

Large Language Models for Code Generation: A Comprehensive Survey of Challenges, Techniques, Evaluation, and Applications

Nam Huynh and Beiyu Lin

School of Computer Science, University of Oklahoma, Norman, OK, USA

Abstract

Large Language Models (LLMs) have demonstrated their remarkable capabilities in numerous fields. This survey focuses on how LLMs empower users, regardless of their technical background, to use human languages to automatically generate executable code. We begin with understanding LLMs' limitations and challenges in automated code generation. Subsequently, we review various fine-tuning techniques designed to enhance both the performance and adaptability of LLMs in code generation tasks. We then review the existing metrics and benchmarks for evaluations to assess model performance based on fine-tuning techniques. Finally, we explore the applications of LLMs (e.g. CodeLlama, GitHub Copilot, ToolGen) in code generation tasks to illustrate their roles and functionalities. This survey provides a comprehensive overview of LLMs for code generation, helps researchers in diverse fields better understand the current state-of-the-art technologies, and offers the potential of effectively leveraging LLMs for code generation tasks.

Index Terms— Large Language Models (LLMs), Code Generation, Machine Learning, Artificial Intelligence (AI).

1 Introduction

In general, data mining requires users with good coding skills and domain knowledge based on extensive training hours [1]. A competent data mining expert needs a foundational education in computer science or related fields along with practical experience in programming languages (PLs), such as Python and R [2]. Coding with a programming language like Python and its powerful libraries like Pandas can make these efforts easier by automating many preprocessing tasks [3]. However, for example, data preprocessing, a critical step in data mining, is still very time consuming. The New York Times states that data preprocessing accounts

for 50–80% of a data scientist's workload [3]. But now, with assistance from LLMs, it would only take several minutes to automatically generate code for data preprocessing. Figure 1 shows an example of using ChatGPT to automatically generate code for a data preprocessing task (replacing missing values). The accessibility, effectiveness, and efficiency of using LLMs in daily tasks show the huge potential of applying LLMs in automatic code generation. We write this survey in the hope of opening the possibilities of Data Mining for everyone.

Today, it is feasible to train a Generative AI (GenAI) model based on vast amounts of rich and diverse data from various resources, such as code repositories, technical forums, and web data on coding aspects [4]. For example, a descendant of GPT-3, OpenAI Codex, has been trained on data from billions of source code, such as code in public GitHub repositories [5]. The comprehensive training on rich data enables LLMs to better understand the context of code comments and function names, as well as to better interpret variable names [4]. For example, since LLMs have also been integrated into Integrated Development Environments (IDEs), such as PyCharm and VSCode to allow programmers to develop their code, LLMs within these environments can comprehend this contextual information and provide suggestions for users [6]. Figure 2 shows an example of using GitHub Copilot to handle a data preprocessing task - splitting the dataset [7].

This survey outlines key aspects of LLMs in code generation by organizing them into four parts: (III) limitations and challenges, (IV) fine-tuning techniques, (V) evaluations, and (VI) applications. These four parts are supported by 38 references (including research papers and technical articles). Each part is further detailed by specific subtopics. Part III, Limits and Challenges of LLMs in automatic code generation, has four topics: (A) resource constraints, (B) syntactic and semantic errors, (C) biases, and (D) security risks. Subsequently, Part IV, Fine-tuning Techniques for better

User Prompt:

Help me generate code for the following task: Remove rows with missing Product_ID, replace missing Price values with the median, forward-fill missing Quantity values, and save the cleaned dataset as cleaned_sales_data.csv.
In addition, record the time you take to generate this code.

ChatGPT 4o Output: Python Code

```
# Import pandas library
import pandas as pd

# Load the dataset
data = pd.read_csv(
    'sales_data.csv')

# Remove rows with
# missing Product_ID
data = data.dropna(subset
    =['Product_ID'])

# Replace missing Price
# values with the median
median_price = data['Price'].median()
data['Price'] = data['Price'].fillna(
    median_price)

# Forward-fill missing
# Quantity values
data['Quantity'] = data['Quantity'].fillna(
    method='ffill')

# Save the cleaned
# dataset
data.to_csv(
    'cleaned_sales_data.csv',
    index=False)

print("Data preprocessing
      completed
      successfully.")
```

Time Taken:

The code generation took approximately 0.000044 seconds (44 microseconds).

Given the comment:

```
# Split the data into features
# and target variable
```

GitHub Copilot generates the following code:

```
X = df.drop(['Date', 'Rented
Bike Count'], axis=1)
y = df['Rented Bike Count']
```

Figure 2: Example of using GitHub Copilot generating code [7]

performance and adaptability of LLMs, covers topics ranging from simple prompt engineering to more complex approaches, such as reinforcement learning and domain-specific dataset fine-tuning. Next, Part V, Evaluations, introduces established metrics and benchmarks that quantify the effectiveness and reliability of these techniques from the second part. Finally, Part VI, LLMs’ applications, examines the practical applications of LLMs in automating specific coding tasks. These tasks represent various aspects of code generation, including foundational tasks like code generation and code completion, to advanced tasks such as advanced code generation and code search, and auxiliary tasks including debugging and code translation. By integrating these findings, this paper consolidates existing knowledge and identifies gaps and opportunities to advance LLMs in code generation. The layout of this survey paper is shown in Table ??, which serves as a roadmap for the discussions presented in the paper.

2 Overview: LLMs for Code Generations

2.1 Large Language Models (LLMs)

LLMs have demonstrated significant breakthroughs in many different fields ranging from healthcare to education by using advanced deep learning architectures and training on extensive, diverse datasets [8]. The deep learning architecture, transformer networks, allows LLMs to understand the meaning of human language [9]. Transformer networks identify relationships within sequential data, such as words in a sentence [10]. A transformer consists of multiple layers, such as self-attention layers and feed-forward layers [10]. These layers can be stacked to create deeper transformers and develop more powerful models. Figure 3 is an illustration of how transformer models work. For example, as the

Figure 1: Example of using ChatGPT 4o to generate code for removing missing values.

Table 1: Summary of Reviewed Studies

Topic	Subtopics	Sources
III. Limits and Challenges	A. Resource Constraints	Training Requirements [26] Memory Efficiency [27] Performance of Smaller Models [28]
	B. Syntactic and Semantic Errors	Common Issues [29] Errors Analysis [30] Errors in Translation [31] ChatGPT Errors [32]
	C. Biases	Multilingual Bias [33] Social Bias [34]
	D. Security Risks	Security Vulnerabilities [35] Systematic Testing [36] Prompting Strategies [37] Vulnerable Data Sources [38]
IV. Fine-tune Techniques	A. Fine-Tuning on Domain-Specific Datasets	Instruction-Tuning for Optimization [40] Parameter-Efficient Fine-Tuning [41] Data Pruning for Fine-Tuning [42]
	B. Feedback	ClarifyGPT Framework [43] RLEF [44] Crowd-Sourced RLHF [45]
	C. Prompt Engineering	Chain-of-Thought Prompting [46] AceCoder Technique [47] Security-Focused Prompting [48]
V. Evaluations	A. Metrics	CodeBLEU [52] pass@k [56] pass-ratio@n [53] ICE-Score [54]
	B. Benchmarks	HumanEval [56] ClassEval [57] SWE-bench [58] BigCodeBench [59]
VI. LLMs' Applications	A. Code Generation and Code Completion	GitHub Copilot [62] Code Llama [63] ToolGen [64]
	B. Advanced Code Generation and Code Search	RepoRift [65] CodeBERT [66] AlphaCode [67]
	C. Code Debugging and Code Translation	GPT-4 [68] Codex [69] Flourine [70]

transformer model receives a user prompt to summarize an article, it analyzes the text and generates a concise summary that only contains the article’s key points.

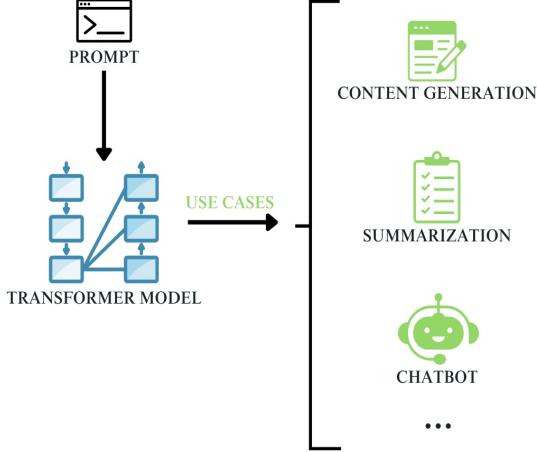


Figure 3: How transformer models work [10]

Training with an extensive and diverse dataset enables LLMs to process many tasks in numerous areas, including healthcare and education [11]. For instance, IntelHealth Company develops personal health LLMs to generate recommendations on personalized health plans for patients based on their medical history combined with clinical data [12]. In education, an interactive textbook - CourseKata using LLMs shows the benefits of training diverse datasets to meet educational needs [13]. Incorporating datasets from textbook materials, student responses, and assessments allows CourseKata to provide students with personalized feedback in real time via the generation of personalized practice questions by offering detailed explanations and adapting to each student’s learning pace [13]. These examples highlight how training on large and diverse datasets enables LLMs to generalize to different domains.

