

---

# SYNTHETIC DATA GENERATION USING LARGE LANGUAGE MODELS: ADVANCES IN TEXT AND CODE

---

**Mihai Nadăș**  
Babeș-Bolyai University  
mihai.nadas@ubbcluj.ro

**Laura Dioșan**  
Babeș-Bolyai University  
laura.diosan@ubbcluj.ro

**Andreea Tomescu**  
KlusAI Labs  
andreea.tomescu@klusai.com

March 19, 2025

## ABSTRACT

Large language models (LLMs) have unlocked new possibilities for generating synthetic training data in both natural language and code domains. By producing artificial but task-relevant examples, these models can significantly augment or even substitute for real-world datasets, particularly in scenarios where labeled data is scarce, expensive, or sensitive. This paper surveys recent advances in leveraging LLMs to create synthetic text and code, highlighting key techniques such as prompt-based generation, retrieval-augmented pipelines, and iterative self-refinement. We examine how these methods can enrich low-resource tasks (e.g. classification, question answering) and facilitate code-centric applications (e.g. instruction tuning, code translation, bug repair) through automated verification of functional correctness. Alongside potential benefits—cost-effectiveness, broad coverage, and controllable diversity—we discuss the accompanying challenges, including factual inaccuracies in generated text, insufficient stylistic or distributional realism, and risks of bias amplification. Proposed mitigation strategies range from filtering and weighting synthetic outputs to reinforcement learning with execution feedback in code domains. We conclude by outlining open research directions, such as automated prompt engineering, cross-modal data synthesis, and robust evaluation frameworks, underscoring the growing importance of LLM-generated synthetic data in accelerating AI development while emphasizing ethical and quality safeguards.

## 1 Introduction

Large Language Models (LLMs) have achieved remarkable success across natural language and code generation tasks, owing in part to training on massive datasets. However, acquiring sufficient high-quality training data remains a bottleneck in many domains [60]. Data scarcity, high annotation costs, and privacy constraints often limit the availability of large supervised corpora. These challenges have spurred growing interest in synthetic data generation, where additional training examples are produced artificially rather than collected from the real world. Recent advances in generative AI – particularly LLMs like Anthropic’s Claude 3.7 Sonnet, DeepSeek’s R1, Meta’s Llama 3, and OpenAI’s GPT-o3 – provide powerful new tools to generate synthetic text and code that mimic real data distributions. This paper surveys and analyzes the latest developments in LLM-driven synthetic data generation for both natural language text and programming code, highlighting techniques, applications, challenges, and future directions.

LLMs can produce human-like text and code, making them attractive data generators for tasks where obtaining real data is costly or infeasible. For example, instead of manually labeling thousands of sentences for a classifier, one can prompt an LLM to create diverse labeled examples [33], or have it generate code snippets to augment code model training. Synthetic data generated by LLMs has shown promise in boosting model performance in low-resource settings, reducing annotation costs, and enabling data augmentation for improved robustness [11]. At the same time, using LLMs as data generators raises questions about quality control, realism, and biases in the generated data.

A comprehensive review of LLM-based approaches for text and code generation is currently lacking, despite their increasing adoption [52]. This paper aims to fill that gap by providing a structured survey of existing techniques and challenges. The focus is specifically on text and code generation, as opposed to other forms of data, due to several

key reasons. First, visual data synthesis has been extensively studied, making an additional review redundant [10, 8]. Second, while synthetic signal data (e.g., audio) is a developing field, it remains outside the scope of this study, though future work could extend these analyses [24]. Lastly, text and code share a common modality, as both are purely symbolic representations that can be generated and processed using the same fundamental methods—unlike images, sounds, or videos, which require distinct encoding formats and specialized equipment [29].

This paper provides a comprehensive analysis of LLM-based synthetic data generation, with contributions summarized as follows:

- **Survey of Techniques:** We review major approaches for LLM-driven data generation in text and code, including prompt-based augmentation, retrieval-augmented generation, self-instruct methods, and reinforcement learning with feedback. We categorize methods by their strategies (e.g., zero-shot vs. few-shot prompting, knowledge integration, iterative refinement) and discuss representative examples from recent literature [3].
- **Advances in Text Data Generation:** We explore how LLMs are used to create synthetic text datasets for tasks such as classification, question answering, and instruction-following. Key studies are examined to understand the impact of synthetic data on model performance, data diversity, and efficiency. We highlight empirical findings (e.g., improvements of 3–26% with synthetic augmentation in low-data regimes) and analyze how prompt design and data curation affect quality [33].
- **Advances in Code Data Generation:** We survey the parallel emergence of LLM-generated code data for training code intelligence models. We discuss techniques for generating code through LLM prompts, executing code to validate correctness, and synthesizing coding instruction data at scale [37]. Examples include synthetic programming problems and solutions, code edits for diversity, and instruction-tuning datasets (e.g., Code Alpaca, WizardCoder) generated by LLMs.
- **Challenges and Mitigations:** We outline the main challenges in using LLMs for synthetic data, such as ensuring data fidelity and factuality, avoiding modeling biases, maintaining diversity without drifting from real data distribution, and preventing *model collapse* from iterative self-training on AI-generated data. We discuss solutions proposed in literature, like filtering and weighting synthetic examples [33], blending synthetic and real data [18], and leveraging feedback (e.g., executors for code, retrieval for facts) to improve quality [17] [28].
- **Future Directions:** We identify open research directions, including more rigorous evaluation metrics for synthetic data quality, techniques for controlling generation (to target rare phenomena or reduce biases), expanding synthetic data generation to multimodal and low-resource language settings, and frameworks for safe and privacy-preserving synthetic data in industry applications. Throughout, we connect recent findings to prior work to provide a well-rounded perspective on how LLM-based synthetic data generation is shaping the future of NLP and software engineering research.

The rest of this paper is organized as follows. Section 2 provides background on synthetic data and early data augmentation approaches. Section 3 and 4 detail advances in synthetic data generation for text and code, respectively. Section 5 discusses overarching challenges and comparative insights. Section 6 suggests promising future research directions, and Section 7 concludes the paper.

## 2 Background and Motivation

### 2.1 Synthetic Data

**Synthetic Data** refers to data that is artificially generated rather than directly collected from real-world events or annotations. In the context of AI/ML, synthetic data is created via algorithms or generative models (such as neural networks) to **mimic the characteristics** of real data [37].

Early uses of synthetic data in machine learning date back decades (e.g., artificially generated images or simulated data for computer vision and robotics) [48]. In Natural Language Processing, traditional *data augmentation* techniques like synonym replacement, random insertion/deletion, and back-translation have long been used to expand text training sets. For example, in machine translation, **back-translation** [48] generates synthetic parallel sentences by translating monolingual target-language data back into the source language, which significantly improved translation quality. Similarly, for text classification, methods like EDA (Easy Data Augmentation) [54] applied random word-level transformations to create variant training sentences, helping reduce overfitting on small datasets. These early augmentation approaches demonstrated that even modest synthetic perturbations can improve model robustness and performance when real data is limited [56].

However, simple rule-based augmentations have limitations. They often do not introduce truly new linguistic patterns or semantic variations, and thus only provide limited diversity. As models grew in size and tasks became more complex, the need for more **diverse and realistic** synthetic data became evident [61]. This paved the way for using **generative models** to create synthetic data.

Early neural approaches included using sequence-to-sequence models or language models to paraphrase text or generate new examples. For instance, unsupervised data augmentation (UDA) techniques combined back-translation with consistency training [56] to leverage unlabeled data, and knowledge distillation methods had high-capacity teacher models label unlabeled examples for student training (a form of synthetic labeling)[22]. These efforts hinted at the potential of **model-generated data** to stand in for human annotations.

The advent of large pre-trained language models brought synthetic data generation to a new level.

**Large Language Models** (LLMs) (with billions of parameters, trained on vast corpora) are capable of producing high-quality, coherent text that often closely resembles human-written text [64]. Importantly, LLMs can be guided via prompts to generate data for specific tasks. For example, given a task description and a few examples (*few-shot prompting*), an LLM can generate additional plausible examples for that task [33]. Even without any examples (*zero-shot*), a well-crafted prompt can elicit relevant outputs for many tasks. This makes LLMs extremely flexible data generators, effectively serving as **universal data augmenters** that can create labeled data for a wide range of problems on demand [61].

Several factors motivate the use of LLMs for synthetic data generation:

- **Data Scarcity and Cost:** In many NLP tasks (and domains like healthcare or finance), annotated data is scarce or expensive to obtain [37]. LLMs offer a cheap and fast alternative to manual labeling. For instance, labeling 3,000 sentences for sentiment analysis was estimated to cost hundreds of dollars and many hours of human work, whereas GPT-3 could generate a similar number of examples in minutes for a fraction of the cost [33]. Table 1 illustrates this cost and time disparity, as reported by [11] for the SST-2 sentiment task. The Stanford Sentiment Treebank (SST) [50] is a widely used dataset in Natural Language Processing (NLP), particularly for sentiment analysis and text classification. It consists of sentences extracted from movie reviews, annotated with sentiment labels at different levels of granularity.  
The SST-2 variant is a binary classification task where each sentence is labeled as either positive or negative sentiment. It is often used to train and evaluate sentiment analysis models, testing their ability to correctly classify text based on emotional tone. Given its structured annotations and real-world applicability, SST-2 serves as a benchmark for evaluating machine learning models' text classification capabilities.
- **Scalability:** Synthetic data can be generated at virtually unlimited scale. Once an effective prompt or generation pipeline is established, one can synthesize thousands or millions of examples to bolster training. This ability to **mass-produce data** is valuable for training data-hungry models and for creating balanced datasets (e.g., generating more minority class examples to address class imbalance). Researchers predict that real text data available for training may plateau in coming decades [39], in which case synthetic data could play a crucial role in sustaining model improvements.
- **Controllability:** Unlike collected data, synthetic data generation can be tailored to specific needs. By adjusting prompts or generation criteria, one can target particular **attributes** in the data. For example, one can ensure coverage of edge cases, generate counterfactual examples to debias a model, or produce data in a specific style or reading level. Synthetic data thus provides **fine-grained control** over dataset composition (e.g., augmenting low-resource languages or rarely seen contexts) [37]. This is difficult to achieve with natural data collection alone.
- **Privacy Preservation:** In scenarios where real data is sensitive (user data, medical records, etc.), sharing or using it directly raises privacy concerns. Synthetic data that captures aggregate patterns of the sensitive data without revealing individual-specific information can serve as a privacy-friendly alternative. For instance, an LLM could be used to generate clinical notes that are stylistically and medically similar to real notes but contain no real patient information [37]. This enables research and model training in domains with stringent privacy requirements, effectively **anonymizing** data via generation.
- **Diversity and Robustness:** LLMs can introduce variations that humans might not think of, potentially surfacing novel phrasings or logic. This can lead to **more robust models**. By training on a wide range of synthetic scenarios (including adversarial or uncommon cases), models may generalize better and be less prone to overfitting idiosyncrasies of a small real dataset [33]. For example, generating paraphrases or semantically equivalent rephrasings of existing sentences can teach a model to handle linguistic variability.

