastive learning to align visual scene elements with linguistic prototypes. It introduces an *Agent-centric Learning Paradigm (ALP)*, where cropped BEV features of scene agents (vehicles, pedestrians) are aligned with their text descriptions via an InfoNCE loss:

$$\mathcal{L}_{VLP} = -\mathbb{E} \left[ \log \frac{\exp(sim(f_{bev}, f_{text})/\tau)}{\sum_{neg} \exp(sim(f_{bev}, f_{neg})/\tau)} \right] \quad (12)$$

This ensures that the perception backbone learns semantically distinct representations for long-tail objects, significantly reducing collision rates in unseen environments. Other approaches, such as **Vi-LAD** (Elnoor et al. 2025), extend this by distilling attention maps to ensure socially compliant navigation behaviors.

**Summary.** The autonomous driving domain exemplifies why Knowledge Distillation has become indispensable in modern vision systems. Unlike NLP tasks where inference can tolerate higher latency, vision applications, particularly safety-critical scenarios like AD, demand real-time decision-making under strict hardware constraints (e.g., <50ms latency on vehicle edge devices). VLMs, while achieving remarkable semantic reasoning capabilities, are fundamentally incompatible with these deployment requirements due to their billion-parameter architectures. KD bridges this gap by distilling the reasoning depth, spatial-temporal awareness, and causal logic of VLMs into compact student models that preserve critical capabilities while achieving orders-of-magnitude speedup. The methodologies reviewed here, including rationale-based alignment, surrogate task learning, and contrastive planning, demonstrate that effective distillation in vision extends beyond simple logit matching and requires transfer of structured spatial knowledge and semantic prototypes. This paradigm has broad implications beyond Autonomous Driving: medical imaging diagnosis, robotic manipulation, and video understanding all face similar trade-offs between model capacity and deployment feasibility, positioning vision-centric KD as a foundational technique for practical AI systems.

## 3.6 Task-Specific KD

Task specific distillation is often combined with instruction tuning, where a large instruction-tuned teacher model distills its task-specific knowledge into a smaller student model while preserving instruction-following capabilities (Wu et al. 2023; Yang et al. 2023; Zhang et al. 2023).

Taori et al. (2023a) and Wei et al. (2021) introduced instruction-tuned LLMs that distill task-specific knowledge through instruction-following datasets. Alpaca fine-tunes Meta’s LLaMA-7B using 52,000 instruction-response pairs generated by OpenAI’s text-davinci-003, mimicking knowledge distillation. Similarly, FLAN fine-tunes a 137B model on over 60 NLP datasets phrased as natural language instructions, enabling better generalization to unseen tasks. Both methods leverage instruction tuning to enhance model performance across diverse domains, outperforming larger models like GPT-3, LaMDA-PT, and GLaM in tasks such as inference, question answering, and translation.

Although LLMs can achieve high performance through instruction tuning, the predefined instructions are often simpler than the complex cases encountered in real-world scenarios (Zhang

et al. 2023; Wang et al. 2022; Ouyang et al. 2022). Therefore, domain-specific distillation, which tailors knowledge transfer for specialized fields (e.g. medical, programming), is crucial for improving model effectiveness. Most solutions to these problems involve designing a domain-specific dataset to fine-tune the model. Zhang et al. (2023) proposed AplaCare, which enhances medical LLMs through a semi-automated instruction fine-tuning pipeline. Starting with a clinician-curated seed dataset, they used GPT-4 to generate diverse medical tasks and ChatGPT to create responses, forming MedInstruct-52k. This dataset fine-tunes LLaMA models, improving instruction-following ability. Their approach achieves significant gains in medical benchmarks.

### 3.7 Theoretical Studies

Theoretical studies of knowledge distillation focus on explaining how knowledge is transferred and why this process is effective. The understanding is built upon several main insights, including (1) the interpretation of the teacher-student paradigm, (2) the convergence and generalization theory, (3) the impact of architectural choices on knowledge transfer, and (4) the role of data in knowledge distillation.

The key aspect of the teacher-student paradigm is the use of soft labels, as introduced in Section 2.1.1. One interpretation is that these soft labels contain dark knowledge, which reveals information about both the teacher’s uncertainty and the relationships between classes (Müller et al. 2019). In addition, several studies consider the soft label as label smoothing regularization, which provides the trade-off between bias and variation when training the student model (Yuan et al. 2020; Zhou et al. 2021). More recent studies provide the systematic Bayesian framework for interpreting the regularization of the knowledge distillation loss as the effect of implicitly imposing a prior distribution on the student model (Menon et al. 2021; Fang et al. 2024). Furthermore, the temperature paper  $\tau$  in Equation (1) can be interpreted as the variance parameter of the prior or as a measure of dispersion within the framework of statistical Bayesian modeling (Fang et al. 2024).

Several theoretical studies have examined the convergence behavior of student models in the context of knowledge distillation. Most research often focuses on simplified model architectures such as linear models, deep linear models, or the Gaussian process (Phuong and Lampert 2019; Mobahi et al. 2020; Borup and Andersen 2021). For more general cases, Menon et al. (2021) establishes a concrete criterion for evaluating the performance of a teacher model. It is proven that knowledge distillation can reduce the prediction variance of the student model, ultimately resulting in a more accurate student model. Some studies have surprisingly indicated that a student model trained via distillation can obtain good generalization bounds even from a teacher model that having poor theoretical bounds (Hsu et al. 2021). This is related to the regularization effect of KD. The model compression of knowledge distillation can also be viewed as a form of regularization of the model space, which leads to better performance on the testing data. Furthermore, Lopez-Paz et al. (2015) unifies the knowledge distillation with privileged information (Vapnik et al. 2015) and proposes the generalized distillation. The benefits of learning with a teacher are shown to arise due to three factors: (1) The capacity of the teacher model is sufficiently small, allowing it to be well approximated by a smaller student model. (2) The teacher’s approximation error relative to the true or optimal decision boundary is smaller than the student’s approximation error. (3) The rate at which the student learns the teacher’s decision boundary is faster than the rate at which the student learns the true function. However, providing a rigorous justification for these conditions remains an open question in the field.

The difference in model capacity between the teacher and the student models, often referred to as the capacity gap, is a crucial factor influencing the effectiveness of distillation (Mirzadeh et al. 2020; Li et al. 2023; Niu et al. 2022). A large capacity gap, where the student might lack the representational power to fully absorb the knowledge learned by the teacher, can sometimes hinder effective knowledge transfer. Conversely, if the capacity gap is too small, the student might already have sufficient capacity to learn the task effectively from the original data, making distillation less beneficial. Zhang et al. (2023) finds that the optimal teacher scale almost consistently follows a linear scaling with the student scale across different language model architectures and data scales in terms of the scaling law.

The choice of the distillation dataset can also influence what aspects of the teacher’s knowledge are learned by the student. Increasing the size of the distillation dataset does not always lead to improved student fidelity to the teacher and can even negatively affect the student’s generalization

performance (Stanton et al. 2021). Phuong and Lampert (2019) studied the effect of data geometry on knowledge distillation in the case of linear and deep linear models. They describe the data geometry by defining the angular alignment between the angular alignment between the data teacher’s weight vector, which has been proven to be a crucial factor in the derived generalization bound.

**Table 4:** Comparative Summary of Representative KD Methods for LLMs.

KD paradigm	Key idea / strengths	Typical limitations	Representative use cases
Rationale-based KD (Hsieh et al. 2023; Chu et al. 2023; Feng et al. 2024)	Transfers chain-of-thought rationales, yielding interpretable students with strong generalization and lower data demand	Needs reliable rationales and metrics; harder to extract rationale knowledge with black-box teachers	Decision-critical domains (clinical or legal reasoning)
Uncertainty-aware KD (Korattikara Balan et al. 2015; Vadera et al. 2020; Malinin et al. 2019; Menon et al. 2021; Fang et al. 2024)	Distils predictive distributions, giving calibrated confidence and robustness to noise	Bayesian sampling / ensemble distillation adds computational cost	Safety-critical or noisy-label tasks (medical diagnosis, risk assessment)
Multi-teacher KD (You et al. 2017; Zhang et al. 2022; Du et al. 2020; Fukuda et al. 2017; Zhu et al. 2021; Tian et al. 2024; Khanuja et al. 2021; Liu et al. 2024; Wadhwa et al. 2025)	Fuses complementary expertise from several teachers to broaden coverage and robustness	High compute cost; reconciling conflicts among heterogeneous teachers	Cross-domain or multilingual students inheriting specialized skills
Dynamic and adaptive KD (Chang et al. 2022; Sun et al. 2021; Li et al. 2024, 2023; Zhang et al. 2019, 2022; Allen-Zhu and Li 2020; Jumper et al. 2021; Yang et al. 2023)	Teacher and student co-evolve, or a model refines itself, enabling continual learning with vast unlabeled data	Possible unstable convergence, bias amplification, and training cost	Rapidly evolving fields or unlabeled-data regimes (e.g., AlphaFold protein prediction)
Task-specific KD (Wu et al. 2023; Yang et al. 2023; Zhang et al. 2023; Taori et al. 2023a; Wei et al. 2021; Zhang et al. 2023; Wang et al. 2022; Ouyang et al. 2022)	Tailors transfer via instruction tuning, giving compact students strong task performance with minimal data	Requires curated instruction data; simple instructions may miss real-world complexity	Domain specific applications, such as medical LLMs, specialist code or legal assistants

### 3.8 Summary and comparison of KD methods

In prior subsections, we examined the principal KD methodologies in detail. Table 4 then offers a concise overview of each method’s core idea, its main strengths and limitations, and representative use cases. This summary serves as a quick reference to guide readers in choosing the most appropriate approach based on model capacity, computational budget, and application domain.

To give a complementary perspective, Table 5 contrasts single-teacher and multi-teacher KD, and domain-specific and general-purpose KD. It outlines each method’s key idea, data requirements, robustness/performance, computational cost/flexibility, and typical application scenarios, helping practitioners identify the approach best suited to their objectives.

## 4 Methodologies and Techniques for Dataset Distillation in LLMs

LLMs are trained on massive corpora that often include redundant, low-quality, or uninformative samples. To improve training efficiency without sacrificing performance, dataset distillation

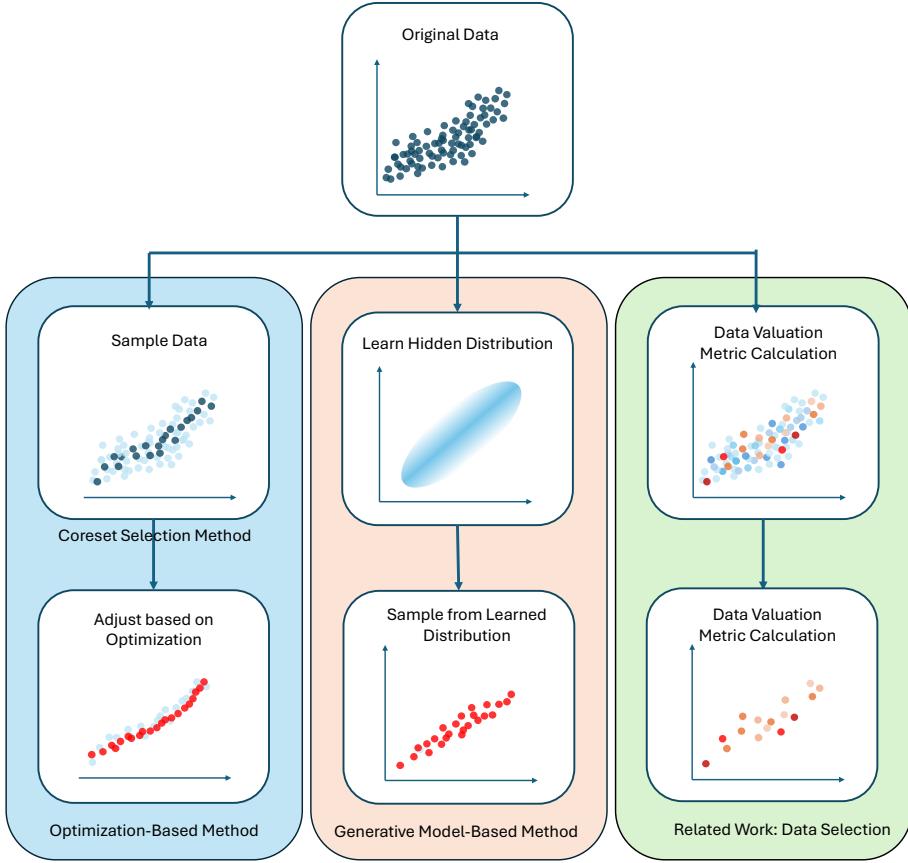
**Table 5:** Comparative overview of single-teacher versus multi-teacher KD and domain-specific versus general-purpose KD.

Aspect	Method A	Method B
<i>Single-teacher vs Multi-teacher KD</i>		
Key idea	Distill knowledge from one expert teacher	Fuse knowledge from an ensemble of multiple teachers
Data requirement	Outputs from a single teacher model	Outputs from several teacher models, increasing diversity
Robustness	Performance tied to one teacher’s biases	Improved generalization by reconciling complementary expertise
Compute cost	Moderate (one forward pass per example)	Higher (multiple forward passes per example)
Use cases	When a single strong teacher is available or compute is constrained	Cross-domain transfer, multilingual KD, ensemble compression
<i>Domain-specific vs General-purpose KD</i>		
Key idea	Tailor distillation on specialized, domain-relevant data	Distill on broad, diverse corpora for wide applicability
Data requirement	High-quality, annotated domain data	Large-scale, heterogeneous datasets covering many topics
Performance	Superior in the target domain	Good across varied tasks but may underperform on niche tasks
Flexibility	Limited outside the trained domain	High adaptability to new tasks with minimal re-distillation
Use cases	Clinical decision support, legal analysis, scientific literature	Open-domain chatbots, general question answering, conversational agents

techniques aim to synthesize compact datasets that retain the essential information of the full corpus. These methods generally fall into two main categories: optimization-based approaches, which directly learn a small set of synthetic samples by optimizing them to replicate the behavior of the full dataset during training; and synthetic data generation, which leverages generative models or heuristics to produce representative training examples. In addition to distillation, we also review data selection strategies, which focus on identifying high-quality subsets from existing datasets to achieve optimal model performance with fewer training samples.

**Table 6:** Advantages and Disadvantages of Different Dataset Distillation Methods.

Method	Advantages	Disadvantages	Use Cases
<b>Optimization-Based Methods</b> (Yu et al. 2023; Wu and Yao 2025; Liu et al. 2024; Zhong et al. 2024)	<i>Adaptive Learning:</i> Continuously refines data selection. <i>Flexibility:</i> Can be tailored to specific tasks.	<i>Computational Demand:</i> Resource-intensive process. <i>Complexity:</i> Requires iterative meta-careful hyper-parameter tuning.	Best GPU when allows iterative optimization.
<b>Generative Model-Based Methods</b> (Sajedi et al. 2024; Li et al. 2024; Kanagavelu et al. 2024; Zhang et al. 2025)	<i>Data Augmentation:</i> Creates diverse data samples. <i>Privacy Preservation:</i> Synthetic data mitigates privacy concerns.	<i>Model Complexity:</i> Training accurate models is challenging. <i>Quality Assurance:</i> Ensuring accurate representation is difficult.	When diversity or privacy is paramount with sufficient compute.
<b>Data Selection</b> (Yu et al. 2023; Yang et al. 2022; Tan et al. 2025; Marion et al. 2023)	<i>Efficiency:</i> Reduces dataset size and training time. <i>Simplicity:</i> Straightforward implementation.	<i>Information Loss:</i> Risk of discarding valuable data. <i>Static Evaluation:</i> May not adapt to evolving distributions.	Ideal for rapid pre-filtering with limited compute.



**Fig. 8:** Taxonomy of Dataset Distillation, including Data Selection, Optimization-based methods, and Generative Model-based Methods. The root plot shows the original dataset. The first column illustrates the data selection and pruning methods. Data points in different colors represent data with different importance scores. The warmer the color, the higher the importance score of the data point. After selection, data points with higher importance scores are preserved. The second column shows the optimization-based methods, which start with a subset of the original data and iteratively adjust the selected data according to an optimization loss function. The third column shows the generative model-based methods. These methods first train a generative model that learns the hidden distribution of the original data and then generates new data from the learned distribution.

#### 4.1 Optimization-Based Dataset Distillation

Optimization-based dataset distillation has emerged as a powerful method to compress large-scale datasets into compact, highly informative subsets for efficiently training LLMs. Given the exponential growth of pretraining corpora, storing and processing full datasets for model training is computationally prohibitive. Optimization-based approaches aim to synthesize a small yet representative dataset that can guide LLM training while preserving the generalization capability of models trained on much larger datasets.

The foundation of optimization-driven dataset distillation was established by Wang et al. (2018), leveraging meta-learning to iteratively refine synthetic images. This foundational work led to multiple refinements focusing on improving data compression and training efficiency. A major advancement was the use of gradient-matching techniques, where synthetic datasets  $\mathcal{D}_s$  are optimized to minimize the difference between the gradients computed on real and synthetic data  $\mathcal{D}_r$ :

$$\min_{\mathcal{D}_s} \sum_t \|\nabla_{\theta_t} \mathcal{L}(\mathcal{D}_s; \theta_t) - \nabla_{\theta_t} \mathcal{L}(\mathcal{D}_r; \theta_t)\|^2. \quad (13)$$

Here,  $\theta_t$  represents the model parameters at step  $t$ . This ensures that training an LLM on  $\mathcal{D}_s$  follows an optimization trajectory similar to training on the full dataset. Techniques such as Dataset Condensation (Zhao et al. 2020) and Differentiable Siamese Augmentation (Zhao and Bilen 2021)

employ this approach, aligning synthetic data with real dataset gradients to improve training efficiency.

Beyond gradient matching, embedding-based and trajectory-based methods further refine dataset compression for LLMs. Embedding-based methods optimize synthetic text sequences to preserve the statistical distribution of real embeddings in LLMs' latent space, ensuring effective knowledge retention with fewer training samples (Zhao and Bilen 2023). Trajectory-based approaches, such as Matching Training Trajectories (Cazenavette et al. 2022), align synthetic text with full dataset training trajectories, enabling long-term consistency in optimization paths.