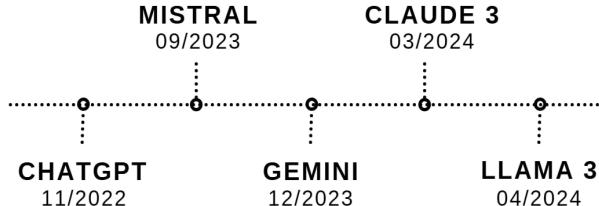


Figure 4: Notable released LLMs timeline

Both deep learning architectures and extensive and diverse datasets trained in LLMs lay the foundations for practical real-world applications in various industries like BERT from Google. These applications show the transformative potential of LLMs and the commitment of leading tech com-

panies to advance their development and unlock new possibilities in various domains. For example, Google leverages BERT, with its ability to extract important information, summarize long texts, and optimize search results, to summarize texts with precision and conciseness [14]. In addition, Microsoft developed an LLM called Turing-NLG to enhance its system for identifying and extracting meaningful information from text (e.g. names, locations, and dates), allowing Microsoft to enhance language understanding, deliver reliable, context-aware information, and improve applications in NLP, search engines, and information retrieval [14]. Furthermore, IBM uses the WATSON NLU (Natural Language Understanding) model, which leverages LLMs to analyze and extract valuable insights from customer feedback, social media posts, etc., allowing them to make decisions based on this information [14]. These real-world applications illustrate how LLMs are revolutionizing industries and unlocking innovative solutions in diverse domains.

2.2 LLMs for Code Generations

To begin with, we will introduce the steps of how LLMs generate code, in terms of before data feed into LLMs, during training, and after training. Before training, data preprocessing plays a crucial step in ensuring that datasets (e.g. open source repositories from GitHub and Overflow) are clean, standardized, and suitable for the chosen models, maximizing LLMs’ ability to learn [15]. During training, LLMs build complicated internal code representations to comprehend their meanings (semantics), structures (syntax), and relationships of various code elements [16]. Furthermore, the generating code process consists of four steps: Understanding the Prompt, Retrieving Relevant Code Patterns, Assembling Code Fragments, and Generating code [16]. In the first step, LLMs analyze the given input by breaking it down to understand the intended functionality, the programming language, and any specified constraints. Secondly, LLMs find their input’s internal representations to match the prompt’s requirements by finding code patterns and structures. Subsequently, they smartly combine retrieved code fragments and modify them to fit the prompt’s context in the third step. Lastly, LLMs generate final code output with many variations or suggestions for users to decide.

With the transformative process of how LLMs generate code, these models have shown rapid development in just a few years. The journey started in 2021 with GitHub Copilot, becoming one of the first widely available code generation tools [17]. From then on, numerous LLMs with code generation capability have been developed. By 2022, more advanced models like Replit Ghostwriter were re-

leased, which allow users to perform tasks, such as code completion, explanation, transformation, generation, and error detection with debugging [17]. In 2023, Bard was announced to support coding in more than 20 programming languages such as C++, Go, Java, JavaScript, TypeScript, and Python [17]. These milestones illustrate the rapid advancements in LLMs with the capability of generating code that changes the way users approach coding tasks.

To further demonstrate the impact of advancements in LLMs on code generation, OpenAI o1, also known as “Strawberry,” shows a significant leap forward in coding performance. Compared to other top-performing models like 3.5 Sonnet, GPT-4o, and Llama 405b in coding challenges using HumanEval benchmark data, OpenAI o1 achieves the highest performance rate of 92.4% which establishes itself as the best coding model according to Vellum[18]. OpenAI o1 - one of the latest OpenAI models - presents its groundbreaking technology of “thinking capacity” [19]. Unlike previous models that focus on the number of parameters, OpenAI o1 has the ability to “think” before responding by creating a long internal chain of thought, similar to how humans brainstorm to respond to problems [20]. Using a large-scale reinforcement learning algorithm, OpenAI o1 can detect and fix its mistakes, break down complicated problems into simpler components as well as attempt a new approach if the old one is not working [21]. OpenAI o1 outperforms previous models like GPT-4o by ranking in the 89th percentile on Codeforces and has the skills to solve PhD-level problems when evaluating on GPQA diamond, a difficult intelligence benchmark [21].

3 Limits and Challenges of Using LLMs for Code Generation

The rapid development and huge potential capabilities of LLMs also raise several significant limitations and challenges. In this section, we discuss four areas: Resource Constraints, Syntactic and Semantic Errors, Bias, and Security Risks. Firstly, the training and deployment of LLMs require immense processing power and memory. For example, as one token defines one unit of text (e.g. a word or a small piece of a sentence), Llama 3.1 models were trained on over 15 trillion tokens, in which Llama 3.1 - 8B required 7 million GPU hours while Llama 3.1 - 405B required approximately 31 million GPU hours. [22]. Secondly, LLM performance, such as accuracy and reliability, would be significantly impacted by syntactic and semantic er-

rors, resulting in failure during execution or program generation with unexpected incorrect output [23]. For example, code generation can contain errors in “if” statements, as these can lead to incorrect branching, such as skipping conditions or executing the wrong code parts [23]. Thirdly, a bias testing framework called Code Bias Score (CBS) revealed that 38.92% of GPT-4’s generated code contained gender bias [24]. Finally, security risks of LLMs may come from their training data with unsanitized open source code as an example that contains vulnerabilities such as memory safety violations or SQL injection risks[25].

3.1 Resource Constraints

Chen et al. [26] point out that training LLMs for generating code is highly resource intensive, as these models require vast computational capacity and memory, especially on Graphics Processing Units (GPUs). For instance, StarCoder - an LLM trained on over 80 PLs from GitHub - is a 15B model trained on 1T tokens. These models demonstrate significant computational demands. Recent trillion-parameter models have stretched the limits of current computational capacity for consuming extensive processing resources and memory. For example, CodeLlama, a coding model from Llama2, is available in the following model sizes: 7B, 13B, 34B, and 70B parameters. All of these, except the 70B model, have been trained on 500B tokens of code and code-related data, whereas the 70B model required 1 trillion tokens as they continue to push the limits on current state-of-the-art computing resources.

Meanwhile, Chavan et al. [27] highlight the critical challenges of optimizing LLMs for faster and more memory-efficient, particularly in resource-constrained environments. For example, loading a LLaMa-70B model requires 140GB of VRAM, which excludes the additional memory needed for inference. This addressed the need for model compression when deploying in a resource-constrained environment. Because of this, quantization, which reduces memory usage by lowering the numerical precision of model weights, is introduced to handle these challenges. However, various methods such as GPTQ, AWQ, OmniQuant demonstrate the trade-offs between memory efficiency and model accuracy. This study presents a performance comparison between these quantization methods in compressing the LLaMA2-7B model. For example, OmniQuant reduces memory usage by lowering the precision to 4-bit, achieving strong performance with a perplexity of 5.97. However, reducing precision further to 3-bit with OmniQuant increases the perplexity to 6.65, indicating a decline in output quality. Similarly, GPTQ with 3-bit precision reduces weight

memory to 2.87 GB but leads to a higher perplexity score (7.36) compared to 4-bit (6.08) and 8-bit (5.86) configurations.

In contrast, Hassid et al. [28] examine the impact of computational constraints on LLM performance by comparing smaller models such as Code Llama of 7B/8B/13B with larger models such as the 70B model under the same resource limits. Interestingly, smaller models of 7B/8B/13B demonstrated better results by showing 5 to 15% performance gains over the 70B model. In a “small budget regime” capped at 32 normalized FLOP units and 64 wall-time units, the study evaluated code generation benchmarks (HumanEval, MBPP) in different model sizes of 7B/8B/13B models against the 70B model. Evaluation tasks include HumanEval with 164 function completions and MBPP with 500 code generation instructions. In addition, CodeLlama 7B/13B achieved a 60% score in only a quarter of the time required by larger models to achieve the same result on the HumanEval benchmark. This finding highlights the efficiency of smaller models and their ability to deliver competitive results with fewer resources in limited environments.

Table 2: Summary of Resource Constraints

Source	Key Ideas	Challenges
Chen et al. [26]	Training inefficiency	A 15B model trained on 1T tokens.
Chavan et al. [27]	Issues for faster and lighter LLMs.	Quantization method trade-offs.
Hassid et al. [28]	Performance comparison between smaller and larger LLMs.	7B/13B models outperform the 70B model.

