On LLMs-Driven Synthetic Data Generation, Curation, and Evaluation: A Survey

Lin Long[1], **Rui Wang**[1], **Ruixuan Xiao**[1]
Junbo Zhao[1], **Xiao Ding**[2], **Gang Chen**[1], **Haobo Wang**[1]*
[1]Zhejiang University, China [2]Harbin Institute of Technology, China
Correspondence: wanghaobo@zju.edu.cn

Abstract

Within the evolving landscape of deep learning, the dilemma of data quantity and quality has been a long-standing problem. The recent advent of Large Language Models (LLMs) offers a data-centric solution to alleviate the limitations of real-world data with synthetic data generation. However, current investigations into this field lack a unified framework and mostly stay on the surface. Therefore, this paper provides an organization of relevant studies based on a generic workflow of synthetic data generation. By doing so, we highlight the gaps within existing research and outline prospective avenues for future study. This work aims to shepherd the academic and industrial communities towards deeper, more methodical inquiries into the capabilities and applications of LLMs-driven synthetic data generation.

1 Introduction

The game-changing emergence of Large Language Models (LLMs) instigated a significant paradigm shift in the field of deep learning (Zhang et al., 2023a; Guo et al., 2023; Bang et al., 2023). Despite these advancements, a large amount of high-quality data remains the foundation for building robust NLP models (Gandhi et al., 2024). To be more specific, here high-quality data typically refers to diverse data that carries rich supervision signals (generally in the form of labels) closely aligned with human intent. However, fulfilling such data reliance with human data can be challenging or even unrealistic sometimes, due to high costs, data scarcity, privacy concerns, etc. (Kurakin et al., 2023). Moreover, several studies (Hosking et al., 2023; Singh et al., 2023; Gilardi et al., 2023) have highlighted that human-generated data, being inherently susceptible to biases and errors, may not even be optimal for model training or evaluation. These considerations necessitate a more serious inquiry into the question: are there other more *effective and scalable* methods of data collection that can overcome the current limitations?

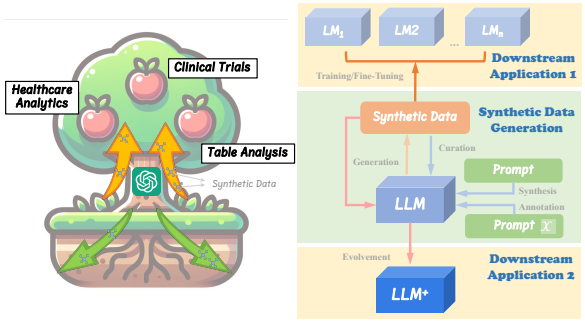

Figure 1: Illustration of the LLMs-based application ecosystem, where synthetic data serves as the flowing nutrients for fruiting (training of small LMs or fine-tuning task-specific LLMs) and rooting (training stronger LLMs or self-improvement).

Given the recent advancements in LLMs, which demonstrate the capability to generate fluent text on par with human output (Hartvigsen et al., 2022; Sahu et al., 2022; Ye et al., 2022a; Tang et al., 2023; Gao et al., 2023a), synthetic data produced by LLMs emerges as a viable alternative or supplement to human-generated data. Specifically, synthetic data is designed to mimic the characteristics and patterns of real-world data (Liu et al., 2024). On the one hand, LLMs, through extensive pretraining, have acquired a vast repository of knowledge and demonstrate exceptional linguistic comprehension (Kim et al., 2022; Ding et al., 2023a), which forms a foundation for generating faithful data. On the other hand, the profound instruction-following capabilities of LLMs allow better controllability and adaptability over the generation process, facilitating the creation of tailored datasets for specific applications with more flexible process designs (Eldan and Li, 2023). These two advantages make LLMs highly promising synthetic data generators.

As a pivotal application of LLMs, synthetic data

*Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2024, pages 11065–11082

generation holds significant importance for the development of deep learning. As shown in Figure 1, LLMs-driven synthetic data generation (Li et al., 2023c; Wang et al., 2021; Seedat et al., 2023) enables the automation of the entire model training and evaluation process with minimal human participation required in the loop (Huang et al., 2023), which allows the advantages of deep learning models to be applied across a broader range of applications. Beyond providing a scalable supply of training and testing data, LLM-driven synthetic data generation also may pave the way for developing next-generation LLMs. Insights from TinyStories (Eldan and Li, 2023) and the *Phi* series (Gunasekar et al., 2023; Li et al., 2023b) emphasize that data quality is crucial for effective model learning, while LLMs empower us to actively "design" what the models learn through data manipulation, significantly enhancing the efficacy and controllability of model training. As of June 2024, there are over 300 datasets on Hugging Face[1] that are tagged as "synthetic", with many mainstream LLMs leveraging high-quality synthetic data for training, including Alpaca (Taori et al., 2023), Vicuna (Zheng et al., 2023), OpenHermes 2.5, and Openchat 3.5 (Wang et al., 2023a).

Though seemingly straightforward, generating synthetic datasets that simultaneously have high correctness and sufficient diversity requires careful process designs and involves a lot of tricks (Gandhi et al., 2024), making LLMs-driven synthetic data generation a non-trivial problem. While most existing works generally target data generation for various tasks (e.g., pre-training (Gunasekar et al., 2023; Li et al., 2023b; Eldan and Li, 2023), fine-tuning (Mukherjee et al., 2023; Mitra et al., 2023; Xu et al., 2023a), evaluation (Feng et al., 2023; Wei et al., 2024)) across different domains (e.g., math (Yu et al., 2023a; Luo et al., 2023a), code (Luo et al., 2023b; Wei et al., 2023b), instruction (Honovich et al., 2023a; Wang et al., 2023d)), they share many common ideas. To address the lack of a unified framework in the emerging field of LLM-driven synthetic data generation and develop a general workflow, this survey investigates recent studies and organizes them according to the topics of generation, curation, and evaluation, which are closely related, as shown in Figure 2. Our primary aim is to provide a comprehensive overview of the current state of the field, identify key areas of focus, and highlight the gaps that remain to be addressed. We hope to bring insights to both the academic and industrial communities and drive further development in LLM-driven synthetic data generation.

[1] https://huggingface.co

2 Preliminaries

2.1 Problem Definition

In this paper, we investigate the challenge of generating high-quality synthetic data using pre-trained LLMs, denoted as $\mathcal{M}$. Rather than creating new datasets from scratch, in more cases, we perform data augmentation with a small number of seed samples or unlabeled inputs, which we denote uniformly as $\mathcal{D}_{\text{sup}}$. Although optional for LLMs-driven synthetic data generation, $\mathcal{D}_{\text{sup}}$ can typically provide valuable supporting information when available. Consequently, the overall generation task can be formulated as:

$$\mathcal{D}_{\text{gen}} \leftarrow \mathcal{M}_p(\mathcal{T}, \mathcal{D}_{\text{sup}}), \tag{1}$$

where $\mathcal{D}_{\text{gen}}$ represents the final generated dataset, and p refers to the prompt used for model inference. $\mathcal{T}$ specifies the generation task, such as rewriting, question answering, annotation, etc. Notably, data annotation as a specialized paradigm of synthetic data generation, has particularly extensive applicability, including RLAIF (Bai et al., 2022) and LLMs-based evaluation (Chen et al., 2023b; Zheng et al., 2023; Kim et al., 2023), which may involve specific challenges and corresponding solution techniques. Due to page limitations, further details about data annotation can be found in Appendix A.

2.2 Requirements of $\mathcal{D}_{\text{gen}}$

Briefly speaking, our goal is to generate data that closely aligns with evaluation metrics. While the standard of high-quality data may vary across different downstream tasks, there are two general requirements that are considered challenging in most existing literature:

- **Faithfulness.** To provide valid supervision, the generated data must first be logically and grammatically coherent. However, the inherent problems of hallucination fat-tailed knowledge distribution of LLMs can introduce significant noise into the generated results, manifesting as factual errors, incorrect labels, or irrelevant content. These issues become more pronounced when generating long, complex, or domain-specific data.