Efficiency improvements in LLM-specific dataset distillation have emerged in response to the immense computational demands of large-scale language model training. RFAD (Loo et al. 2022) employs random feature approximation to reduce computational overhead, while FRePo (Zhou et al. 2022) uses model-pooling techniques to mitigate overfitting in LLM training. DREAM (Liu et al. 2023) further enhances efficiency by replacing random text sampling with representative selection, significantly reducing distillation iterations.

#### *Challenges and Future Directions.*

Optimization-based dataset distillation for LLMs struggles to reconcile the competing demands of dataset compression and model generalization. Scaling these methods to modern LLMs introduces a critical bottleneck: preserving diverse linguistic patterns and factual knowledge in highly compressed synthetic datasets, which often prioritize common features at the expense of rare or nuanced content. Techniques like training trajectory matching face inherent limitations in maintaining alignment with full-dataset optimization paths over prolonged pretraining, as even minor discrepancies compound into significant divergence. Furthermore, evaluation frameworks remain inadequate, focusing narrowly on efficiency or task-specific accuracy while neglecting essential LLM capabilities such as factual coherence, reasoning, and cross-task adaptability. Future research should address several concrete gaps: First, developing hierarchical trajectory matching that aligns synthetic data at multiple temporal scales (token-level, sequence-level, and epoch-level) rather than relying solely on gradient matching, with specific investigation into optimal weighting schemes for different trajectory components. Second, creating factual density metrics that quantify how well synthetic datasets preserve low-frequency but critical knowledge (e.g., historical dates, scientific constants, cultural references) compared to high-frequency linguistic patterns, potentially through knowledge graph-based validation. Third, establishing compositional reasoning benchmarks specifically designed for distilled datasets, testing whether models trained on synthetic data can perform multi-step logical inference and maintain consistency across related facts. Finally, integrating knowledge-aware constraints into distillation objectives could help preserve rare linguistic and factual patterns in compressed datasets, bridging the gap between efficiency and generalization.

## 4.2 Synthetic Data Generation for Dataset Distillation

While optimization-based methods refine existing datasets, generative-model-based dataset distillation leverages large pre-trained generative models to synthesize highly compact but expressive training corpora for LLMs. Rather than directly optimizing dataset representations, these methods aim to learn a distribution over linguistic knowledge and generate synthetic text sequences that retain the essential structure and diversity of real data.

A key distinction of generative approaches is their use of latent space representations to encode knowledge. Instead of optimizing individual data points, these methods sample latent variables  $\mathbf{z}$  from a latent distribution  $p(\mathbf{z})$  and map them to synthetic data using a learned generative function  $G(\mathbf{z})$ :

$$\mathbf{x}_s = G(\mathbf{z}), \quad \mathbf{z} \sim p(\mathbf{z}). \quad (14)$$

Methods such as GLaD (Cazenavette et al. 2023) employ GAN-based priors to synthesize datasets that maintain high representational fidelity while being computationally efficient. It trains a generator  $G$  in an adversarial setting against a discriminator  $D$ , where the objective is:

$$\min_G \max_D \mathbb{E}_{\mathbf{x}_r \sim p(\mathbf{x}_r)} [\log D(\mathbf{x}_r)] + \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z})} [\log(1 - D(G(\mathbf{z})))]. \quad (15)$$

This adversarial process ensures that the generated synthetic dataset is indistinguishable from real data.

Besides, HaBa (Liu et al. 2022) and KFS (Lee et al. 2022) introduce latent-code-based synthesis, where a set of learned latent vectors and decoders reconstruct informative synthetic images. Unlike optimization-based methods, which refine dataset representations explicitly, generative models produce synthetic samples by learning underlying data distributions. One key advantage of generative approaches is their ability to capture complex structural dependencies that might be difficult to encode explicitly in optimization-based methods.

#### *Challenges and Future Directions.*

Despite their strengths, generative approaches face unique challenges. Unlike optimization-based methods that explicitly control dataset refinement, generative models risk distribution drift, where synthetic text deviates from real-world linguistic structures. Additionally, generative dataset distillation for LLMs requires extensive model inference, making it computationally expensive for large-scale training. One promising future direction is developing efficient statistical drift detection frameworks that continuously monitor synthetic text quality during generation, with automatic generation stopping or correction when drift exceeds predefined thresholds. Computational amortization methods are a promising cost-saving direction, such as cached intermediate representations or progressive distillation in which smaller models synthesize preliminary synthetic samples that are post-processed by larger models, possibly cutting the cost of inference drastically. Additionally, hybrid optimization-generation pipelines provide opportunities to combine optimization-based methods with generative-based methods, where gradient-based methods identify optimal data characteristics (e.g., optimal token frequency distributions, key factual dependencies) which then guide generative models through structured prompting or fine-tuning.

### 4.3 Related Works: Data Selection

While dataset distillation methods focus on synthesizing compact, informative training subsets, data selection addresses a closely related goal: optimizing training efficiency for LLMs by strategically identifying and selecting high-quality data (Albalak et al. 2024). Data selection can be used at various stages of LLM training, including data collection, data cleaning, data deduplication, selecting coresets, and fine-tuning task-specific data. At the preprocessing stage, entropy-based filtering methods remove low-information or highly uncertain text, while perplexity-based filtering excludes content that is either too simple or too complex for the model. During the data selection phase, coresets selection techniques identify representative samples, and data attribution methods prioritize examples based on their relevance to specific tasks. In this subsection, we review three major categories of data selection methods: data filtering, coresets selection, and data attribution.

#### 4.3.1 Data Filtering

When preparing data to train a LLM, ensuring that the dataset is clean and diverse is critical (Albalak et al. 2024). The first step of filtering uses simple but effective methods, including rule-based filtering (Penedo et al. 2023; Rae et al. 2021), keyword matching, and filtering spam data using the Fast Text model (Yao et al. 2020), which aims to strip personal information, inappropriate words, and offensive content. RefinedWeb (Penedo et al. 2023) manually designed heuristic rules, like in-document repetition, URL ban list, and page length.

Following this, a secondary filter is applied to remove similar documents or sentences, which helps reduce redundancy by identifying near-duplicate content through techniques such as Jaccard similarity (Khan et al. 2024):

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}, \quad (16)$$

where  $A$  is the and  $B$  is the set of the  $n$ -grams. An  $n$ -gram is a sequence of  $n$  consecutive tokens from the text. Jaccard similarity measures the similarity between two text documents, which can be viewed as a bag of  $n$ -grams. Finally, advanced semantic analysis is performed using model embedding similarity filtering. In this step, embeddings for each document or sentence are generated using a pre-trained language model, and semantic similarities are calculated (commonly using cosine similarity) (Abbas et al. 2023):

$$\text{cosine similarity}(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\| \|\mathbf{b}\|}, \quad (17)$$

where  $\mathbf{a}$  and  $\mathbf{b}$  are embeddings learned by the model. By setting a similarity threshold, data points that are semantically too similar are identified and one of them is removed. For document-level filtering, one path is to calculate the similarity between documents and remove one document that is above the threshold. SemDeDup (Abbas et al. 2023) deduplicates based on the similarity of the text data’s embeddings from pre-trained language models. However, the semantically-based deduplication methods requires training LLM and is prohibitively expensive. Thus, it is more efficient to employ LSHBloom (Khan et al. 2024), which identifies duplicated documents based on the Jaccard similarity calculated on the raw text content of each document.

For filtering high-quality data, perplexity-based filtering leverages a language models’ own prediction confidence to assess the quality of text data. Perplexity is defined as the exponential of the average negative log-likelihood of the predicted words. For a sequence  $w_1, w_2, \dots, w_N$ , it is computed as

$$\text{Perplexity} = \exp\left(-\frac{1}{N} \sum_{i=1}^N \log P(w_i | w_1, \dots, w_{i-1})\right). \quad (18)$$

A higher perplexity indicates that the model is less confident in its predictions and indicates the data has lower quality (Wenzek et al. 2020). FastText filter (Yan 2024) trains a classifier or scoring model to decide which data to include. Entropy-based filtering methods eliminate low-information or highly uncertain text. By calculating the cross-entropy loss from models trained on in-domain datasets and general-purpose datasets, Moore-Lewis selection (Moore and Lewis 2010; Axelrod et al. 2011) is proposed to identify in-domain data by selecting sentences that are much more predictable (i.e., lower cross-entropy) under the in-domain model compared to the out-of-domain model.

#### 4.3.2 Coreset Selection

Coreset selection methods aim at reducing the computational burden of training machine learning models by condensing large datasets into representative subsets (Albalak et al. 2024) and have been used in linear regression (Li et al. 2024; Ma et al. 2015), and nonparametric regression (Meng et al. 2020; Zhang et al. 2018; Sun et al. 2020), deep learning (Mirzasoleiman et al. 2020; Fang et al. 2025; Borsos et al. 2020), and network data analysis (Wu et al. 2023; Liu and Liu 2024). Let  $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$  be the training data and  $y_i$  can be categorical or continuous for classification and regression. Coreset selection chooses a subset  $S \subseteq \mathcal{D}$  of the full dataset  $\mathcal{D}$  so that a cost function computed over  $S$  closely approximates that computed  $\mathcal{D}$  for all model parameters. One general formulation is:

$$S^* = \arg \min_{S \subseteq \mathcal{D}, |S|=k} \sup_{\theta \in \Theta} \left| \frac{1}{|\mathcal{D}|} \sum_{\mathbf{x} \in \mathcal{D}} f(\mathbf{x}; \theta) - \frac{1}{|S|} \sum_{\mathbf{x} \in S} f(\mathbf{x}; \theta) \right| \quad (19)$$

where  $f(\mathbf{x}; \theta)$  is the cost function evaluated at data point  $\mathbf{x}$  with model parameters  $\theta$ ,  $\Theta$  is the set of all parameters over which the cost is evaluated, and  $k$  is the size of the coresset.

In instructor tuning, Zhang et al. (2024) presents a coresset selection approach that leverages gradient information from training samples. This method involves clustering data based on gradient similarities and selecting representative samples from each cluster. Similarly, Low-rank gradiEnt Similarity Search (LESS) also identifies the most influential data subsets on a target validation example using cosine similarity between the gradient features of the training samples and target examples (Xia et al. 2024).

In the alignment step, the coresset has to be aligned with human instructions. Their methodology focuses on evaluating data samples across three key dimensions: complexity, quality, and diversity (Liu et al.). They start by selecting data samples with higher scores of complexity and quality of instruction-response pairs. To ensure diversity, they select samples that are sufficiently different based on the cosine distance of the embeddings.

In task-specific fine-tuning, STAFF (Zhang et al.) addresses the challenges of computational overhead and data relevance. The method leverages a smaller model from the same family as the target LLM to speculate on data importance scores efficiently. These speculative scores are then verified on the target LLM to accurately identify and allocate more selection budget to important regions while maintaining coverage of easier regions.

### 4.3.3 Data Attribution

Data attribution methods select data by quantifying the value of individual data points. The concept of data Shapley ([Ghorbani and Zou 2019](#)) and its variants ([Wang et al. 2025, 2024](#)) provide a tool for explaining the importance of each data point. The Shapley value, defined based on the prediction and the performance score of the predictor trained on data  $\mathcal{D}$ , quantifies its average marginal contribution to all possible subsets of the dataset:

$$\phi_i = C \sum_{S \subseteq \mathcal{D} - \{i\}} \frac{V(S \cup \{i\}) - V(S)}{\binom{n-1}{|S|}}, \quad (20)$$

where  $V(S)$  is the performance metric on the training subset  $S$  and  $C$  is a constant. Due to the computational complexity of calculating exact Shapley values, the authors propose approximation methods, including Monte Carlo sampling and Gradient-based approximations. [Wang et al. \(2025\)](#) addressed the computational challenges associated with traditional data Shapley calculations by introducing in-run data Shapley. The approach leverages the iterative nature of training algorithms, using first- and second-order Taylor expansions to approximate the impact of each data point on the model's performance. However, data Shapley values may mislead data valuation, especially in the presence of strong correlations among data points. [Wang et al. \(2024\)](#) proposed refinements by grouping similar data points and computing Shapley values for clusters rather than individual points to account for redundancy.

In addition to calculating a static influential score, the temporal dependence of the training data quantifies the impact of omitting a particular data point at a specific training iteration ([Wang et al. 2025](#)) by computing the dot product between the data point's embedding and the gradient of the test example's loss.

### 4.3.4 Challenges and Future Directions

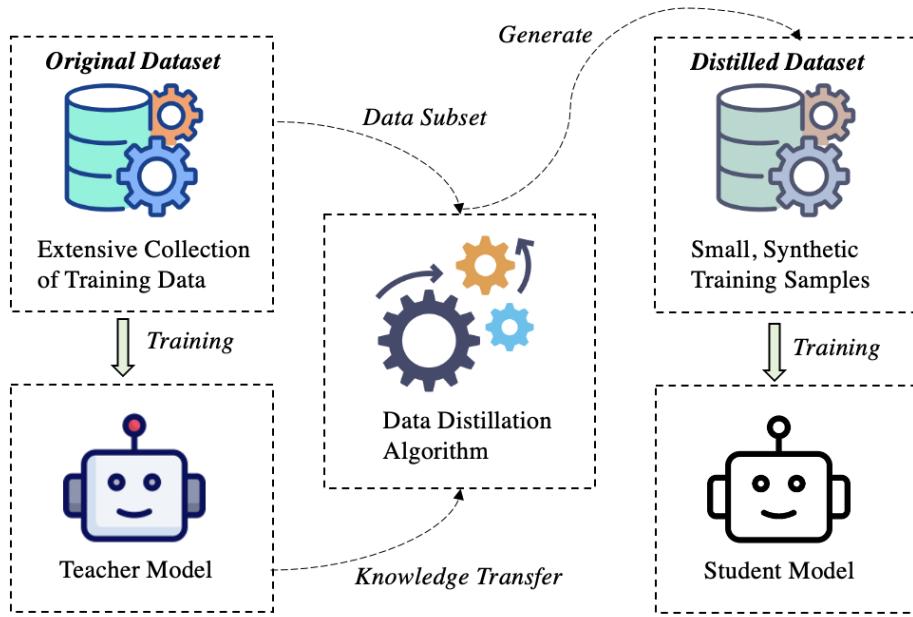
The current state of data selection methods reflects a balance between efficiency, accuracy, and adaptability. While methods like Jaccard similarity offer simplicity, they lack semantic depth. Jaccard similarity, a measure based on the size of the intersection divided by the size of the union of two sets, is often used to remove redundant data by identifying similar text samples. In the context of LLMs, it might be employed as a utility function to filter out duplicate or highly similar training examples, aiming to reduce dataset size without losing information. However, a key limitation is its inability to capture semantic meaning. For instance, two sentences with different wordings but similar meanings (e.g., “The cat is on the mat” vs. “A feline rests atop a rug”) may have low Jaccard similarity in [Equation 4.3.1](#), leading to their retention despite redundancy, or vice versa, potentially excluding valuable paraphrased data. This can result in suboptimal data selection, particularly for tasks requiring nuanced understanding, and highlights the need for semantic-aware measures.

Advanced techniques like LESS, STAFF, and Data Shapley address specific needs but are constrained by computational costs and practical limitations. Many methods, including LESS, STAFF, and data attribution methods like Data Shapley, involve training or evaluating models on subsets, which can be computationally expensive, especially for large datasets and models. For instance, STAFF's selection process can take over ten hours on powerful devices, driven by multiple epochs of training on the target model to evaluate data scores and regions, adding to computational cost. LESS requires a preliminary training phase to obtain useful gradient features, increasing computational complexity and cost. LLMs often involve massive datasets, and the dynamic nature of training means data importance can change over time, making static selection methods less effective. In-Run Data Shapley's ([Wang et al. 2025, 2024](#)) ability to capture training dynamics is a step forward, but adaptation remains a challenge. To overcome the challenges, we suggest the following future research directions to enhance semantic understanding, reduce computational cost, and adapt to dynamic training dynamics, thereby fully leveraging the potential of LLMs. To save computational time and cost, we suggest using more efficient ways of data selection, subsampling based on the metric calculated by the original dataset ([Meng et al. 2017; Ma et al. 2022; Li and Meng 2020; Li et al. 2023; Wu et al. 2023](#)) instead of metrics by training the model. Thus, we can reduce the computational cost by saving the model training time.

## 5 Integration of Knowledge Distillation and Dataset Distillation

In this section, we examine the integration of KD and DD to mitigate the computational and scalability challenges inherent to modern LLMs. KD transfers nuanced reasoning skills, such as chain-of-thought capabilities, from large teacher models to compact students, while DD generates minimal yet representative datasets that preserve critical knowledge patterns. By combining these approaches, we demonstrate how their synergy reduces dependency on large-scale data, improves computational efficiency, and maintains advanced functionalities in distilled models. This integration establishes a cohesive framework for balancing model compression with sustainable data utilization.

### 5.1 Knowledge Transfer via Dataset Distillation



**Fig. 9:** An illustration of Knowledge Transfer via Data Distillation. The process illustrates how a teacher model trained on the original large dataset transfers its knowledge through distillation processes to create a small, synthetic dataset. This distilled dataset enables the efficient training of student models while maintaining performance.

While meta-learning and bi-level optimization-based dataset distillation have demonstrated effectiveness across various small-scale benchmarks, they often suffer from two critical issues: (1) overfitting to the training dynamics during the distillation phase, and (2) limited scalability as the dataset size or model architecture complexity increases. This is primarily because optimization-based distillation methods rely on unrolling and storing the computation graph of multiple gradient steps during the distillation phase, an approach that becomes increasingly prohibitive in modern experimental settings. Moreover, their dependence on specific gradient steps leads to overfitting to the teacher architecture.