3.2 Syntactic and Semantic Errors

Wang et al. [29] provide an in-depth analysis of common syntactic and semantic errors in code generated by six prominent LLMs-CodeGen-16B, InCoder-1.3B, GPT-3.5, GPT-4, SantaCoder, and StarCoder by evaluating their 557 incorrect code snippets across 164 Python tasks from the HumanEval dataset. Syntactic errors, such as missing or incorrectly structured code blocks, were two common problems with the six LLMs generating 40% of these errors. This indicated that even widely used LLMs struggle with fundamental code structure. Simple errors, including “if” errors, in-

correct function names, and incorrect function arguments, were also common and easy to correct. The most common semantic issues from the code generated by these six LLMs are incorrect logical flow and flawed conditional statements. In contrast to the other four models (CodeGen-16B, InCoder-1.3B, SantaCoder, and StarCoder), GPT-3.5 and GPT-4 show better performance for generating code with fewer missing steps.

Based on these findings, Dou et al. [30] conduct a study on both syntactic and semantic errors in LLM-generated code by analyzing seven different models such as GPT-4, GPT-3.5, Claude-3, Llama-3, Phi-3, StarCoder-2, and DeepSeekCoder on three benchmarks: HumanEval+, MBPP+, and APPS+. They observed that syntactic errors, including incorrect syntax structure, indentation errors, and missing library import, are relatively rare, accounting for less than 10% of the total errors among all models. In contrast, semantic errors, such as misunderstanding task requirements, logic errors, hallucinations, and input/output format issues, constitute the largest category of errors. DeepSeekCoder, Llama-3, Phi-3, and GPT-3.5 have proportions of semantic error that exceed 50% on the APPS+ benchmark, showing their struggles with intricate logic and conditional structures. Additionally, as the complexity of the benchmark grows, semantic errors increase proportionally, highlighting the challenges LLMs face in accurately interpreting and executing complex tasks.

Extending the scope to translation tasks, Pan et al. [31] categorize translation errors in LLM code into 15 types, covering both syntactic and semantic issues. Syntactic errors, often involving misalignment with target language-specific requirements, and semantic errors, which affect the logical consistency of translated code, were common. The study used real-world projects like Apache Commons CLI and Python Click to evaluate LLMs’ effectiveness in code translation by categorizing translation errors and assessing the resulting syntactic and semantic issues across multiple benchmarks, such as HumanEval, MBPP, and APPS. In particular, 30.5% of translation errors resulted from syntactic and semantic misalignments between the source and target languages, and 24.3% of these errors involved unmet target language requirements. This study underscores the challenges LLMs face in code translation tasks, where nearly 80% of the issues arise from such discrepancies.

Finally, Liu et al. [32] analyze ChatGPT’s correctness on code generation by examining their semantic and syntactic errors that impact code reliability and quality. The study was analyzed on 4,066 code snippets generated in Java and Python in 2,033 programming tasks, revealing that both types of errors affect the compilation and runtime

errors of the generated program. The result demonstrates that Illegal Index and Type Mismatch errors are the most common semantic errors in ChatGPT-generated code. Illegal Index errors account for 46.4% of the 97 runtime errors in Java, while Type Mismatch errors are more frequent in Python because of its dynamic typing system. Furthermore, for semantic errors, 1,930 snippets (47%) exhibited maintainability issues, such as inconsistent variable use and improper loop handling, affecting readability and reusability. This breakdown underlines semantic and syntactic issues that lead not only to runtime errors but also to a higher demand for manual correction to achieve functional code.

3.3 Biases

Wang et al. [33] explore multilingual bias in LLM code generation, including Multi-NL bias and Multi-PL bias. The paper studied multilingual bias using three popular LLMs, such as StarCoder, CodeLlama, and DeepSeek-Coder, while evaluating it on the Pass@k metric. For bias in multi-NL, the results showed that LLMs exhibit a significant performance gap when generating code from different language instructions like English and Chinese across different PLs (e.g. Python, Java, C++, etc.). Using Chinese instructions led to an average Pass@1 rate drop of 17.2% for base models and 14.3% for instruction-tuned models in Python, with CodeLlama-34B experiencing more severe bias as its Java code generation dropped by 37.8%. For bias in multi-PL, the results showed various LLMs' performance in generating code in different PLs. Base models achieved the highest Pass@1 rate in Python for outperforming C++ and Java by 5.7% and 11.3%, respectively.

Expanding the discussion to social biases, Liu et al. [34] investigate the severity of these biases in the generation by LLM code. Their experiments were performed on different LLMs such as Codex, InCoder, and CodeGen with different sizes to evaluate social biases in code using three metrics, such as Code Bias Score (CBS), UnFairness Score (UFS), and the standard deviation of the frequency for all valid demographics (e.g., ethnicity, religion, and gender). The results revealed that models such as Codex and InCoder generated harmful codes in which certain ethnicities or religions were associated with the derogatory term “disgusting” by expressing prejudice against “Islam” and “Muslim”. Furthermore, Codex, with over 100 billion parameters, achieved the highest code generation quality (Pass@1: 47.03%) but also demonstrated the most severe biases (CBS: 82.64%), highlighting a troubling trade-off between performance and fairness. Similarly, as the sizes of CodeGen model increase from 350M to 6.1B, their performance im-

proves from 12.76% to 26.13% on the Pass@1 but reveals a sharp increase in the bias of CBS escalating from 9.36% to 62.65%.

3.4 Security Risks

Islam et al. [35] introduce security vulnerabilities in LLMs including three main technical issues data quality, model design, and evaluation practices. LLMs show their disadvantage in producing 10% more vulnerable code than human developers. Data quality issues, including incorrect labeling and data leakage, as indicated by datasets such as MVD and Devign, were observed to trigger the generation of false positives or false negatives in vulnerability detection. In addition, models designed only for supervised fine-tuning, such as VulRepair, mostly generate non-functional code due to scarce syntax and functionality checks. Lastly, for evaluation, the common metrics used to evaluate these models, such as BLEU and Exact Match, are not indicative enough for the security and functionality of the generated code.

Based on this analysis, He [36] explores recent efforts to evaluate code security by LLMs from systematic testing to user studies. Initially, the author discusses a popular security risk called “Out-of-Bounds Write” (CWE-787), which can allow attackers to exploit computer memory for criminal activities by writing malicious information. Recent efforts to assess the security of LLM-generated code include systematic evaluations using Common Weakness Enumeration (CWE), focusing on how Copilot handles various vulnerabilities across different prompts, weaknesses, and programming domains. Copilot’s response to the scenarios of the diversity of prompt and domain shows that around 40% of the generated code is vulnerable from a security standpoint. In addition, a security-driven user study examines code written by student programmers with LLM’s assistance. The user study found that while LLM-assisted code generation introduced some vulnerabilities, the overall impact on security was small. AI-assisted students produced security-critical bugs about 10% more often than non-assisted students.

Furthermore, Black et al. [37] investigate the security issues with LLM-generated code that arise from challenges in balancing security and correctness based on prompting strategies, model selection, and the degree of randomness allowed in responses. CWE-22, which is a directory traversal, and CWE-190, which is an integer overflow, have been two of the common vulnerabilities used as benchmarks to evaluate generated programs. In CWE-22 (directory traversal), the task is to generate programs that write files to specified paths. The results show that GPT-3.5 generated code that

allowed filenames with “..”, enabling unauthorized access to parent directories. In CWE-190 (integer overflow), the task required generating programs to handle numerical operations safely. The results show that Claude Opus initially used standard int types that failed to handle large numbers such as $2 * 9,999,999,999$, resulting in incorrect output.

Finally, Wang et al. [38] highlight the security risks in the LLM code generated in their training and during their generation process. Firstly, LLM training using unsanitized data from open source such as GitHub can lead to potential risks of inadvertently embedding security vulnerabilities in generated code. For example, the 2022 Open Source Security and Risk Analysis (OSSRA) report highlights that 81% of the 2,049 codebases analyzed had at least one vulnerability, with 49% contained high-risk vulnerabilities. Therefore, these models are prone to propagating vulnerabilities during the code generation process, potentially resulting in flawed outputs that are highly susceptible to exploitation and malicious attacks. To better illustrate this, Copilot generates insecure code in about 40% of the cases, whereas ChatGPT showed that of the 21 programs, only 5 were initially secure.

4 Fine-Tuning Techniques for Enhancing LLM Performance in Code Generation

To handle the limitations and challenges of code generation by LLMs, fine-tuning has become an important strategy to enhance LLMs’ capabilities. Fine-tuning allows users to put the pre-trained LLMs into use more for specialized applications with significantly improved performance while preserving the remaining knowledge. For example, Google revealed that the fine-tuning of sentiment analysis boosted the accuracy of LLMs by 10% [39]. Because of that, this section explores three fine-tuning techniques: prompt engineering, which optimizes LLM outputs by crafting effective input instructions; feedback refinement, which reduces errors by incorporating corrections; and domain-specific dataset tuning, which improves LLM performance in specialized areas. Together, these techniques mitigate specific weaknesses within LLMs and open the way to more effective and robust applications involving code generation.