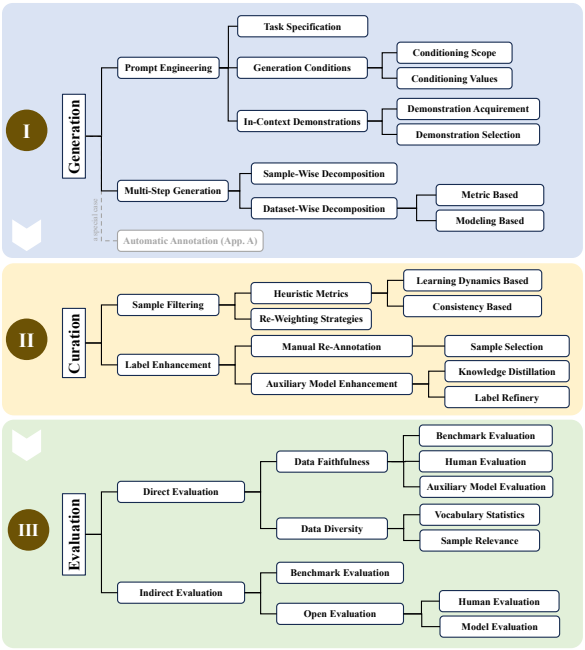

Figure 2: A taxonomy of LLMs-driven synthetic data generation, curation, and evaluation.

- **Diversity.** Diversity captures the variation among the generated data, reflecting differences in text length, topic, or even writing style. It is crucial for generating synthetic samples that mimic the diversified nature of real-world data, thereby preventing overfitting and bias during model training or evaluation. Nevertheless, due to the inherent biases of LLMs, uncontrolled generated content often tends to be monotonous, limiting its applicability in downstream tasks.

These two requirements are the focal points of most current research efforts. In the subsequent workflow, we will introduce how different methods address these issues.

3 Generic Workflow

Existing studies on LLMs-driven synthetic data generation generally incorporate three main topics: generation, curation, and evaluation. Various approaches are employed within these aspects to collaboratively achieve optimal data generation.

3.1 Data Generation

In this section, we systematically summarize some common practices for synthetic data generation with LLMs, which can be roughly divided into prompt engineering and multi-step generation. An overall illustration is provided in Figure 3.

3.1.1 Prompt Engineering

One of the greatest advantages of LLMs for synthetic data generation is their instruction-following capability, which contributes to great controllability (Wang et al., 2023c; Radford et al., 2019). Therefore, many approaches try to guide LLMs with heuristic prompts to enhance the faithfulness and diversity of the synthetic data (Liu et al., 2024).

Empirically, an effective prompt generally contains three key elements: *task specification* e_{task}, *generation conditions* $e_{\text{condition}}$, and *in-context demonstrations* e_{demo}, which are then collectively wrapped with a template E into the form of natural instruction:

$$p(\mathcal{T}, \mathcal{D}) \leftarrow E(e_{\text{task}}, e_{\text{condition}}, e_{\text{demo}}). \quad (2)$$

As shown above, both the generation task $\mathcal{T}$ and the support dataset $\mathcal{D}$ will affect the design of p. Next, we will proceed to detail how each part of the prompt should be appropriately designed to accommodate various scenarios.

Task Specification. In traditional crowdsourced annotation scenarios, the recruited workers are commonly offered a codebook that specifies the necessary contexts, such as task purpose, data explanation, and other background knowledge, so that they can better understand their jobs (Gilardi et al., 2023). Similarly, such task specification is crucial for setting the right context for LLMs-driven data generation, which can also include role-play (Li et al., 2023c), format clarification, knowledge augmentation (Xu et al., 2023b; Sudalairaj et al., 2024), etc. Evidence shows that a simple prologue such as "suppose you are a `{xxx}`" can significantly improve the LLMs' performance by setting up a proper scenario for data generation and allowing the LLMs to better take on the roles (Li et al., 2023c). More formally, Yoo et al. (2021) defines the task specification with a triplet of text type, label type, and label-token verbalizer. Such a description header is particularly important when extra domain expertise is demanded to address issues like terminology complexities in both context understanding and data generation. Consequently, Xu et al. (2023b) leverages external knowledge graphs and LLMs to obtain domain topics for context-informed prompting, which effectively enhances the faithfulness and complexity of generated data.

Conditional Prompting. As mentioned in Section 2.2, a pivotal challenge in using LLMs for

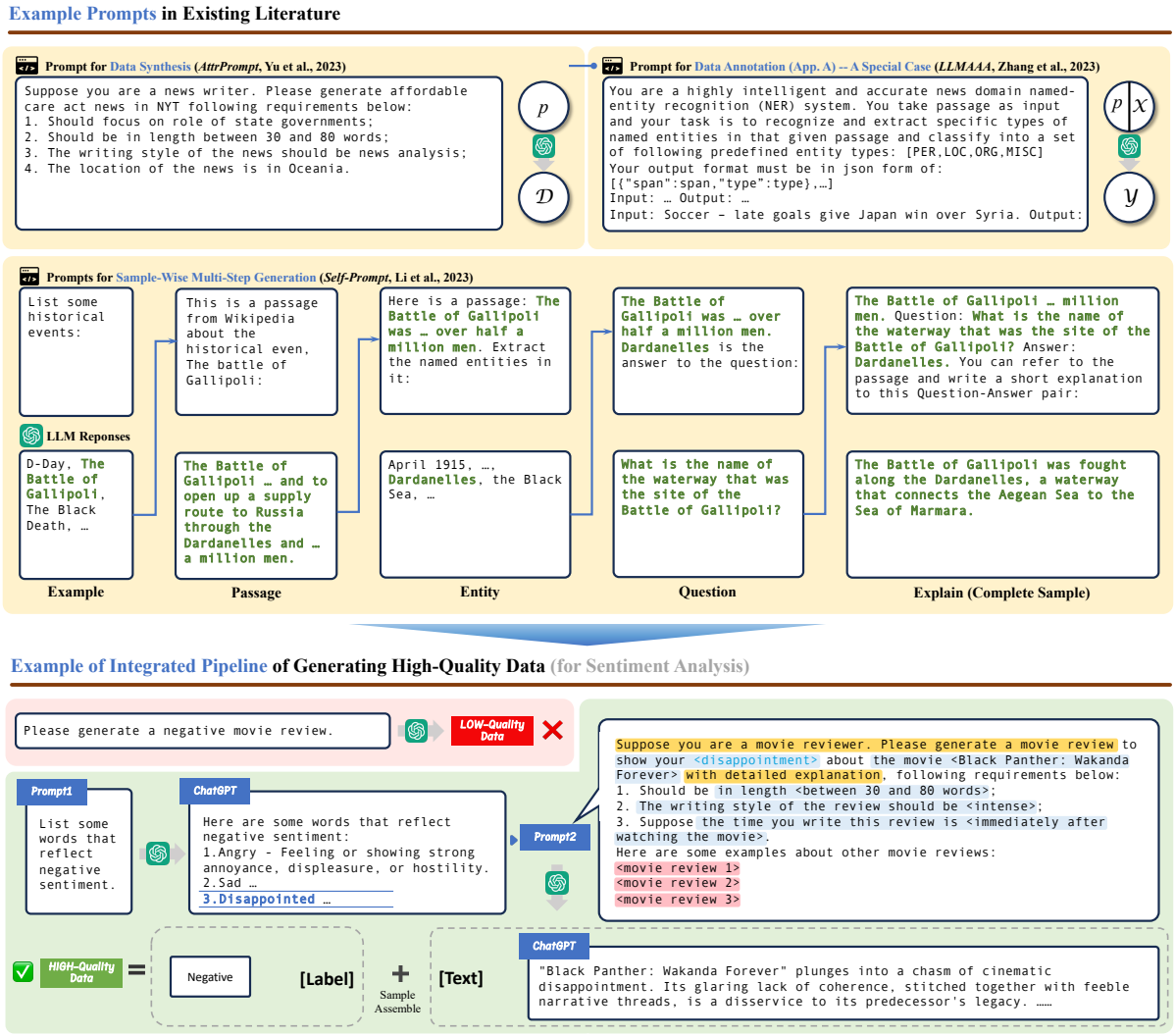

Figure 3: A toy example of effective synthetic data generation. The corresponding fields for task specification, conditions, and in-context demonstrations are highlighted, while < > marks the switchable contents.