Recent research, building on these insights while addressing prior limitations, has developed integrated knowledge transfer frameworks that combine KD and DD techniques. Figure 9 illustrates this unified methodology. More specifically, [Yin et al. \(2023\)](#) recently proposed SRe2L, a method designed to disentangle meta-learning and bi-level optimization by reframing dataset distillation as data-frugal knowledge distillation. SRe2L follows a three-stage process: (1) pretraining the teacher

model on a large-scale dataset, (2) leveraging model inversion (Yin et al. 2020) to synthesize a coresnet, and (3) transferring the teacher’s soft labels and the coresnet to the student model. The objective of SRe2L is to learn a compact synthetic dataset  $C_{\text{syn}}$ , composed of synthetic input-label pairs  $(\tilde{\mathbf{x}}, \tilde{y})$ , that encapsulates essential information from the original dataset  $\mathcal{D}$ , which consists of real samples  $(\mathbf{x}, y)$ . The learning process involves training a neural network  $\varphi_\theta$  on the synthetic data to minimize the expected classification loss:

$$\theta_{C_{\text{syn}}} = \arg \min_{\theta} L_C(\theta), \quad \text{where } L_C(\theta) = \mathbb{E}_{(\tilde{\mathbf{x}}, \tilde{y}) \in C_{\text{syn}}} [\ell(\varphi_{\theta_{C_{\text{syn}}}}(\tilde{\mathbf{x}}), \tilde{y})], \quad (21)$$

where  $\ell(\cdot)$  denotes the training loss function, implemented as the soft-label cross-entropy loss using teacher outputs obtained during the relabel stage:

$$\ell(\varphi_\theta(\tilde{\mathbf{x}}), \tilde{y}) = -\tilde{y} \cdot \log(\varphi_\theta(\tilde{\mathbf{x}})), \quad (22)$$

where  $\tilde{y}$  is the soft label predicted by a teacher network trained on the original dataset.

To ensure that the synthetic dataset induces model training dynamics similar to those of the original data, SRe2L further constrains the optimization by minimizing the worst-case generalization gap:

$$\sup_{(\mathbf{x}, y) \sim \mathcal{D}} |\ell(\varphi_{\theta_T}(\mathbf{x}), y) - \ell(\varphi_{\theta_{C_{\text{syn}}}}(\mathbf{x}), y)| \leq \epsilon, \quad (23)$$

where  $\theta_T$  are the parameters of a model trained on the full dataset  $\mathcal{D}$ , and  $\epsilon$  is a small bound on the acceptable discrepancy in generalization performance between the two models. By decoupling dataset optimization and neural network training, SRe2L reduces computational costs while improving scalability. These synthesized samples, along with the soft labels, are then used to perform knowledge distillation on the student model.

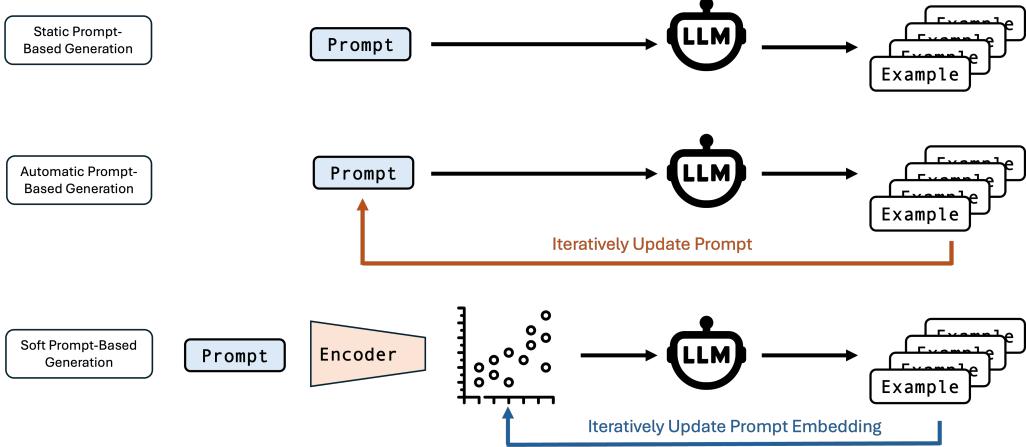
Following SRe2L, several knowledge distillation-based dataset distillation approaches have emerged. Shao et al. (2024) introduced a method that employs diverse architectures during the distillation phase to enhance generalizability between architectures. Sun et al. (2024) proposed an approach that prunes image pixels by selecting the most informative patches and assembling them into a collage. Additionally, Yin and Shen (2023) introduced advanced cropping techniques to further refine the quality of distilled samples. Moreover, although knowledge-distillation-based dataset distillation has proven highly effective in mitigating the computational complexity of dataset distillation, challenges remain, particularly in storing and communicating data and soft-label correspondence for training downstream students (Xiao and He 2024).

Overall, knowledge transfer via data distillation represents a promising direction that bridges dataset distillation and knowledge distillation paradigms. By reformulating the problem as transferring knowledge from teacher to student through synthetic data, these methods effectively address the computational bottlenecks of traditional optimization-based approaches while maintaining competitive performance. This integration of concepts offers a more scalable and adaptable framework for knowledge compression that continues to evolve as researchers develop more sophisticated techniques for sample synthesis and information preservation.

## 5.2 Prompt-Based Synthetic Data Generation for LLM Distillation

Another emerging trend is prompt-based synthetic data generation, which synergistically integrates KD and DD. Here, large-scale teacher LLMs generate compact, high-quality training sets through strategically designed prompts, transferring their knowledge into synthetic data while enabling efficient student model training. This unified approach has proven effective for low-resource adaptation, task-specific augmentation, and self-distillation. The methodologies can be categorized into three main types: static prompt-based generation, automatic prompt optimization, and soft prompt-based frameworks.

Static Prompt-Based Generation involves using fixed, manually crafted prompts to direct LLMs in producing synthetic data. Researchers design prompts that elicit desired responses from the model, generating data that aligns with specific tasks or domains. For example, PromDA (Wang et al. 2022) and GAL (He et al. 2022) use manually crafted prompts to augment data for natural language understanding tasks in low-resource settings.



**Fig. 10:** Illustrations of Different Types of Prompt-based Synthetic Data Generation. All three types of methods generate task-related samples by providing a prompt to a well-trained LLM. Static prompt-based generation directly inputs the prompt into the target LLM. Automatic prompt-based generation iteratively updated the input prompt to get more complete and diverse samples. Soft prompt-based generation encodes the prompt into an embedding space and iteratively updates the embedding of the prompt, which converts the discrete prompt into a continuous value.

Automatic Prompt Optimization refers to techniques that iteratively refine prompts to enhance the quality and relevance of the generated data. This process involves using algorithms to adjust prompts based on feedback from the LLM’s outputs, aiming to maximize certain performance metrics. For instance, [Deng et al. \(2022\)](#) introduces a reinforcement learning approach to automatically optimize discrete text prompts for language models. [Pryzant et al. \(2023\)](#) presents ProTeGi, a method for automatic prompt optimization using gradient-based techniques and beam search strategies.

Soft Prompt-Based Frameworks utilize learnable embeddings, known as soft prompts, to steer LLMs without altering their internal parameters. Unlike traditional text-based prompts, soft prompts are continuous vectors optimized during training to induce the desired model behavior. This method enables more nuanced control over the generated data and can be particularly effective in producing structured or domain-specific content. For example, DiffLM ([Zhou et al. 2024](#)) introduces a controllable data synthesis framework that leverages diffusion language models to enhance the quality of synthetic data generation, particularly for structured formatted data like tabular and code data. SoftSRV framework ([DeSalvo et al. 2024](#)) introduces a novel approach by employing soft prompts - trainable vectors - to steer frozen pre-trained LLMs toward generating targeted synthetic text sequences. This method allows for the creation of domain-specific synthetic data without extensive manual prompt engineering, enhancing the adaptability and efficiency of synthetic data generation.

These methodologies highlight the versatility of prompt-based data generation in leveraging LLMs to create synthetic datasets tailored to specific needs. By employing static prompts, automatic optimization, or soft prompt-based frameworks, researchers can effectively generate data that enhances model training and performance across various applications.

## 6 Evaluation and Metrics for LLM Distillation Techniques

### 6.1 Evaluation

Evaluating distillation techniques for LLMs demands a rigorous framework that measures performance, efficiency, robustness, and knowledge transfer efficacy. This ensures distilled models are systematically assessed on their ability to retain task effectiveness while balancing computational and data efficiency.

To provide concrete reference points for evaluating distillation methods, we incorporate widely-used standardized benchmarks such as GLUE ([Wang et al. 2021](#)) for natural language understanding tasks, GSM8K ([Cobbe et al. 2021a](#)) and MATH ([Hendrycks et al. 2021](#)) for mathematical reasoning evaluation, and MMLU ([Wilkins and Rodriguez 2024](#)) for comprehensive

knowledge assessment. These benchmarks enable meaningful comparison across different distillation approaches and provide standardized evaluation protocols. Typical performance retention rates show student models achieving 90-95% of teacher performance on these benchmarks while reducing model size by 5-10x, demonstrating the practical trade-offs between performance and efficiency in the field.

### ***Performance Metrics.***

A key approach to evaluating distilled models is to assess their performance, with a primary focus on accuracy and generalization. Common evaluation metrics include perplexity for language modeling, exact match (EM) and F1-score for question answering, and BLEU or ROUGE scores for text generation. These metrics aim to quantify how effectively the student model preserves the predictive capabilities of the teacher model while minimizing performance degradation. Additionally, similarity measures such as cosine similarity and KL divergence between the teacher's and student's outputs provide insight into the extent of knowledge transfer.

Recent research in the LLM literature has introduced several evaluation metrics specifically designed for large language models. One such metric is the MAUVE score, which quantifies the divergence between the probability distributions of human-generated and model-generated text (Pillutla et al. 2021). It is formulated based on the Jensen–Shannon divergence, providing a principled measure of distributional similarity.

$$\text{MAUVE} = \exp(-D_{\text{JS}}(P_{\text{human}} \| P_{\text{model}})), \quad (24)$$

where  $D_{\text{JS}}$  denotes the Jensen–Shannon divergence between the distribution  $P_{\text{human}}$  of human texts and  $P_{\text{model}}$  of model outputs.

Another metric, BERTScore, measures the similarity between generated and reference texts by computing the cosine similarity between contextual embeddings (Zhang et al. 2019). Its formulation can be written as

$$\text{BERTScore} = \frac{1}{T} \sum_{t=1}^T \max_{s \in S} \cos(\mathbf{e}_t, \mathbf{e}_s), \quad (25)$$

where  $\mathbf{e}_t$  and  $\mathbf{e}_s$  represent the embedding vectors of tokens in the candidate and reference sentences respectively,  $T$  is the number of tokens in the candidate text, and  $S$  is the set of tokens in the reference text. Other metrics, such as Self-BLEU, which quantifies repetition by computing the BLEU score of a generated text against itself, and distinct- $n$ , which is defined as the number of unique  $n$ -grams normalized by the total number of  $n$ -grams, address critical aspects of text quality like diversity and consistency. Consistency metrics, measured as the agreement rate, compute the agreement rate of multiple model responses to similar prompts to ensure stability in open-ended generation tasks.

For higher-order capabilities, evaluation extends to specialized benchmarks. For example, reasoning ability is measured through mathematical problem-solving (e.g., GSM8K (Cobbe et al. 2021a), MATH (Hendrycks et al. 2021)) and logical deduction tasks (Liu et al. 2020), while natural language generation is assessed via summarization fidelity, question answering accuracy, and retrieval-augmented generation performance in search engine contexts (Min et al. 2023). For a systematic taxonomy of LLM evaluation methodologies, we refer to the comprehensive survey by Chang et al. (2024).

### ***Complexity and Efficiency***

Distillation aims to reduce the computational footprint of LLMs. Evaluation in this aspect involves measuring inference speed, memory consumption, and model size. Metrics such as FLOPs (floating point operations), latency, and peak memory usage provide quantitative benchmarks for comparing different distillation techniques. Specific efficiency metrics include inference latency measurements showing reductions from hundreds of milliseconds to tens of milliseconds per query, and memory consumption comparisons demonstrating decreases from gigabytes to hundreds of megabytes for model storage and inference. Efficiency gains are particularly relevant for edge and low-resource deployment scenarios where computational constraints are stringent.

### **Robustness and Uncertainty.**

Ensuring the robustness of distilled models under adversarial conditions and distribution shifts is critical for reliable deployment. Robustness evaluation involves stress-testing models against domain shifts, adversarial perturbations, and noisy inputs. The Attack Success Rate (ASR) measures the efficacy of adversarial attacks in fooling the model (Wang et al. 2023, 2021). For a dataset  $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$  containing  $N$  input-output pairs and an attack method  $\mathcal{A}$  that generates adversarial examples  $\mathcal{A}(x)$ , ASR is defined as:

$$ASR = \sum_{(x,y) \in \mathcal{D}} \frac{I[f(\mathcal{A}(x)) \neq y]}{I[f(x) = y]}, \quad (26)$$

where  $I$  is the indicator function,  $f$  is the model under test, and the denominator counts samples correctly classified before the attack. To evaluate prompt-based robustness, the Performance Drop Rate (PDR) quantifies relative degradation after adversarial modifications to prompts (Zhu et al. 2023). For an original prompt  $P$ , adversarial prompt  $A(P)$ , and task-specific evaluation metric  $\mathcal{M}$  (e.g., accuracy or F1-score), PDR is calculated as:

$$PDR = 1 - \frac{\sum_{(x,y) \in \mathcal{D}} \mathcal{M}[f([A(P), x], y)]}{\sum_{(x,y) \in \mathcal{D}} \mathcal{M}[f([P, x], y)]} \quad (27)$$

where higher PDR indicates greater vulnerability to prompt attacks.

Calibration and uncertainty estimation ensure reliable confidence scores. The Expected Calibration Error (ECE) partitions model predictions into  $M$  equally spaced confidence bins  $B_1, \dots, B_M$  and measures the discrepancy between accuracy and confidence within each bin:

$$ECE = \sum_{m=1}^M \frac{|B_m|}{N} |\text{acc}(B_m) - \text{conf}(B_m)|, \quad (28)$$

where  $|B_m|$  is the number of samples in bin  $m$ ,  $\text{acc}(B_m)$  is the bin's accuracy, and  $\text{conf}(B_m)$  is the average predicted confidence (Guo et al. 2017; Tian et al. 2023). Entropy-based uncertainty estimates prediction stability using  $\text{Uncertainty} = -\sum_y p(y|x) \log p(y|x)$ , where higher entropy reflects lower confidence in outputs.

## 6.2 Quantifying Distillation Level in LLMs

A critical challenge in LLMs development lies in quantifying how much knowledge is distilled from teacher models, particularly since unregulated distillation risks model homogenization, identity leakage, and robustness degradation.

One approach is to compute the divergence between the probability distributions of teacher and student models using KL divergence or Jensen-Shannon divergence. These metrics help in determining the extent of information retained post-distillation. Another method involves evaluating feature representations extracted from different layers of the teacher and student models. Cosine similarity and centered kernel alignment provide a way to compare internal representations, offering insights into structural similarities between models. Additionally, task-specific evaluation can reveal how distillation affects downstream performance.

Recent work (Lee et al. 2025) advances distillation quantification through Identity Consistency Evaluation (ICE) and Response Similarity Evaluation (RSE). These methods systematically measure identity leakage (e.g., Qwen-Max-0919 falsely attributing its development to Claude in 32% of cases) and output homogenization (e.g., DeepSeek-V3 achieving RSE > 4.1 against GPT-4o) to assess distillation efficacy. These methods highlight critical trade-offs: while distillation improves efficiency, it risks robustness degradation and reduced model diversity.

## 7 Applications and Use Cases of Distillation

This section surveys knowledge distillation techniques across specialized domains, including medical and healthcare, education, and bioinformatics, demonstrating their transformative impact in optimizing domain-specific AI systems.

### 7.1 Medical and Healthcare

Recent advancements in KD for LLMs have enabled more efficient and specialized applications in healthcare. Distillation techniques allow for the compression of large models into smaller, task-specific ones while preserving essential capabilities, making them practical for clinical deployment. Below, we highlight recent research contributions in clinical decision support, medical summarization, patient interaction, and drug discovery.

#### 7.1.1 Clinical Decision Support

ClinRaGen ([Niu et al. 2024](#)) exemplifies how KD can be harnessed to build efficient clinical decision support systems by equipping small language models (SLMs) with LLM-level reasoning. The framework first retrieves disease-specific medical knowledge and uses an LLM (e.g., ChatGPT) to generate structured rationales from multimodal EHR data that combine textual clinical notes and time-series lab results and serve as high-quality distillation targets. Through a three-phase sequential distillation process, including medical note-based rationale learning, knowledge-augmented attention for lab test rationale, and full multimodal integration, the student SLM is distilled into a compact 80M-parameter model, which is more than 2,000x smaller than the 175B-parameter teacher LLM and 80x smaller than a fine-tuned 7B-parameter LLaMA model, yet it achieves excellent diagnostic performance on MIMIC-III and MIMIC-IV while training in under half the time. In rationale quality evaluations with GPT-4 and human judges, ClinRaGen ranks second for readability and correctness—matching the larger LLaMA3 model—and outperforms other 7B- and 60M-parameter models in consistency and clinical soundness. A larger variant, ClinRaGen\* (793M parameters), further boosts F1 by over 1.5%, highlighting the favorable trade-off between scale and performance.

[Ding et al. \(2024\)](#) proposes CKLE, a framework for ICU health event prediction through KD from general LLMs into multimodal electronic health record (EHR) models. Their approach transfers knowledge from a text-based teacher LLM to a smaller student model that processes both clinical text and structured patient data using cross-modal contrastive objectives. The distilled model improved prediction accuracy for heart failure and hypertension by up to 4.48% over state-of-the-art models while addressing privacy and deployment constraints through local, smaller models with LLM-level insights. The significance of this approach is underscored by similar real-world deployments at major medical centers, such as Mount Sinai’s Advanced Alert Monitor program, which has demonstrated that AI-based prediction systems can save more than 500 lives per year when properly integrated into clinical workflows.