4.1 Fine-Tuning on Domain-Specific Datasets

Ma et al. [40] propose LLaMoCo, the first instruction-tuning framework designed to adapt LLMs for the optimization of code generation. This

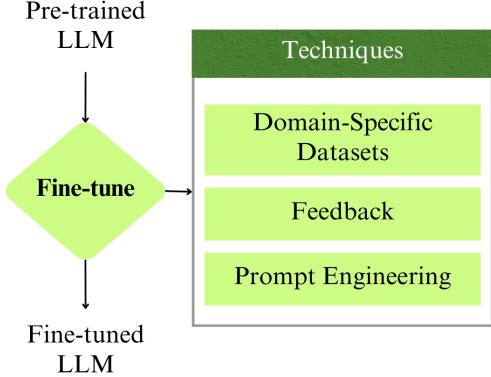


Figure 5: LLMs-Based Fine-tuning process

framework establishes a comprehensive instruction set that contains well-described problem prompts and effective optimization codes. Then, it proposes a new two-phase learning strategy that includes a contrast learning-based warm-up procedure before the instruction-tuning phase to boost the convergence behavior during model fine-tuning. The results of the experiment showed that LLaMoCo significantly improved the performance of LLMs as a fine-tuned CodeGen (350M) model of LLaMoCo demonstrated superior optimization performance compared to GPT-4 Turbo on both synthetic and realistic problem sets, showing less error for 4.168% and 79.483% and better performance for 87.227% and 59.174%, respectively. In addition, LLaMoCo boosted CodeLlama 7B model performance from 29.717% to 81.843%.

Furthermore, Weyssow et al. [41] review four popular Parameter-Efficient Fine-Tuning-LoRA, IA3, Prompt Tuning, and Prefix Tuning. These techniques ensure fine-tuning of LLMs by updating only a subset of model parameters rather than all the parameters. Thus, LLM focuses on task-specific data while maintaining good resource usage. Then, this paper compares these four techniques with ICL and traditional full fine-tuning on code generation tasks using Python datasets like CoNaLa or CodeAlpaca. The results indicated that the fine-tuned LLMs consistently perform significantly better with PEFT compared to ICL as LoRA fine-tuning LLMs improved 25.4% and 22.8% (150% and 29.8%) in EM@10 and CodeBLEU respectively in CoNaLa (CodeAlpacaPy) dataset. QLoRA reduces memory usage by allowing fine-tuning of LLMs with up to 34B parameters. This investigation emphasizes the potential of PEFT techniques in efficiently fine-tuning LLMs to task-specific data to generate code.

Complementing these approaches, Tsai et al. [42] introduce a novel approach to fine-tuning LLMs for code generation by integrating data pruning methods. The paper explores the use of clustering algorithms (KMeans, Agglomerative Cluster-

ing, HDBSCAN) and pruning metrics (Diversity Metric, Density Metric) to reduce the size of training data selectively while maintaining the accuracy and functionality of the generated code as there are significant redundancies in training data. HumanEval(+) and MBPP(+) datasets are used to evaluate pruning methods and highlight performance improvements. Surprisingly, the results show that pruning in a small portion of the training data can lead to performance improvements of up to 2.7% in HumanEval and 3.5% in MBPP. Remarkably, using data pruning on only 1% of the data can result in a 4.1% improvement compared to the base model, achieving performance nearly equivalent to training with the entire dataset.

4.2 Feedback

Mu et al. [43] present a novel framework - ClarifyGPT - that can identify and clarify ambiguous user requirements to improve LLM-based code generation. ClarifyGPT can perform a code consistency check to detect ambiguity and generate targeted clarifying questions to refine unclear input. Consequently, it generates the solution code from the received response. Therefore, this framework plays an important role in improving the interpretability of the code generated by LLMs. Furthermore, it helps users better understand generated code from interaction and provides more clarification of their intentions. Using two publicly available benchmarks: MBPP-sanitized and MBPP-ET for evaluation, ClarifyGPT improved the average performance of GPT-4 and ChatGPT from 68.02% to 75.75% and from 58.55% to 67.22%, respectively.

Furthermore, Gehring et al. [44] discuss Reinforcement Learning from Execution Feedback (RLEF), a method to improve LLMs in code synthesis by using feedback from code execution to iteratively refine outputs. The process includes three steps: generating code, receiving feedback from test cases, and updating the model through reinforcement learning using Proximal Policy Optimization (PPO). In experiments on competitive programming tasks such as those in CodeContests, models that are trained with RLEF achieved a solve rate of 37.5% in the test set for the standalone Llama 3.1 70B model. These significantly outperform the previous state-of-the-art AlphaCodium at 29%. The method also reduced samples by an order of magnitude compared to the RLEF approach. This approach generalizes well to other benchmarks like HumanEval+ and MBPP+, where feedback is used for the grounding of the output of LLMs, especially on multi-turn code generation tasks.

Finally, Wong et al. [45] introduce a new method that combines crowd-sourced computation

and reinforcement learning from human feedback (cRLHF), to improve code generation in LLMs. This aims to maximize code quality using multiple user feedback. As the traditional method - RLHF contains biases and misses important insights that limit LLMs' potential, cRLHF collects feedback data from different sources and uses Bayesian inference to align and combine the feedback data into one belief that gives more objective assessments without complicated reward modeling. The framework fine-tunes LLMs by using aggregated feedback to improve code correctness and quality. The results show significant improvements in LLMs of different sizes when the cRLHF method is applied. In the HumanEval benchmark evaluation, the success rate for CodeGen-2.7B improved from 39.8% to 45.4% and from 17.3% to 20.0% for the smaller model CodeGen-350M.

4.3 Prompt Engineering

Sun et al. [46] apply the “Chain-of-Thought” prompting technique to generate “solution plans” for complex programming challenges to develop a framework called CodePLAN. This framework is designed to infuse the reasoning capabilities of LLMs in smaller models to enhance their code generation performance. CodePLAN uses multi-task learning to train smaller models on both code generation and solution plan generation simultaneously. It uses backward reasoning and plan sampling strategies to improve solution plan quality. The higher quality of the solution plan may lead to more accurate code generation outputs. The framework considers LLMs as “teachers” to provide solution plans that distill into smaller models considered “students”. This allows them to develop solution plans independently at inference time. Experiments demonstrated that this approach significantly improves the code generation abilities of smaller models by more than 130% in performance using the pass@1 metric on the APPS benchmark.

Expanding on prompting techniques, Li et al. [47] develop a novel approach named AceCoder to improve LLM’s performance in code generation. It is designed to perform two major challenges of code generation: requirement understanding and code implementation. This method performs code generation in three steps: example retrieval, prompt construction, and code generation. First, the retriever selects similar programs based on language input, whereas the selector selects non-redundant programs based on prioritizing non-overlapping information. Second, the technique identifies a combination of chosen examples, their preliminary artifacts in the form of test cases, and input requirements to construct a prompt. Finally, the LLM uses the constructed prompt to generate test cases that

yield the final source code. AceCoder was evaluated on three LLMs, such as CodeGeeX, CodeGen, and InCoder, using three public benchmarks using Pass@k. It follows that AceCoder has surpassed the state-of-the-art prompting techniques in improving Pass@1 by up to 56.4% in MBPP, 70.7% in MBJP, and 88.4% in MBJSP and has proven to be effective in different LLM sizes and languages.

Lastly, Tony et al. [48] explore the impact of different prompting techniques on the security of the code generated by LLMs from NL instructions. These techniques were implemented in the GPT-3, GPT-3.5, and GPT-4 models. The authors investigated some of these techniques using a dataset of 150 NL prompts related to security-relevant coding tasks. 15 different explored prompting techniques are classified into 5 categories depending on their common characteristics, such as root techniques, refinement-based techniques, decomposition-based techniques, reasoning-based techniques, and priming techniques. For instance, refinement-based techniques focus on improving model outputs through iterative refinement, feedback loops, or self-assessment, including methods such as Recursive Criticism and Improvement (RCI), Self-refine, and Progressive Hint prompting. The results indicated that RCI performed best for both GPT-3.5 and GPT-4, while zero-shot prompting performed best out of these techniques for GPT-3. The persona / memetic proxy yielded the poorest performance, generating the most security weaknesses across all models.

5 Evaluation Metrics and Benchmarks for Assessing LLM-Generated Code

As numerous LLMs with code generation capability have been developing - a crucial tool for programmers of all skill levels, evaluating these models is essential to ensure their dependability and efficiency in meeting users' needs. While significant efforts have been dedicated to the performance evaluation of LLMs, most of these research questions remain unanswered, such as: "Are the evaluations and comparisons fair and are the differences significant?" or "Do findings from performance evaluation truly reflect the usability of LLMs as practical programming tools?" [49]. This section will discuss two key aspects of evaluation: benchmarks and performance metrics.

Before we discuss these aspects in depth, it is essential to clarify the terms "benchmarks" and "metrics". LLM evaluation metrics are criteria used to quantify the performance of LLM systems in aspects such as correctness of the answers, se-

mantic similarity, and hallucination [50]. On the other hand, benchmarks are constructed from evaluation datasets and metrics where test cases create an evaluation dataset [51]. Figure 6 illustrates the LLM benchmark structure that includes the integration of metrics within this evaluation framework.