synthetic data generation is ensuring sufficient diversity, as directly prompting the LLMs to produce data for certain tasks often results in highly repetitive outputs, even with a high decoding temperature (Gandhi et al., 2024; Liu et al., 2024). Addressing this problem, a widely adopted strategy is conditional prompting, which explicitly and concretely communicates to the LLMs the specific type of data desired. The core of conditional prompting involves delineating the targeted data through the formulation of a series of condition-value pairs:

$$e_{\text{condition}} = \{(c_1, v_1), (c_2, v_2), \cdots, (c_n, v_n)\}, \quad (3)$$

which effectively characterizes the desired attributes and characteristics of the synthetic data. With different combinations of such attributes, we can automatically achieve a degree of "artificially defined" diversity in the generated samples (Gunasekar et al., 2023; Li et al., 2023b; Eldan and Li, 2023). Conditional prompting not only allows better control over the diversity and coverage of the generated dataset but also refines the content to a narrower, more focused scope that is more likely to align with our specific expectations and requirements (Li et al., 2023c). Current research on conditional prompting primarily centers on the following two subjects:

1) **Conditioning Scope.** As the backbone of $e_{\text{condition}}$, conditioning scope defined by $\{c_1, \cdots, c_n\}$ delineates the dimensions that we utilize to characterize our target data. Early studies (Gao et al., 2023a; Ye et al., 2022a,b) employed a basic output-conditional prompting strategy, utilizing the specific label asso-

ciated with the classification task as the conditioning variable. The rationale behind this was primarily to maintain class balance and coverage. However, such a strategy is unsuitable for data lacking explicit category labels. Subsequent work by Yu et al. (2023b) argues that conditional-prompting with finer-grained attributes (e.g., topics, length, and style (Xu et al., 2023b)), can lead to more diversified generation due to the vast number of possible attribute combinations, being also applicable to open-ended data. Additionally, Eldan and Li (2023) also condition each generation on the task of incorporating three randomly chosen words into the generated story. This approach was also proven to significantly enhance the diversity of the generated data, shifting the focus from the heuristic features of the output to a more structured and targeted conditioning mechanism by adding "creative randomness" to the prompt (Eldan and Li, 2023).

2) **Conditioning Values.** After defining the conditioning scope, we then need to assign concrete values to each condition. Despite the seemingly straightforward strategy of sampling from the known classes or labels (Ye et al., 2022a), there are cases where such an instance pool is unavailable. Addressing this problem, Josifoski et al. (2023) actively retrieves the conditioning instances from external knowledge graphs, while Xu et al. (2023b); Ding et al. (2023b) leverage the LLMs to generate diversified instances for conditional prompting. Specifically, Ding et al. (2023b) construct a concept tree to delve into different subtopics, ensuring the coverage of sampled conditioning values, which then contributes to more diverse generated data. Moreover, the prompt template E can also be considered a special type of condition. It has been demonstrated that incorporating templates with a certain level of randomness throughout the generation process can enhance the diversity of the generated contents (Meng et al., 2022).

In-Context Learning. Due to the inherent bias of LLMs, it remains challenging to elicit favorable responses from the LLMs with merely task specification and conditional prompting. In this case, a straightforward yet effective strategy is to provide several demonstrations, which can serve as a form of implicit human guidance. Research has shown that, owing to LLMs' remarkable in-context learning (ICL) capabilities, a few exemplars can provide them with insights into the patterns exhibited in real-world data, thereby significantly improving the faithfulness of generated data (Li et al., 2023c). In the few-shot setting, where labeled samples are available in the support set $\mathcal{D}_{\text{sup}}$, these samples can be directly utilized as demonstrations for ICL. However, in scenarios where no ground truth data is available, approaches like Self-Instruct (Wang et al., 2023e) and Self-Prompting (Li et al., 2022) instead leverage ICL with synthetic demonstrations generated by LLMs. This allows the models to learn from their own predictions or other teacher models, even in the absence of labeled data.

However, given the constraint of prompt length and data inconsistency, the quality of in-context samples significantly affects the effectiveness of in-context learning. Sudalairaj et al. (2024) argue that randomly selecting in-context examples from the pool of seed samples, as done in Self-Instruct (Wang et al., 2023e), results in a lack of diversity and quality in the generated data. To address this issue, Sudalairaj et al. (2024) opt for selecting examples that concentrate on specific aspects to better stimulate the long tail of knowledge inherent in LLMs. Liu et al. (2022b) and Su et al. (2023) prioritize consistent samples as demonstrative examples based on their cosine similarity in the embedding space. Alternatively, Ye et al. (2022b) selects the most informative samples using quantified influence scores to steer the generation process. To enhance the informativeness of in-context examples, He et al. (2023) prompts LLMs to provide an explanation for each sample before integrating it into the prompt. This approach not only offers valuable additional information but also aligns well with the subsequent Chain-of-Thought generation.

3.1.2 Multi-Step Generation

In the previous paragraphs, we have introduced some common prompting strategies, which are typically designed for a specific generation task $\mathcal{T}$. However, in most cases, due to the lack of enough reasoning abilities, it is unrealistic to expect the LLMs to generate the entire desired dataset within a single reference, especially when targeting data with complex structures or semantics (Cui and Wang, 2023). In addressing this problem, a common strategy is multi-step generation, through which the overall generation process is manually decomposed into a chain of simpler sub-tasks $\mathcal{T}_{1:k}$,

to force the LLMs to produce data in a step-by-step manner as scheduled:

$$\mathcal{D}_i \leftarrow \mathcal{M}^i_{p_i}(\mathcal{T}_i, \mathcal{D}_{0:i-1}),\ i = 1, 2, \cdots, k, \quad (4)$$

where $\mathcal{D}_0 = \mathcal{D}_{\text{sup}}$. Each intermediate output $\mathcal{D}_i$ is generated using model $\mathcal{M}^i$, prompted by p_i, for a sub-task $\mathcal{T}_i$. These outputs can then potentially be used in subsequent generations. By manually scheduling the generation procedure, we implicitly align the reasoning paths of LLMs with human prior knowledge. Specifically, there are two common strategies for task decomposition: *sample-wise* and *dataset-wise* decomposition, which mainly aim at enhancing the quality of synthetic data at different scales.

Sample-Wise Decomposition. A typical use-case of multi-step generation is for addressing the challenges of long-text processing and logical reasoning when dealing with multi-text data such as dialogues and entity-relation triplets. In such cases, a straightforward approach is to divide the sample into smaller chunks and generate only a portion of each sample at a time (Li et al., 2022; Ye et al., 2023; Wang et al., 2023e). In this way, $\mathcal{D}_{1:k}$ can be considered as different parts of $\mathcal{D}_{\text{gen}}$:

$$\mathcal{D}_{\text{gen}} = (\mathcal{D}_1, \mathcal{D}_2, \cdots, \mathcal{D}_k). \quad (5)$$

Notably, as shown in Eq. 4, each iteration of the generation process can be conditioned on the previously generated contents. For example, Ding et al. (2023b) prompts the LLMs to alternate between acting as the assistant and the user, replying to each other based on the context, ultimately producing a complete conversation transcript. In this way, the coherence among each internal component $\mathcal{D}_i$ can be pointedly reinforced with separated instructions, thus making it easier for the model to follow the requirements and generate more faithful data. It should be noted that $D_{1:k}$ may not necessarily form part of the final D_{gen}, instead, explicitly outputting some intermediate reasoning steps can also improve the generation of complex data (Bai et al., 2022; He et al., 2023). Chain-of-Thought (CoT) prompting stands out as one of the most popular strategies for improving the faithfulness of LLM-generated content (Wei et al., 2022). Nevertheless, current research on the exploration of such latent metadata is still insufficient, leaving sample-wise task decomposition from a reasoning perspective an open problem for future studies.