[Hasan et al. \(2024\)](#) introduces OptimCLM, a comprehensive compression framework combining ensemble learning, knowledge distillation, pruning, and quantization for clinical BERT models. Their method uses an ensemble of domain-specific BERTs (DischargeBERT and COReBERT) to teach compact student models (TinyBERT and BERT-PKD) via black-box distillation. For hospital outcome tasks including length of stay, mortality, diagnosis, and procedure prediction, the student models achieved up to 22.8 $\times$  model size compression and 28.7 $\times$  speedup with minimal performance loss (under 5% decrease in AUROC), demonstrating that knowledge distillation can preserve accuracy while dramatically reducing computational requirements.

#### 7.1.2 Medical Summarization

[Tariq et al. \(2024\)](#) presents a novel application of knowledge distillation to develop a patient-centric radiology report summarization system. They leverage a 13B-parameter LLaMA model as a teacher to generate noisy layman summaries for approximately 7K chest CT reports, and then fine-tune a 770M-parameter T5 student model on this weakly labeled data, reducing model size by over 17 $\times$  while maintaining high fidelity. Compared to LLaMA zero-shot outputs, the distilled student cuts hallucinations from 18% to 6% and missing information from 17% to 4%. Expert radiologists and

lay users rate the student’s summaries as factually accurate and substantially more understandable, improving patient comprehension by 63%, all with far lower computational cost, demonstrating how distillation can yield efficient, reliable clinical decision support tools.

While explicit applications of KD for medical summarization are still emerging, large LLMs such as GPT-4 have shown impressive ability in this domain, in some cases producing summaries non-inferior to physicians. The success of LLMs in summarizing clinical text suggests significant opportunities to distill these capabilities into lightweight summarizers for electronic health records, representing a promising area for future research.

### 7.1.3 Patient Interaction (Q&A and Matching)

[Nievias et al. \(2024\)](#) presents Trial-LLAMA, an open-source LLAMA model fine-tuned for patient–trial matching. They generate a synthetic dataset of 2,000 patient–trial reasoning examples using GPT-4 and then supervise-fine-tune the LLAMA-2-70B model on this data. The resulting Trial-LLAMA model matches or surpasses proprietary GPT-3.5 baselines across key aggregate metrics such as NDCG, Precision, and AUROC, and achieves competitive criterion-level accuracy in both explicit and implicit eligibility judgments. By relying on synthetic data and open-source architectures, this work demonstrates a cost-effective, privacy-conscious approach to deploying large-model reasoning in clinical decision support.

[Sutanto et al. \(2024\)](#) addresses few-shot medical question answering by distilling knowledge from large LLMs into smaller models. Their approach uses a teacher LLM to generate synthetic multiple-choice questions with answer likelihood scores, creating a rich training set for a DeBERTa-v3 student model. On the MMLU benchmark, which includes medical exam questions, the distilled student achieved 39.3% accuracy compared to 28.9% for a baseline trained on few-shot data—a significant 10% absolute improvement. This demonstrates how LLM-generated data and knowledge distillation can substantially enhance a compact model’s medical reasoning capabilities.

[Yagnik et al. \(2024\)](#) compares general versus medical-specific distilled language models for open-ended medical question answering. Their evaluation indicates that medical-domain distilled models, such as BioMedBERT variants, can rival larger general models on expert question answering, reinforcing the value of distilling domain knowledge into compact chatbots for health-related queries.

### 7.1.4 Clinical Information Extraction and Knowledge Curation

[Gu et al. \(2023\)](#) leverages GPT-3.5 and GPT-4 as teachers to train a compact biomedical model for extracting adverse drug events (ADEs) from text. Instead of relying on costly manual labels, they prompt GPT-3.5 to structure raw biomedical text, then distill its outputs into a PubMedBERT student via self-supervised learning. Remarkably, the distilled PubMedBERT (with  $1,000\times$  fewer parameters) outperformed its teacher: achieving 6 points higher F1 than GPT-3.5 and even 5 points higher than GPT-4 on standard ADE extraction. Similar gains were shown for other information extraction tasks, demonstrating that distillation can transfer LLM capabilities into small, domain-specific models that are both efficient and high-performing.

[Vedula et al. \(2024\)](#) explores using LLMs as labelers for clinical text, distilling their knowledge into BERT-based models for clinical named entity recognition tasks. By generating pseudo-labels on clinical notes using GPT-4-level models and medical ontologies as teachers, they trained a BioBERT student that achieved F1 scores close to a fully supervised model while running  $12\times$  faster and at 1% of the cost of the large LLM in inference. This demonstrates that distillation can deliver nearly the same accuracy as an LLM on clinical text extraction while being dramatically more efficient in speed and cost—enabling practical deployment in hospital data pipelines.

### 7.1.5 Drug Discovery

[Zhang et al. \(2025\)](#) introduced KEDRec-LM, an instruction-tuned LLM for explainable drug recommendation. The model leverages knowledge from biomedical literature, drug repurposing databases, and clinical trials. The authors use a teacher LLM to generate domain-specific rationales, which are then distilled into a student model. The distilled model effectively synthesizes biomedical knowledge to support drug–disease relationship analysis and hypothesis generation.

By constructing a large training corpus integrating a drug repurposing knowledge graph, clinical trial data, and PubMed articles, then distilling this multi-source knowledge into a generative LLM, KEDRec-LM achieves state-of-the-art performance on drug selection tasks. On the expRxRec dataset, it achieved F1 scores up to 88.05%.

### 7.1.6 Multimodal and Vision-Language Applications

Ge et al. (2025) introduces ClinKD, a framework to boost medical visual question answering (Med-VQA) performance by injecting prior knowledge via distillation. The model builds on a vision-language architecture capable of analyzing both images and text questions. A teacher model incorporating Med-CLIP visual features and a large multimodal transformer is used to teach the student model about image-text alignment and medical concepts through a distillation phase. ClinKD achieved state-of-the-art accuracy on the Med-GRIT 270k benchmark for medical image QA, demonstrating superior ability to answer questions about radiology images, including pinpointing anatomical regions. By distilling cross-modal knowledge, the approach addresses issues like visual hallucination and improves the grounding of answers in medical images, highlighting how vision-language LLMs in healthcare can be made more compact and accurate through targeted knowledge distillation.

These studies collectively illustrate the transformative role of distillation techniques in adapting LLMs for healthcare applications. By reducing computational costs while maintaining high performance, knowledge-distilled models hold promise for scalable and interpretable AI-driven medical solutions that can be deployed in real-world clinical settings.

### 7.1.7 Key Lessons Learned

Distillation enables compact clinical models to rival large LLMs. Previous examples show that even with tens, hundreds, or even thousands of times fewer parameters, student models can still achieve comparable results using distillation techniques, which makes AI more deployable in resource-constrained clinical settings by reducing computational costs while preserving diagnostic accuracy. However, distillation in healthcare must contend with fragmented, imbalanced datasets that can undermine student robustness, and biases inherent in teachers or synthetic examples can propagate and even amplify disparities in underrepresented patient groups (Cross et al. 2024). Federated distillation offers a privacy-preserving alternative by exchanging distilled outputs rather than raw records, but it faces challenges with heterogeneous site distributions and potential information leakage (Salman et al. 2025). Moreover, the memory and compute demands of loading large teachers during offline distillation often exceed the capacities of typical hospital infrastructure (Templin et al. 2024).

## 7.2 Education

LLMs have demonstrated remarkable capabilities in automated scoring, lesson planning, and other generative tasks in education (Zhai 2022; Lee and Zhai 2024). However, their practical deployment in real-time educational settings faces significant challenges, particularly when considering resource-constrained environments such as schools with low-end devices, mobile tablets, and laptops with limited computational power (Selwyn 2019; Hinton et al. 2015). The primary obstacles include:

- **High Computational Requirements:** LLMs such as BERT-based models require significant processing power, often necessitating specialized hardware such as GPUs or TPUs. The necessity for such resources limits the accessibility of these models in standard school environments, where computational resources are minimal (Hinton et al. 2015).
- **Large Model Size:** The memory footprint of LLMs, often in the range of hundreds of megabytes, presents difficulties in deploying them on edge devices or cloud-limited school networks. Latif et al. (2024) report that their teacher LLM has approximately 114 million parameters, making direct deployment infeasible on low-resource devices.
- **Inference Latency:** The complexity of LLMs results in longer processing times, making real-time feedback and scoring challenging (Moon et al. 2010). Latency is a critical factor in automated assessment, where immediate results are needed for adaptive learning and student engagement.

To address these challenges, researchers have explored model compression techniques such as KD to create smaller yet efficient models for educational assessments (Hinton et al. 2015). The goal is to retain the accuracy and predictive capabilities of the teacher model while significantly reducing computational overhead. This is achieved by training the student model using the soft labels generated by the teacher model instead of hard class labels. Latif et al. (2024) proposes a KD-based approach specifically designed for educational assessments. Their methodology involves the utilization of a fine-tuned BERT model as the teacher model for scoring student responses in science and mathematical reasoning assessments. Their study demonstrates that the KD approach achieves a drastic reduction in model size; for example, the student model is 4,000 times smaller than the teacher model, making it feasible for deployment on low-end educational devices. The method has reduced inference time, such as the KD model, which is 10 times faster in inference compared to the teacher model, ensuring real-time scoring capabilities. Finally, the high scoring accuracy achieved by the distilled student model with a mere 2-3% accuracy reduction compared to the teacher model, while outperforming existing state-of-the-art distilled models such as TinyBERT (Jiao et al. 2019).

Other studies have also explored the distillation of LLMs specifically tailored for educational contexts. For instance, Dan et al. (2023) introduced EduChat, a large-scale LLM-based chatbot designed to facilitate intelligent education. EduChat utilizes KD to improve conversational efficiency, providing scalable and personalized educational interactions. Similarly, Qu et al. (2024) developed CourseGPT-ZH, an educational LLM incorporating KD and prompt optimization, significantly enhancing performance in pedagogical settings by efficiently distilling domain-specific knowledge into more manageable model sizes. Additionally, Baladón et al. (2023) fine-tuned open-source LLMs specifically for generating teacher-like responses, underscoring the practical applications of KD in real-world educational scenarios.

Another representative example is EduChat, developed by Dan et al. (2023), which demonstrates large-scale real-world deployment of distilled LLMs in education. EduChat is a conversational AI platform tailored for use in Chinese primary and secondary schools, supporting subjects such as Chinese, mathematics, English, physics, chemistry, and history. In its initial deployment, EduChat was piloted across over 200 schools, serving a student population of more than 50,000. The developers applied KD to shrink their base LLM by about 80 percent compared with the teacher model. Memory use dropped from more than 40 GB to less than 8 GB, allowing the chatbot to run smoothly on standard school servers and many high-end tablets. This size reduction allowed EduChat to deliver response times averaging under 1.5 seconds per query, even during peak classroom use, and to support thousands of simultaneous student sessions without performance degradation. In practice, the distilled EduChat model was able to provide real-time, curriculum-aligned explanations, quiz generation, and homework help across a variety of subjects, greatly expanding the reach of AI-powered tutoring in resource-constrained environments. However, field deployment also highlighted certain limitations: while the distilled model handled factual and procedural questions with high reliability, it sometimes struggled with more open-ended, creative, or cross-disciplinary prompts. This finding underscores the ongoing challenge of transferring deep reasoning abilities through distillation.

Beyond direct educational assessment, KD approaches have demonstrated potential in facilitating a deeper understanding of scientific concepts, which can significantly benefit education. The Darwin series, introduced by Xie et al. (2023, 2024), developed domain-specific LLMs tailored for natural sciences and materials science education, respectively. These models leverage KD to optimize domain-specific knowledge representation, improving accessibility and comprehension in educational contexts. Dagdelen et al. (2024) further advanced this idea by leveraging structured information extraction from scientific texts through LLMs, illustrating how sophisticated knowledge distillation techniques can enhance educational content delivery and learner engagement.

The success of knowledge distillation in automated educational assessment paves the way for scalable AI adoption in schools. By leveraging KD, LLM-powered assessment systems can be deployed on widely available hardware, reducing reliance on expensive cloud-based solutions. This democratizes AI-driven education, making advanced assessment tools accessible to a broader range of learners and institutions (Selwyn 2019). Future research can further enhance KD strategies by incorporating adaptive loss functions, multimodal distillation techniques, and task-specific optimizations to improve performance across diverse educational applications (Fang et al. 2025). By

refining KD methodologies, educational AI can become more practical, efficient, and equitable in real-world deployment scenarios.

**Key Lessons Learned.** Across these cases, several lessons emerge. First, KD enables the deployment of high-performing educational AI on common hardware by drastically reducing model size and inference time, democratizing access to advanced tools. Second, there is often a small but measurable drop in performance, which may affect particularly complex or nuanced tasks. This makes the choice of what to distill, and how, especially significant. Third, model distillation is not just about efficiency: by making LLMs available in more settings, it can transform how feedback and assessment are delivered, making real-time, personalized education feasible at scale. Future work should focus on minimizing knowledge loss in distillation and optimizing models for diverse educational needs, including subject-specific tasks and multimodal learning contexts ([Fang et al. 2025](#)).

### 7.3 Bioinformatics

Knowledge Distillation (KD) in bioinformatics serves as a pivotal strategy to compress large neural architectures into efficient models tailored for specific biological tasks. Broadly, these applications fit into four practical domains: protein representation learning, genomic regulatory modeling, drug-target affinity prediction, and biomedical named entity recognition. Below, we present a cohesive overview organized by usage, highlighting multiple studies per category to illustrate both scope and depth.

#### 7.3.1 Protein Representation Learning

In protein representation learning, knowledge distillation tackles the scalability challenge posed by large protein language models (PLMs). The typical pipeline starts with selecting several high-capacity PLMs, such as ESM, ProtTrans, or ProtT5, as teacher models, which separately learn rich representations of protein sequences. A compact student network is then trained using multiple objectives: a soft-label loss to replicate teacher predictions, a feature-matching loss to align intermediate activations, and occasionally, meta-learning components that adaptively weight teacher contributions based on each protein’s characteristics. For example, [Shang et al. \(2024\)](#) demonstrates that a distilled student maintained within  $\pm 1.5\%$  accuracy of the teacher ensemble while achieving 70% faster inference. [Wang et al. \(2022\)](#) in Contact-Distil translates ProtT5 embeddings into a smaller TCN-BiLSTM architecture to preserve contact map fidelity, particularly in data-scarce regimes. [Zhang et al. \(2023\)](#) introduces meta-teacher selection to dynamically influence the student based on task difficulty or protein family, boosting adaptability. Additionally, KD-MultiSucc ([Tran et al. 2025](#)) synthesizes cross-species teachers for better post-translational site prediction, showing that even highly specialized embeddings benefit from multi-teacher fusion.

#### 7.3.2 Genomic Regulatory Modeling

Regulatory genomics demands models that combine interpretability and scalability. Ensemble methods deliver high accuracy, but their computational cost is prohibitive. The distilled pipelines begin by training an ensemble of neural networks on regulatory elements—such as transcription factor binding sites or chromatin states—then distill this ensemble into a compact student by minimizing both mean-squared error for predictions and KL-divergence for ensemble uncertainty. In some designs, auxiliary uncertainty losses, like aleatoric variance, further enhance the quality of calibration. [Zhou et al. \(2024\)](#) used this strategy to retain ensemble-level confidence in a lightweight model suitable for genome-scale scans, reducing inference time by over 50%. It also included the aleatoric uncertainty estimation to the student, leading to better attribution maps and uncertainty allocation. These pipelines highlight both the efficiency gains and the interpretive clarity achievable via KD, albeit with caution—ensemble reduction occasionally dulls edge-case detection, underscoring the need for future work on preserving ensemble diversity during distillation.

#### 7.3.3 Drug–Target Affinity Prediction

In computational drug discovery, KD facilitates rapid screening of vast compound libraries. The pipeline integrates diverse teacher models—such as graph neural networks trained on drug structures, sequence-based protein encoders, or descriptor-based predictors, and uses an attention,

guided fusion loss to distill rich feature representations into a student model. The objective is calibrated to recover teacher outputs while enforcing compactness. Lu et al. (2023) showed that a student model could emulate complex multi-modal teachers with negligible accuracy loss and 60–80% faster inference. Follow-up models like FusionDTA (Yuan et al. 2022) and ViDTA (Li et al. 2024) introduced graph attention mechanisms and virtual nodes to better model drug–protein interactions during distillation. These pipelines highlight KD’s role in creating scalable, high-throughput affinity predictors, though they also reveal persistent gaps in out-of-distribution chemical generalization, motivating the need for adaptive teacher selection and domain-augmented distillation.

### 7.3.4 Biomedical Named Entity Recognition (BioNER)

BioNER systems benefit from KD through model compression and label noise mitigation. The typical pipeline starts with using a large pretrained teacher (e.g., BERT or BioBERT) to generate soft labels or to correct noisy annotations. A student model, often either a BiLSTM-CRF or a distilled TinyBERT, is then trained using a combination of standard cross-entropy and KD-guided losses. Zhou et al. (2021) refined noisy labels via teacher predictions before distilling into a compact student, leading to better recall and faster inference. Jiao et al. (2019) adapted TinyBERT for biomedical corpora, achieving transformer-level accuracies with a much smaller footprint. More recent methods like ProTeGi (Mehmood et al. 2023) extended KD to train non-transformer student models for protein–gene relation extraction, enabling deployment on edge devices.

### 7.3.5 Key Lessons Learned

Across the bioinformatics cases reviewed, several key lessons emerge. First, knowledge distillation enables the deployment of complex biological models, such as large protein language models or genomic ensembles, on standard hardware by significantly reducing model size and inference time. This facilitates broader access to high-performance tools, particularly in environments with limited computational resources. Second, while distillation often preserves overall accuracy, it can introduce subtle but important performance drops, especially in tasks involving rare entities, noisy data, or complex biological patterns. This highlights the importance of choosing appropriate teacher models and designing domain-aware distillation strategies. Third, KD is not only a tool for efficiency, but it can also reshape how biological data is processed and analyzed, enabling real-time sequence annotation, scalable genome analysis, and faster experimental feedback. Future research should prioritize minimizing information loss during distillation and tailoring models to diverse biological contexts, including low-resource species, noisy experimental conditions, and multi-omics integration.