Table 1: Cost and time comparison: Human annotation vs LLM-generated data (for 3,000 samples in SST-2 sentiment task). Synthetic data generation by GPT-3 dramatically reduces both the time and monetary cost, though resulting model accuracy is slightly lower than using human-curated data. [11]

Data Source	Samples (SST-2)	Approx. Cost	Time Required	Model Accuracy (on SST-2)
Human-Labeled Data	3,000	\$221–\$300	~1000 minutes	88%
GPT-3 (Synthetic)	3,000	\$14.37	~46 minutes	(n/a for 3k)
GPT-3 (Synthetic)	6,000	\$28.74 (est.)	~92 minutes	76%

*Note:* With 6,000 GPT-3 generated samples, a classifier achieved 76% accuracy, compared to 88% with 3,000 human-labeled samples. This highlights a quality gap, but also the improvement gained by doubling synthetic data volume.

## 2.2 Prompt-Based Data Augmentation with LLMs

Prompting techniques are at the core of LLM-driven text data generation. By providing an LLM with a carefully designed prompt, possibly including instructions and examples, one can guide it to produce new data instances for a target task [33]. Several prompting strategies exist:

- **Zero-Shot Generation:** The prompt contains only a task instruction or description, and the LLM must generate an example from scratch [64]. For instance, the instruction *"Generate a new sentence expressing positive sentiment about a product"* demonstrates this method. Zero-shot prompts rely on the model’s general knowledge and often yield reasonable but sometimes generic outputs. Zero-shot generation is straightforward but may produce limited diversity if used repeatedly with the same prompt.
- **One-Shot Generation:** The prompt provides the task description plus one example (input-output pair) to illustrate the format. The LLM then creates a similar example. This single demonstration can anchor the style of outputs (closer to the given example), which can improve relevance at the cost of some diversity [12].
- **Few-Shot Generation:** The prompt includes multiple examples (e.g., 3–5) of the task. The LLM then generates additional examples in line with the given ones. Few-shot prompting often yields higher-quality, task-specific outputs since the model can mimic patterns from the examples, making it particularly useful when the task has a specific structure or label space [43]. However, as with one-shot, there is a risk of **repeating patterns** from the limited context, reducing diversity.
- **Topic/Controlled Generation:** An advanced variant introduced by Yu et al. [62] is *zero-shot topic generation*. Here, one first uses the LLM to generate a list of relevant topics or scenarios for the task, then prompts the LLM with a selected topic to generate a specific example. This approach injects controlled diversity: each generation focuses on a different topic, leading to a more varied dataset. For instance, for sentiment classification, the LLM might generate topics like “restaurant review”, “electronics product review”, “movie review”, etc., then produce a synthetic example for each. This was found to increase the diversity of synthetic datasets and often improved model training results.

In summary, the use of LLMs for data generation is driven by the promise of **cheaper, faster, and more flexible data creation**. Nonetheless, it comes with the responsibility of ensuring that this synthetic data is as **effective and reliable** as real data. The following sections explore how researchers are harnessing LLMs to generate synthetic text and code, and what strategies are used to maximize utility while mitigating downsides.

## 3 Synthetic Data Generation for Text Tasks

Using LLMs to generate synthetic text data has become increasingly popular for tasks like text classification, question answering, summarization, and dialog systems. We discuss prominent approaches and findings in LLM-based text data generation, from simple prompt-driven augmentation to complex pipelines that ensure quality and diversity. We also incorporate insights from recent empirical studies that evaluate the impact of synthetic data on downstream task performance.

### 3.1 Various approaches for Prompt-Based Synthetic Text Data Generation

Prompt-based generation has been empirically shown to produce useful training data. A case study by Zhuoyan Li et al. [35] used GPT-3.5 to generate text classification data under various prompting methods. They evaluated the quality by training classifiers on the synthetic data and testing on real validation sets. A key finding was that in low-resource settings, synthetic data significantly boosts performance. Specifically, when only 100 real training examples

were available, augmenting with 100 additional GPT-generated examples yielded 3%–26% absolute improvement in accuracy/F1 across several classification tasks. This gain was most pronounced in tasks where the model was severely underfit due to data paucity (e.g., a 26% jump in a news classification task with few real samples) [3]. However, as the amount of real data increases, the marginal benefit of synthetic data diminishes. With 1000 real examples, Yu et al. [60] observed only minor improvements (<5%) from adding synthetic data, and in some tasks no improvement. This suggests that synthetic augmentation is most useful for low-data regimes, and that purely synthetic data still has a quality gap compared to large real datasets [60]. In fact, prior works noted that models trained on an equal number of synthetic vs. real examples usually perform worse with the synthetic ones [60, 51]. Thus, synthetic data is better viewed as a supplement to small real datasets, rather than a wholesale replacement for large real corpora (except in zero-data scenarios) [39, 61].

Another insight from prompt-based generation studies is **the trade-off between diversity and relevance**. Zero-shot prompting can produce a wide range of outputs if the model isn’t constrained by examples, but some outputs may be off-target or less relevant [11]. Few-shot prompting, on the other hand, produces on-target outputs but risks many duplicates or very similar phrasing since the model mimics the few given examples [24]. Yu et al. [39] found that the zero-shot topic method, which explicitly varied the prompt for each generation via different topics, yielded the most diverse synthetic datasets and often the best downstream performance in the synthetic-only setting [27]. Datasets with higher diversity (covering varied subtopics) tended to train more robust models [62]. Conversely, a plain zero-shot method (same prompt each time) gave the least diverse data and worst performance on several tasks [62]. Interestingly, when synthetic data was mixed with real data, the importance of maximizing diversity dropped, as the real data provided a baseline of variation. In such cases, even one-shot or few-shot generated data (which may be conservative in diversity) helped by aligning well with real examples [62]. The best practice emerging is to use prompt designs that encourage diversity (via randomization or controlled variation) especially if relying heavily on synthetic data, while ensuring the generated examples remain realistic and on-task (using in-context examples or instructions to ground the model) [15].

**Advanced Prompting Strategies** Researchers have also developed **iterative prompting** strategies to refine synthetic data quality. For instance, *Self-Instruct* [53] is an approach where after an initial round of generation, one identifies difficult cases where a model still makes mistakes (misclassified examples) and then *iteratively prompts the LLM to generate more data similar to those hard cases*. By focusing on the model’s errors, this method aims to reduce the gap between the synthetic data distribution and the true data distribution, essentially *targeting the weaknesses* of the model. Such feedback loops can yield synthetic data that incrementally pushes model performance on challenging aspects of the task. Another approach, *SunGen* [16], acknowledges that not all LLM outputs are perfect and some synthetic examples may be noisy or low-quality. SunGen applies a **weighted loss** during model training, giving lower weight to potentially noisy synthetic examples so they do not harm the model [16]. In essence, it learns with a down-weighted penalty on data suspected to be hallucinated or label-noisy. This kind of selective training can make the model more robust to imperfect synthetic data.

Table 2 summarizes a few representative methods and studies in LLM-based text data generation, highlighting their target tasks and key contributions.

Table 2: Selected LLM-based synthetic text data generation approaches and their key features.

Approach / Study	Task / Domain	Key Idea and Contribution
<b>A Survey of Data Augmentation</b> [52]	<i>General NLP</i> (various tasks)	Comprehensive survey of traditional and neural text augmentation methods, laying groundwork for modern LLM-based augmentation. Highlights early techniques (synonym replacement, back-translation) and their impact.
<b>Is GPT-3 a Good Data Annotator?</b> [11]	<i>Text Classification</i> (SST-2, etc.)	Empirical study on using GPT-3 to <b>annotate data</b> . Found GPT-3 can generate labeled data cheaply and quickly, yielding reasonably high model accuracy, though not fully matching human data quality. Suggests LLM-generated labels are a cost-effective alternative to manual annotation.
<b>GPT3Mix</b> [58]	<i>Text Classification</i> (low-resource)	Proposed mixing GPT-3 generated examples with real data to improve low-resource classification. Uses prompt-based generation to create new examples that are then <b>mixed</b> into training. Demonstrated improved accuracy on several classification benchmarks with limited real data.