Dataset-Wise Decomposition. In Section 3.1.1 we have introduced how to generate data with specified properties. However, generating a series of such data that can eventually form a dataset with good diversity and domain coverage requires long-term scheduling. To this end, dataset-wise task decomposition dynamically adjusts the conditions used at each stage of multi-step generation to ensure the overall dataset grows in the right direction:

$$\mathcal{D}_{\text{gen}} = \bigcup_{i=1}^{k} \mathcal{D}_i. \quad (6)$$

Specifically, S3 (Wang et al., 2023b) targets the most frequently mislabeled categories at each iteration, according to the performance of the downstream model trained on previously generated data. Similarly, Honovich et al. (2023b); Shao et al. (2023) utilize a generate-then-expand paradigm, to enhance the diversity of the overall dataset accordingly. Some other methods also leverage specific data structures to model the pathways of data generation. For example, Explore-Instruct (Wan et al., 2023) models the domain space as a tree structure and continually refines the generated data along with tree traversal to promote both the specialization and domain coverage of the generated data.

3.2 Data Curation

After the preceding steps, one may excessively generate overflowing and theoretically unlimited data $\mathcal{D}_{\text{gen}}$. However, these datasets often comprise a considerable portion of noisy, worthless, or even toxic samples, which primarily stems from two causes. Firstly, LLMs can inevitably produce corrupted samples with incorrect labels due to the hallucination problem. Secondly, ineffective prompts containing ambiguous descriptions can trick the model into generating irrelevant or redundant samples. Consequently, directly utilizing these low-quality data without proper processing may have a significant negative impact.

To address this, plenty of data curation approaches have been studied, which mainly fall into two dominant groups of *high-quality sample filtering* and *label enhancement* as elaborated below.

3.2.1 High-Quality Sample Filtering

Sample filtering aims to weed out undesired low-quality samples and obtain a more helpful subset $\mathcal{D}_{\text{curated}} \subset \mathcal{D}_{\text{gen}}$. These methods typically design *heuristic criteria* or *re-weighting functions* to rerank samples for filtering, as shown in Figure 4.

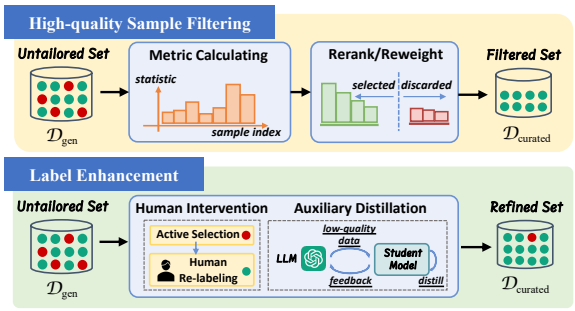

Figure 4: Two dominant approaches of data curation.

Heuristic Metrics. For methods based on heuristic metrics, the key step is to design appropriate criteria based on the learning dynamics, such as confidence score (Seedat et al., 2023), influence function (Ye et al., 2022b), and generation ability (Meng et al., 2022). SuperGen (Meng et al., 2022) employs the estimated generation probability to identify samples most related to the desired label. Seedat et al. (2023) discard samples with both low confidence and low uncertainty. Some other methods assume that clean samples are prone to hold similar predictions under different conditions and employ cross-condition consistency for filtering. Specifically, such consistency can be between LLM and downstream classifier (Yu et al., 2023c), between multiple executions (Ye et al., 2023), or between neighboring data points (Seedat et al., 2023). More recently, Chen et al. (2023b) leverage the powerful text understanding capabilities of large models to assess the quality of different samples and filter out the ones with scores lower than a specific threshold. Their findings indicate that Alpagasus (Chen et al., 2023b), trained on a much smaller but curated dataset, notably surpasses the original Alpaca (Taori et al., 2023) across several benchmarks, underscoring the critical role of data curation.

Sample Re-Weighting. On the other hand, reweighting methods believe all data are valuable but with varying importance. Thus, they assign larger weights to correctly annotated or influential samples during downstream utilization (Zhang et al., 2023b; Gao et al., 2023a; Meng et al., 2023). For instance, SunGen (Gao et al., 2023a) proposes an adaptive bi-level re-weighting algorithm without human annotations. FewGen (Meng et al., 2023) designs a discriminative meta-learning objective to adjust sample weights and demarcate the nuanced differences between different labels.

3.2.2 Label Enhancement

Label enhancement methods strive to rectify the potentially erroneous annotations in generated samples. Due to confirmation bias, it is unrealistic for LLMs to identify their own mistakes. To address this, recent works either rely on *human intervention* or incorporate a student model for *human-free knowledge distillation*.

Human Intervention. A straightforward strategy for label refinery is to include human efforts to re-annotate the corrupted samples (Chung et al., 2023a; Wang et al., 2021; Pangakis et al., 2023). Wang et al. (2021) proposed to actively select samples with the lowest confidence for human re-labeling. Pangakis et al. (2023) and Liu et al. (2022a) further emphasize the importance of human review and suggest comparing annotations from humans and LLMs guided by the same codebook. Despite the simplicity, these methods can lead to considerable labeling costs and can be unrealistic in practical deployment.

Auxiliary Model. To reduce the labeling cost, a more pragmatic human-free paradigm is developed which involves auxiliary student models for knowledge distillation and label refinery (Xiao et al., 2023; Zhao et al., 2023a; Saad-Falcon et al., 2023). These methods rely on the weakly supervised ability of student models and hypothesize that a student distilled from the LLM teacher can produce superior labels. The seminal work FreeAL (Xiao et al., 2023) proposes a collaborative framework, where a student model is leveraged to distill the high-quality task-related knowledge from the weak annotations and in return feedback LLMs for label refinery. MCKD (Zhao et al., 2023a) designs a multistage distillation pipeline with data-split training and cross-partition labeling to avoid overfitting on noisy labels. With the expanding abilities and availability of LLMs, the incorporation of auxiliary student models will play a more crucial role as a cost-effective alternative to human intervention.

3.3 Data Evaluation

Before the employment of generated data, it is important to evaluate the quality and application effectiveness of the data, to ensure its value to downstream tasks. The current mainstream evaluation methods can be roughly divided into two categories: *direct* and *indirect*, which evaluate the quality of $\mathcal{D}_{\text{gen}}$ individually and through its effectiveness on downstream tasks, respectively.

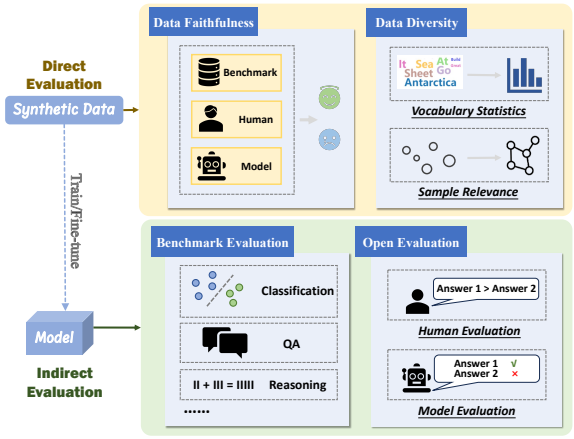

Figure 5: Direct and indirect methods of data evaluation.

3.3.1 Direct Evaluation

Data Faithfulness. Ideally, automatic evaluation of the LLMs' generation results can be easily realized with ground truths from existing datasets, if available (Zhu et al., 2023). However, for open-ended data, human-based evaluation is necessitated. A straightforward idea is to provide some generated samples to human experts, who will then determine whether they are correct, according to which we can estimate the overall generation quality (Wang et al., 2023e). Theoretically, the larger the sample size, the more accurate the estimation results will be, but the labor it costs will correspondingly get higher. To this end, a reliable auxiliary model can be leveraged for a more comprehensive yet cost-effective evaluation of the generated data in replace of human experts (Chung et al., 2023b). Considering that most models can only process contents of limited length, appropriate information extraction can reduce the burden of the auxiliary model and contribute to a more precise prediction of whether a sample contains factual errors (Lee et al., 2022).

Data Diversity. The quantification of data diversity primarily employs vocabulary statistics and sample relevance calculations. Vocabulary statistics (Yu et al., 2023b), such as vocabulary size and N-gram frequency, provide a straightforward and intuitive approach. However, they struggle to capture the semantic information of a dataset. The calculation of sample relevance compensates for this limitation effectively. The most common measures of sample correlation are based on cosine similarity (Wang et al., 2023b) and sample distance (Chung et al., 2023b) calculations, which can better capture the contextual and semantic diversity of the dataset. Furthermore, these metrics can also be leveraged to select in-context demonstrations e_{demo} (Wang et al., 2023e) that are more dissimilar with the previously generated samples, thereby leading to more diversified generation results.