## 8 Open Challenges and Future Directions

Despite significant advances in KD and DD, several core challenges, especially those that are unique to or significantly amplified in the context of LLMs, remain unresolved. These challenges present exciting opportunities for future research in this rapidly evolving field. In Sections 3 and 4, we discussed the limitations and future directions for specific KD and DD approaches. Here, we summarize the major open questions and challenges, aiming to provide guidance for essential and promising directions in future research.

### *Preserve the deeper context and reasoning knowledge.*

Traditional distillation methods (e.g., logit or layer matching) often fail to retain emergent large-scale capabilities such as chain-of-thought reasoning and in-context learning (Hinton et al. 2015; Gou et al. 2021). Excessive compression or selective sampling further disrupts long-range dependencies required for tasks like narrative coherence and coreference resolution (Li et al. 2024; Bender et al. 2021). These challenges are compounded by language’s inherent diversity across domains, dialects, and specialized fields (Gururangan et al. 2020). Without explicit safeguards, nuanced linguistic elements, including rare idioms, archaic phrases, or technical jargon, risk being discarded, diminishing the adaptability and domain-specific utility of distilled models compared to general-purpose LLMs (Wang et al. 2024). Future research should prioritize methodologies that explicitly

extract and preserve deeper contextual relationships and reasoning mechanisms, such as chain-of-thought dynamics and in-context learning signals (Hsieh et al. 2023; Feng et al. 2024). These strategies must also integrate domain-aware frameworks to retain linguistic diversity, from common usage to specialized vocabulary, ensuring distilled models maintain the structural and contextual fidelity of their larger counterparts.

#### ***Distillation cost.***

The exponential growth of LLMs, now reaching trillions of parameters, has amplified the computational and logistical burden of knowledge distillation. Iterative updates to these models, whether through fine-tuning or new data additions, force repeated distillation cycles that demand significant hardware resources and training time (Xu et al. 2024). At the same time, compressing massive datasets without sacrificing linguistic diversity and rare patterns remains intractable with traditional techniques, such as gradient matching or Jaccard similarity, which either scale poorly or compromise semantic fidelity (Zhao and Bilen 2021; Khan et al. 2024). Balancing efficiency with the retention of nuanced knowledge is further complicated by static, heuristic-driven approaches that cannot adapt to ever-evolving models and data. Such methods risk discarding critical context, specialized vocabulary, and domain-specific phenomena crucial for robust downstream performance (Zhou et al. 2023; Wang et al. 2024). Looking ahead, the path forward requires the development of adaptive, lightweight frameworks that unify semantic-aware data compression, such as embedding-guided subsampling, and modular knowledge extraction (Wang et al. 2025, 2024). By selectively targeting critical reasoning layers and avoiding full retraining, these strategies can reduce redundant computations and achieve real-time alignment with evolving model parameters and data distributions. Advances in dynamic trajectory alignment and hybrid generative-optimization approaches offer promising avenues for bridging efficiency gaps. Ultimately, success in distillation efficiency will hinge on designing methods that account for the sheer scale of models and corpora while preserving the depth and diversity essential to building robust, generalizable student models.

#### ***Trustworthy distillation.***

The rapid advancement of LLMs has amplified concerns about trustworthiness, including bias amplification and fairness degradation during distillation (Rejeleene et al. 2024; Augenstein et al. 2023; Liu et al. 2023). When distillation is applied, a student model may inadvertently inherit and even intensify harmful artifacts from teacher models, particularly when compression discards nuanced contextual safeguards. Massive training corpora can also embed sensitive or private information that persists through distillation, raising additional ethical and legal concerns (Zhou et al. 2024; Dong et al. 2024; Shokri et al. 2017). Hallucination is another serious challenge, as smaller models may further increase the likelihood of fabricating unfounded responses (Huang et al. 2025). A promising avenue to address some of these concerns is uncertainty quantification. By incorporating confidence-aware mechanisms (e.g., BKD described in Section 3’s uncertainty-aware KD part (Fang et al. 2024)), student models can flag high-uncertainty responses, prompting users or downstream systems to apply additional checks or override questionable results. Existing studies have also attempted to solve the trustworthiness problem by learning concentrated or reliable knowledge (Yang et al. 2024; Shum et al. 2024). Integrating these techniques with rigorous data governance frameworks, which emphasize real-time monitoring and ethical oversight, can ensure distilled models maintain accuracy, fairness, and security as LLMs evolve.

#### ***Fairness and bias amplification in distillation.***

The process of KD and DD can inadvertently amplify biases present in teacher models and training data, creating significant ethical and practical concerns for deployed systems. This challenge manifests through several distinct mechanisms that require systematic understanding and mitigation strategies. In KD, the students can amplify biased patterns as compression optimizes for aggregate rather than individual fairness. Distillation losses can incentivize shortcuts through demographic correlations while sacrificing teacher models’ subtle protection, developing less complex, more biased decision boundaries (Augenstein et al. 2024; Li et al. 2023). This issue is more severe in self-distillation, as it is trained through feedback cycles in which biased outputs train the next iteration, continually amplifying the bias (Arazo et al. 2020; Allen-Zhu and Li 2020). In DD, data selection or synthesis regularly underrepresents minority groups. Classic measures such

as gradient matching favor high-frequency patterns at the expense of the rare instances essential for equitable representation, generating datasets lacking minority instances along with required linguistic diversity for equitable behavior.

Overcoming these challenges entails fairness-aware distillation processes under the constraints of demographic parity or equalized odds. A key open question is how to set learning objectives that collectively maximize efficiency, fairness, and predictive performance without degenerating into insignificant solutions. Some promising directions involve multi-objective optimization strategies that actively control the balance between parameter compression and group-wise error gaps. Additionally, current evaluations typically focus on aggregate metrics while neglecting fairness assessments. Future research should develop comprehensive protocols measuring bias amplification and fairness degradation across demographic groups, including specialized benchmarks and monitoring systems for distilled models. There is a need to formulate sound distillation methodologies that balance fairness as they achieve efficiency gains because these frameworks operate in high-stake environments, which affect human well-being.

#### ***Dynamic evolution of teacher models and training corpora.***

The rapid evolution of LLMs introduces two intertwined challenges for distillation: the dynamic nature of teacher models (KD) and the continuous expansion of training corpora (DD). First, teacher models are frequently updated through methods such as fine-tuning, domain adaptation, or reinforcement learning. This makes it impractical to repeatedly distill from the original large-scale dataset, necessitating incremental KD techniques that integrate new capabilities (e.g., emergent reasoning skills) while preserving previously acquired knowledge. Second, training corpora themselves evolve as datasets grow to include new domains, languages, or specialized content. DD methods must, therefore, adapt to retain rare or foundational data, ensuring linguistic diversity and structural coherence without discarding critical information ([Jayasuriya et al. 2025](#); [Kirkpatrick et al. 2017](#); [Gururangan et al. 2020](#); [Gao et al. 2020](#)).

A pressing practical research question is how to develop incremental KD that effectively integrates newly emergent abilities, such as long-form reasoning, measured on benchmarks such as BIG-Bench Lite, into an existing student model without reprocessing the original training set. Important requirements include creating lightweight, sequential knowledge integration strategies. To measure such techniques, a concrete experimental setup involves: (1) distilling a 2024-trained, GPT-4-level teacher model into a smaller student (e.g., 7B parameters), and (2) updating this student exclusively using the teacher's 2025-acquired data. Retention is then measured on 2024-specific knowledge benchmarks (e.g., MMLU), while adaptation is assessed on 2025-specific capability benchmarks (e.g., new BIG-Bench Lite tasks). On the DD side, methods must retain low-resource or time-stamped data that will never re-enter future crawls. Promising approaches for balancing historical and novel knowledge include metadata-aware sampling ([He et al. 2024](#); [Liu et al. 2024](#); [Blowe et al. 2024](#)). Ultimately, a unified framework is needed to jointly optimize student model updates and data curation strategies. This would allow for the robust estimation of retained knowledge based on controlled experiments as well as ensure distilled models scale efficiently with emerging LLMs and adaptive data environments.

#### ***Non-Traditional Knowledge Distillation Frameworks.***

Transitioning beyond static one-teacher, one-student paradigms introduces challenges in stability, scalability, and knowledge coherence, while also opening new research directions. Simultaneous teacher-student training may lead to unstable convergence due to conflicting objectives, motivating the design of tailored loss functions or staged curricula ([Li et al. 2023, 2024](#); [Zhang et al. 2018](#)). A practical question is how to structure joint training protocols that promote synergy rather than interference, such as gradually exposing the student to different teacher knowledge. In self-distillation, the risk of amplifying biases from uncalibrated self-generated labels raises the need to assess mitigation techniques, including filtering low-confidence predictions and tracking fairness metrics over time ([Arazo et al. 2020](#); [Allen-Zhu and Li 2020](#)).

Multi-teacher frameworks pose distinct challenges such as managing conflicting outputs from domain-specialized models. A key question is how to evaluate and preserve reasoning fidelity in the student across diverse tasks. For instance, reasoning benchmarks like HotpotQA ([Yang et al. 2018](#)) or GSM8K ([Cobbe et al. 2021b](#)) can reveal whether the student retains step-by-step reasoning

abilities. Partial teacher availability due to privacy or deployment constraints also motivates exploration of sequential or asynchronous distillation setups. Computational overhead further increases with teacher count, encouraging research into sparse teacher selection. Finally, tracing individual teacher contributions remains difficult. Approaches such as influence-function analysis or leave-one-teacher-out ablations could help identify knowledge origins, building on early efforts like Wadhwa et al. (2025). Addressing these challenges with interpretable and efficient methods will help make non-traditional KD frameworks both robust and scalable.

#### ***Architectural Mismatch.***

A central hurdle arises from the difference in architectures between teacher and student models. Expansive Transformer-based LLMs, often optimized for long-context modeling, have distinct inductive biases that smaller or alternative architectures (e.g., non-autoregressive models or mixture-of-experts) may not share (Tay et al. 2022; Raian et al. 2024). Traditional KD methods can struggle under these conditions, as the student model may lack the structural capacity or design features to fully absorb the teacher’s capabilities (Fedus et al. 2022; Gu et al. 2017). This challenge extends to DD, where synthetic or compressed data produced for a specific LLM architecture may fail to transfer effectively to models with different configurations. Attempting to create a single distilled dataset for multiple downstream purposes, from general pretraining to specialized instruction-tuning, only amplifies this problem (Sucholutsky and Schonlau 2021; Li et al. 2023).

Moving forward, three research questions can steer progress toward architecture-aware distillation. First, what alignment metrics best quantify fidelity of transferred reasoning steps when teacher and student architectures diverge, and how can they be integrated into distillation objectives? Second, what loss functions can be re-weighted adaptively to consider a student’s shifting capacity, such that token-level imitation is not prioritized more than transferring higher-level patterns? Third, what principles should guide the generation or selection of distilled datasets so that their linguistic and structural properties remain informative across heterogeneous model families? Addressing these questions will help close the performance gap between heterogeneous teachers and students while preserving the efficiency gains that motivate distillation in the first place.

#### ***Knowledge Distillation and Foundation Agents***

Knowledge distillation is a promising approach for overcoming key deployment challenges faced by LLM-based agent systems. Agents designed for tasks such as reasoning, planning, tool manipulation, and interaction with complex environments typically require substantial computational resources (Liu et al. 2025; Li 2025). By transferring these sophisticated capabilities from large teacher models into more compact student models, knowledge distillation facilitates the development of efficient yet effective agents suitable for deployment on resource-constrained hardware. This method is especially valuable for embodied agents operating in real-time contexts, where minimizing latency directly impacts performance (Liu et al. 2024), or when deploying agents on edge devices with limited computational capacities.

A crucial consideration during distillation is selectively preserving essential elements such as planning pathways, reasoning chains, and tool-usage patterns, which collectively enable agents to perform effectively in complex tasks (Liu et al. 2025). Additionally, multi-agent systems may benefit from targeted distillation strategies that integrate specialized knowledge from multiple teacher agents (Chen et al. 2024) into a unified and efficient student model capable of versatile operations. However, one notable technical challenge is ensuring the distilled agent maintains the nuanced interplay between decision-making processes and real-world actions—elements that underpin the emergent capabilities characteristic of high-performing autonomous agents. Future research could explore hybrid distillation methods, combining traditional knowledge distillation techniques with reinforcement learning fine-tuning, thus ensuring robustness and adaptability of the distilled agents in dynamic environments.

#### ***Evaluation Gaps.***

Existing evaluation frameworks for distillation often fail to capture the diverse capabilities expected of modern LLMs. For KD, existing benchmarks prioritize narrow metrics, such as task accuracy, while overlooking critical dimensions like coherence, ethical alignment, and zero-shot generalization (Chang et al. 2024; Liu et al. 2023). In DD, common metrics like token overlap or perplexity fail

to evaluate whether distilled datasets preserve deeper reasoning abilities, such as chain-of-thought logic or emergent behaviors inherent to large-scale pertaining (Qin et al. 2025). Additionally, current distilled model evaluations lack systematic bias and fairness assessments across groups. Standard benchmarks measure only aggregate performance, ignoring whether distillation maintains equitable outcomes for underrepresented populations. This oversight is critical since compression may disproportionately impair performance on minority cases and sensitive attributes.

Addressing these gaps requires developing holistic evaluation protocols that measure nuanced LLM capabilities, including contextual reasoning, adaptability to unseen tasks, and alignment with ethical guidelines, ensuring distilled models and datasets retain the sophistication of their source systems (Dong et al. 2024; Chang et al. 2024). It is also important to incorporate evaluation frameworks with fairness-specific metrics alongside traditional performance assessments.

#### *Practical Barriers to Adoption.*

While knowledge distillation and data distillation have shown impressive empirical results, their widespread adoption remains limited, particularly among academic researchers and practitioners with constrained resources, mainly due to the following reasons.

**Computational Requirements.** Many recent KD/DD methods, especially those involving large LLMs, are developed using substantial computational infrastructure, often involving multi-GPU or TPU clusters and extended training times. For example, DistilBERT required 8 V100 GPUs running continuously for 90 hours to distill knowledge from BERT-base (Sanh et al. 2019). Vicuna training costs around \$300 for the 13B model (Chiang et al. 2023), demonstrating the financial investment required even for moderately-sized models. More efficient single-GPU approaches like T5 distillation can compress models in 16 hours on a single A100 (Nawrot 2023), but still represent significant computational investments for many research groups. Historical examples like GPT-2 distillation, with training costs of \$256 per hour, demonstrate the substantial financial barriers to reproducing state-of-the-art distillation methods. Additionally, extracting knowledge from teacher LLMs through paid APIs is costly. Building a comprehensive distilled dataset can require thousands of queries, and - depending on prompt complexity, call volume, and dataset size - the total bill can quickly rise into the thousands or even tens of thousands of dollars. In Table 7, we list some examples of KD of current LLMs, illustrating the substantial computational and financial requirements across different approaches. These requirements pose significant barriers for smaller research groups or institutions.

**Code and Data Availability.** Although a few works, such as DistilBERT, provide publicly available implementations and pretrained models, many others do not release code, model checkpoints, or distilled datasets. This lack of transparency hinders reproducibility, limits benchmarking, and increases the burden of implementation.

**Reproducibility Challenges.** Even when code is available, differences in seemingly minor aspects, such as initialization schemes, optimizer settings, or data preprocessing, can lead to large variations in performance. This is particularly true for DD methods, where the sensitivity to hyperparameters and training schedules is often underreported.

## 9 Conclusion

In this survey, we examined the intersection of Knowledge Distillation (KD) and Dataset Distillation (DD) for large language models (LLMs), demonstrating how these complementary approaches address critical challenges in computational efficiency, data scalability, and retention of advanced capabilities. Through an analysis of methodological innovations across both fields, including multi-teacher architectures, rationale-based guidance, adaptive teacher-student co-evolution, and advanced data filtering or synthesis techniques, we present a comprehensive framework for compressing model size and training data while preserving essential LLM strengths, such as contextual reasoning, cross-domain generalization, and linguistic diversity.

Despite rapid progress, several open challenges remain. Preserving emergent abilities in smaller models calls for more sophisticated approaches to capture the nuanced distributions and structural dependencies present in teacher networks. Similarly, generating or filtering distilled data to balance coverage of rare phenomena with overall efficiency demands hybrid or adaptive strategies that continuously align with evolving teacher models and tasks. Ethical and trustworthiness considerations, including mitigating biases and preventing the inheritance of harmful hallucinations, further

**Table 7:** Hardware requirements, training times for distilled models across different LLM scales.

Method	Teacher → Student	Dataset & Technique	Hardware Requirements	Training Time / Cost
<b>DistilBERT</b>	BERT-base (110M) → DistilBERT (66M)	BookCorpus + English Wikipedia • Classical knowledge distillation • Soft-logit loss + MLM loss	8× V100 16GB GPUs Very large batch sizes (4K examples)	90 hours (3.75 days)
<b>T5 Distillation</b>	T5-Large (770M) → T5-Small (60M)	C4 dataset (text-to-text tasks) • Sequence-to-sequence distillation • Multi-task learning	1× A100 GPU (efficient single-GPU setup)	16 hours (0.67 days)
<b>Alpaca-LoRA</b>	LLaMA-7B → 6-layer student	Stanford Alpaca + Evol-Instruct (52K) • LoRA/QLoRA (rank-8, 4-bit) • Parameter-efficient fine-tuning	1× RTX 4090 24GB ( $\leq 8$ GB GPU RAM usage)	6 hours (single GPU)
<b>Vicuna</b>	LLaMA-13B → Vicuna-13B	ShareGPT conversations (70K) • Supervised fine-tuning • Multi-turn dialogue training	4× A100 40GB GPUs Context length expanded to 2048	2 days estimated
<b>GPT-4 Distillation</b>	GPT-4 (1.76T est.) → Research models	API-based distillation • Reasoning model replication • Fast iteration cycles	Standard GPUs	\$450, 19 hrs

underscore the importance of carefully calibrated distillation protocols and uncertainty-aware techniques.