Continuation of Table 2		
Approach / Study	Task / Domain	Key Idea and Contribution
WANLI [36]	<i>Natural Language Inference</i>	Created a <b>synthetic NLI dataset</b> without human labeling. Used an LLM to generate challenging NLI examples (with label inference) focusing on corners of the feature space. Models trained on WANLI plus standard data showed improved generalization to tougher NLI examples.
InPars [1]	<i>Information Retrieval (QA)</i>	Used a T5-based LLM to generate <b>search queries</b> from passages (inverse of typical QA). These synthetic query–passage pairs augmented the training of retriever models, significantly boosting performance on web search and QA retrieval tasks. Demonstrates using LLMs to generate realistic information-seeking queries.
Unnatural Instructions [23]	<i>Instruction-Following (general NLP)</i>	Produced a large dataset of <b>synthetic instruction-output pairs</b> by prompting GPT-3 with task instructions (not from users, hence "unnatural"). Aimed to supplement scarce human-written instructions for training instruction-following models. This work was foundational for models like <b>Alpaca</b> and similar contemporaries.
Self-Instruct [53]	<i>Instruction &amp; Multi-task data</i>	A bootstrapping algorithm where a pretrained LM generates new tasks and solutions from a few seed examples. Initially applied to create <b>Stanford Alpaca</b> dataset (52k instructions) using text-davinci-003. This method showed that an LLM can recursively teach itself new tasks, yielding high-quality multi-turn instruction data without manual writing. Inspired many subsequent works in synthetic instruction generation for both text and code.
Empirical Study [35]	<i>Text Classification (6 tasks)</i>	Systematic analysis of GPT-3.5 generated data for classification. Showed <b>significant gains (3–26%)</b> with synthetic augmentation at 100 real samples, diminishing returns as data increases. Compared prompting methods: finding that <b>topic-guided generation</b> improves diversity and performance. Provided practical recommendations on how much data to generate and which prompting strategy to use in various scenarios.
AugGPT: Leveraging ChatGPT for Text Data Augmentation [9]	<i>Few-shot text classification (Amazon Reviews, PubMed20K, Symptoms Dataset)</i>	AugGPT rephrases sentences into multiple diverse yet semantically consistent versions, improving training data quality. Evaluation shows AugGPT significantly improves model accuracy compared to traditional augmentation methods (e.g., synonym replacement, back-translation). Findings suggest ChatGPT-generated synthetic data is a scalable, high-quality augmentation strategy, making it a cost-effective alternative to manual data expansion in NLP.
Self-Ask [47]	<i>Multi-hop Question Answering (QA), Reasoning</i>	Introduced an LLM-based self-asking strategy where models recursively break down complex questions into simpler sub-questions. Demonstrated improved reasoning performance on multi-step inference tasks by leveraging self-generated sub-questions for enhanced accuracy.
CoT [13]	<i>Scientific QA, Logical Reasoning</i>	Examined how Chain-of-Thought (CoT) prompting improves LLM reasoning capabilities. Proposed Sci-CoT, a two-stage framework where intermediate rationales are generated separately from answer inference, improving efficiency and model performance.
End of Table		

### 3.2 Quality and properties of generated data

While prompt-based generation can yield large quantities of data, ensuring the **quality and usefulness** of that data is crucial. A major concern is that LLMs may produce *incorrect or nonsensical outputs*, especially in knowledge-intensive tasks or if asked to generate beyond their training distribution. Two common strategies to improve quality are: (1) integrating external knowledge or constraints into generation, and (2) post-generation filtering or selection.

- Retrieval-Augmented Generation - One way to ground LLM outputs in reality is to provide relevant facts from a knowledge base or corpus during the prompt. This has given rise to *retrieval-augmented data augmentation* [3]. For example, for a question-answer pair generation, instead of having the LLM make up an answer which might be incorrect, one can retrieve a passage from Wikipedia and prompt the LLM to generate a question

whose answer is contained in that passage (and maybe also generate the answer). This way, the synthetic QA pair is guaranteed to be grounded in a real passage, improving factual correctness. Studies have shown that introducing such external knowledge significantly enhances the **faithfulness** of generated data. The LLM essentially uses the retrieval as additional context, reducing hallucination. Retrieval-based methods have been used for tasks like question answering, dialogue (retrieving profiles or factual info to ground responses), and even for generating fact-rich classification data.

- **Post-Processing and Filtering** - After generation, it is common to apply filters or selection criteria to remove low-quality synthetic data. Simple heuristics include removing exact duplicates, checking for prompt leakage (e.g., the LLM copying the input example in a few-shot prompt verbatim), or enforcing output format correctness [61]. More advanced techniques train a **critic or classifier** to judge the quality of synthetic examples [11]. For instance, one could train a classifier to distinguish real vs. synthetic data and then drop synthetic examples that are too easily identified as fake (assuming those are less realistic). Another approach is to only keep synthetic examples on which a strong model (or the LLM itself) predicts the label with high confidence, on the intuition that obvious or easy synthetic examples are likely correct. Ding et al. asked whether GPT-3 is a good annotator [11]; part of answering that is identifying when GPT-3’s generated label or text is likely wrong. They and others have used *agreement among multiple model outputs* or consistency checks as signals: if an LLM produces contradictory answers when rephrasing the query, the data point might be unreliable. A notable approach to boost quality is to incorporate **human feedback or evaluation in the loop**. While the goal is to avoid heavy human labor, having humans quickly review a sample of synthetic data can provide an estimate of quality and help calibrate automated filters. Some studies employed crowdworkers to rate the correctness of synthetic QA pairs or the grammaticality of synthetic sentences, and then used those ratings to train automatic filters (e.g., a model to predict the score) [42]. This semi-automatic process aims to capture human judgment at scale.

**Distribution Alignment** Another subtle issue is aligning the synthetic data distribution with the real data distribution of the task. Synthetic data that is too *out-of-distribution* may not actually help the model perform better on real data. For example, an LLM might generate very pedantic or overly complex sentences that, while correct, do not resemble the style of text the model will see at test time. Techniques like the *Step-by-Step* method mentioned earlier explicitly target alignment by focusing on misclassified real examples [49]. Additionally, some works (e.g., [63] in the context of topic-guided generation) aim to ensure the synthetic data covers the same topical space as the real data. If real data topics are unknown (in zero-shot cases), one might use broad coverage generation and then filter to those that seem plausible.

**Synthetic QA and Reasoning** An emerging area is using LLMs to generate not just answers but also **explanations or reasoning chains** for QA tasks. For instance, to train models that perform multi-hop reasoning, one can prompt an LLM to generate a question, the correct answer, and a step-by-step explanation (chain-of-thought). This produces synthetic training data for both the task and the format of reasoning. However, ensuring the correctness of the chain-of-thought is challenging; often an LLM might reach a correct answer via a flawed explanation, or vice versa. Therefore, researchers sometimes constrain generation by verifying final answers or using tools (like calculators, knowledge bases) in the loop to confirm intermediate facts.

### 3.3 Final remarks

In summary, synthetic text data generation with LLMs is a **balancing act between quantity and quality**. With only prompting, it is easy to get a lot of data quickly, but careful design is needed to ensure that data is diverse, relevant, and correct. Augmenting generation with retrieval, iteratively refining prompts based on model errors, and filtering outputs are all effective ways to improve the final dataset. Empirical evidence suggests that these measures can make synthetic data *nearly as effective as real data* in some scenarios, especially for augmenting small human-curated datasets.

We now turn to the code domain, where similar ideas are being applied, with some unique twists due to the nature of code.

## 4 Synthetic Data Generation for Code Tasks

Generative models for programming code – often referred to as **Code LLMs** when specialized – have also benefited from synthetic data generation [57, 46]. The code domain introduces distinct challenges and opportunities. Unlike free-form text, code has a rigid syntax and an execution semantics, allowing for automatic verification of correctness (a property rarely available for natural language) [60]. This section reviews how LLMs are leveraged to generate synthetic code data for training models on tasks such as code completion, code translation, bug fixing, and coding QA [31, 25].

We highlight methods that exploit execution feedback [19], the creation of large synthetic coding instruction datasets [2], and strategies to handle language-specific or low-resource coding scenarios [38].

#### 4.1 Various approaches for Prompt-Based Synthetic Code Generation

Large language models capable of coding (e.g., OpenAI Codex, CodeGen, Code-Llama, StarCoder) are typically trained on massive corpora of existing code from repositories. However, researchers have started to **supplement real code corpora with synthetic code** generated by LLMs to address specific gaps or to fine-tune models on particular tasks [6]. Synthetic code data generation can take multiple forms, as creation of instruction-based datasets, code translation and refactoring, generating problem’s specifications and solutions.

##### 4.1.1 Instruction-Following approaches

One major use of LLMs is to create QA or instruction-based datasets for code. For example, **Code Alpaca** is a synthetic dataset of 20K coding instruction-following examples (instruction and solution) generated by applying the Self-Instruct method to code tasks [4]. Starting from a handful of **seed prompts** (like “Write a Python function to ...”), an LLM (ChatGPT) was used to generate new programming instructions and their solutions, across various domains (file I/O, algorithms, web development, etc.). This dataset was used to fine-tune smaller models (similar to how the Alpaca dataset was used for text [4, 26]), producing code assistants without needing OpenAI’s proprietary data.

**WizardCoder** [40] extended this idea by using an *evolutionary strategy* to increase instruction complexity. They iteratively prompted ChatGPT to produce slightly more complex variations of coding tasks and solutions (Code Evol-Instruct), aiming to enrich the diversity and difficulty of the synthetic data beyond what straightforward prompting would yield.

Another, **Magicator** [55], focused on open-source code: it collected code snippets from GitHub and then generated diverse instruction prompts that would lead to those code snippets as answers, yielding 75K synthetic **instruction-output pairs**. The dataset was generated using ChatGPT as the LLM, which was prompted to create programming instructions corresponding to the collected code snippets. All these efforts show a trend of using LLMs to generate *specialized fine-tuning data for code LLMs*, analogous to how synthetic instruction data is used to fine-tune general LLMs for better following user requests.

##### 4.1.2 Code Translation and Refactoring

LLMs can be used to generate parallel code pairs, for example translating code from one programming language to another. *Synthetic parallel data* is valuable in training code translation models (transpilers) or enhancing multilingual code understanding. A study by Lachaux et al. [29] on TransCoder had initially used unsupervised translation; now with LLMs, one can prompt a model like GPT-4 to translate a given code snippet from Python to Java, etc., creating high-quality pairs.