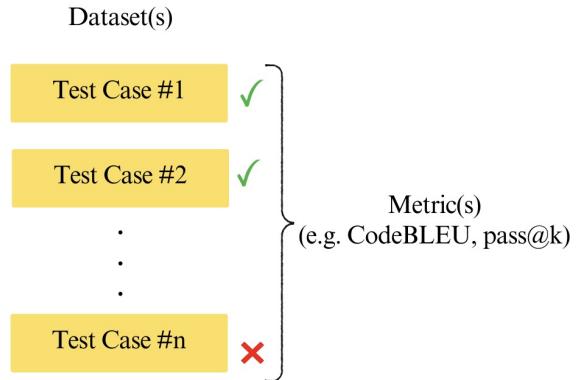


Figure 6: LLMs System Benchmark [51]

5.1 Metrics

One common metric for evaluating code generation by LLMs is CodeBLEU [52]. Compared to the generally adopted BLEU metric for NL evaluation, which lacks key syntactical and semantic characteristics of codes, CodeBLEU was designed to incorporate both traditional n-gram matching and syntactic and semantic matching. Specifically, the n-gram match weights different n-grams differently, the syntactic match plugs in AST information by aligning subtrees, while the semantic match measures similarities of code based on the analysis of its data-flow structure. CodeBLEU combines these elements (including weighted n-gram match, syntactic AST match, and semantic data-flow match) into a comprehensive evaluation metric. The experiments are tested in three coding tasks text-to-code (Java), code translation (from Java to C#), and code refinement (Java). The results demonstrate that CodeBLEU has a better correlation with human evaluation scores compared to traditional metrics like BLEU and perfect accuracy in all three tasks.

Another popularly used metric for LLMs generated code is pass@k [56]. Pass@k, with the goal of addressing the shortcomings of traditional text similarity metrics, is designed to assess the functional correctness of generated code samples. It presents the probability that at least one of the top k samples passes unit tests. This metric contains variations, such as pass@1, pass@10, pass@100, etc. Pass@1 provides the likelihood of correctness at the first attempt, while pass@10 and pass@100 assess the model's performance on much larger sample

sets to provide a comprehensive view of its ability to generate valid solutions. The following formula is used to calculate pass@k to handle all problems with E defining the expected value of the problems, n for the total number of samples, and k for the number of top samples to consider:

$$\text{pass}@k := \mathbb{E}_{\text{problems}} \left[1 - \frac{C(n - c, k)}{C(n, k)} \right]$$

Yeo et al. [53] propose a new metric called the pass-ratio@n that measures precision based on the accuracy granularity through the pass rate of the test cases. As LLMs can generate different solutions across inferences, considering that n inferences are made, the average pass-ratio across the n solutions is used to mitigate bias. For each solution i ($0 < i \leq n$), the pass-ratio is calculated by the following formula:

$$\text{pass-ratio}_i = \left(\frac{\# \text{ of test cases passed at code } i}{\# \text{ of test cases}} \right)^2$$

And the pass-ratio@n demonstrates the mean pass-ratio of n generated code.

$$\text{pass-ratio}@n = \frac{\sum_{i=1}^n \text{pass-ratio}_i}{n}$$

The metric pass-ratio@n was tested on three coding problems and five generated coding solutions for each problem from LLM inference to compare to pass@k. The results show that the pass-ratio@n can provide more granular insight than the pass@k metrics. In a coding problem where none of the solutions passed all test cases, pass-ratio@5 got a partial score of 61%, while pass@k recorded 0%.

In addition, Zhuo [54] introduces another new metric called ICE-Score by instructing LLMs for code assessments. The ICE-Score metric includes two key components: (1) the task definition, evaluation criteria, and structured evaluation steps, and (2) the provided problem along with the generated code snippet for assessment. Unlike traditional metrics such as BLEU, reference-based methods relying on human-written test suites, ICE-Score uses LLMs' capabilities to assess generated code on two aspects: usefulness and functional correctness. By entering the task problems and their generated code, ICE-Score outputs the corresponding assessments. The results outputted for both aspects include "Nearly Useless" or "Totally Useless", and "Functional Incorrect" or "Functional Correct".

5.2 Benchmarks

Table 3 represents the information of several notable benchmarks, which will be discussed in this benchmark section from Symflower [55].

Table 3: Benchmark Features Summary

Benchmarks	Number of Tasks	PLs	Release Date
HumanEval	164	Python	July 2021
CLASSEVAL	100	Python	August 2023
SWE-bench	2,294	Python	October 2023
BigCodeBench	1,140	Python	June 2024

The HumanEval [56] dataset is a benchmark designed to evaluate LLMs in code generation. It includes 164 programming challenges as each problem contains "function signatures, docstrings, body, and unit tests" to evaluate functional correctness. On average, there are 7.7 tests per problem. Traditional metrics like BLEU can measure the similarity of texts. Unfortunately, these measures are not suitable for evaluating code generation because functional correctness is much more important. This was addressed by introducing the pass@k metric, which came with the HumanEval dataset and helped to assess this functional correctness. This metric measures the probability that at least one of the top k-generated code samples passes the unit tests to provide a more practical evaluation of the generated code. HumanEval and pass@k have become critical factors in testing LLM coding capabilities to provide more meaningful and valuable test results.

In contrast, ClassEval [57] is a benchmark designed to evaluate LLMs on the more challenging coding tasks, such as class-level code generation, unlike existing benchmarks like HumanEval that focus only on simple scenarios, such as function-level or statement-level. The benchmark consists of 100 manually constructed Python coding tasks, including 100 classes and 412 methods. The study experimented with the evaluation of 11 state-of-the-art LLMs in class-level generated code using three different code generation strategies, including holistic generation, incremental generation, and compositional generation. First, existing LLMs perform significantly worse in class-level code generation compared to standalone method-level benchmarks like HumanEval. Secondly, GPT-4 and GPT-3.5 consistently outperform other models; and models such as WizardCoder, Instruct-StarCoder, and Instruct-CodeGen have similar performance. Lastly, the

best strategy for GPT-4 and GPT-3.5 is to generate code for the entire class, while using a method-by-method strategy will be a better choice for other models.

Furthermore, SWE-bench [58] is a benchmark designed to evaluate LLMs for the study of capabilities in real-world software engineering settings. As an evaluation framework, this benchmark comprises 2,294 tasks from Github issues and their related pull requests across 12 well-known Python repositories. Solving issues in the SWE-bench often requires a comprehensive understanding and coordination with changes across various functions, classes, and even multiple files at the same time, requiring models to interact with execution environments, handle extensive contexts, and carry out complex reasoning beyond typical code generation tasks. In addition, the evaluation experiment revealed that both advanced proprietary models and the fine-tuned model of the paper, SWE-Llama, can handle only the most simple issues. The best-performing model, Claude 2, can solve only 1.96% of the tasks with the BM25 retriever. Therefore, SWE-bench reflects the real-world coding environments to create solutions immediately applicable in open source software development.

Finally, BigCodeBench [59] is a new benchmark designed to evaluate LLMs on tackling practical and complex programming tasks to ensure no data contamination. Due to many library and function calls in real-world software development, concerns are raised for HumanEval, which is a simpler benchmark and is affected by contamination and overfitting problems. BigCodeBench includes 1,140 function-level tasks that require LLMs to use various libraries and compose multiple function calls. On average, there are 5.6 test cases per task with branch coverage of 99%. The benchmark tests performance using the Pass@1 metric with greedy decoding to measure the percentage of tasks correctly solved by the first generated code against curated test cases. For experiments, BigCodeBench ensures the quality of the task through collaboration between GPT-4 and 20 human experts, refining the tasks with test cases in a sandbox environment. The tasks are further evaluated by other LLMs and cross-checked by 7 experts, with the resulting average human performance at 97%.

6 LLMs’ Applications in Code Generation and Development

LLMs have transformed code generation in software development as these models provide user assistance in a variety of coding tasks, such as code com-

pletion, code translation, etc. In addition to that, each model has unique strengths and developers must understand and manipulate them effectively in their workflows. Selecting a suitable model for the task can give users the advantages of enhancing productivity, streamlining processes, reducing errors, and maximizing its potential. For example, OpenAI Codex benefits users whose workflow depends on GitHub due to the integration of Codex with GitHub Copilot [60]. Furthermore, for developers working on AWS, CodeWhisperer provides domain-specific insights and customized recommendations that position it as one of the top LLMs for cloud computing-focused development [60]. Moreover, tools have been designed to augment LLMs’ capabilities. For example, CodeAgent - an LLM-based agent framework - was developed to assist and allow LLMs to handle complicated programming tasks [61]. The following sections demonstrate several aspects of the broader code generation topic and are categorized into three groups of tasks: (A) code generation and code completion (foundational tasks), (B) code generation and code search (advanced tasks), and (C) debugging and code translation (auxiliary tasks).