Looking ahead, addressing these issues will require novel evaluation frameworks that go beyond standard accuracy metrics to assess deeper language understanding, compositional reasoning, and alignment with societal values. As LLMs are deployed in ever more diverse and critical applications, the continued integration of KD and DD, together with robust checks on model safety and domain appropriateness, represents a promising avenue for sustainable, resource-efficient NLP systems. By combining theoretical insights with real-world validations, future research can help ensure that distillation remains an empowering rather than constraining force for next-generation LLMs.

## Declarations

**Funding.** This work was partially supported by the U.S. National Science Foundation (NSF) [DMS-1925066, DMS-1903226, DMS-2124493, DMS-2311297, DMS-2319279, DMS-2318809, DMS-2101104, DMS-2138854], the National Institutes of Health (NIH) [R01GM152814], and the Institute of Education Sciences (IES) [R305C240010].

**Author contribution.** The survey scope and overall structure were defined by LF and XY. All authors contributed to the literature search and manuscript writing. PM, TL, and WZ supervised the project, refined the survey outline, and provided critical revisions.

**Conflict of interest.** The authors declare that there is no Conflict of interest.

**Ethics approval and consent to participate.** Not applicable.

**Data/Code/Materials availability.** Not applicable.

## References

- Augenstein, I., Baldwin, T., Cha, M., Chakraborty, T., Ciampaglia, G.L., Corney, D., DiResta, R., Ferrara, E., Hale, S., Halevy, A., et al.: Factuality challenges in the era of large language models. arXiv preprint arXiv:2310.05189 (2023)

- Augenstein, I., Baldwin, T., Cha, M., Chakraborty, T., Ciampaglia, G.L., Corney, D., DiResta, R., Ferrara, E., Hale, S., Halevy, A., et al.: Factuality challenges in the era of large language models and opportunities for fact-checking. *Nature Machine Intelligence* **6**(8), 852–863 (2024)
- Albalak, A., Elazar, Y., Xie, S.M., Longpre, S., Lambert, N., Wang, X., Muennighoff, N., Hou, B., Pan, L., Jeong, H., et al.: A survey on data selection for language models. *Transactions on Machine Learning Research* (2024)
- Axelrod, A., He, X., Gao, J.: Domain adaptation via pseudo in-domain data selection. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pp. 355–362 (2011)
- Allen-Zhu, Z., Li, Y.: Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. *CoRR* **abs/2012.09816** (2020) [2012.09816](https://arxiv.org/abs/2012.09816)
- Arazo, E., Ortego, D., Albert, P., O'Connor, N.E., McGuinness, K.: Pseudo-labeling and confirmation bias in deep semi-supervised learning. In: 2020 International Joint Conference on Neural Networks (IJCNN), pp. 1–8 (2020). IEEE
- Abbas, A., Tirumala, K., Simig, D., Ganguli, S., Morcos, A.S.: Semdedup: Data-efficient learning at web-scale through semantic deduplication. *arXiv preprint arXiv:2303.09540* (2023)
- Allen-Zhu, Z., Li, Y.: Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. *arXiv preprint arXiv:2012.09816* (2020)
- Borup, K., Andersen, L.N.: Even your teacher needs guidance: Ground-truth targets dampen regularization imposed by self-distillation. *Advances in Neural Information Processing Systems* **34**, 5316–5327 (2021)
- Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S.: On the dangers of stochastic parrots: Can language models be too big? In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 610–623 (2021)
- Blei, D.M., Kucukelbir, A., McAuliffe, J.D.: Variational inference: A review for statisticians. *Journal of the American statistical Association* **112**(518), 859–877 (2017)
- Borsos, Z., Mutny, M., Krause, A.: Coresets via bilevel optimization for continual learning and streaming. *Advances in neural information processing systems* **33**, 14879–14890 (2020)
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. *Advances in neural information processing systems* **33**, 1877–1901 (2020)
- Binici, K., Pham, N.T., Mitra, T., Leman, K.: Preventing catastrophic forgetting and distribution mismatch in knowledge distillation via synthetic data. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 663–671 (2022)
- Baladón, A., Sastre, I., Chiruzzo, L., Rosá, A.: Retuyt-inco at bea 2023 shared task: Tuning open-source llms for generating teacher responses. In: Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023), pp. 756–765 (2023)
- Blowe, R., Vandersteen, A., Winterbourne, D., Hathersage, J., Ravenscroft, M.: Semantic entanglement modeling for knowledge distillation in large language models. *Authorea Preprints* (2024)
- Chu, Z., Chen, J., Chen, Q., Yu, W., He, T., Wang, H., Peng, W., Liu, M., Qin, B., Liu, T.: A survey of chain of thought reasoning: Advances, frontiers and future. *arXiv preprint arXiv:2309.15402* (2023)

- Cross, J.L., Choma, M.A., Onofrey, J.A.: Bias in medical ai: Implications for clinical decision-making. *PLOS Digital Health* **3**(11), 0000651 (2024)
- Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., et al.: Training verifiers to solve math word problems. arXiv preprint arXiv:2110.14168 (2021)
- Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., Schulman, J.: Training verifiers to solve math word problems. arXiv preprint arXiv:2110.14168 (2021)
- Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., et al.: Vicuna: An open-source chatbot impressing gpt-4 with 90%\* chatgpt quality. See <https://vicuna.lmsys.org> (accessed 14 April 2023) **2**(3), 6 (2023)
- Chang, X., Lee, S.Y.M., Zhu, S., Li, S., Zhou, G.: One-teacher and multiple-student knowledge distillation on sentiment classification. In: Calzolari, N., Huang, C.-R., Kim, H., Pustejovsky, J., Wanner, L., Choi, K.-S., Ryu, P.-M., Chen, H.-H., Donatelli, L., Ji, H., Kurohashi, S., Paggio, P., Xue, N., Kim, S., Hahm, Y., He, Z., Lee, T.K., Santus, E., Bond, F., Na, S.-H. (eds.) Proceedings of the 29th International Conference on Computational Linguistics, pp. 7042–7052. International Committee on Computational Linguistics, Gyeongju, Republic of Korea (2022). <https://aclanthology.org/2022.coling-1.614/>
- Chen, J.C.-Y., Saha, S., Stengel-Eskin, E., Bansal, M.: Magdi: Structured distillation of multi-agent interaction graphs improves reasoning in smaller language models. arXiv preprint arXiv:2402.01620 (2024)
- Cui, J., Wang, R., Si, S., Hsieh, C.-J.: Scaling up dataset distillation to imagenet-1k with constant memory. In: ICML, pp. 6565–6590 (2023)
- Cazenavette, G., Wang, T., Torralba, A., Efros, A.A., Zhu, J.-Y.: Dataset distillation by matching training trajectories. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4750–4759 (2022)
- Cazenavette, G., Wang, T., Torralba, A., Efros, A.A., Zhu, J.-Y.: Generalizing dataset distillation via deep generative prior. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3739–3748 (2023)
- Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., et al.: A survey on evaluation of large language models. ACM transactions on intelligent systems and technology **15**(3), 1–45 (2024)
- Dasgupta, A., Drineas, P., Harb, B., Kumar, R., Mahoney, M.W.: Sampling algorithms and coresets for  $\ell_1$ - $\ell_p$  regression. SIAM Journal on Computing **38**(5), 2060–2078 (2009)
- Dagdelen, J., Dunn, A., Lee, S., Walker, N., Rosen, A.S., Ceder, G., Persson, K.A., Jain, A.: Structured information extraction from scientific text with large language models. Nature Communications **15**(1), 1418 (2024)
- DeSalvo, G., Kagy, J.-F., Karydas, L., Rostamizadeh, A., Kumar, S.: No more hard prompts: Softsrv prompting for synthetic data generation. arXiv preprint arXiv:2410.16534 (2024)
- Dan, Y., Lei, Z., Gu, Y., Li, Y., Yin, J., Lin, J., Ye, L., Tie, Z., Zhou, Y., Wang, Y., et al.: Educhat: A large-scale language model-based chatbot system for intelligent education. arXiv preprint arXiv:2308.02773 (2023)
- Dong, Y., Mu, R., Jin, G., Qi, Y., Hu, J., Zhao, X., Meng, J., Ruan, W., Huang, X.: Building guardrails for large language models. arXiv preprint arXiv:2402.01822 (2024)

- Dong, Y., Mu, R., Zhang, Y., Sun, S., Zhang, T., Wu, C., Jin, G., Qi, Y., Hu, J., Meng, J., et al.: Safeguarding large language models: A survey. arXiv preprint arXiv:2406.02622 (2024)
- Deng, Z., Russakovsky, O.: Remember the past: Distilling datasets into addressable memories for neural networks. *Advances in Neural Information Processing Systems* **35**, 34391–34404 (2022)
- Deng, M., Wang, J., Hsieh, C.-P., Wang, Y., Guo, H., Shu, T., Song, M., Xing, E.P., Hu, Z.: Rlprompt: Optimizing discrete text prompts with reinforcement learning. arXiv preprint arXiv:2205.12548 (2022)
- Ding, M., Xu, Y., Rabbani, T., Liu, X., Gravelle, B., Ranadive, T., Tuan, T.-C., Huang, F.: Calibrated dataset condensation for faster hyperparameter search. arXiv preprint arXiv:2405.17535 (2024)
- Ding, S., Ye, J., Hu, X., Zou, N.: Distilling the knowledge from large-language model for health event prediction. arXiv preprint arXiv:2403.03242 (2024)
- Du, S., You, S., Li, X., Wu, J., Wang, F., Qian, C., Zhang, C.: Agree to disagree: Adaptive ensemble knowledge distillation in gradient space. *advances in neural information processing systems* **33**, 12345–12355 (2020)
- Elnoor, M., Weerakoon, K., Seneviratne, G., Liang, J., Rajagopal, V., Manocha, D.: Vi-LAD: Vision-Language Attention Distillation for Socially-Aware Robot Navigation in Dynamic Environments (2025). <https://arxiv.org/abs/2503.09820>
- Fang, L., Chen, Y., Zhong, W., Ma, P.: Bayesian knowledge distillation: A bayesian perspective of distillation with uncertainty quantification. In: Forty-first International Conference on Machine Learning (2024)
- Fang, L., Latif, E., Lu, H., Zhou, Y., Ma, P., Zhai, X.: Efficient multi-task inferencing: Model merging with gromov-wasserstein feature alignment. arXiv preprint arXiv:2503.09774 (2025)
- Feng, K., Li, C., Zhang, X., Zhou, J., Yuan, Y., Wang, G.: Keypoint-based progressive chain-of-thought distillation for llms. arXiv preprint arXiv:2405.16064 (2024)
- Feng, B., Mei, Z., Li, B., Ost, J., Ghilotti, F., Girgis, R., Majumdar, A., Heide, F.: VERDI: VLM-Embedded Reasoning for Autonomous Driving (2025). <https://arxiv.org/abs/2505.15925>
- Fang, L., Meng, C., Zhao, L., Wang, T., Liu, T., Zhong, W., Ma, P.: Spot: An active learning algorithm for efficient deep neural network training. *Big Data Mining and Analytics* **8**(5), 1060–1074 (2025)
- Fukuda, T., Suzuki, M., Kurata, G., Thomas, S., Cui, J., Ramabhadran, B.: Efficient knowledge distillation from an ensemble of teachers. In: Interspeech, pp. 3697–3701 (2017)
- Fedus, W., Zoph, B., Shazeer, N.: Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research* **23**(120), 1–39 (2022)
- Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., et al.: The pile: An 800gb dataset of diverse text for language modeling. arXiv preprint arXiv:2101.00027 (2020)
- Gu, J., Bradbury, J., Xiong, C., Li, V.O., Socher, R.: Non-autoregressive neural machine translation. arXiv preprint arXiv:1711.02281 (2017)
- Ge, H., Hao, L., Xu, Z., et al.: Clinkd: Cross-modal clinic knowledge distiller for multi-task medical images. arXiv preprint arXiv:2503.01034 (2025)
- Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., Smith, N.A.: Don't stop pretraining: Adapt language models to domains and tasks. arXiv preprint

arXiv:2004.10964 (2020)

- Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International Conference on Machine Learning, pp. 1321–1330 (2017). PMLR
- Gou, J., Yu, B., Maybank, S.J., Tao, D.: Knowledge distillation: A survey. International Journal of Computer Vision **129**(6), 1789–1819 (2021)
- Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025)
- Ghorbani, A., Zou, J.: Data shapley: Equitable valuation of data for machine learning. In: International Conference on Machine Learning, pp. 2242–2251 (2019). PMLR
- Gu, Y., Zhang, S., Usuyama, N., et al.: Distilling large language models for biomedical knowledge extraction: A case study on adverse drug events. arXiv preprint arXiv:2311.09602 (2023)
- Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., Steinhardt, J.: Measuring mathematical problem solving with the math dataset. arXiv preprint arXiv:2103.03874 (2021)
- He, J., Guo, H., Zhu, K., Zhao, Z., Tang, M., Wang, J.: Seekr: Selective attention-guided knowledge retention for continual learning of large language models. arXiv preprint arXiv:2411.06171 (2024)
- Hu, L., Huang, T., Xie, H., Gong, X., Ren, C., Hu, Z., Yu, L., Ma, P., Wang, D.: Semi-supervised Concept Bottleneck Models (2025). <https://arxiv.org/abs/2406.18992>
- Hsu, D., Ji, Z., Telgarsky, M., Wang, L.: Generalization bounds via distillation. arXiv preprint arXiv:2104.05641 (2021)
- Hsieh, C.-Y., Li, C.-L., Yeh, C.-K., Nakhost, H., Fujii, Y., Ratner, A., Krishna, R., Lee, C.-Y., Pfister, T.: Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. arXiv preprint arXiv:2305.02301 (2023)
- He, X., Nassar, I., Kiros, J., Haffari, G., Norouzi, M.: Generate, annotate, and learn: Nlp with synthetic text. Transactions of the Association for Computational Linguistics **10**, 826–842 (2022)
- Hadi, M.U., Qureshi, R., Shah, A., Irfan, M., Zafar, A., Shaikh, M.B., Akhtar, N., Wu, J., Mirjalili, S., et al.: A survey on large language models: Applications, challenges, limitations, and practical usage. Authorea Preprints **3** (2023)
- Hasan, M.J., Rahman, F., Mohammed, N.: Optimclm: Optimizing clinical language models for predicting patient outcomes. arXiv preprint arXiv:2403.02559 (2024)
- Hinton, G., Vinyals, O., Dean, J., et al.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
- Hegde, D., Yasarla, R., Cai, H., Han, S., Bhattacharyya, A., Mahajan, S., Liu, L., Garrepalli, R., Patel, V.M., Porikli, F.: Distilling Multi-modal Large Language Models for Autonomous Driving (2025). <https://arxiv.org/abs/2501.09757>
- Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., et al.: A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Transactions on Information Systems **43**(2), 1–55 (2025)
- Jumper, J., Evans, R., Pritzel, A., et al.: Highly accurate protein structure prediction with alphafold. Nature **596**, 583–589 (2021) <https://doi.org/10.1038/s41586-021-03819-2>
- Jayasuriya, D., Tayebati, S., Ettori, D., Krishnan, R., Trivedi, A.R.: Sparc: Subspace-aware prompt

- adaptation for robust continual learning in llms. arXiv preprint arXiv:2502.02909 (2025)
- Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., Wang, F., Liu, Q.: Tinybert: Distilling bert for natural language understanding. arXiv preprint arXiv:1909.10351 (2019)
- Korattikara Balan, A., Rathod, V., Murphy, K.P., Welling, M.: Bayesian dark knowledge. Advances in neural information processing systems **28** (2015)
- Khanuja, S., Johnson, M., Talukdar, P.: Mergedistill: Merging pre-trained language models using distillation. arXiv preprint arXiv:2106.02834 (2021)
- Kim, J.-H., Kim, J., Oh, S.J., Yun, S., Song, H., Jeong, J., Ha, J.-W., Song, H.O.: Dataset condensation via efficient synthetic-data parameterization. In: ICML (2022)
- Kong, X., Li, L., Dou, M., Chen, Z., Wu, Y., Guo, G.: Quantum-enhanced llm efficient fine tuning. arXiv preprint arXiv:2503.12790 (2025)
- Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., *et al.*: Overcoming catastrophic forgetting in neural networks. Proceedings of the national academy of sciences **114**(13), 3521–3526 (2017)
- Keswani, M., Ramakrishnan, S., Reddy, N., Balasubramanian, V.N.: Proto2Proto: Can you recognize the car, the way I do? (2022). <https://arxiv.org/abs/2204.11830>
- Kang, I., Ram, P., Zhou, Y., Samulowitz, H., Seneviratne, O.: On learning representations for tabular data distillation. arXiv preprint arXiv:2501.13905 (2025)
- Khan, A., Underwood, R., Siebenschuh, C., Babuji, Y., Ajith, A., Hippe, K., Gokdemir, O., Brace, A., Chard, K., Foster, I.: Lshbloom: Memory-efficient, extreme-scale document deduplication. arXiv preprint arXiv:2411.04257 (2024)
- Kanagavelu, R., Walia, M., Wang, Y., Fu, H., Wei, Q., Liu, Y., Goh, R.S.M.: Medsynth: Leveraging generative model for healthcare data sharing. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 654–664 (2024). Springer
- Li, J., Beeching, E., Tunstall, L., Lipkin, B., Soletskyi, R., Huang, S., Rasul, K.e.a.: Numinamath: The largest public dataset in ai4maths with 860k pairs of competition math problems and solutions. Hugging Face repository **13** (2024)
- Liu, J., Cui, L., Liu, H., Huang, D., Wang, Y., Zhang, Y.: Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. arXiv preprint arXiv:2007.08124 (2020)
- Li, L., Dong, P., Li, A., Wei, Z., Yang, Y.: Kd-zero: Evolving knowledge distiller for any teacher-student pairs. Advances in Neural Information Processing Systems **36**, 69490–69504 (2023)
- Li, Y., Du, M., Song, R., Wang, X., Wang, Y.: A survey on fairness in large language models. arXiv preprint arXiv:2308.10149 (2023)
- Latif, E., Fang, L., Ma, P., Zhai, X.: Knowledge distillation of llms for automatic scoring of science assessments. In: International Conference on Artificial Intelligence in Education, pp. 166–174 (2024). Springer
- Liu, Y., Gu, J., Wang, K., Zhu, Z., Jiang, W., You, Y.: Dream: Efficient dataset distillation by representative matching. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 17314–17324 (2023)
- Li, M., Guo, Z., Wu, Y., Guo, P., Shi, Y., Hu, S., Wan, W., Hu, S.: Vidta: Enhanced drug-target affinity prediction via virtual graph nodes and attention-based feature fusion. In: 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 42–47 (2024). IEEE