3.3.2 Indirect Evaluation

Benchmark Evaluation. The performance of downstream models trained on the generated data can also reflect the generation quality to some extent (Yu et al., 2023b; Chung et al., 2023b). Specifically, the impact of synthetic data can be evaluated from multiple dimensions except for the specialized capabilities of the downstream models. For example, TruthfulQA enables the assessment of a model's ability to identify true claims (Sun et al., 2023); NIV2 is employed to evaluate a model's language comprehension and reasoning abilities across multiple tasks (Wang et al., 2023e).

Open Evaluation. For open-ended benchmarks, evaluation by humans or auxiliary models is necessitated due to the absence of standardized answers. To fully leverage the preference outputs of the auxiliary models, multiple evaluation strategies have been designed, such as response ranking (Xu et al., 2023a), four-level rating system (Wang et al., 2023e) and Elo scores (Bai et al., 2022). To further reduce evaluation costs, Sun et al. (2023); Xu et al. (2023a) utilize the automatic evaluation framework based on GPT-4 proposed by Vicuna for evaluation. However, general LLMs (e.g., ChatGPT) sometimes lack enough knowledge for domain-specific tasks, which hinders them to provide effective evaluation (Bran et al., 2023). Therefore, collecting human assessment data to fine-tune open-source models for evaluation purposes is an important practice in real-world scenarios (He et al., 2023).

4 Future Directions

4.1 Complex Task Decomposition

Current multi-step generation algorithms depend on the model's understanding of task requirements, requiring it to perform complex logical reasoning with limited information. However, in real-world complex scenarios, this limited information may not adequately support effective decision-making. For instance, the generation of mathematical problem-solution pairs entails multiple reasoning steps and may necessitate the utilization of calculator tools for validation. To date, there remains a lack of systematic investigation on how

to activate the reasoning and planning capabilities of LLMs for autonomous synthetic data generation. Inspired by prevalent LLMs-based agents like HuggingGPT (Shen et al., 2023) and MetaGPT (Hong et al., 2023), we believe it would also be quite valuable to develop a data generation *agent* for industrial applications.

4.2 Knowledge Enhancement

Recent research has found that LLMs' knowledge is long-tailed and biased (Navigli et al., 2023; Fei et al., 2023). Lacking specific domain knowledge, LLMs tend to generate biased, monotonous, and even unfaithful data. Though we have introduced how to mildly guide the data generation with task specification and conditional prompting in the previous sections, such methods still hold strong limitations and are not conducive to scalable implementation. Instead, we believe that developing automated condition controls directly on mature domain knowledge bases will significantly improve the efficiency of knowledge enhancement. For example, we can establish certain links between the LLMs and external knowledge graphs (Ji et al., 2022) or retrieve augmentation from the website (Gao et al., 2023b), which is helpful for the definition, decomposition, and reasoning of data features throughout the entire generation process. Additionally, with enhanced domain knowledge, we may also better assess the quality of generated data or even develop automatic evaluation systems. Overall, we believe that knowledge-driven data generation will be a key focus for future studies.

4.3 Synergy between Large & Small LMs

In Section 3.2, we introduced the use of small domain-specific models for data curation. In particular, FreeAL (Xiao et al., 2023) has shown the feasibility of low-cost data curation with integrated collaboration between large and small models. The idea of leveraging real-time feedback provided by automated performance evaluation during the data generation process to guide the corresponding adjustments in the following generation hints at an important research direction. However, the exploitation of small LMs at the current stage is simply based on prediction confidence. In the future, we are looking forward to seeing more diversified collaboration modes between large and small models to improve the quality of generated data, e.g., usage of various output information, new design of collaborative architectures, and so on.

4.4 Human-Model Collaboration

Data, as the source of model intelligence, theoretically cannot be generated completely without human intervention. Otherwise, wild synthetic data that carries noisy, toxic information can easily "poison" a model, even resulting in mode collapse. Due to the inherent bias of LLMs, they can hardly be self-aware of the bias in their generated data and finally deviate from our intentions. Thus, designing a human-friendly interactive system to involves a few necessary human knowledge for annotation and verification is vital and irreplaceable. To date, there is still a lack of a generic framework to standardize and systematize the human-machine collaboration involved in the data production process.

We believe that an appropriate design of such a system must be based on a thorough understanding of the strengths and limitations of human intervention, and should follow the human-centered principle. To achieve sustainable and efficient human involvement, we need comprehensive consideration of various factors such as feasibility, cost, and even labor psychology. For specific examples: (i)-readability and interpretability of the information provided by the LLMs should be ensured to reduce obstacles to human understanding; (ii)-upstream knowledge enrichment or filtering should be carried out to improve the efficiency of human resource utilization and reduce consumption on tasks with low cost-effectiveness; (iii)-incorporating enjoyable interactive features can not only mitigate the negative impact of mechanical data processing tasks on humans but also attract a broader audience.

5 Conclusion

In this paper, we present a systematic review of advancements in synthetic data generation propelled by Large Language Models (LLMs). We aim to offer guidance to enterprises and organizations on effectively building their domain-specific datasets using LLMs. In the meantime, we endeavor to provide insights into the challenges and opportunities within this field, while also proposing potential directions for future research. We hope that our work can promote the rapid production of large amounts of data in various fields and push the limits of data-centric AI. We also envision a fantastic future, where an LLMs community, endowed with human-like abilities such as bionics and communication, may be constructed to generate data for its own self-improvement.

Limitations

In this paper, we survey existing studies on LLMs-driven synthetic data generation, curation, and evaluation, proposing a generic workflow for real-world practice. Synthetic data generation is a broad topic that involves data and models of various modals, including vision and speech. Due to the page limit, we mainly focus on the objective of text data and LLMs-driven approaches, while leaving investigations in other fields for future work. We will also keep paying attention to the latest work and add more related approaches with more detailed analysis.

Ethics Statement

We believe that our proposed workflow of LLMs-driven synthetic data generation, curation, and evaluation can benefit both researchers who are interested in data-centric AI and industrial producers who are facing data problems. However, the malicious use of such synthetic data also raises ethical concerns that should arouse our vigilance.

Acknowledgements

This work is supported by the Pioneer R&D Program of Zhejiang (No. 2024C01035), NSFC under Grants (No. 62206247), and the Fundamental Research Funds for the Central Universities (No. 226-2024-00049).

References

Tiago A. Almeida, José María Gómez Hidalgo, and Akebo Yamakami. 2011. Contributions to the study of SMS spam filtering: new collection and results. In *Proceedings of the 2011 ACM Symposium on Document Engineering, Mountain View, CA, USA, September 19-22, 2011*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosiute, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemí Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022. Constitutional AI: harmlessness from AI feedback. *CoRR*, abs/2212.08073.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *CoRR*, abs/2302.04023.

Parikshit Bansal and Amit Sharma. 2023. Large language models as annotators: Enhancing generalization of NLP models at minimal cost. *CoRR*, abs/2306.15766.

Max Bartolo, Alastair Roberts, Johannes Welbl, Sebastian Riedel, and Pontus Stenetorp. 2020. Beat the AI: Investigating adversarial human annotation for reading comprehension. *Transactions of the Association for Computational Linguistics*, 8:662–678.

Andres M Bran, Sam Cox, Andrew D White, and Philippe Schwaller. 2023. Chemcrow: Augmenting large-language models with chemistry tools. *arXiv preprint arXiv:2304.05376*.

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. 2020. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 38–45, Online. Association for Computational Linguistics.

Derek Chen, Celine Lee, Yunan Lu, Domenic Rosati, and Zhou Yu. 2023a. Mixture of soft prompts for controllable data generation. In *EMNLP*, pages 14815–14833. Association for Computational Linguistics.

Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, and Hongxia Jin. 2023b. Alpagasus: Training A better alpaca with fewer data. *CoRR*, abs/2307.08701.

John Chung, Ece Kamar, and Saleema Amershi. 2023a. Increasing diversity while maintaining accuracy: Text data generation with large language models and human interventions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 575–593, Toronto, Canada. Association for Computational Linguistics.