Similarly, LLMs can perform **code refactoring**: generating a cleaned-up or optimized version of a given code. [25] introduced a **pipeline to improve code dataset** quality by systematically refactoring code – renaming variables, formatting, adding comments – to make multiple versions of the same program that are functionally identical. These transformed versions act as synthetic data that can help a model learn to generalize across different coding styles and improve readability. The CodeLLaMa-7B model was used for these transformations, as the study found that fine-tuning CodeLLaMa-7B on the cleaned-up versions of programs improved performance by up to 30% compared to fine-tuning on unprocessed datasets.

Another approach, **LintSeq** [46], generates synthetic data by taking real code and creating a sequence of incremental edits (like commit changes) to reach a goal. The CodeLLaMa-7B model was used to generate these edits, systematically modifying existing code while preserving its functionality. Models trained on these edit sequences learned to produce more **diverse solutions** when asked to generate code, as they implicitly saw many ways to modify and improve code [6]. Such data is particularly useful for code generation tasks where there are multiple valid solutions (e.g., competitive programming can have many different correct programs for the same problem).

##### 4.1.3 Problem Generation and Solution Synthesis

Another line of work uses LLMs to generate entire **programming problems** (specifications) and their solutions. This is akin to how one might generate QA pairs for text; here we generate challenge prompts for coding. For instance, an LLM might be asked: “Create a coding challenge about graph traversal” and then also asked to produce a reference solution code. These synthetic challenges can augment datasets like competitive programming problems, which are otherwise limited by human-curated problem sets.



Microsoft’s **AlphaCode** [34] and DeepMind’s **AlphaDev** [41] efforts, while primarily focused on solving problems, hint at generating many candidate solutions and sometimes even new problems to test on. AlphaCode was developed using a fine-tuned version of CodeGen [44], while AlphaDev leveraged reinforcement learning with AlphaZero-inspired techniques to discover more efficient sorting algorithms.

A related concept is generating **buggy code** and fixes: by intentionally introducing bugs into correct code using an LLM (or vice versa), one can create training data for bug detection and repair systems. In a recent preprint, “*LLM-itation is the Sincerest Form of Data*” [32], researchers generated synthetic buggy student code submissions for programming education by **prompting GPT-4 to imitate typical mistakes** students make [32]. They found that the distribution of errors (test case failures) in the synthetic buggy code was statistically similar to real student code, indicating that LLMs can realistically model the space of **incorrect solutions**. This synthetic data can help create better automated tutors and grading systems without needing to collect tons of real student submissions (which are privacy-sensitive).

Similarly, for automated code repair, one can generate (buggy code  $\rightarrow$  fixed code) pairs: [19] proposed a self-improvement strategy where a model generates its own code output, then attempts to identify errors (perhaps via unit tests or linting) and repair them, producing (initial, fixed) code pairs for training another model [37]. The approach utilized GPT-3 as the base LLM for generating buggy and corrected code samples, which were verified using a Python interpreter before being incorporated into training data.

Table 3 provides examples of LLM-driven synthetic data initiatives in the code domain and their focus.

Table 3: Selected LLM-based synthetic code data generation approaches.

Begin of Table				
Approach / System	Code Task Domain	Key Idea and Contribution	Input Type	Benchmark/Dataset Used
<b>CodeRL</b> [31]	Generation of problem’s specification and solution – Code generation with RL	Uses an actor-critic reinforcement learning setup on code generation: the LLM generates code, an <i>execution feedback</i> (did the code run correctly?) serves as a reward, and the model is refined with this feedback.	Code2Code	HumanEval[5], MultiPL-E[21]
<b>Self-Code-Improve</b> [19]	Code translation and refactoring – Code optimization / repair	LLM generates code and its own chain-of-thought for solving a task, then analyzes the code for errors or inefficiencies using heuristic prompts (simulated self-debugging).	Code2Code	No specific benchmark mentioned, but aligns with execution-based debugging tasks.
<b>InterCode</b> [57]	Generation of problem’s specification and solution – Interactive code generation (RL)	Framework combining LLMs with an interactive environment: the LLM produces code as actions and receives execution feedback as observations.	Code2Code	No specific benchmark mentioned, aligns with execution-based evaluations.
<b>Reflexion</b> [49]	Generation of problem’s specification and solution – Self-refinement (general)	A strategy in which an LLM first outputs a solution, then refines it based on feedback, creating pairs of (initial attempt, refined attempt).	Code2Code	HumanEval[5]
<b>Code Alpaca</b> [4]	Creation of instruction-based datasets – Code instruction following	A 20K dataset of synthetic coding instructions and answers generated via Self-Instruct using ChatGPT.	Text2Code	No specific benchmark, aligns with instruction-based evaluations.
<b>WizardCoder</b> [40]	Creation of instruction-based datasets – Complex code instructions	Built on Code Alpaca, WizardCoder applied an evolutionary approach to generate more complex and diverse instructions.	Text2Code	HumanEval[5], CoderEval[59]
<b>Magicode – OSS-Instruct</b> [55]	Creation of instruction-based datasets – Code instruction tuning	Created 75K synthetic Q&A pairs from open-source code.	Text2Code	No specific benchmark, aligns with instruction-tuned datasets.

Continuation of Table 3				
Approach / System	Code Task Domain	Key Idea and Contribution	Input Type	Benchmark/Dataset Used
<b>Test-Translating for Low-Resource</b> [2]	Code translation and refactoring – Multilingual coding	Proposed generating code in low-resource programming languages by translating high-resource code and verifying correctness.	Code2Code	No specific benchmark mentioned, aligns with multilingual dataset evaluations.
<b>NL2SQL Augmentation</b> [38]	Creation of instruction-based datasets – Text-to-SQL parsing	Combines strong and weak LLMs to generate diverse SQL queries.	Text2Code	No specific benchmark mentioned, aligns with text-to-SQL evaluation.
<b>PanGu-Coder</b> [7]	Creation of instruction-based datasets – Code generation	A large-scale LLM trained by Huawei for code generation.	Text2Code	No specific benchmark mentioned.
<b>CodeGen</b> [44]	Creation of instruction-based datasets – Code generation and completion	Open-source model designed for text-to-code generation across multiple programming languages.	Text2Code	HumanEval[5], MultiPL-E[21]
<b>InCoder</b> [14]	Creation of instruction-based datasets – Code completion and synthesis	Generative model for code completion, predicting and generating code snippets autoregressively.	Text2Code	HumanEval[5], DS-1000[30]
<b>AlphaCode</b> [34]	Generation of problem’s specification and solution – Competitive programming	Developed by Microsoft, AlphaCode generates multiple solutions per problem and ranks them by correctness.	Text2Code	HumanEval[5], CodeEval[59]
<b>Codex</b> [5]	Creation of instruction-based datasets – General code generation	OpenAI’s Codex powers GitHub Copilot, pretraining on both natural language and source code.	Text2Code	HumanEval[5], MultiPL-E[21]
<b>TransCoder</b> [29]	Code translation and refactoring – Code translation	An unsupervised machine translation model for programming languages.	Code2Code	No specific benchmark, aligns with multilingual model evaluations.
<b>LintSeq</b> [46]	Code translation and refactoring – Incremental code modifications	Generates synthetic data by taking real code and creating incremental edits.	Code2Code	No specific benchmark, aligns with execution and edit-based dataset evaluations.
<b>AlphaDev</b> [41]	Generation of problem’s specification and solution – Optimized code synthesis	DeepMind’s reinforcement learning-based model for discovering more efficient sorting algorithms.	Code2Code	AiXBench[20]
<b>LLM-itation (Buggy Code Generation)</b> [32]	Creation of instruction-based datasets – Buggy code and fix generation	Uses LLMs (GPT-4) to generate synthetic buggy student code submissions for tutoring.	Text2Code	No specific benchmark, aligns with bug-detection models and educational datasets.
End of Table				

## 4.2 Quality and properties of generated data

**Execution Feedback and Validation** A key advantage in the code domain is that the correctness of generated data can often be *automatically checked*. This is typically done by running the code (if possible) or using tests and compilers. Execution feedback serves multiple roles:

- **Filter Correct Solutions:** When generating code solutions, one can run each candidate on a set of unit tests. Only those that pass all tests are kept as synthetic data. This ensures the synthetic set consists of functionally correct programs. As a result, a model trained on this data learns from mostly correct examples, which is important because learning from incorrect outputs could mislead it. Some projects use an LLM to propose dozens of solutions to a coding task and then filter by execution to build a high-quality dataset of solutions.
- **Learn from Failure:** Conversely, the patterns in *failed executions* can also be informative. As seen in Reflexion and related methods [49], if a code fails, the error messages or partial progress can be fed back into the generation loop. This can yield a synthetic sequence: (attempt1, error, attempt2 fixed, error2, ..., final correct). Training on such sequences can teach a model to perform iterative refinement. Moreover, analyzing where the LLM’s outputs tend to fail (e.g., always off-by-one in a loop, or common syntax mistakes) highlights where model logic is weak. One can then prompt the LLM to specifically generate examples around those tricky cases, turning model failure modes into new training data to correct those modes.
- **Reward Modeling:** Execution results provide a natural reward signal for reinforcement learning. CodeRL [31], for instance, uses whether the code executes correctly as a reward to fine-tune the model via RL (reinforcement learning). Instead of supervised training on static data, the model *explores* code solutions and is reinforced to produce outputs that pass tests. Over time, the distribution of generated code shifts towards higher correctness. This approach effectively generates and uses synthetic data on the fly in a closed loop.
- **Measuring Difficulty:** By using execution feedback, one can also stratify the synthetic data by difficulty. Solutions that required multiple attempts or where most LLMs failed except one might indicate a hard problem. These can be labeled as such and a model can be trained to handle varying difficulty by exposing it to a mix (or by curriculum learning from easier synthetic tasks to harder ones).