6.1 Code Completion and Code Generation

Xu [62] introduces an advanced code autocompletion tool, GitHub Copilot, from the collaboration between GitHub’s vast software development resource and OpenAI’s groundbreaking AI development. GitHub Copilot uses deep learning models (recurrent neural networks (RNNs)) and transformers (Transformer) which help the model to learn the code’s syntactic and semantic structure and developers’ coding habits. In addition, it is capable of generating context-sensitive code suggestions from the training data of multiple open source code libraries and developers’ code contributions. GitHub Copilot analyzes the code context from developers, including the currently written code (functions, classes, and methods) to understand developers’ demands. Subsequently, based on given inputs, it generates code suggestions in real time and can continuously modify them depending on what developers need. This allows developers to learn from the coding community’s knowledge and experience as well as enhances the reusability, quality, and efficiency of the code. Furthermore, another advantage to mention is that GitHub Copilot can save developers a lot of time and effort by reducing manual code written from repetitive or boilerplate tasks.

Moreover, Meta AI [63] releases a state-of-the-art LLM - Code Llama for code generation built from LLaMA 2 architecture, which excels in code completion tasks. Three available models are Code

Llama - the base model, Code Llama - Python for optimizing Python Programming, and Code Llama - Instruct fine-tuned for better understanding and responding to NL instructions. There are four Code Llama sizes for 7B, 13B, 34B, and 70B parameters, respectively, as each model is trained on 500B code tokens, and related data with the 70B model are specifically trained on 1T tokens. The base and instruct models with 7B and 13B are trained using fill-in-the-middle (FIM) capability. This enables models to insert code within existing code to support tasks like code completion. Additionally, smaller 7B and 13B models are quicker and more suitable for tasks requiring low latency like real-time code completion. Code Llama was tested for its ability to complete code with given doc-strings using the HumanEval benchmark. The results show that Code Llama surpassed open source, code-focused LLMs and outperformed Llama 2 as model 34B achieved 53.7% on HumanEval, matching ChatGPT in performance.

Lastly, Wang et al. [64] develop ToolGen to improve repository-level code generation in the LLM generation process by integrating autocompletion tools. It includes two phases: Offline Trigger Insertion and Model Fine-tuning, and Online Tool-integrated Code Generation. ToolGen tackles problems in code generation for dependency errors, such as undefined-variable and no-member errors by manipulating autocompletion tools to fill repository-level dependencies. In the experiments, ToolGen was applied to three different LLMs- CodeGPT, CodeT5, and CodeLlama, and tested on two datasets CodeSearchNet and CoderEval to evaluate similarity-based and dependency-based effectiveness, and execution-based effectiveness, respectively. The results demonstrated models' improvement by enhancing 31.4% to 39.1% in Dependency Coverage, and 44.9% to 57.7% in Static Validity Rate for the three LLMs, while maintaining competitive performances on metrics BLEU-4, CodeBLEU, Edit Similarity, and Exact Match. In addition, ToolGen improved CodeT5 and Code Llama by 40% and 25%, respectively, and maintained the same pass rate for CodeGPT.

6.2 Code Search and Advanced Code Generation

Code search is an essential task in software development practices, allowing developers to efficiently create solutions to problems. An LLM-assisted tool that can enhance this task is RepoRift. Jain et al. [65] introduce this advanced code search application RepoRift designed to improve code snippet retrieval using LLMs with Retrieval Augmented Generation (RAG). It enhances user queries by injecting more context from GitHub repositories to

address issues like ambiguity and Vocabulary Mismatch Problems. RepoRift utilizes a multi-stream ensemble architecture that refines the search results by doing multiple comparisons and generating the most relevant snippets. For evaluation in the CodeSearchNet dataset, RepoLift significantly outperformed other code search methods by successfully achieving a success rate of 78.2% and 34.6% at Success@10 and Success@1, respectively. Furthermore, it delivers high accuracy using minimal preprocessing of the evaluation set and efficiently manages queries in different forms.

Extending code search capabilities, Feng et al. [66] present CodeBERT which is a bimodal pre-trained model designed to understand and generate NL and PL code, such as Python, Java, etc. With the Transformer-based neural architecture and training on a hybrid objective function combined with the pretraining task of replaced token detection, this allows codeBERT to leverage both bimodal (NL-PL pairs) and (only NL or PL) unimodal data. Based on these, CodeBERT shows its strong potential in code search. For evaluation, CodeBERT is tested using a dataset for NL-PL probing including NL code search in a zero-shot scenario and compared with an NL-based pre-trained model called RoBERTa. In an experiment on the CodeSearchNet corpus, CodeBERT performed better and more consistently than RoBERTa. Moreover, on the documentation generation task in six PLs, CodeBERT outperformed RoBERTa by achieving a 1.3 BLEU score gain and state-of-the-art performance.

Switching to advanced code generation, Li et al. [67] develop a model named AlphaCode to handle competitive programming problems that require advanced problem solving skills. It is initially pre-trained on selected GitHub code and fine-tuned on a curated dataset of competitive programming problems like CodeContests. The approach is to automatically generate millions of code examples, filter them according to their execution results, and cluster them, after which a small number of high-quality submissions are manually selected. For evaluation, using simulation on the Codeforces platform, AlphaCode's performance reached the top 54.3% among more than 5,000 human competitors. To improve this model, DeepMind developed a new dataset for training and evaluation called CodeContests. It combines data from multiple sources, where training data predate the evaluation problems, adds additional tests for accuracy, and has the evaluation of submissions done in a competitive programming-like setting. This results in 34.2% of the long-held competitive problems from CodeContests being solved by the best model. Finally, for the model's good and reliable performance, the paper found the following critical com-

ponents: a high-quality competitive programming dataset, efficient transformer models, and large-scale sampling.

6.3 Code Translation and Code Debugging

Hou and Ji [68] discuss the fact in a study that GPT-4 is the top-performing LLM in generating programming code that outperforms other LLMs, such as Gemini Ultra and Claude 2. It has gained success with various forms of programming tasks, including assisting in writing code, learning from coding error messages, and code translation. In a LeetCode and GeeksforGeeks coding contest between human programmers and LLMs, the GPT-4 success rate reached over 90% for tasks that only more than 20% human participants could solve. These showed that GPT-4 has the ability to be a reliable coding assistant. Furthermore, using prompt strategies, GPT-4 demonstrated its ability to learn from past errors by salvaging over 60% of easy and medium tasks after failing in the first attempt. Finally, for the task of translating the correct Python3 code to multiple different languages, it translated most of the tasks accurately. Surprisingly, in several medium tasks, it even tackled the programming task correctly despite giving the incorrect original Python3 code, proving it a reliable tool for code translation.

Furthermore, Prenner et al. [69] investigate Codex’s ability to detect and fix bugs, which are essential tasks for automated program repair. Codex, built on GPT-3 architecture, has shown great potential in generating code from NL descriptions. In this paper, Codex’s ability to fix software bugs was evaluated on the QuixBugs benchmark, which contains 40 bugs in both Java and Python, then comparing its performance with three APR approaches, such as CoCoNut, DeepDebug, and CURE. The results show that Codex performed the tasks surprisingly competitively, especially in Python, with 50% more bugs fixed compared to Java despite not being trained on Automatic Program Repair (APR). Codex outperformed both CoCoNuT and DeepDebug in Python and even outperformed CoCoNut in Java. Additionally, Codex performance was also tested using different prompt strategies for bug localization and repair, revealing that prompts can significantly impact Codex’s capability of fixing bugs effectively.

Finally, a notable application that can greatly assist LLMs in code translation is Flourine. Flourine [70], which is an end-to-end translation tool, ensures translation validation based on cross-language differential fuzzing without requiring any test case to check input-output similarity for the original and translated code. Flourine implements the feedback

strategy that provides input to the LLMs, allowing it to correct the identified counterexample. Experiments were carried out on 8160 code translations of 408 code samples, four feedback strategies, and five LLMs, including GPT4, Claude 3, Claude 2.1, Gemini Pro, and Mixtral. Benchmarks are collected as real-world projects from GitHub, using C and GO as the source code. The results revealed that the most successful LLM can achieve up to 47% of the benchmarks.

7 Conclusion

This survey provides an overview of the recent landscape of LLMs for automatic code generation. To begin with, we point out the limits and challenges LLM has faced, such as resource constraints; syntactic and semantic errors; biases; and security risks, highlighting factors that need to be mitigated. Subsequently, we discuss various fine-tuning techniques, including prompt engineering, reinforcement learning, and domain-specific dataset tuning which are essential approaches to handle the issues and enhance model performance and adaptability. We then examine the importance of evaluation metrics and benchmarks, as they are critical for assessing the effectiveness and reliability of the models, techniques, and their generated code to guide future development. Finally, we explore the significant potential of LLMs in many different coding tasks, including code generation, completion, search, debugging, and translation, which significantly boost productivity and efficiency for users in writing code.