John Joon Young Chung, Ece Kamar, and Saleema Amershi. 2023b. Increasing diversity while maintaining accuracy: Text data generation with large language models and human interventions. In *ACL*, pages 575–593. Association for Computational Linguistics.

Wanyun Cui and Qianle Wang. 2023. Ada-instruct: Adapting instruction generators for complex reasoning. *CoRR*, abs/2310.04484.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.

Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Boyang Li, Shafiq Joty, and Lidong Bing. 2023a. Is GPT-3 a good data annotator? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11173–11195, Toronto, Canada. Association for Computational Linguistics.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023b. Enhancing chat language models by scaling high-quality instructional conversations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3029–3051, Singapore. Association for Computational Linguistics.

Ronen Eldan and Yuanzhi Li. 2023. Tinystories: How small can language models be and still speak coherent english? *CoRR*, abs/2305.07759.

Yu Fei, Yifan Hou, Zeming Chen, and Antoine Bosselut. 2023. Mitigating label biases for in-context learning. In *ACL*, pages 14014–14031. Association for Computational Linguistics.

Shangbin Feng, Vidhisha Balachandran, Yuyang Bai, and Yulia Tsvetkov. 2023. FactKB: Generalizable factuality evaluation using language models enhanced with factual knowledge. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 933–952, Singapore. Association for Computational Linguistics.

Saumya Gandhi, Ritu Gala, Vijay Viswanathan, Tongshuang Wu, and Graham Neubig. 2024. Better synthetic data by retrieving and transforming existing datasets. *CoRR*, abs/2404.14361.

Jiahui Gao, Renjie Pi, Yong Lin, Hang Xu, Jiacheng Ye, Zhiyong Wu, Weizhong Zhang, Xiaodan Liang, Zhenguo Li, and Lingpeng Kong. 2023a. Self-guided noise-free data generation for efficient zero-shot learning. In *ICLR*. OpenReview.net.

Tianyu Gao, Xu Han, Hao Zhu, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. 2019. FewRel 2.0: Towards more challenging few-shot relation classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6250–6255, Hong Kong, China. Association for Computational Linguistics.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. 2023b. Retrieval-augmented generation for large language models: A survey. *CoRR*, abs/2312.10997.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd-workers for text-annotation tasks. *CoRR*, abs/2303.15056.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. 2023. Textbooks are all you need. *CoRR*, abs/2306.11644.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *CoRR*, abs/2301.07597.

Kelvin Han and Claire Gardent. 2023. Multilingual generation and answering of questions from texts and knowledge graphs. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13740–13756, Singapore. Association for Computational Linguistics.

Yucheng Han, Chi Zhang, Xin Chen, Xu Yang, Zhibin Wang, Gang Yu, Bin Fu, and Hanwang Zhang. 2023. Chartllama: A multimodal LLM for chart understanding and generation. *CoRR*, abs/2311.16483.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326, Dublin, Ireland. Association for Computational Linguistics.

Xingwei He, Zhenghao Lin, Yeyun Gong, A-Long Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, and Weizhu Chen. 2023. Annollm: Making large language models to be better crowdsourced annotators. *CoRR*, abs/2303.16854.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the MATH dataset. In *Proc. of NeurIPS*.

Sirui Hong, Xiawu Zheng, Jonathan Chen, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, and Chenglin Wu. 2023. Metagpt: Meta programming for multi-agent collaborative framework. *CoRR*, abs/2308.00352.

Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. 2023a. Unnatural instructions: Tuning language models with (almost) no human labor. In

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 14409–14428, Toronto, Canada. Association for Computational Linguistics.

Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. 2023b. Unnatural instructions: Tuning language models with (almost) no human labor. In *ACL*, pages 14409–14428. Association for Computational Linguistics.

Tom Hosking, Phil Blunsom, and Max Bartolo. 2023. Human feedback is not gold standard. *CoRR*, abs/2309.16349.

Zhiqiang Hu, Lei Wang, Yihuai Lan, Wanyu Xu, Ee-Peng Lim, Lidong Bing, Xing Xu, Soujanya Poria, and Roy Lee. 2023. LLM-adapters: An adapter family for parameter-efficient fine-tuning of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5254–5276, Singapore. Association for Computational Linguistics.

Jiaxin Huang, Shixiang Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2023. Large language models can self-improve. In *EMNLP*, pages 1051–1068. Association for Computational Linguistics.

Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and Philip S. Yu. 2022. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE Trans. Neural Networks Learn. Syst.*, 33(2):494–514.

Martin Josifoski, Marija Sakota, Maxime Peyrard, and Robert West. 2023. Exploiting asymmetry for synthetic training data generation: Synthie and the case of information extraction. In *EMNLP*, pages 1555–1574. Association for Computational Linguistics.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. 2023. Prometheus: Inducing fine-grained evaluation capability in language models. *CoRR*, abs/2310.08491.

Su Young Kim, Hyeon-Jin Park, Kyuyong Shin, and Kyung-Min Kim. 2022. Ask me what you need: Product retrieval using knowledge from GPT-3. *CoRR*, abs/2207.02516.

Jan Kocon, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydlo, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, Anna Kocon, Bartlomiej Koptyra, Wiktoria Mieleszczenko-Kowszewicz, Piotr Milkowski, Marcin Oleksy, Maciej Piasecki, Lukasz Radlinski, Konrad Wojtasik, Stanislaw Wozniak, and Przemyslaw Kazienko. 2023. Chatgpt: Jack of all trades, master of none. *Inf. Fusion*, 99:101861.

Anastasia Kritharoula, Maria Lymperaiou, and Giorgos Stamou. 2023. Large language models and multimodal retrieval for visual word sense disambiguation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13053–13077, Singapore. Association for Computational Linguistics.

Alexey Kurakin, Natalia Ponomareva, Umar Syed, Liam MacDermed, and Andreas Terzis. 2023. Harnessing large-language models to generate private synthetic text. *arXiv preprint arXiv:2306.01684*.

Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. 2019. An evaluation dataset for intent classification and out-of-scope prediction. In *Proc. of EMNLP*.

Nayeon Lee, Wei Ping, Peng Xu, Mostofa Patwary, Pascale N Fung, Mohammad Shoeybi, and Bryan Catanzaro. 2022. Factuality enhanced language models for open-ended text generation. *NeurIPS*.

Bryan Li and Chris Callison-Burch. 2023. PAXQA: Generating cross-lingual question answering examples at training scale. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 439–454, Singapore. Association for Computational Linguistics.

Junlong Li, Zhuosheng Zhang, and Hai Zhao. 2022. Self-prompting large language models for open-domain QA. *CoRR*, abs/2212.08635.

Minzhi Li, Taiwei Shi, Caleb Ziems, Min-Yen Kan, Nancy Chen, Zhengyuan Liu, and Diyi Yang. 2023a. CoAnnotating: Uncertainty-guided work allocation between human and large language models for data annotation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1487–1505, Singapore. Association for Computational Linguistics.

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023b. Textbooks are all you need II: phi-1.5 technical report. *CoRR*, abs/2309.05463.

Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. 2023c. Synthetic data generation with large language models for text classification: Potential and limitations. In *EMNLP*, pages 10443–10461. Association for Computational Linguistics.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2022a. WANLI: Worker and AI collaboration for natural language inference dataset creation.

In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6826–6847, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023. Improved baselines with visual instruction tuning. *CoRR*, abs/2310.03744.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022b. What makes good in-context examples for gpt-3? In *DeeLIO@ACL*, pages 100–114. Association for Computational Linguistics.

Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, and Andrew M. Dai. 2024. Best practices and lessons learned on synthetic data for language models. *CoRR*, abs/2404.07503.

Yuzhe Lu, Sungmin Hong, Yash Shah, and Panpan Xu. 2023. Effectively fine-tune to improve large multimodal models for radiology report generation. *CoRR*, abs/2312.01504.

Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, and Dongmei Zhang. 2023a. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. *CoRR*, abs/2308.09583.

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

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022. Generating training data with language models: Towards zero-shot language understanding. In *NeurIPS*.

Yu Meng, Martin Michalski, Jiaxin Huang, Yu Zhang, Tarek Abdelzaher, and Jiawei Han. 2023. Tuning language models as training data generators for augmentation-enhanced few-shot learning. In *ICML*, pages 24457–24477. PMLR.