Using execution signals is somewhat analogous to *fact-checking* in text generation – just as a fact-based QA might check an answer against a source text, a code generator checks a program against test cases. The difference is the test oracle for code is often definitive (pass/fail), whereas in language tasks, factual correctness can be fuzzy. This makes synthetic code generation uniquely powerful: one can generate vast amounts of code and reliably keep only the correct and relevant ones, yielding a high-precision training set [37].

**Addressing Biases and Data Coverage in Code** Synthetic code generation also needs to consider diversity and bias, albeit in a different sense than text. **Bias in code data** might refer to over-representation of certain patterns (e.g., always using a certain library) or style (e.g., code that always uses list comprehensions versus loops). LLMs might learn these biases from training data and also reflect them in synthetic outputs. To counter this, some strategies include:

- **Generating the same functionality in multiple ways** (as the refactoring approaches do) to prevent the model from thinking there’s only a single correct coding style. For instance, generate both iterative and recursive solutions for a problem if both are valid.
- **Ensuring comment diversity** and variable naming diversity in synthetic code, so the model doesn’t overly rely on specific token sequences that might correlate with certain outputs.
- **Covering edge cases:** If real code data is mostly clean, an LLM might not naturally produce code with unusual constructs. But for robustness, one might want synthetic examples that include, say, very deeply nested loops, or unusual combinations of language features. This can be prompted explicitly (e.g., “Write a program to X that uses recursion and bit manipulation”).
- **Bias towards older or newer practices:** Code corpora often contain outdated patterns. An LLM might produce deprecated syntax if trained on old data. Synthetic generation can intentionally produce updated code (for example, use Python f-strings instead of older % formatting) by instructing it accordingly, thus helping models stay modern. Conversely, it could also simulate legacy code if needed for training a model to understand older code.

In terms of data coverage, one challenge is the long tail of programming tasks. Real code datasets might have many examples of common tasks (sorting, web requests) but few of niche ones (like bit-level hacks or graphics shaders). LLMs can be directed to generate scenarios for these tail tasks, effectively filling gaps. Some surveys like *Mastering the Craft of Data Synthesis for CodeLLMs* categorize over 50 recent works across 23 topics in code synthetic data [6], indicating a rich taxonomy of methods addressing different areas (from data structure manipulations to API usage scenarios). It shows that the community is actively ensuring that synthetic data methods cover a wide spectrum of programming needs.

### 4.3 Final remarks

Empirical evidence suggests that synthetic code data, when used appropriately, can significantly improve code-focused models. Several open-source code LLMs (e.g., variations of Code-Llama, StarCoder) have integrated synthetic instruction tuning data and reported better benchmark results in code generation and completion tasks after fine-tuning on such data. One analysis noted that models which used primarily real code for pre-training but **synthetic data for fine-tuning (especially for instruction following)** often outperform those fine-tuned on limited human data, due to the **abundance and diversity** of the synthetic data. For instance, Qwen-2.5-Coder (hypothetical name from context) favored a pipeline where the base model is trained on GitHub code and then **instruction-tuned on a large synthetic set**, finding this more efficient to achieve alignment than gathering human-written instructions.

In specialized tasks like SQL query generation (NL2SQL [38]), using LLMs to generate extra training pairs has advanced the state of the art [6]. By generating varied NL questions for existing databases and ensuring their correctness (through execution on the database or a SQL validator), researchers created training data that taught models to handle more linguistic variations and database schemas, yielding higher accuracy on text-to-SQL benchmarks.

Overall, synthetic data generation for code is quickly becoming an indispensable part of the toolkit for building code AI systems. The ability to get infinite “free” coding examples, combined with the power to verify them, allows for creative training regimes that were not possible with just the raw code found in the wild. As with text, though, caution is needed to ensure synthetic code is **representative, correct, and useful** for the target tasks.

## 5 Challenges and Considerations

Despite the progress and successes of LLM-based synthetic data generation, several challenges persist. We outline the key issues and research considerations when applying these techniques, applicable to both text and code domains (with some domain-specific nuances where appropriate).

### 5.1 Quality Assurance and Factuality

Ensuring the *factual correctness* (for text) and *functional correctness* (for code) of synthetic data is paramount. LLMs are prone to **hallucinations** – producing incorrect facts or invalid code confidently. If not filtered out, these errors can propagate into the models trained on such data, causing them to learn spurious patterns. For text data, hallucinated content might teach a model incorrect knowledge or bias; for code, learning from non-working code could directly degrade a model’s ability to generate runnable programs.

*Mitigations:* As discussed, integrating retrieval to ground text generation [3] and execution to validate code [37] are primary ways to tackle this. Another line of defense is using **evaluation metrics** or critics. For example, using perplexity under a strong language model as a proxy for quality: if a GPT-4 model finds a GPT-3.5-generated sentence to be very unlikely, that sentence might be of low quality. Similarly, static analyzers or linters can catch obvious issues in code (like syntax errors or insecure practices) to filter those out. Human review of a subset can also estimate the precision of synthetic data; if it’s too low, one might tighten generation criteria. Recent research also develops metrics for *faithfulness* of generated content, which could be employed to score and select synthetic examples.

### 5.2 Distribution Shift and Realism

Synthetic data, no matter how well produced, may not perfectly follow the distribution of real data, especially user-generated data. Models trained heavily on synthetic data might become **overfit to synthetic patterns**. For instance, an LLM might generate cleaner and more coherent text than the often noisy real-world text (full of typos, slang, etc.), and a classifier trained on such immaculate synthetic text might struggle on messy real text. In code, an LLM might prefer a certain style (e.g., always using for loops) and underrepresent other valid styles, causing a model to be biased.

*Mitigations:* Mixing synthetic data with real data is a common practice to alleviate this [33]. Maintaining a core of real examples can *anchor* the model in reality, while the synthetic portion provides augmentation. Some works explicitly try to *match the distributions*: e.g., generate synthetic data that is indistinguishable from real data according to some classifier (an adversarial approach). In the extreme, **Generative Adversarial Networks (GANs)** style training could be considered, though with text that is tricky to do in practice. Another mitigation is **diversity enhancement** – ensuring synthetic data itself is diverse enough so that the model doesn’t get a narrow view. Using multiple LLMs or multiple prompts to generate a variety of styles can help. Also, curriculum strategies (train first on some real, then gradually add synthetic) might prevent the model from skewing too far.

A related concern is **model bias amplification**. If the LLM used for generation has certain biases (e.g., it might generate more male characters in stories, or prefer certain political slants), those carry into the synthetic data and can even amplify if the data is used to train a smaller model which then is used to generate more data, and so on. Careful prompt design (e.g., instructing the LLM to ensure diverse representation) and post-hoc balancing of the synthetic dataset are needed to counteract this.

### 5.3 Evaluation of Synthetic Data

Evaluating the utility of synthetic data is non-trivial. Ultimately, the value of synthetic data lies in how much it improves (or at least maintains) model performance on *real* data tasks. Thus, one straightforward evaluation is the **downstream task performance** of a model trained on synthetic versus real data. This was the approach used by multiple studies – train models on synthetic data and compare their validation set accuracy to those trained on gold data. If performance is close, the synthetic data is deemed effective. However, this requires training potentially many models for comparison and can be resource-intensive. Also, it doesn’t pinpoint *why* synthetic data might be failing if there is a gap.

Other evaluation criteria include:

- **Diversity Metrics:** e.g., distinct n-grams, embedding distribution, or topic coverage in the synthetic set versus real set. High diversity and good coverage often correlate with better generalization. But too much diversity (including off-topic data) could hurt, so it’s a balance.
- **Realism Metrics:** e.g., train a classifier to distinguish synthetic from real and see its accuracy. If the classifier struggles, the synthetic data is very realistic. This is analogous to the discriminator in a GAN. Some researchers use perplexity of a real-data-trained language model on synthetic vs real data to measure how *surprising* the synthetic set is. Ideally, synthetic data should be as unsurprising as real data from the perspective of a model that knows real data distribution.
- **Error Analysis:** Manually examining a sample of synthetic data for errors, irrelevance, or unnaturalness provides qualitative insight. For code, checking a random subset for correctness and style issues (with an expert or using automated analysis) can estimate overall quality.

### 5.4 Scale and Cost Trade-offs

While generating data is cheaper than manual labeling, it still incurs cost (especially if using an API like OpenAI’s) and time for large volumes. Deciding *how much synthetic data to generate* is an open question. Generating more might improve coverage but with diminishing returns and increased redundancy [33]. In [35], after a certain point, adding more GPT-generated examples didn’t improve the classifier further for some tasks. There is also the question of using more expensive models (like GPT-4) to generate presumably higher-quality data versus using cheaper ones (GPT-3.5) for more quantity. A few works have begun to compare generation quality from different LLMs; early indications are that larger models do give better data (more nuanced, fewer mistakes), but perhaps one can achieve similar results by filtering a larger set from a smaller model. Strategically, one might generate a bulk of data with a fast model and then have a powerful model or humans filter/evaluate it.

From a research perspective, sharing and reproducibility of synthetic data is tricky. If each user just prompts ChatGPT and gets data, it might be slightly different. There are efforts like **DataDreamer** [45] that aim to create standardized tools for synthetic data generation with LLMs, so that datasets can be reproduced and experiments repeated consistently. This is essential for scientific rigor.

### 5.5 Ethical and Legal Considerations

Synthetic data, despite being artificial, can still pose ethical issues. If the LLM’s training data contained biases, hate speech, or private information, these could surface in synthetic outputs. There have been instances of LLMs leaking bits of memorized training data verbatim. If such data is re-used, it could violate privacy or copyright. Researchers must ensure that no sensitive or copyrighted content is inadvertently **regenerated** in synthetic datasets [37]. Techniques to detect verbatim regeneration (like Google’s duplex scanning or embedding similarity to known datasets) are used to filter outputs that are too close to any single training example. This is especially important if synthetic data is meant to be openly shared – one must ensure it’s not just a regurgitation of proprietary data.