References

- [1] Discover Data Science, “How to become a Data Mining Specialist – A Complete Career Guide,” *Discover Data Science*. Available: https://www.discoverdatasience.org/career-information/data-mining-specialist/?utm_source=chatgpt.com.
- [2] edX, “Learn data mining with online courses and programs,” *edX*. Available: https://www.edx.org/learn/data-mining?utm_source=chatgpt.com.
- [3] O. Samuel, “How to Use Pandas for Data Cleaning and Preprocessing,” *freeCodeCamp*, 2024. Available: https://www.freecodecamp.org/news/data-cleaning-and-preprocessing-with-pandasbdvhj/?utm_source=chatgpt.com.
- [4] K. Ketan, “Large Language Models for Code Generation,” *Medium*, 2023. Available: https://medium.com/@ketan_ketan/large-language-models-for-code-generation-5a2f3a2e3a2c.

- //blog.fabrichq.ai/large-language-models-for-code-generation-f95f93fe7de4.
- [5] OpenAI, “OpenAI Codex,” OpenAI, 2021. Available: <https://openai.com/index/openai-codex/>.
- [6] O. Mendelevitch, “Large Language Models for Code Generation - Part 1,” *Vectara*, 2023. Available: <https://www.vectara.com/blog/large-language-models-llms-for-code-generation-part-1>.
- [7] E. Anello, “How to Use GitHub Copilot: Use Cases and Best Practices,” *DataCamp*, 2024. Available: <https://www.datacamp.com/tutorial/github-copilot-a-complete-guide-for-beginners>.
- [8] IBM, “Large language models,” IBM. Available: <https://www.ibm.com/topics/large-language-models>.
- [9] AWS, “What is LLM (Large Language Model)?” AWS. Available: <https://aws.amazon.com/what-is/large-language-model/#:~:text=Large%20language%20models%2C%20also%20known,decoder%20with%20self%2Dattention%20capabilities>.
- [10] Nvidia, “Large Language Models Explained,” Nvidia. Available: <https://www.nvidia.com/en-us/glossary/large-language-models/>.
- [11] Analytics Insight, “Exploring Large Language Models: Foundations and Applications,” *Analytics Insight*, 2024. Available: <https://www.analyticsinsight.net/llm/exploring-large-language-models-foundations-and-applications>.
- [12] L. Price, “Large language models: What is driving the hype behind LLMs in healthcare?” *Nelson Advisors*, 2023. Available: <https://www.healthcare.digital/single-post/large-language-models-what-is-driving-the-hype-behind-llm-s-in-healthcare>.
- [13] J. D. Baierl, “Applications of Large Language Models in Education: Literature Review and Case Study,” *UCLA*, 2023. Available: <https://escholarship.org/uc/item/6kf0r28s>.
- [14] Daivi, “7 Top Large Language Model Use Cases And Applications,” *ProjectPro*, 2024. Available: https://www.projectpro.io/article/large-language-model-use-cases-and-applications/887#mcetoc_1h6mcnr1022.
- [15] A. Bleiweiss and N. Luo, “Mastering LLM Techniques: Data Preprocessing,” *Nvidia Technical Blog*, 2024. Available: <https://developer.r.nvidia.com/blog/mastering-llm-techniques-data-preprocessing/>.
- [16] AI Verse Info, “How does a Large Language Model (LLM) write Code,” *AI Verse Info*, 2024. Available: https://aiverseinfo.com/how-llm-writes-code/?amp=1&fbclid=IwZXh0bgNhZWOCMTEAAR0CD-t2AHBiDS2dca56gLKHuXx6pb6AsAi2jiOyVZ96HDTumoYAoqTmJPU_aem_EzXu_6J4h8wi11f3wPAwrQ.
- [17] M. Heller, “LLMs and the rise of the AI code generators,” *InfoWorld*, 2023. Available: <https://www.infoworld.com/article/2338500/llms-and-the-rise-of-the-ai-code-generators.html>.
- [18] Vellum, “LLM Leaderboard,” Vellum. Available: <https://www.vellum.ai/llm-leaderboard>.
- [19] AI/ML API, “GPT o1: Real-World Applications and Ultimate Prompt Guide,” *AI/ML API*, 2024. Available: <https://aimlapi.com/blog/gpt-o1-real-world-applications-and-ultimate-prompt-guide#:~:text=What%20Makes%20GPT%20AI%20o1,o1%20does%20the%20heavy%20lifting>.
- [20] V. Chhetri, “Why OpenAI’s new AI model, code-named Strawberry, can be good and bad at the same time,” *Tech Funding News*, 2024. Available: <https://techfundingnews.com/why-openais-new-ai-model-code-named-strawberry-can-be-good-and-bad-at-the-same-time/>.
- [21] OpenAI, “Learning to Reason with LLMs,” OpenAI. Available: <https://openai.com/index/learning-to-reason-with-llms/>.
- [22] P. Schmid, O. Sanseviero, A. Bartolome, L. von Werra, D. Vila, V. Srivastav, M. Sun, and P. Cuenca, “Llama 3.1 - 405B, 70B & 8B with multilinguality and long context,” *Hugging Face*, 2024. Available: <https://huggingface.co/blog/llama31>.
- [23] D. Cleary, “Using LLMs for Code Generation: A Guide to Improving Accuracy and Addressing Common Issues,” *PromptHub*, 2024. Available: <https://www.promphub.us/blog/using-llms-for-code-generation-a-guide-to-improving-accuracy-and-addressing-common-issues>.
- [24] D. Huang, Q. Bu, J. Zhang, X. Xie, J. Chen, and H. Cui, “Bias Testing and Mitigation in LLM-based Code Generation,” *arXiv preprint arXiv:2309.14345*, 2023. Available: <https://arxiv.org/abs/2309.14345>.

- [25] H. Hajipour, K. Hassler, T. Holz, L. Schönherr, and M. Fritz, “CodeLMSec Benchmark: Systematically Evaluating and Finding Security Vulnerabilities in Black-Box Code Language Models,” *arXiv preprint arXiv:2302.04012*, 2023. Available: <https://arxiv.org/abs/2302.04012>.
- [26] L. Chen, N. K. Ahmed, A. Dutta, A. Bhattacharjee, S. Yu, Q. I. Mahmud, W. Abebe, H. Phan, A. Sarkar, B. Butler, N. Hasabnis, G. Oren, V. A. Vo, J. P. Munoz, T. L. Willke, T. Mattson, and A. Jannesari, “The Landscape and Challenges of HPC Research and LLMs,” *arXiv preprint arXiv:2402.02018*, 2024. Available: <https://arxiv.org/pdf/2402.02018.pdf>.
- [27] A. Chavan, R. Magazine, S. Kushwaha, M. Debbah, and D. Gupta, “Faster and Lighter LLMs: A Survey on Current Challenges and Way Forward,” *arXiv preprint arXiv:2402.01799*, 2024. Available: <https://arxiv.org/pdf/2402.01799.pdf>.
- [28] M. Hassid, T. Remez, J. Gehring, R. Schwartz, and Y. Adi, “The Larger the Better? Improved LLM Code-Generation via Budget Reallocation,” *arXiv preprint arXiv:2404.00725*, 2024. Available: <https://arxiv.org/html/2404.00725v1>.
- [29] Z. Wang, Z. Zhou, D. Song, Y. Huang, S. Chen, L. Ma, and T. Zhang, “Where Do Large Language Models Fail When Generating Code?” *arXiv preprint arXiv:2406.08731*, 2024. Available: <https://arxiv.org/pdf/2406.08731.pdf>.
- [30] S. Dou, H. Jia, S. Wu, H. Zheng, W. Zhou, M. Wu, M. Chai, J. Fan, C. Huang, Y. Tao, Y. Liu, E. Zhou, M. Zhang, Y. Zhou, Y. Wu, R. Zheng, M. Wen, R. Weng, J. Wang, X. Cai, T. Gui, X. Qiu, Q. Zhang, and X. Huang, “What’s Wrong with Your Code Generated by Large Language Models? An Extensive Study,” *arXiv preprint arXiv:2407.06153*, 2024. Available: <https://arxiv.org/html/2407.06153v1>.
- [31] R. Pan, A. R. Ibrahimzada, R. Krishna, D. Sankar, L. P. Wassi, M. Merler, B. Sobolev, R. Pavuluri, S. Sinha, and R. Jabbarvand, “Lost in Translation: A Study of Bugs Introduced by Large Language Models while Translating Code,” *arXiv preprint arXiv:2308.03109*, 2023. Available: <https://arxiv.org/abs/2308.03109>.
- [32] Y. Liu, T. Le-Cong, R. Widyasari, C. Tanithamthavorn, L. Li, X.-B. D. Le, and D. Lo, “Refining ChatGPT-Generated Code: Characterizing and Mitigating Code Quality Issues,” *ACM Journals*, 2024. Available: <https://dl.acm.org/doi/full/10.1145/3643674>.
- [33] C. Wang, Z. Li, C. Gao, W. Wang, T. Peng, H. Huang, Y. Deng, S. Wang, and M. R. Lyu, “Exploring Multi-Lingual Bias of Large Code Models in Code Generation,” *arXiv preprint arXiv:2404.19368*, 2024. Available: <https://arxiv.org/abs/2404.19368>.
- [34] Y. Liu, X. Chen, Y. Gao, Z. Su, F. Zhang, D. Zan, J.-G. Lou, P.-Y. Chen, and T.-Y. Ho, “Uncovering and Quantifying Social Biases in Code Generation,” *arXiv preprint arXiv:2305.15377*, 2023. Available: <https://arxiv.org/abs/2305.15377>.
- [35] N. Islam, J. Khouri, A. Seong, E. Bou-Hardb, and P. Najafirad, “Enhancing Source Code Security with LLMs: Demystifying the Challenges and Generating Reliable Repairs,” *arXiv preprint arXiv:2409.00571*, 2024. Available: <https://arxiv.org/abs/2409.00571>.
- [36] He, “Large language models for code writing: Security assessment,” *Medium*, 2023. Available: <https://medium.com/@researchgraph/large-language-models-for-code-writing-security-assessment-f305f9f01ce9>.
- [37] G. S. Black, B. P. Rimal, and V. M. Vaidyan, “Balancing Security and Correctness in Code Generation: An Empirical Study on Commercial Large Language Models,” *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2024. Available: <https://ieeexplore.ieee.org.ezproxy.lib.ou.edu/document/10658990>.
- [38] J. Wang, X. Luo, L. Cao, H. He, H. Huang, J. Xie, A. Jatowt, and Y. Cai, “Is Your AI-Generated Code Really Secure? Evaluating Large Language Models on Secure Code Generation with CodeSecEval,” *arXiv preprint arXiv:2407.02395*, 2024. Available: <https://arxiv.org/html/2407.02395v1>.
- [39] Turing, “Finetuning large language models: An in-depth guide,” *Turing*. Available: <https://www.turing.com/resources/fine-tuning-large-language-models>.
- [40] Z. Ma, H. Guo, J. Chen, G. Peng, Z. Cao, Y. Ma, and Y.-J. Gong, “LLaMoCo: Instruction Tuning of Large Language Models for Optimization Code Generation,” *arXiv preprint arXiv:2403.01131*, 2024. Available: <https://arxiv.org/pdf/2403.01131v1.pdf>.
- [41] M. Weyssow, X. Zhou, K. Kim, D. Lo, and H. Sahraoui, “Exploring Parameter-Efficient