- Loo, N., Hasani, R., Amini, A., Rus, D.: Efficient dataset distillation using random feature approximation. *Advances in Neural Information Processing Systems* **35**, 13877–13891 (2022)
- Li, Y., Han, S.C., Dai, Y., Cao, F.: Chulo: Chunk-level key information representation for long document processing. arXiv preprint arXiv:2410.11119 (2024)
- Li, X., He, S., Wu, J., Yang, Z., Xu, Y., Jun, Y., Liu, H., Liu, K., Zhao, J.: Mode-cotd: Chain-of-thought distillation for complex reasoning tasks with mixture of decoupled lora-experts. In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pp. 11475–11485 (2024)
- Li, X.: A review of prominent paradigms for llm-based agents: Tool use, planning (including rag), and feedback learning. In: Proceedings of the 31st International Conference on Computational Linguistics, pp. 9760–9779 (2025)
- Liu, C., Kang, Y., Zhao, F., Kuang, K., Jiang, Z., Sun, C., Wu, F.: Evolving knowledge distillation with large language models and active learning. arXiv preprint arXiv:2403.06414 (2024)
- Liu, Z., Liu, L.: Graphsaid: Graph sampling via attention based integer programming method. In: Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems, pp. 1256–1264 (2024)
- Lee, H.B., Lee, D.B., Hwang, S.J.: Dataset condensation with latent space knowledge factorization and sharing. arXiv preprint arXiv:2208.10494 (2022)
- Liu, Z., Liu, Q., Li, Y., Liu, L., Shrivastava, A., Bi, S., Hong, L., Chi, E.H., Zhao, Z.: Wisdom of committee: Distilling from foundation model to specialized application model. arXiv preprint arXiv:2402.14035 (2024)
- Liu, W., Liu, P., Ma, J.: DSDrive: Distilling Large Language Model for Lightweight End-to-End Autonomous Driving with Unified Reasoning and Planning (2025). <https://arxiv.org/abs/2505.05360>
- Li, L., Li, G., Togo, R., Maeda, K., Ogawa, T., Haseyama, M.: Generative dataset distillation: Balancing global structure and local details. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7664–7671 (2024)
- Li, X., Lin, L., Wang, S., Qian, C.: Unlock the power: Competitive distillation for multi-modal large language models. arXiv preprint arXiv:2311.08213 (2023)
- Liu, B., Li, X., Zhang, J., Wang, J., He, T., Hong, S., Liu, H., Zhang, S., Song, K., Zhu, K., et al.: Advances and challenges in foundation agents: From brain-inspired intelligence to evolutionary, collaborative, and safe systems. arXiv preprint arXiv:2504.01990 (2025)
- Li, T., Meng, C.: Modern subsampling methods for large-scale least squares regression. *International Journal of Cyber-Physical Systems (IJCPS)* **2**(2), 1–28 (2020)
- Lopez-Paz, D., Bottou, L., Schölkopf, B., Vapnik, V.: Unifying distillation and privileged information. arXiv preprint arXiv:1511.03643 (2015)
- Lu, R., Wang, J., Li, P., Li, Y., Tan, S., Pan, Y., Liu, H., Gao, P., Xie, G., Yao, X.: Improving drug-target affinity prediction via feature fusion and knowledge distillation. *Briefings in Bioinformatics* **24**(3), 145 (2023)
- Liu, Y., Wu, Z., Lu, Z., Nie, C., Wen, G., Hu, P., Zhu, X.: Noisy node classification by bi-level optimization based multi-teacher distillation. arXiv preprint arXiv:2404.17875 (2024)
- Liu, S., Wang, K., Yang, X., Ye, J., Wang, X.: Dataset distillation via factorization. *Advances in neural information processing systems* **35**, 1100–1113 (2022)

- Li, M., Yu, J., Li, T., Meng, C.: Core-elements for classical linear regression. arXiv preprint arXiv:2206.10240 **1398**, 1399–1400 (2023)
- Li, M., Yu, J., Li, T., Meng, C.: Core-elements for large-scale least squares estimation. *Statistics and Computing* **34**(6), 190 (2024)
- Liu, Y., Yao, Y., Ton, J.-F., Zhang, X., Guo, R., Cheng, H., Klochkov, Y., Taufiq, M.F., Li, H.: Trustworthy llms: a survey and guideline for evaluating large language models' alignment. arXiv preprint arXiv:2308.05374 (2023)
- Lee, G.-G., Zhai, X.: Using chatgpt for science learning: A study on pre-service teachers' lesson planning. *IEEE Transactions on Learning Technologies* (2024)
- Lee, S., Zhou, J., Ao, C., Li, K., Du, X., He, S., Liu, J., Yang, M., Wen, Z., Ni, S.: Distillation quantification for large language models. arXiv preprint arXiv:2501.12619 (2025)
- Latif, E., Zhou, Y., Guo, S., Gao, Y., Shi, L., Nayaaba, M., Lee, G., Zhang, L., Bewersdorff, A., Fang, L., et al.: A systematic assessment of openai o1-preview for higher order thinking in education. arXiv preprint arXiv:2410.21287 (2024)
- Liu, W., Zeng, W., He, K., Jiang, Y., He, J.: What makes good data for alignment? a comprehensive study of automatic data selection in instruction tuning. In: The Twelfth International Conference on Learning Representations
- Liu, S., Zhang, B., Huang, Z.: Benchmark real-time adaptation and communication capabilities of embodied agent in collaborative scenarios. arXiv preprint arXiv:2412.00435 (2024)
- Li, Z., Zhu, H., Lu, Z., Yin, M.: Synthetic data generation with large language models for text classification: Potential and limitations. arXiv preprint arXiv:2310.07849 (2023)
- Li, M., Zhou, F., Song, X.: Bild: Bi-directional logits difference loss for large language model distillation. arXiv preprint arXiv:2406.13555 (2024)
- MacKay, D.J.: A practical bayesian framework for backpropagation networks. *Neural computation* **4**(3), 448–472 (1992)
- Mirzasoleiman, B., Cao, K., Leskovec, J.: Coresets for robust training of deep neural networks against noisy labels. *Advances in Neural Information Processing Systems* **33**, 11465–11477 (2020)
- Ma, P., Chen, Y., Zhang, X., Xing, X., Ma, J., Mahoney, M.W.: Asymptotic analysis of sampling estimators for randomized numerical linear algebra algorithms. *Journal of Machine Learning Research* **23**(177), 1–45 (2022)
- Mobahi, H., Farajtabar, M., Bartlett, P.: Self-distillation amplifies regularization in Hilbert space. *Advances in Neural Information Processing Systems* **33**, 3351–3361 (2020)
- Mirzadeh, S.I., Farajtabar, M., Li, A., Levine, N., Matsukawa, A., Ghasemzadeh, H.: Improved knowledge distillation via teacher assistant. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 5191–5198 (2020)
- Mehmood, T., Gerevini, A.E., Lavelli, A., Olivato, M., Serina, I.: Distilling knowledge with a teacher's multitask model for biomedical named entity recognition. *Information* **14**(5), 255 (2023)
- Maekawa, A., Kosugi, S., Funakoshi, K., Okumura, M.: Dilm: distilling dataset into language model for text-level dataset distillation. arXiv preprint arXiv:2404.00264 (2024)
- Müller, R., Kornblith, S., Hinton, G.E.: When does label smoothing help? *Advances in neural information processing systems* **32** (2019)

- Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W.-t., Koh, P.W., Iyyer, M., Zettlemoyer, L., Hajishirzi, H.: Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. arXiv preprint arXiv:2305.14251 (2023)
- Moore, R.C., Lewis, W.: Intelligent selection of language model training data. In: Proceedings of the ACL 2010 Conference Short Papers, pp. 220–224 (2010)
- Moon, T., Li, L., Chu, W., Liao, C., Zheng, Z., Chang, Y.: Online learning for recency search ranking using real-time user feedback. In: Proceedings of the 19th ACM International Conference on Information and Knowledge Management, pp. 1501–1504 (2010)
- Michel, P., Levy, O., Neubig, G.: Are sixteen heads really better than one? Advances in neural information processing systems **32** (2019)
- Malinin, A., Mlodozeniec, B., Gales, M.: Ensemble distribution distillation. arXiv preprint arXiv:1905.00076 (2019)
- Ma, P., Mahoney, M.W., Yu, B.: A statistical perspective on algorithmic leveraging. The Journal of Machine Learning Research **16**(1), 861–911 (2015)
- Menon, A.K., Rawat, A.S., Reddi, S., Kim, S., Kumar, S.: A statistical perspective on distillation. In: International Conference on Machine Learning, pp. 7632–7642 (2021). PMLR
- Marion, M., Üstün, A., Pozzobon, L., Wang, A., Fadaee, M., Hooker, S.: When less is more: Investigating data pruning for pretraining llms at scale. arXiv preprint arXiv:2309.04564 (2023)
- Meng, C., Wang, Y., Zhang, X., Mandal, A., Zhong, W., Ma, P.: Effective statistical methods for big data analytics. In: Handbook of Research on Applied Cybernetics and Systems Science, pp. 280–299. IGI Global Scientific Publishing, Hershey, PA, USA (2017)
- Meng, C., Zhang, X., Zhang, J., Zhong, W., Ma, P.: More efficient approximation of smoothing splines via space-filling basis selection. Biometrika **107**(3), 723–735 (2020)
- Nawrot, P.: nanoT5: Fast & simple pre-training and fine-tuning of T5 models with limited resources. In: Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023). Association for Computational Linguistics, Singapore (2023). <https://doi.org/10.18653/v1/2023.nlposs-1.11> . <https://aclanthology.org/2023.nlposs-1.11>
- Nievas, M., Basu, A., Wang, Y., Singh, H.: Distilling large language models for matching patients to clinical trials. Journal of the American Medical Informatics Association **31**(9), 1953–1963 (2024)
- Niu, Y., Chen, L., Zhou, C., Zhang, H.: Respecting transfer gap in knowledge distillation. Advances in Neural Information Processing Systems **35**, 21933–21947 (2022)
- Niu, S., Ma, J., Bai, L., Wang, Z., Xu, Y., Song, Y., Yang, X.: Multimodal clinical reasoning through knowledge-augmented rationale generation. arXiv preprint arXiv:2411.07611 (2024)
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., *et al.*: Training language models to follow instructions with human feedback. Advances in neural information processing systems **35**, 27730–27744 (2022)
- Prabhakar, S.N., Deshwal, A., Mishra, R., Kim, H.: Distilnas: neural architecture search with distilled data. IEEE Access **10**, 124990–124998 (2022)
- Pryzant, R., Iter, D., Li, J., Lee, Y.T., Zhu, C., Zeng, M.: Automatic prompt optimization with “gradient descent” and beam search. arXiv preprint arXiv:2305.03495 (2023)
- Park, W., Kim, D., Lu, Y., Cho, M.: Relational knowledge distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3967–3976 (2019)

- Phuong, M., Lampert, C.: Towards understanding knowledge distillation. In: International Conference on Machine Learning, pp. 5142–5151 (2019). PMLR
- Penedo, G., Malartic, Q., Hesslow, D., Cojocaru, R., Alobeidli, H., Cappelli, A., Pannier, B., Almazrouei, E., Launay, J.: The refinedweb dataset for falcon llm: Outperforming curated corpora with web data only. Advances in Neural Information Processing Systems **36**, 79155–79172 (2023)
- Pillutla, K., Swayamdipta, S., Zellers, R., Thickstun, J., Welleck, S., Choi, Y., Harchaoui, Z.: Mauve: Measuring the gap between neural text and human text using divergence frontiers. Advances in Neural Information Processing Systems **34**, 4816–4828 (2021)
- Pan, C., Yaman, B., Nesti, T., Mallik, A., Allievi, A.G., Velipasalar, S., Ren, L.: Vlp: Vision language planning for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14760–14769 (2024)
- Qu, G., Chen, Q., Wei, W., Lin, Z., Chen, X., Huang, K.: Mobile edge intelligence for large language models: A contemporary survey. IEEE Communications Surveys & Tutorials (2025)
- Qin, Z., Dong, Q., Zhang, X., Dong, L., Huang, X., Yang, Z., Khademi, M., Zhang, D., Awadalla, H.H., Fung, Y.R., et al.: Scaling laws of synthetic data for language models. arXiv preprint arXiv:2503.19551 (2025)
- Qu, Z., Yin, L., Yu, Z., Wang, W., et al.: Coursegpt-zh: An educational large language model based on knowledge distillation incorporating prompt optimization. arXiv preprint arXiv:2405.04781 (2024)
- Ramamurthy, P., Aakur, S.N.: Isd-qa: Iterative distillation of commonsense knowledge from general language models for unsupervised question answering. In: 2022 26th International Conference on Pattern Recognition (ICPR), pp. 1229–1235 (2022). <https://doi.org/10.1109/ICPR56361.2022.9956441>
- Rae, J.W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, F., Aslanides, J., Henderson, S., Ring, R., Young, S., et al.: Scaling language models: Methods, analysis & insights from training gopher. arXiv preprint arXiv:2112.11446 (2021)
- Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: Fitnets: Hints for thin deep nets. arXiv preprint arXiv:1412.6550 (2014)
- Raiaan, M.A.K., Mukta, M.S.H., Fatema, K., Fahad, N.M., Sakib, S., Mim, M.M.J., Ahmad, J., Ali, M.E., Azam, S.: A review on large language models: Architectures, applications, taxonomies, open issues and challenges. IEEE access **12**, 26839–26874 (2024)
- Rejeleene, R., Xu, X., Talbert, J.: Towards trustable language models: Investigating information quality of large language models. arXiv preprint arXiv:2401.13086 (2024)
- Sun, S., Cheng, Y., Gan, Z., Liu, J.: Patient knowledge distillation for bert model compression. arXiv preprint arXiv:1908.09355 (2019)
- Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108 (2019)
- Selwyn, N.: Should Robots Replace Teachers?: AI and the Future of Education. John Wiley & Sons, Hoboken, NJ, USA (2019)
- Shi, H., Gao, J., Ren, X., Xu, H., Liang, X., Li, Z., Kwok, J.T.: SparseBERT: Rethinking the Importance Analysis in Self-attention (2021). <https://arxiv.org/abs/2102.12871>
- Sorscher, B., Geirhos, R., Shekhar, S., Ganguli, S., Morcos, A.: Beyond neural scaling laws: beating

power law scaling via data pruning. *Advances in Neural Information Processing Systems* **35**, 19523–19536 (2022)

Sun, L., Gou, J., Yu, B., Du, L., Tao, D.: Collaborative teacher-student learning via multiple knowledge transfer. arXiv preprint arXiv:2101.08471 (2021)

Stanton, S., Izmailov, P., Kirichenko, P., Alemi, A.A., Wilson, A.G.: Does knowledge distillation really work? *Advances in neural information processing systems* **34**, 6906–6919 (2021)

Sajedi, A., Khaki, S., Liu, L.Z., Amjadian, E., Lawryshyn, Y.A., Plataniotis, K.N.: Data-to-model distillation: Data-efficient learning framework. In: European Conference on Computer Vision, pp. 438–457 (2024). Springer

Song, R., Liu, D., Chen, D.Z., Festag, A., Trinitis, C., Schulz, M., Knoll, A.: Federated learning via decentralized dataset distillation in resource-constrained edge environments. In: 2023 International Joint Conference on Neural Networks (IJCNN), pp. 1–10 (2023). IEEE

Sachdeva, N., McAuley, J.: Data distillation: A survey. *Transactions on Machine Learning Research* (2023)

Shang, J., Peng, C., Ji, Y., Guan, J., Cai, D., Tang, X., Sun, Y.: Accurate and efficient protein embedding using multi-teacher distillation learning. arXiv preprint arXiv:2405.11735 (2024)

Sucholutsky, I., Schonlau, M.: Soft-label dataset distillation and text dataset distillation. In: 2021 International Joint Conference on Neural Networks (IJCNN), pp. 1–8 (2021). IEEE

Shin, D., Shin, S., Moon, I.-C.: Frequency domain-based dataset distillation. *Advances in Neural Information Processing Systems* **36**, 70033–70044 (2023)

Shridhar, K., Stolfo, A., Sachan, M.: Distilling reasoning capabilities into smaller language models. *Findings of the Association for Computational Linguistics: ACL 2023*, 7059–7073 (2023)

Shokri, R., Stronati, M., Song, C., Shmatikov, V.: Membership inference attacks against machine learning models. In: 2017 IEEE Symposium on Security and Privacy (SP), pp. 3–18 (2017). IEEE

Sutanto, P., Santoso, J., Setiawan, E.I., Wibawa, A.P.: Llm distillation for efficient few-shot multiple choice question answering. arXiv preprint arXiv:2403.00633 (2024)

Sun, P., Shi, B., Yu, D., Lin, T.: On the diversity and realism of distilled dataset: An efficient dataset distillation paradigm. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9390–9399 (2024)

Shum, K., Xu, M., Zhang, J., Chen, Z., Diao, S., Dong, H., Zhang, J., Raza, M.O.: First: Teach a reliable large language model through efficient trustworthy distillation. arXiv preprint arXiv:2408.12168 (2024)

Shao, S., Yin, Z., Zhou, M., Zhang, X., Shen, Z.: Generalized large-scale data condensation via various backbone and statistical matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16709–16718 (2024)

Salman, H., Zaki, C., Charara, N., Guehis, S., Pradat-Peyre, J.-F., Nasser, A.: Knowledge distillation in federated learning: a comprehensive survey. *Discover Computing* **28**(1), 145 (2025)

Sun, X., Zhong, W., Ma, P.: An asymptotic and empirical smoothing parameters selection method for smoothing spline ANOVA models in large samples. *Biometrika* **108**(1), 149–166 (2020) <https://doi.org/10.1093/biomet/asaa047> <https://academic.oup.com/biomet/article-pdf/108/1/149/36441097/asaa047.pdf>