On the flip side, synthetic data can enhance privacy if done right (as noted in motivation), but one should not assume synthetic equals safe by default. For instance, an LLM might produce a very realistic fake patient record. If that record accidentally matches an actual person’s data (due to the model memorizing and outputting it), privacy is breached. The



chance is low but non-zero with large models. So, thorough checks or using models fine-tuned to avoid such leakage (some research is happening in this area) are needed.

**Model collapse** is a more theoretical but increasingly relevant concern: this is the idea that if models are trained on their own outputs repeatedly, the quality degenerates over generations. Think of it as making a copy of a copy of a copy... eventually, the content might become nonsensical. Some studies in 2023-2024 (e.g., [18]) have investigated this by simulating multiple rounds of training a model on its predecessor’s outputs. They did find that if each new model is trained purely on the previous model’s synthetic data (replacing real data entirely), performance drops – confirming the *model collapse* effect [18]. However, they also found an encouraging result: if each generation *keeps the real data and just adds some synthetic*, the performance can be maintained or even improved without collapse. In other words, **accumulating** synthetic data on top of real data avoids the degenerative feedback loop. This suggests that in practice, as long as we don’t throw away our original human datasets, we can iteratively augment with model-generated data with less fear of collapse. Nonetheless, vigilance is needed if we foresee a future where a model might be trained on predominantly AI-created corpora (e.g., lots of websites start hosting AI-written content).

## 5.6 Task-Specific Nuances

Different tasks might face unique issues with synthetic data. For example, in dialogue systems, synthetic conversations generated by an LLM might not capture the quirks of human dialogue (like frequent topic changes, misunderstandings, or the richness of multi-turn coherence). Also, user simulators (for task-oriented dialogue) need to behave realistically; an LLM generating both sides of a conversation might become too cooperative, not simulating the true difficulty of dealing with real users. Ensuring that synthetic dialogues are challenging and varied is an ongoing challenge.

In summarization, an LLM can generate an infinite number of summaries for an article, but many will be similar and might not add value beyond what a few reference summaries provide. The challenge is to generate summaries that highlight different possible interpretations or focus points of the article, thereby teaching a summarization model to handle ambiguity or preference.

In code, tasks like generating **malicious code** (for cybersecurity training) or **esoteric language code** have their own quirks. An LLM might not know much about a very niche programming language, so generating synthetic data for it might require a different approach (like transpiling known code from a similar language).

## 5.7 Human Acceptance

Finally, a more social challenge: will practitioners trust models trained on synthetic data? There can be a perception that data made by AI is not "real" and thus the model might not work in the real world. Building confidence requires transparency about how the synthetic data was created and evaluated. If synthetic data is to be used in high-stakes areas (medical, legal), one might need regulatory approval or at least proof that using it does not degrade outcomes. Demonstrating equivalence to real data through rigorous tests is important for adoption.

In summary, while synthetic data generation with LLMs is a powerful approach, it must be applied with careful consideration of these challenges. Each challenge is an active area of research, and progress is being made. Next, we look at some of the promising directions for future research that address these challenges and push the boundaries of synthetic data generation.

## 6 Future Directions

The intersection of LLMs and synthetic data is a rapidly evolving field. Based on current trends and open problems, we highlight several future research directions and opportunities:

- **Unified Frameworks and Taxonomies:** As the field grows, there is a need for comprehensive frameworks that unify various techniques. Recent surveys [52] have started organizing approaches (e.g., by stages of the LLM lifecycle or by data generation vs. curation vs. evaluation), but an agreed-upon taxonomy would aid researchers in identifying gaps. A framework that describes, for instance, the pipeline from prompt design, generation, filtering, to integration into model training could help systematically compare methods and build upon each other’s insights. Projects like *On LLMs-Driven Synthetic Data Generation, Curation, and Evaluation: A Survey* [39] propose generic workflows – these should be expanded and possibly standardized.
- **Automated Prompt Engineering:** The quality of synthetic data heavily depends on prompt quality. We may see more research into **automated prompt optimization** for data generation. Techniques like prompt tuning (learnable prompts) or using evolutionary algorithms to evolve prompts that yield better synthetic data

could reduce manual trial-and-error. For example, a future system might start with a basic prompt and then automatically adjust it (or try many variants) to maximize downstream validation performance of a model trained on the generated data – essentially tuning the prompt for optimal data utility.

- **Active Learning with LLM Generators:** Integrate LLM data generation into an active learning loop. A model being trained could identify which additional data points would most help it (areas of uncertainty) and request an LLM to generate examples in those areas, possibly with some human vetting. This combines the best of active learning (selecting informative examples) and synthetic generation (actually producing them on demand). Some initial work in this direction might leverage uncertainty sampling to drive prompt topics.
- **Cross-Modal Synthetic Data:** While we focused on text and code, extending LLM synthetic generation to other modalities is intriguing. For instance, using language models to generate image captions or descriptions (text) paired with images, effectively yielding synthetic image-text pairs for vision-language tasks. Conversely, models like GPT-4 (with vision) could generate descriptions of images that don't exist yet (which could then be turned into images via diffusion models – a multi-step synthetic data pipeline). Similarly, generating synthetic data for speech (transcripts with certain properties) could help train speech models. The concept of **simulators** for the world (e.g., using GPT to simulate dialogues or environments) could also aid reinforcement learning by generating scenarios for agents to train on.
- **Enhancing Real Data with Synthetic Variants:** Instead of generating entirely new data, an alternative is to generate *variants of existing real data*. For example, given a sentence, generate a paraphrase that changes some attributes (sentiment, tense, etc.). This can be more controlled than free generation. Future research might focus on *attribute-conditioned generation*: one explicitly sets what aspects to vary and what to keep. This can create counterfactual datasets for causal analysis or fairness testing (e.g., change demographic references in a sentence to test model bias).
- **Theoretical Understanding:** More theoretical work is needed to understand *why and when* synthetic data helps or hurts. Some initial theoretical analysis addresses model collapse dynamics [18] and error accumulation. But a theory for something like "How does the synthetic data distribution interact with the learning algorithm's generalization error?" would be valuable. Perhaps PAC-learning or domain adaptation theory could be applied: synthetic data can be seen as coming from a distribution that approximates the real distribution; theories of domain shift or distribution blending might provide guarantees or insights. Understanding the limits of augmentation (diminishing returns) in a mathematical sense could guide practitioners on how much is enough.
- **Quality Evaluation Metrics:** Developing better automatic metrics for synthetic data quality will be important. Metrics that correlate strongly with downstream performance would allow tuning generation without having to do full training runs. For example, a metric that combines diversity and fidelity (maybe using pre-trained model embeddings to judge if a synthetic set covers the real set's embedding space well) could be useful. In code, metrics beyond just "pass rate" might consider code elegance or complexity, depending on what we want the model to learn.
- **Human-in-the-Loop and Expert Knowledge:** Another future direction is incorporating expert rules or human insights directly into generation. An expert could specify domain constraints that the LLM must adhere to when generating data (kind of like guardrails). Or interactive generation where a human quickly corrects or annotates LLM outputs and those corrections feed back into producing better data (perhaps using reinforcement learning from human feedback, RLHF, but applied to data generation policy).
- **Safety Filtering and Bias Mitigation:** Building on ethical concerns, research will likely produce more sophisticated *safety filters* specifically for synthetic data. For instance, if generating a dataset of dialogues, ensuring none of the dialogues turn toxic inadvertently. We might see tools that can enforce certain bias mitigations – e.g., a constraint solver that ensures the synthetic dataset has equal representation of male/female or different ethnic names in certain contexts, to prevent training data bias. LLMs themselves can be asked to *self-debias* outputs (there is some work on prompting models to be mindful of bias). Future synthetic data generators could come with a suite of such self-checks or adjustments.
- **Domain-Specific Synthetic Data Generators:** While generic LLMs are used now, we might have specialized generators fine-tuned specifically for data generation in a domain. For example, a medical note generator LLM (trained on public medical text) that is particularly good at producing plausible medical records for research, or a legal text generator that can produce contracts or case descriptions for legal NLP. These domain-specific LLMs could produce more realistic synthetic data for those fields than a general model like GPT-4. Investing in such specialized synthetic data models might become a trend for companies that need data in domains where real data is hardest to get.
- **Benchmarking and Competitions:** We anticipate the rise of benchmarks specifically for synthetic data generation methods. For instance, a challenge might provide a small real dataset and require participants

to produce a synthetic dataset (via an LLM or other methods) that best improves a model’s performance on a hidden test set. This would directly measure how effective different generation strategies are. Similarly, competitions for code might provide some base code and encourage generation of the best augmentation set. These would spur innovation and provide comparative evaluations of techniques.

In essence, the future of LLM-based synthetic data generation is likely to involve more **automation, integration, and domain-awareness**. As LLMs themselves continue to improve (becoming more factual, less biased, more controllable), the quality of synthetic data we can get from them will also improve. This symbiotic relationship—better models → better data → even better models—could accelerate AI development. But careful research is needed to ensure we avoid pitfalls (like models reinforcing their own errors). By addressing current challenges and exploring these future directions, we can unlock the full potential of synthetic data in both NLP and code-related fields.

## 7 Conclusion

Synthetic data generation using large language models has emerged as a powerful technique to address data scarcity and enhance model training in both natural language and programming language domains. We have reviewed the state-of-the-art advances in this area, highlighting how LLMs can generate high-quality text and code data that, when used judiciously, lead to significant performance gains on downstream tasks.

In the **text domain**, LLM-generated data has proven especially useful for low-resource scenarios, delivering substantial improvements in tasks like classification and QA when human-labeled data is limited [33]. Techniques such as prompt-based augmentation (zero-shot, few-shot, etc.), retrieval augmentation for grounding facts [3], and iterative refinement have pushed synthetic text data closer in effectiveness to real data.