Arindam Mitra, Luciano Del Corro, Shweti Mahajan, Andrés Codas, Clarisse Simões, Sahaj Agrawal, Xuxi Chen, Anastasia Razdaibiedina, Erik Jones, Kriti Aggarwal, Hamid Palangi, Guoqing Zheng, Corby Rosset, Hamed Khanpour, and Ahmed Awadallah. 2023. Orca 2: Teaching small language models how to reason. *CoRR*, abs/2311.11045.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. Orca: Progressive learning from complex explanation traces of GPT-4. *CoRR*, abs/2306.02707.

Roberto Navigli, Simone Conia, and Björn Ross. 2023. Biases in large language models: Origins, inventory, and discussion. *ACM J. Data Inf. Qual.*, 15(2):10:1–10:21.

Seokjin Oh, Su Ah Lee, and Woohwan Jung. 2023. Data augmentation for neural machine translation using generative language model. *CoRR*, abs/2307.16833.

Nicholas Pangakis, Samuel Wolken, and Neil Fasching. 2023. Automated annotation with generative AI requires validation. *CoRR*, abs/2306.00176.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Runchu Tian, Ruobing Xie, Jie Zhou, Mark Gerstein, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2023. Toolllm: Facilitating large language models to master 16000+ real-world apis. *CoRR*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Jon Saad-Falcon, Omar Khattab, Keshav Santhanam, Radu Florian, Martin Franz, Salim Roukos, Avirup Sil, Md Sultan, and Christopher Potts. 2023. UDAPDR: Unsupervised domain adaptation via LLM prompting and distillation of rerankers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11265–11279, Singapore. Association for Computational Linguistics.

Gaurav Sahu, Pau Rodriguez, Issam Laradji, Parmida Atighehchian, David Vazquez, and Dzmitry Bahdanau. 2022. Data augmentation for intent classification with off-the-shelf large language models. In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 47–57, Dublin, Ireland. Association for Computational Linguistics.

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

Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. 2023. Synthetic prompting: Generating chain-of-thought demonstrations for large language models. In *ICML*, volume 202 of *Proceedings of Machine Learning Research*, pages 30706–30775. PMLR.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. Hugginggpt: Solving AI tasks with chatgpt and its friends in huggingface. *CoRR*, abs/2303.17580.

Avi Singh, John D. Co-Reyes, Rishabh Agarwal, Ankesh Anand, Piyush Patil, Xavier Garcia, Peter J. Liu, James Harrison, Jaehoon Lee, Kelvin Xu, Aaron Parisi, Abhishek Kumar, Alex Alemi, Alex Rizkowsky, Azade Nova, Ben Adlam, Bernd Bohnet, Gamaleldin F. Elsayed, Hanie Sedghi, Igor Mordatch, Isabelle Simpson, Izzeddin Gur, Jasper Snoek, Jeffrey Pennington, Jiri Hron, Kathleen Kenealy, Kevin Swersky, Kshiteej Mahajan, Laura Culp, Lechao Xiao, Maxwell L. Bileschi, Noah Constant, Roman Novak, Rosanne Liu, Tris Warkentin, Yundi Qian, Yamini Bansal, Ethan Dyer, Behnam Neyshabur, Jascha Sohl-Dickstein, and Noah Fiedel. 2023. Beyond human data: Scaling self-training for problem-solving with language models. *CoRR*, abs/2312.06585.

Ryan Smith, Jason A. Fries, Braden Hancock, and Stephen H. Bach. 2022. Language models in the loop: Incorporating prompting into weak supervision. *CoRR*, abs/2205.02318.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Santilli, Andreas Stuhlmüller, Andrew M. Dai, Andrew La, Andrew K. Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakas, and et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *CoRR*.

Hongjin Su, Jungo Kasai, Chen Henry Wu, Weijia Shi, Tianlu Wang, Jiayi Xin, Rui Zhang, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Tao Yu. 2023. Selective annotation makes language models better few-shot learners. In *ICLR*. OpenReview.net.

Shivchander Sudalairaj, Abhishek Bhandwaldar, Aldo Pareja, Kai Xu, David D. Cox, and Akash Srivastava. 2024. LAB: large-scale alignment for chatbots. *CoRR*, abs/2403.01081.

Zhiqing Sun, Yikang Shen, Qinhong Zhou, Hongxin Zhang, Zhenfang Chen, David D. Cox, Yiming Yang, and Chuang Gan. 2023. Principle-driven self-alignment of language models from scratch with minimal human supervision. *CoRR*, abs/2305.03047.

Ruixiang Tang, Xiaotian Han, Xiaoqian Jiang, and Xia Hu. 2023. Does synthetic data generation of llms help clinical text mining? *CoRR*, abs/2303.04360.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Fanqi Wan, Xinting Huang, Tao Yang, Xiaojun Quan, Wei Bi, and Shuming Shi. 2023. Explore-instruct: Enhancing domain-specific instruction coverage through active exploration. In *EMNLP*, pages 9435–9454. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. 2023a. Openchat: Advancing open-source language models with mixed-quality data. *arXiv preprint arXiv:2309.11235*.

Ruida Wang, Wangchunshu Zhou, and Mrinmaya Sachan. 2023b. Let's synthesize step by step: Iterative dataset synthesis with large language models by extrapolating errors from small models. In *EMNLP (Findings)*, pages 11817–11831. Association for Computational Linguistics.

Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Want to reduce labeling cost? GPT-3 can help. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4195–4205, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yidong Wang, Zhuohao Yu, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xingxu Xie, Wei Ye, Shi-Bo Zhang, and Yue Zhang. 2023c. Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization. *ArXiv*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023d. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

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

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak,

Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. 2022. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Fusheng Wei, Robert Keeling, Nathaniel Huber-Fliflet, Jianping Zhang, Adam Dabrowski, Jingchao Yang, Qiang Mao, and Han Qin. 2023a. Empirical study of LLM fine-tuning for text classification in legal document review. In *IEEE Big Data*, pages 2786–2792. IEEE.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and Quoc V. Le. 2024. Long-form factuality in large language models. *CoRR*, abs/2403.18802.

Yuxiang Wei, Zhe Wang, Jiawei Liu, Yifeng Ding, and Lingming Zhang. 2023b. Magicoder: Source code is all you need. *CoRR*, abs/2312.02120.

Le Xiao and Xiaolin Chen. 2023. Enhancing LLM with evolutionary fine tuning for news summary generation. *CoRR*, abs/2307.02839.

Ruixuan Xiao, Yiwen Dong, Junbo Zhao, Runze Wu, Minmin Lin, Gang Chen, and Haobo Wang. 2023. FreeAL: Towards human-free active learning in the era of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14520–14535, Singapore. Association for Computational Linguistics.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023a. Wizardlm: Empowering large language models to follow complex instructions. *CoRR*, abs/2304.12244.

Ran Xu, Hejie Cui, Yue Yu, Xuan Kan, Wenqi Shi, Yuchen Zhuang, Wei Jin, Joyce C. Ho, and Carl J. Yang. 2023b. Knowledge-infused prompting: Assessing and advancing clinical text data generation with large language models. *CoRR*, abs/2311.00287.

Jiacheng Ye, Jiahui Gao, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong. 2022a. ZeroGen: Efficient zero-shot learning via dataset generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11653–11669, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jiacheng Ye, Jiahui Gao, Zhiyong Wu, Jiangtao Feng, Tao Yu, and Lingpeng Kong. 2022b. ProGen: Progressive zero-shot dataset generation via in-context feedback. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3671–3683, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jiacheng Ye, Chengzu Li, Lingpeng Kong, and Tao Yu. 2023. Generating data for symbolic language with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8418–8443, Singapore. Association for Computational Linguistics.

Kang Min Yoo, Dongju Park, Jaewook Kang, Sang-Woo Lee, and Woo-Myoung Park. 2021. Gpt3mix: Leveraging large-scale language models for text augmentation. In *EMNLP*, pages 2225–2239. Association for Computational Linguistics.

Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T. Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. 2023a. Metamath: Bootstrap your own mathematical questions for large language models. *CoRR*, abs/2309.12284.

Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander Ratner, Ranjay Krishna, Jiaming Shen, and Chao Zhang. 2023b. Large language model as attributed training data generator: A tale of diversity and bias. *CoRR*, abs/2306.15895.