Tay, Y., Dehghani, M., Bahri, D., Metzler, D.: Efficient transformers: A survey. *ACM Computing Surveys* **55**(6), 1–28 (2022)

- Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., Hashimoto, T.B.: Alpaca: A strong, replicable instruction-following model. Stanford Center for Research on Foundation Models. <https://crfm.stanford.edu/2023/03/13/alpaca.html> **3**(6), 7 (2023)
- Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., Hashimoto, T.B.: Stanford alpaca: An instruction-following llama model. Stanford, CA, USA (2023)
- Tian, Y., Han, Y., Chen, X., Wang, W., Chawla, N.V.: Beyond answers: Transferring reasoning capabilities to smaller llms using multi-teacher knowledge distillation. arXiv preprint arXiv:2402.04616 (2024)
- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
- Tian, K., Mitchell, E., Zhou, A., Sharma, A., Rafailov, R., Yao, H., Finn, C., Manning, C.D.: Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. arXiv preprint arXiv:2305.14975 (2023)
- Tran, T.-X., Nguyen, T.-T., Le, N.-Q.-K., Nguyen, V.-N.: Kd\_multisucc: incorporating multi-teacher knowledge distillation and word embeddings for cross-species prediction of protein succinylation sites. Biology Methods and Protocols **10**(1), 041 (2025)
- Tariq, A., Purkayastha, S., Dai, T.: Patient centric summarization of radiology findings using large language models. arXiv preprint arXiv:2402.01562 (2024)
- Templin, T., Perez, M.W., Sylvia, S., Leek, J., Sinnott-Armstrong, N.: Addressing 6 challenges in generative ai for digital health: a scoping review. PLOS digital health **3**(5), 0000503 (2024)
- Tan, H., Wu, S., Huang, W., Zhao, S., Qi, X.: Data pruning by information maximization. In: The Thirteenth International Conference on Learning Representations (2025)
- Vedula, K.S., Gupta, A., Swaminathan, A., et al.: Distilling large language models for efficient clinical information extraction. arXiv preprint arXiv:2403.04138 (2024)
- Vapnik, V., Izmailov, R., et al.: Learning using privileged information: similarity control and knowledge transfer. J. Mach. Learn. Res. **16**(1), 2023–2049 (2015)
- Vadera, M., Jalaian, B., Marlin, B.: Generalized bayesian posterior expectation distillation for deep neural networks. In: Conference on Uncertainty in Artificial Intelligence, pp. 719–728 (2020). PMLR
- Wayve: Lingo-1: Exploring natural language for autonomous driving. Wayve Technical Report (2023)
- Wei, J., Bosma, M., Zhao, V.Y., Guu, K., Yu, A.W., Lester, B., Du, N., Dai, A.M., Le, Q.V.: Finetuned language models are zero-shot learners. arXiv preprint arXiv:2109.01652 (2021)
- Wu, S., Cheng, H., Cai, J., Ma, P., Zhong, W.: Subsampling in large graphs using ricci curvature. In: International Conference on Learning Representations (2023)
- Wang, W., Chen, W., Luo, Y., Long, Y., Lin, Z., Zhang, L., Lin, B., Cai, D., He, X.: Model compression and efficient inference for large language models: A survey. arXiv preprint arXiv:2402.09748 (2024)
- Wang, Y., Chen, Z., Yang, D., Sun, Y., Qi, L.: Self-Cooperation Knowledge Distillation for Novel Class Discovery (2024). <https://arxiv.org/abs/2407.01930>
- Wang, Q., Chen, J., Zhou, Y., Li, Y., Zheng, L., Wang, S., Li, Z., Cui, S.: Contact-distil: Boosting low homologous protein contact map prediction by self-supervised distillation. In: Proceedings

of the AAAI Conference on Artificial Intelligence, vol. 36, pp. 4620–4627 (2022)

Wang, J., Hu, X., Hou, W., Chen, H., Zheng, R., Wang, Y., Yang, L., Huang, H., Ye, W., Geng, X., et al.: On the robustness of chatgpt: An adversarial and out-of-distribution perspective. arXiv preprint arXiv:2302.12095 (2023)

Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N.A., Khashabi, D., Hajishirzi, H.: Self-instruct: Aligning language models with self-generated instructions. arXiv preprint arXiv:2212.10560 (2022)

Wenzek, G., Lachaux, M.-A., Conneau, A., Chaudhary, V., Guzmán, F., Joulin, A., Grave, É.: Ccnet: Extracting high quality monolingual datasets from web crawl data. In: Proceedings of the Twelfth Language Resources and Evaluation Conference, pp. 4003–4012 (2020)

Wu, F., Li, Z., Li, Y., Ding, B., Gao, J.: Fedbiot: Llm local fine-tuning in federated learning without full model. Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 3345–3355 (2024)

Wang, J.T., Mittal, P., Song, D., Jia, R.: Data shapley in one training run. In: International Conference on Learning Representations (2025)

Wilkins, J., Rodriguez, M.: Higher performance of mistral large on mmlu benchmark through two-stage knowledge distillation (2024)

Wadhwa, S., Shaib, C., Amir, S., Wallace, B.C.: Who taught you that? tracing teachers in model distillation. arXiv preprint arXiv:2502.06659 (2025)

Wang, J.T., Song, D., Zou, J., Mittal, P., Jia, R.: Capturing the temporal dependence of training data influence. In: International Conference on Learning Representations (2025)

Welling, M., Teh, Y.W.: Bayesian learning via stochastic gradient langevin dynamics. In: Proceedings of the 28th International Conference on Machine Learning (ICML-11), pp. 681–688 (2011)

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., et al.: Emergent abilities of large language models. arXiv preprint arXiv:2206.07682 (2022)

Wu, M., Waheed, A., Zhang, C., Abdul-Mageed, M., Aji, A.F.: Lamini-lm: A diverse herd of distilled models from large-scale instructions. arXiv preprint arXiv:2304.14402 (2023)

Wang, Y., Xu, C., Sun, Q., Hu, H., Tao, C., Geng, X., Jiang, D.: Promda: Prompt-based data augmentation for low-resource nlu tasks. arXiv preprint arXiv:2202.12499 (2022)

Wang, B., Xu, C., Wang, S., Gan, Z., Cheng, Y., Gao, J., Awadallah, A.H., Li, B.: Adversarial glue: A multi-task benchmark for robustness evaluation of language models. arXiv preprint arXiv:2111.02840 (2021)

Wu, O., Yao, R.: Data optimization in deep learning: A survey. IEEE Transactions on Knowledge and Data Engineering (2025)

Wang, J.T., Yang, T., Zou, J., Kwon, Y., Jia, R.: Rethinking data shapley for data selection tasks: Misleads and merits. In: International Conference on Machine Learning (2024)

Wang, T., Zhu, J.-Y., Torralba, A., Efros, A.A.: Dataset distillation. arXiv preprint arXiv:1811.10959 (2018)

Wang, F., Zhang, Z., Zhang, X., Wu, Z., Mo, T., Lu, Q., Wang, W., Li, R., Xu, J., Tang, X., et al.: A comprehensive survey of small language models in the era of large language models: Techniques, enhancements, applications, collaboration with llms, and trustworthiness. arXiv

preprint arXiv:2411.03350 (2024)

Xiao, L., He, Y.: Are large-scale soft labels necessary for large-scale dataset distillation? arXiv preprint arXiv:2410.15919 (2024)

Xu, X., Li, M., Tao, C., Shen, T., Cheng, R., Li, J., Xu, C., Tao, D., Zhou, T.: A survey on knowledge distillation of large language models. arXiv preprint arXiv:2402.13116 (2024)

Xia, M., Malladi, S., Gururangan, S., Arora, S., Chen, D.: Less: Selecting influential data for targeted instruction tuning. In: International Conference on Machine Learning, pp. 54104–54132 (2024). PMLR

Xie, T., Wan, Y., Huang, W., Yin, Z., Liu, Y., Wang, S., Linghu, Q., Kit, C., Grazian, C., Zhang, W., et al.: Darwin series: Domain specific large language models for natural science. arXiv preprint arXiv:2308.13565 (2023)

Xie, T., Wan, Y., Liu, Y., Zeng, Y., Wang, S., Zhang, W., Grazian, C., Kit, C., Ouyang, W., Zhou, D., et al.: Darwin 1.5: Large language models as materials science adapted learners. arXiv preprint arXiv:2412.11970 (2024)

Yan, K.: Optimizing an english text reading recommendation model by integrating collaborative filtering algorithm and fasttext classification method. *Heliyon* **10**(9) (2024)

Yuan, W., Chen, G., Chen, C.Y.-C.: Fusiondta: attention-based feature polymerizer and knowledge distillation for drug-target binding affinity prediction. *Briefings in Bioinformatics* **23**(1), 506 (2022)

Yagnik, N., Jhaveri, J., Sharma, V., Pila, G.: Medlm: Exploring language models for medical question answering. arXiv preprint arXiv:2401.11389 (2024)

Yang, K., Klein, D., Celikyilmaz, A., Peng, N., Tian, Y.: Rlcd: Reinforcement learning from contrastive distillation for lm alignment. In: The Twelfth International Conference on Learning Representations (2024)

Yu, R., Liu, S., Wang, X.: Dataset distillation: A comprehensive review. *IEEE transactions on pattern analysis and machine intelligence* **46**(1), 150–170 (2023)

Yu, R., Liu, S., Ye, J., Wang, X.: Teddy: Efficient large-scale dataset distillation via taylor-approximated matching. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 1–17 (2024)

Yang, W., Lin, Y., Zhou, J., Wen, J.: Enabling large language models to learn from rules. arXiv preprint arXiv:2311.08883 (2023)

Yin, H., Molchanov, P., Alvarez, J.M., Li, Z., Mallya, A., Hoiem, D., Jha, N.K., Kautz, J.: Dreaming to distill: Data-free knowledge transfer via deepinversion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8715–8724 (2020)

Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W.W., Salakhutdinov, R., Manning, C.D.: Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 2369–2380 (2018)

Yin, Z., Shen, Z.: Dataset distillation in large data era. arXiv preprint arXiv:2311.18838 (2023)

Yin, Z., Shen, Z.: Dataset distillation via curriculum data synthesis in large data era. *Transactions on Machine Learning Research* (2024)

Yuan, L., Tay, F.E., Li, G., Wang, T., Feng, J.: Revisiting knowledge distillation via label smoothing regularization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3903–3911 (2020)

- Yang, S., Xie, Z., Peng, H., Xu, M., Sun, M., Li, P.: Dataset pruning: Reducing training data by examining generalization influence. arXiv preprint arXiv:2205.09329 (2022)
- Yin, Z., Xing, E., Shen, Z.: Squeeze, recover and relabel: Dataset condensation at imagenet scale from a new perspective. Advances in Neural Information Processing Systems **36**, 73582–73603 (2023)
- You, S., Xu, C., Xu, C., Tao, D.: Learning from multiple teacher networks. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1285–1294 (2017)
- Yao, T., Zhai, Z., Gao, B.: Text classification model based on fasttext. In: 2020 IEEE International Conference on Artificial Intelligence and Information Systems (ICAIIS), pp. 154–157 (2020). IEEE
- Yang, Z., Zeng, X., Zhao, Y., et al.: Alphafold2 and its applications in the fields of biology and medicine. *Sig Transduct Target Ther* **8**, 115 (2023) <https://doi.org/10.1038/s41392-023-01381-z>
- Zhao, B., Bilen, H.: Dataset condensation with differentiable siamese augmentation. In: International Conference on Machine Learning, pp. 12674–12685 (2021). PMLR
- Zhao, B., Bilen, H.: Dataset condensation with distribution matching. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 6514–6523 (2023)
- Zhang, L., Bao, C., Ma, K.: Self-distillation: Towards efficient and compact neural networks. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(8), 4388–4403 (2022) <https://doi.org/10.1109/TPAMI.2021.3067100>
- Zhong, X., Chen, B., Fang, H., Gu, X., Xia, S.-T., Yang, E.-H.: Going beyond feature similarity: Effective dataset distillation based on class-aware conditional mutual information. arXiv preprint arXiv:2412.09945 (2024)
- Zhang, H., Chen, D., Wang, C.: Confidence-aware multi-teacher knowledge distillation. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4498–4502 (2022). IEEE
- Zhang, H., Chen, D., Wang, C.: Adaptive multi-teacher knowledge distillation with meta-learning. In: 2023 IEEE International Conference on Multimedia and Expo (ICME), pp. 1943–1948 (2023). IEEE
- Zhang, S., Dong, L., Li, X., Zhang, S., Sun, X., Wang, S., Li, J., Hu, R., Zhang, T., Wu, F., et al.: Instruction tuning for large language models: A survey. arXiv preprint arXiv:2308.10792 (2023)
- Zhai, X.: Chatgpt user experience: Implications for education. Available at SSRN 4312418 (2022)
- Zhu, Z., Hong, J., Zhou, J.: Data-free knowledge distillation for heterogeneous federated learning. In: International Conference on Machine Learning, pp. 12878–12889 (2021). PMLR
- Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675 (2019)
- Zhou, H., Liu, Z., Lang, C., Xu, Y., Lin, Y., Hou, J.: Improving the recall of biomedical named entity recognition with label re-correction and knowledge distillation. BMC bioinformatics **22**(1), 295 (2021)
- Zhao, G., Li, G., Qin, Y., Yu, Y.: Improved distribution matching for dataset condensation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7856–7865 (2023)

- Zhou, Y., Lyu, K., Rawat, A.S., Menon, A.K., Rostamizadeh, A., Kumar, S., Kagy, J.-F., Agarwal, R.: Distillspec: Improving speculative decoding via knowledge distillation. arXiv preprint arXiv:2310.08461 (2023)
- Zhou, C., Liu, P., Xu, P., Iyer, S., Sun, J., Mao, Y., Ma, X., Efrat, A., Yu, P., Yu, L., *et al.*: Lima: Less is more for alignment. Advances in Neural Information Processing Systems **36**, 55006–55021 (2023)
- Zhao, B., Mopuri, K.R., Bilen, H.: Dataset condensation with gradient matching. arXiv preprint arXiv:2006.05929 (2020)
- Zhou, Y., Nezhadarya, E., Ba, J.: Dataset distillation using neural feature regression. Advances in Neural Information Processing Systems **35**, 9813–9827 (2022)
- Zhao, L., Qian, X., Guo, Y., Song, J., Hou, J., Gong, J.: Mskd: Structured knowledge distillation for efficient medical image segmentation. Computers in Biology and Medicine **164**, 107284 (2023)
- Zhang, J., Qin, Y., Pi, R., Zhang, W., Pan, R., Zhang, T.: Tagcos: Task-agnostic gradient clustered coresnet selection for instruction tuning data. CoRR (2024)
- Zhou, J., Rizzo, K., Tang, Z., Koo, P.K.: Uncertainty-aware genomic deep learning with knowledge distillation. bioRxiv (2024)
- Zhou, H., Song, L., Chen, J., Zhou, Y., Wang, G., Yuan, J., Zhang, Q.: Rethinking soft labels for knowledge distillation: A bias-variance tradeoff perspective. arXiv preprint arXiv:2102.00650 (2021)
- Zhang, L., Song, J., Gao, A., Chen, J., Bao, C., Ma, K.: Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 3712–3721 (2019). <https://doi.org/10.1109/ICCV.2019.00381> . <https://api.semanticscholar.org/CorpusID:159041406>
- Zhong, X., Sun, S., Gu, X., Xu, Z., Wang, Y., Zhang, M., Chen, B.: Efficient dataset distillation via diffusion-driven patch selection for improved generalization. arXiv preprint arXiv:2412.09959 (2024)
- Zhao, W., Sui, X., Guo, J., Hu, Y., Deng, Y., Zhao, Y., Qin, B., Che, W., Chua, T.-S., Liu, T.: Trade-offs in large reasoning models: An empirical analysis of deliberative and adaptive reasoning over foundational capabilities. arXiv preprint arXiv:2503.17979 (2025)
- Zhang, C., Song, D., Ye, Z., Gao, Y.: Towards the law of capacity gap in distilling language models. arXiv preprint arXiv:2311.07052 (2023)
- Zhang, X., Tian, C., Yang, X., Chen, L., Li, Z., Petzold, L.R.: Alpacare: Instruction-tuned large language models for medical application. arXiv preprint arXiv:2310.14558 (2023)
- Zhou, Y., Wang, X., Niu, Y., Shen, Y., Tang, L., Chen, F., He, B., Sun, L., Wen, L.: Diffim: Controllable synthetic data generation via diffusion language models. arXiv preprint arXiv:2411.03250 (2024)
- Zhu, K., Wang, J., Zhou, J., Wang, Z., Chen, H., Wang, Y., Yang, L., Ye, W., Zhang, Y., Zhenqiang Gong, N., et al.: Promptbench: Towards evaluating the robustness of large language models on adversarial prompts. arXiv e-prints, 2306 (2023)
- Zhang, Y., Xiang, T., Hospedales, T.M., Lu, H.: Deep mutual learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4320–4328 (2018)
- Zhang, X., Xie, R., Ma, P.: Statistical leveraging methods in big data. Handbook of Big Data Analytics, 51–74 (2018)

Zhang, Z., You, W., Wu, T., Wang, X., Li, J., Zhang, M.: A survey of generative information extraction. In: Proceedings of the 31st International Conference on Computational Linguistics, pp. 4840–4870 (2025)

Zhou, W., Zhu, X., Han, Q.-L., Li, L., Chen, X., Wen, S., Xiang, Y.: The security of using large language models: A survey with emphasis on chatgpt. IEEE/CAA Journal of Automatica Sinica (2024)

Zhang, X., Zhai, J., Ma, S., Shen, C., Li, T., Jiang, W., Liu, Y.: Staff: Speculative coreset selection for task-specific fine-tuning. In: The Thirteenth International Conference on Learning Representations

Zhang, K., Zhu, R., Ma, S., et al.: Kedrec-lm: A knowledge-distilled explainable drug recommendation llm. arXiv preprint arXiv:2501.01558 (2025)