In the **code domain**, LLMs have unlocked new possibilities by generating code snippets, programming instructions, and even whole problem solutions, facilitating better training of code models. The ability to verify code correctness via execution [37] provides a strong advantage, allowing the curation of large-scale, correct-by-construction synthetic code datasets (e.g., Code Alpaca, WizardCoder) that have propelled open-source code models to approach the competency of their proprietary counterparts.

Throughout this survey, we have also underscored the **practical considerations**: synthetic data must be used with care to ensure quality and avoid pitfalls like distribution shift, bias amplification, or model collapse from feedback loops. Encouraging findings, such as the avoidance of model collapse by combining synthetic with real data [18], give confidence that these pitfalls can be managed with thoughtful strategies. The community has developed numerous methods (from filtering heuristics to advanced RL techniques) to maximize the signal and minimize the noise in synthetic datasets [33].

Our discussion on challenges reflects that this field is still evolving. Issues of factual accuracy, diversity, and evaluation metrics remain open for continued research. However, the rapid progress in just the last two years — with surveys cataloging dozens of new works [6] — indicates a vibrant research momentum. We are likely to see even more innovative solutions, like integrated human-in-the-loop generation and domain-specific synthetic data models, being realized soon.

In conclusion, LLM-based synthetic data generation stands as a **promising paradigm shift** in how we obtain training data for AI. By leveraging the generative power of advanced models, we reduce our reliance on large hand-curated datasets and open up opportunities to develop AI systems in domains and languages that suffer from data scarcity. This democratizes the development of AI, enabling customization and improvement even when real data is lacking.

As we refine these techniques, backed by rigorous evaluations and ethical safeguards, synthetic data generation will become an increasingly standard part of the machine learning toolkit—complementing real data, accelerating development, and perhaps one day even supplanting the need for certain kinds of manual data collection altogether.

The synergy between learning from data and generating data to learn from encapsulates a fascinating frontier for AI research, one that blurs the line between model and data and pushes us toward more autonomous, self-improving AI systems.

## References

- [1] Luiz Bonifacio, Hugo Abonizio, Marzieh Fadaee, and Rodrigo Nogueira. InPars: Unsupervised Dataset Generation for Information Retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR ’22, pages 2387–2392, New York, NY, USA, July 2022. Association for Computing Machinery.

- [2] Federico Cassano, John Gouwar, Francesca Lucchetti, Claire Schlesinger, Anders Freeman, Carolyn Jane Anderson, Molly Q. Feldman, Michael Greenberg, Abhinav Jangda, and Arjun Guha. Knowledge Transfer from High-Resource to Low-Resource Programming Languages for Code LLMs, September 2024. arXiv:2308.09895 [cs].
- [3] Yaping Chai, Haoran Xie, and Joe S. Qin. Text Data Augmentation for Large Language Models: A Comprehensive Survey of Methods, Challenges, and Opportunities, January 2025. arXiv:2501.18845 [cs] version: 1.
- [4] Sahil Chaudhary. sahil280114/codealpaca, February 2025. original-date: 2023-03-22T11:28:54Z.
- [5] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating Large Language Models Trained on Code, July 2021. arXiv:2107.03374 [cs].
- [6] Meng Chen, Philip Arthur, Qianyu Feng, Cong Duy Vu Hoang, Yu-Heng Hong, Mahdi Kazemi Moghaddam, Omid Nezami, Thien Nguyen, Gioacchino Tangari, Duy Vu, Thanh Vu, Mark Johnson, Krishnaram Kenthapadi, Don Dharmasiri, Long Duong, and Yuan-Fang Li. Mastering the Craft of Data Synthesis for CodeLLMs, February 2025. arXiv:2411.00005 [cs].
- [7] Fenia Christopoulou, Gerasimos Lampouras, Milan Gritta, Guchun Zhang, Yinpeng Guo, Zhongqi Li, Qi Zhang, Meng Xiao, Bo Shen, Lin Li, Hao Yu, Li Yan, Pingyi Zhou, Xin Wang, Yuchi Ma, Ignacio Iacobacci, Yasheng Wang, Guangtai Liang, Jiansheng Wei, Xin Jiang, Qianxiang Wang, and Qun Liu. PanGu-Coder: Program Synthesis with Function-Level Language Modeling, July 2022. arXiv:2207.11280 [cs].
- [8] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. DiffEdit: Diffusion-based semantic image editing with mask guidance, October 2022. arXiv:2210.11427 [cs].
- [9] Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Yihan Cao, Zihao Wu, Lin Zhao, Shaochen Xu, Wei Liu, Ninghao Liu, Sheng Li, Dajiang Zhu, Hongmin Cai, Lichao Sun, Quanzheng Li, Dinggang Shen, Tianming Liu, and Xiang Li. AugGPT: Leveraging ChatGPT for Text Data Augmentation, March 2023. arXiv:2302.13007 [cs].
- [10] Prafulla Dhariwal and Alexander Nichol. Diffusion Models Beat GANs on Image Synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pages 8780–8794. Curran Associates, Inc., 2021.
- [11] Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Boyang Li, Shafiq Joty, and Lidong Bing. Is GPT-3 a Good Data Annotator? In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11173–11195, Toronto, Canada, July 2023. Association for Computational Linguistics.
- [12] Nan Du, Yanping Huang, Andrew M. Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, Barret Zoph, Liam Fedus, Maarten P. Bosma, Zongwei Zhou, Tao Wang, Emma Wang, Kellie Webster, Marie Pellat, Kevin Robinson, Kathleen Meier-Hellstern, Toju Duke, Lucas Dixon, Kun Zhang, Quoc Le, Yonghui Wu, Zhifeng Chen, and Claire Cui. GLaM: Efficient Scaling of Language Models with Mixture-of-Experts. In *Proceedings of the 39th International Conference on Machine Learning*, pages 5547–5569. PMLR, June 2022. ISSN: 2640-3498.
- [13] Caoyun Fan, Jidong Tian, Yitian Li, Wenqing Chen, Hao He, and Yaohui Jin. Chain-of-Thought Tuning: Masked Language Models can also Think Step By Step in Natural Language Understanding. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14774–14785, Singapore, December 2023. Association for Computational Linguistics.
- [14] Daniel Fried, Armen Aghajanyan, Jessy Lin, Sida Wang, Eric Wallace, Freda Shi, Ruiqi Zhong, Wen-tau Yih, Luke Zettlemoyer, and Mike Lewis. InCoder: A Generative Model for Code Infilling and Synthesis, April 2023. arXiv:2204.05999 [cs].
- [15] Michal Gal and Orla Lynskey. Synthetic Data: Legal Implications of the Data-Generation Revolution. *SSRN Electronic Journal*, 2023.
- [16] Jiahui Gao, Renjie Pi, Yong Lin, Hang Xu, Jiacheng Ye, Zhiyong Wu, Weizhong Zhang, Xiaodan Liang, Zhenguo Li, and Lingpeng Kong. Self-Guided Noise-Free Data Generation for Efficient Zero-Shot Learning, February 2023. arXiv:2205.12679 [cs].

- [17] Jonas Gehring, Kunhao Zheng, Jade Copet, Vegard Mella, Quentin Carbonneaux, Taco Cohen, and Gabriel Synnaeve. RLEF: Grounding Code LLMs in Execution Feedback with Reinforcement Learning, February 2025. arXiv:2410.02089 [cs].
- [18] Matthias Gerstgrasser, Rylan Schaeffer, Apratim Dey, Rafael Rafailov, Henry Sleight, John Hughes, Tomasz Korbak, Rajashree Agrawal, Dhruv Pai, Andrey Gromov, Daniel A. Roberts, Diyi Yang, David L. Donoho, and Sanmi Koyejo. Is Model Collapse Inevitable? Breaking the Curse of Recursion by Accumulating Real and Synthetic Data, April 2024. arXiv:2404.01413 [cs].
- [19] Patrick Haluptzok, Matthew Bowers, and Adam Tauman Kalai. Language Models Can Teach Themselves to Program Better, April 2023. arXiv:2207.14502 [cs].
- [20] Yiyang Hao, Ge Li, Yongqiang Liu, Xiaowei Miao, He Zong, Siyuan Jiang, Yang Liu, and He Wei. AixBench: A Code Generation Benchmark Dataset, July 2022. arXiv:2206.13179 [cs].
- [21] Dan Hendrycks, Steven Basart, Saurav Kadavath, Mantas Mazeika, Akul Arora, Ethan Guo, Collin Burns, Samir Puranik, Horace He, Dawn Song, and Jacob Steinhardt. Measuring Coding Challenge Competence With APPS, November 2021. arXiv:2105.09938 [cs].
- [22] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the Knowledge in a Neural Network, March 2015. arXiv:1503.02531 [stat].
- [23] Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. Unnatural Instructions: Tuning Language Models with (Almost) No Human Labor. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14409–14428, Toronto, Canada, July 2023. Association for Computational Linguistics.
- [24] Yue Huang, Siyuan Wu, Chujie Gao, Dongping Chen, Qihui Zhang, Yao Wan, Tianyi Zhou, Chaowei Xiao, Jianfeng Gao, Lichao Sun, and Xiangliang Zhang. DataGen: Unified Synthetic Dataset Generation via Large Language Models. In *Proceedings of the Thirteenth International Conference on Learning Representations (ICLR 2025)*, October 2024.
- [25] Naman Jain, Tianjun Zhang, Wei-Lin Chiang, Joseph E. Gonzalez, Koushik Sen, and Ion Stoica. LLM-Assisted Code Cleaning For Training Accurate Code Generators, November 2023. arXiv:2311.14904 [cs].
- [26] Juyong Jiang, Fan Wang, Jiasi Shen, Sungju Kim, and Sunghun Kim. A Survey on Large Language Models for Code Generation, November 2024. arXiv:2406.00515 [cs].
- [27] Bahjat Kawar, Noam Elata, Tomer Michaeli, and Michael Elad. GSURE-Based Diffusion Model Training with Corrupted Data, June 2024. arXiv:2305.13128 [eess].
- [28] M. Abdul Khaliq, P. Chang, M. Ma, B. Pflugfelder, and F. Miletic. RAGAR, Your Falsehood Radar: RAG-Augmented Reasoning for Political Fact-Checking using Multimodal Large Language Models, July 2024. arXiv:2404.12065 [cs].
- [29] Marie-Anne Lachaux, Baptiste Roziere, Lowik Chanussot, and Guillaume Lample. Unsupervised Translation of Programming Languages, September 2020. arXiv:2006.03511 [cs].
- [30] Yuhang Lai, Chengxi Li, Yiming Wang, Tianyi Zhang, Ruiqi Zhong, Luke Zettlemoyer, Wen-Tau Yih, Daniel Fried, Sida Wang, and Tao Yu. DS-1000: A Natural and Reliable Benchmark for Data Science Code Generation. In *Proceedings of the 40th International Conference on Machine Learning*, pages 18319–18345. PMLR, July 2023. ISSN: 2640-3498.
- [31] Hung Le, Yue Wang, Akhilesh Deepak Gotmare, Silvio Savarese, and Steven Chu Hong Hoi. CodeRL: Mastering Code Generation through Pretrained Models and Deep Reinforcement Learning. *Advances in Neural Information Processing Systems*, 35:21314–21328, December 2022.
- [32] Juho Leinonen, Paul Denny, Olli Kiljunen, Stephen MacNeil, Sami Sarsa, and Arto Hellas. LLM-itation is the Sincerest Form of Data: Generating Synthetic Buggy Code Submissions for Computing Education, November 2024. arXiv:2411.10455 [cs].
- [33] Yinheng Li, Rogerio Bonatti, Sara Abdali, Justin Wagle, and Kazuhito Koishida. Data Generation Using Large Language Models for Text Classification: An Empirical Case Study, July 2024. arXiv:2407.12813 [cs] version: 2.
- [34] Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, Thomas Hubert, Peter Choy, Cyprien de Masson d’Autume, Igor Babuschkin, Xinyun Chen, Po-Sen Huang, Johannes Welbl, Sven Gowal, Alexey Cherepanov, James Molloy, Daniel J. Mankowitz, Esme Sutherland Robson, Pushmeet Kohli, Nando de Freitas, Koray Kavukcuoglu, and Oriol Vinyals. Competition-Level Code Generation with AlphaCode. *Science*, 378(6624):1092–1097, December 2022. arXiv:2203.07814 [cs].