Yue Yu, Yuchen Zhuang, Rongzhi Zhang, Yu Meng, Jiaming Shen, and Chao Zhang. 2023c. ReGen: Zero-shot text classification via training data generation with progressive dense retrieval. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11782–11805, Toronto, Canada. Association for Computational Linguistics.

Chaoning Zhang, Chenshuang Zhang, Sheng Zheng, Yu Qiao, Chenghao Li, Mengchun Zhang, Sumit Kumar Dam, Chu Myaet Thwal, Ye Lin Tun, Le Luang Huy, Dong Uk Kim, Sung-Ho Bae, Lik-Hang Lee, Yang Yang, Heng Tao Shen, In So Kweon, and Choong Seon Hong. 2023a. A complete survey on generative AI (AIGC): is chatgpt from GPT-4 to GPT-5 all you need? *CoRR*, abs/2303.11717.

Jieyu Zhang, Bohan Wang, Xiangchen Song, Yujing Wang, Yaming Yang, Jing Bai, and Alexander Ratner. 2022. Creating training sets via weak indirect supervision. In *ICLR*. OpenReview.net.

Ruoyu Zhang, Yanzeng Li, Yongliang Ma, Ming Zhou, and Lei Zou. 2023b. LLMaAA: Making large language models as active annotators. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13088–13103, Singapore. Association for Computational Linguistics.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Proc. of NeurIPS*.

Jiachen Zhao, Wenlong Zhao, Andrew Drozdov, Benjamin Rozonoyer, Md. Arafat Sultan, Jay-Yoon Lee, Mohit Iyyer, and Andrew McCallum. 2023a. Multi-stage collaborative knowledge distillation from large language models. *CoRR*, abs/2311.08640.

Zilong Zhao, Robert Birke, and Lydia Chen. 2023b. Tabula: Harnessing language models for tabular data synthesis. *arXiv preprint arXiv:2310.12746*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena.

Yiming Zhu, Peixian Zhang, Ehsan ul Haq, Pan Hui, and Gareth Tyson. 2023. Can chatgpt reproduce human-generated labels? A study of social computing tasks. *CoRR*, abs/2304.10145.

A Data Annotation

In the main text, we introduced a series of techniques for general data synthesis. Though annotation can be considered a special type of synthesis with the input of a particular sample as the synthesis condition, there are also approaches specifically suitable for data annotation. Among them, *selective annotation* is one of the most important practices. Selective annotation represents an optimal tradeoff between expensive and precise human annotation and economic but relatively rough LLMs-based annotation(Wang et al., 2021; Kocon et al., 2023).

The key to selective annotation is to define a "cost-effective" sample distribution between humans and LLMs. (Zhang et al., 2023b; Bansal and Sharma, 2023) covers some common selection strategies for LLMs-based annotation, including random selection, maximum entropy selection, least confidence selection and $k$means selection for thorough comparisons. Results show that uncertainty-based methods, i.e. maximal entropy and least confidence, perform significantly better than the random baseline, with faster convergence and better performance of the downstream model trained on the annotated data. (Li et al., 2023a) also utilizes uncertainty to estimate LLMs' annotation capability to effectively allocate the annotation work among humans and LLMs. (Su et al., 2023) instead proposes a novel unsupervised, graph-based selective annotation method named vote-k, to select diverse and representative examples to annotate.

B Tuning Techniques

Another large body of research pertains to the *tuning techniques*, such as model fine-tuning (Zhao et al., 2023b; Sun et al., 2023; Meng et al., 2023; Kurakin et al., 2023) and soft prompting (Chen et al., 2023a), which have already been heavily studied in other fields and can be detailedly referred in (Hu et al., 2023; Lu et al., 2023; Wei et al., 2023a; Xiao and Chen, 2023). Despite their effectiveness in improving the generation performance, most of the existing approaches are established on the accessibility of the LLMs, while their application on black-box models remains to be further explored.

C Applications

LLM-driven synthetic data generation has served as a new alternative to traditional human-dependent data collection and demonstrated great potential in various applications, including general tasks, domain-specific tasks, and multimodal tasks.

Generic Tasks. With the exploding capabilities of LLMs, this generation pipeline has been adopted in a wide range of basic NLP studies, including text classification (Ye et al., 2022b; Yu et al., 2023c; Sahu et al., 2022), named entity recognition (Xiao et al., 2023), question answering (Li and Callison-Burch, 2023), relationship extraction (He et al., 2023), and natural language inference (Zhang et al., 2023b). These studies further underpin diverse applications, such as sentiment recognition (Gao et al., 2023a; Ye et al., 2022b), online translation (Oh et al., 2023), stance detection (Li et al., 2023a) and spam identification (Smith et al., 2022).

Domain-specific Tasks. Some domain-specific tasks also impose significant demands on this pipeline, where human annotation can be extremely expensive and impractical, such as medical diagnosis (Tang et al., 2023), drug discovery (Xiao et al., 2023), clinical trial extraction (Xu et al., 2023b), industrial advertisement (Zhang et al., 2022) and tabular data analysis (Seedat et al., 2023).

Multimodal Tasks. Stemming from the simplicity and low cost, this generation paradigm has also exhibited significant promise in multimodal tasks, including text-image retrieval (Kritharoula et al., 2023), chat understanding (Han et al., 2023), visual question answering (Han and Gardent, 2023), and multimodal instruction tuning (Liu et al., 2023).

Type	Benchmark Dataset	Subdataset Quantity	Partial Subdataset	Task	Ability	Domain/Data Source
Classification	SMS spam (Almeida et al., 2011; Li et al., 2023c)	1	SMS spam	Text Classification	Spam Detection	SMS
	AG News (Zhang et al., 2015; Li et al., 2023c)	1	AG News	Text Classification	Topic Classification	News
	IMDb (Maas et al., 2011; Li et al., 2023c; Wang et al., 2023b)	1	IMDb	Text Classification	Binary Sentiment Classification	Review
	GoEmotions (Demszky et al., 2020; Li et al., 2023c)	1	GoEmotions	Text Classification	Sentiment Classification	Reddit Comments
	CLINC150 (Larson et al., 2019; Sahu et al., 2022)	1	CLINC150	Text Classification	Intent Detection	Human Annotation
	BANKING77 (Casanueva et al., 2020; Sahu et al., 2022)	1	BANKING77	Text Classification	Intent Detection	Bank
	FewRel (Gao et al., 2019; Li et al., 2023c)	1	FewRel	Text Classification	Relation Classification	Wikipedia
	GLUE (Wang et al., 2018, 2023b)	7	QNLI	Natural Language Inference	Recognizing Textual Entailment	Wikipedia
			RTE	Natural Language Inference	Recognizing Textual Entailment	News and Wikipedia
QA	AdversarialQA (Bartolo et al., 2020; Wang et al., 2023b)	1	AdversarialQA	Question Answering	Reading Comprehension	Wikipedia
	TruthfulQA (Lin et al., 2022; Sun et al., 2023)	1	TruthfulQA	Question Answering	Honestness	Hard Data
Reasoning	MATH (Hendrycks et al., 2021; Wan et al., 2023)	1	MATH	mathematical reasoning	Complex Reasoning	Math
	ToolBench (Qin et al., 2023)	**1**	**ToolBench**	**Trajectory Planning**	**Tool manipulation**	**Tool**
-	NIV2 (Wang et al., 2022, 2023e)	1616	-	-	Language Understanding & Reasoning	Benchmark Collection/Human Annotation
-	BIG-bench (Srivastava et al., 2022; Sun et al., 2023)	204	-	-	Language Understanding & Reasoning	Human Annotation

Table 1: Representative benchmark dataset for assessing models trained with generated data. The dataset generated based on LLM is highlighted in bold.

D Benchmark Datasets

In Table 1, we summarize representative benchmark datasets for evaluating models trained through data generation. Among them, ToolBench (Qin et al., 2023) is generated by LLMs and is commonly employed to evaluate the performance of LLMs in tool usage proficiency. In most classification task evaluations (Li et al., 2023c; Wang et al., 2023b; Sahu et al., 2022), LLMs are infrequently used as test models; instead, small language models trained on generated data are often used, followed by testing on existing benchmarks.