- [35] Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. Synthetic Data Generation with Large Language Models for Text Classification: Potential and Limitations, October 2023. arXiv:2310.07849 [cs].
- [36] Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. WANLI: Worker and AI Collaboration for Natural Language Inference Dataset Creation. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6826–6847, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.
- [37] Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, and Andrew M. Dai. Best Practices and Lessons Learned on Synthetic Data, August 2024. arXiv:2404.07503 [cs].
- [38] Xinyu Liu, Shuyu Shen, Boyan Li, Peixian Ma, Runzhi Jiang, Yuyu Luo, Yuxin Zhang, Ju Fan, Guoliang Li, and Nan Tang. A Survey of NL2SQL with Large Language Models: Where are we, and where are we going?, August 2024. arXiv:2408.05109 [cs] version: 1.
- [39] Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. On LLMs-Driven Synthetic Data Generation, Curation, and Evaluation: A Survey, June 2024. arXiv:2406.15126 [cs].
- [40] Ziyang Luo, Can Xu, Pu Zhao, Qingfeng Sun, Xiubo Geng, Wenxiang Hu, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. WizardCoder: Empowering Code Large Language Models with Evol-Instruct, June 2023. arXiv:2306.08568 [cs].
- [41] Daniel J. Mankowitz, Andrea Michi, Anton Zhernov, Marco Gelmi, Marco Selvi, Cosmin Paduraru, Edouard Leurent, Shariq Iqbal, Jean-Baptiste Lespiau, Alex Ahern, Thomas Köppe, Kevin Millikin, Stephen Gaffney, Sophie Elster, Jackson Broshear, Chris Gamble, Kieran Milan, Robert Tung, Minjae Hwang, Taylan Cemgil, Mohammadamin Barekatain, Yujia Li, Amol Mandhane, Thomas Hubert, Julian Schrittwieser, Demis Hassabis, Pushmeet Kohli, Martin Riedmiller, Oriol Vinyals, and David Silver. Faster sorting algorithms discovered using deep reinforcement learning. *Nature*, 618(7964):257–263, June 2023. Publisher: Nature Publishing Group.
- [42] Matthew Maufe, James Ravenscroft, Rob Procter, and Maria Liakata. A Pipeline for Generating, Annotating and Employing Synthetic Data for Real World Question Answering, November 2022. arXiv:2211.16971 [cs].
- [43] Sewon Min, Xinxu Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the Role of Demonstrations: What Makes In-Context Learning Work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics.
- [44] Erik Nijkamp, Bo Pang, Hiroaki Hayashi, Lifu Tu, Huan Wang, Yingbo Zhou, Silvio Savarese, and Caiming Xiong. CodeGen: An Open Large Language Model for Code with Multi-Turn Program Synthesis, February 2023. arXiv:2203.13474 [cs].
- [45] Ajay Patel, Colin Raffel, and Chris Callison-Burch. DataDreamer: A Tool for Synthetic Data Generation and Reproducible LLM Workflows, May 2024. arXiv:2402.10379 [cs].
- [46] Ulyana Piterbarg, Lerrel Pinto, and Rob Fergus. Training Language Models on Synthetic Edit Sequences Improves Code Synthesis, February 2025. arXiv:2410.02749 [cs] version: 3.
- [47] Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. Measuring and Narrowing the Compositionality Gap in Language Models, October 2023. arXiv:2210.03350 [cs].
- [48] Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving Neural Machine Translation Models with Monolingual Data, June 2016. arXiv:1511.06709 [cs].
- [49] Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language Agents with Verbal Reinforcement Learning, October 2023. arXiv:2303.11366 [cs].
- [50] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In David Yarowsky, Timothy Baldwin, Anna Korhonen, Karen Livescu, and Steven Bethard, editors, *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA, October 2013. Association for Computational Linguistics.
- [51] Brandon Trabucco, Kyle Doherty, Max Gurinas, and Ruslan Salakhutdinov. Effective Data Augmentation With Diffusion Models, May 2023. arXiv:2302.07944 [cs].
- [52] Ke Wang, Jiahui Zhu, Minjie Ren, Zeming Liu, Shiwei Li, Zongye Zhang, Chenkai Zhang, Xiaoyu Wu, Qiqi Zhan, Qingjie Liu, and Yunhong Wang. A Survey on Data Synthesis and Augmentation for Large Language Models, October 2024. arXiv:2410.12896 [cs].

- [53] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-Instruct: Aligning Language Models with Self-Generated Instructions. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada, July 2023. Association for Computational Linguistics.
- [54] Jason Wei and Kai Zou. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks, August 2019. arXiv:1901.11196 [cs].
- [55] Yuxiang Wei, Zhe Wang, Jiawei Liu, Yifeng Ding, and Lingming Zhang. Magicoder: Empowering Code Generation with OSS-Instruct, June 2024. arXiv:2312.02120 [cs].
- [56] Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V. Le. Unsupervised Data Augmentation for Consistency Training, November 2020. arXiv:1904.12848 [cs].
- [57] John Yang, Akshara Prabhakar, Karthik Narasimhan, and Shunyu Yao. InterCode: Standardizing and Benchmarking Interactive Coding with Execution Feedback, October 2023. arXiv:2306.14898 [cs].
- [58] Kang Min Yoo, Dongju Park, Jaewook Kang, Sang-Woo Lee, and Woomyeong Park. GPT3Mix: Leveraging Large-scale Language Models for Text Augmentation, November 2021. arXiv:2104.08826 [cs].
- [59] Hao Yu, Bo Shen, Dezhi Ran, Jiaxin Zhang, Qi Zhang, Yuchi Ma, Guangtai Liang, Ying Li, Qianxiang Wang, and Tao Xie. CoderEval: A Benchmark of Pragmatic Code Generation with Generative Pre-trained Models. In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering, ICSE '24*, pages 1–12, New York, NY, USA, February 2024. Association for Computing Machinery.
- [60] Xiao Yu, Zexian Zhang, Feifei Niu, Xing Hu, Xin Xia, and John Grundy. What Makes a High-Quality Training Dataset for Large Language Models: A Practitioners’ Perspective. In *Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering, ASE '24*, pages 656–668, New York, NY, USA, October 2024. Association for Computing Machinery.
- [61] Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander J. Ratner, Ranjay Krishna, Jiaming Shen, and Chao Zhang. Large Language Model as Attributed Training Data Generator: A Tale of Diversity and Bias. *Advances in Neural Information Processing Systems*, 36:55734–55784, December 2023.
- [62] Yue Yu, Yuchen Zhuang, Rongzhi Zhang, Yu Meng, Jiaming Shen, and Chao Zhang. ReGen: Zero-Shot Text Classification via Training Data Generation with Progressive Dense Retrieval. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11782–11805, Toronto, Canada, July 2023. Association for Computational Linguistics.
- [63] Yu Zhang, Yunyi Zhang, Martin Michalski, Yucheng Jiang, Yu Meng, and Jiawei Han. Effective Seed-Guided Topic Discovery by Integrating Multiple Types of Contexts, January 2023. arXiv:2212.06002 [cs].
- [64] Arkaitz Zubiaga. Natural language processing in the era of large language models. *Frontiers in Artificial Intelligence*, 6, January 2024. Publisher: Frontiers.