- Fine-Tuning Techniques for Code Generation with Large Language Models,” *arXiv preprint arXiv:2308.10462*, 2023. Available: <https://arxiv.org/pdf/2308.10462.pdf>.
- [42] Y. Tsai, M. Liu, and H. Ren, “Code Less, Align More: Efficient LLM Fine-tuning for Code Generation with Data Pruning,” *arXiv preprint arXiv:2407.05040*, 2024. Available: <https://arxiv.org/pdf/2407.05040.pdf>.
- [43] F. Mu, L. Shi, S. Wang, Z. Yu, B. Zhang, C. Wang, S. Liu, and Q. Wang, “ClarifyGPT: Empowering LLM-based Code Generation with Intention Clarification,” *arXiv preprint arXiv:2310.10996*, 2023. Available: <https://arxiv.org/pdf/2310.10996.pdf>.
- [44] J. Gehring, K. Zheng, J. Copet, V. Mella, T. Cohen, and G. Synnaeve, “RLEF: Grounding Code LLMs in Execution Feedback with Reinforcement Learning,” *arXiv preprint arXiv:2410.02089*, 2024. Available: <https://arxiv.org/pdf/2410.02089.pdf>.
- [45] M. F. Wong and C. W. Tan, “Aligning Crowd-sourced Human Feedback for Code Generation with Bayesian Inference,” *IEEE*, 2024. Available: <https://ieeecaai.org/2024/wp-content/pdfs/540900a152/540900a152.pdf>.
- [46] Z. Sun, C. Lyu, B. Li, Y. Wan, H. Zhang, G. Li, and Z. Jin, “Enhancing Code Generation Performance of Smaller Models by Distilling the Reasoning Ability of LLMs,” *arXiv preprint arXiv:2403.13271*, 2024. Available: <https://arxiv.org/pdf/2403.13271v1.pdf>.
- [47] J. Li, Y. Zhao, Y. Li, G. Li, and Z. Jin, “AceCoder: Utilizing Existing Code to Enhance Code Generation,” *arXiv preprint arXiv:2303.17780*, 2023. Available: <https://arxiv.org/pdf/2303.17780.pdf>.
- [48] C. Tony, N. E. Díaz Ferreyra, M. Muttas, S. Dhiff, and R. Scandariato, “Prompting Techniques for Secure Code Generation: A Systematic Investigation,” *arXiv preprint arXiv:2407.07064*, 2024. Available: <https://arxiv.org/pdf/2407.07064.pdf>.
- [49] D. G. Paul, H. Zhu, and I. Bayley, “Benchmarks and Metrics for Evaluations of Code Generation: A Critical Review,” *arXiv preprint arXiv:2406.12655*, 2024. Available: <https://arxiv.org/html/2406.12655v1.html>.
- [50] J. Ip, “LLM Evaluation Metrics: The Ultimate LLM Evaluation Guide,” *Confident AI*, 2024. Available: <https://www.confident-ai.com/blog/llm-evaluation-metrics-everything-you-need-for-llm-evaluation>.
- [51] J. Ip, “Evaluating LLM Systems: Essential Metrics, Benchmarks, and Best Practices,” *Confident AI*, 2024. Available: <https://www.confident-ai.com/blog/evaluating-llm-systems-metrics-benchmarks-and-best-practices>.
- [52] S. Ren, D. Guo, S. Lu, L. Zhou, S. Liu, D. Tang, N. Sundaresan, M. Zhou, A. Blanco, and S. Ma, “CodeBLEU: a Method for Automatic Evaluation of Code Synthesis,” *arXiv preprint arXiv:2009.10297*, 2020. Available: <https://arxiv.org/pdf/2009.10297.pdf>.
- [53] S. Yeo, Y.-S. Ma, S. C. Kim, H. Jun, and T. Kim, “Framework for evaluating code generation ability of large language models,” *Wiley Online Library*, 2024. Available: <https://onlinelibrary.wiley.com/doi/10.4218/etrij.2023-0357>.
- [54] T. Y. Zhuo, “ICE-Score: Instructing Large Language Models to Evaluate Code,” *arXiv preprint arXiv:2304.14317*, 2024. Available: <https://arxiv.org/abs/2304.14317>.
- [55] Symflower, “Comparing LLM Benchmarks,” *Symflower*, 2024. Available: <https://symflower.com/en/company/blog/2024/comparing-llm-benchmarks/>.
- [56] Z. Wang, “HumanEval: Decoding the LLM Benchmark for Code Generation,” *Deepgram*, 2023. Available: <https://deepgram.com/learn/humaneval-llm-benchmark>.
- [57] X. Du, M. Liu, K. Wang, H. Wang, J. Liu, Y. Chen, J. Feng, C. Sha, X. Peng, and Y. Lou, “ClassEval: A Manually-Crafted Benchmark for Evaluating LLMs on Class-level Code Generation,” *arXiv preprint arXiv:2308.01861*, 2023. Available: <https://arxiv.org/pdf/2308.01861.pdf>.
- [58] C. E. Jimenez, J. Yang, A. Wettig, S. Yao, K. Pei, O. Press, and K. Narasimhan, “SWE-bench: Can Language Models Resolve Real-World GitHub Issues?” *arXiv preprint arXiv:2310.06770*, 2023. Available: <https://arxiv.org/pdf/2310.06770.pdf>.
- [59] T. Y. Zhuo, J. Liu, Q. Liu, B. Hui, N. Muenninghoff, D. Fried, H. de Vries, L. von Werra, and C. Fourrier, “BigCodeBench: The Next Generation of HumanEval,” *Hugging Face*, 2024. Available: <https://huggingface.co/blog/leaderboard-bigcodebench>.
- [60] Onome, “Top LLMs for Coding All Developers Should Know About,” *AutoGPT*, 2024. Available: <https://autogpt.net/top-llms-for-coding-all-developers-should-know-about>.

- oding-all-developers-should-know-about /#:~:text=CodeQwen1.,working%20with%20d iverse%20programming%20languages.
- [61] K. Zhang, J. Li, G. Li, X. Shi, and Z. Jin, “CodeAgent: Enhancing Code Generation with Tool-Integrated Agent Systems for Real-World Repo-level Coding Challenges,” *arXiv preprint arXiv:2401.07339*, 2024. Available: <https://arxiv.org/abs/2401.07339>.
- [62] H. Xu, “Github Copilot - A Groundbreaking Code Autocomplete Tool,” *Research Gate*, 2023. Available: https://www.researchgate.net/publication/376406939_Github_Copilot_-_A_Groundbreaking_Code_Autocomplete_Tool.
- [63] Meta, “Introducing Code Llama, a state-of-the-art large language model for coding,” *Meta AI*, 2023. Available: <https://ai.meta.com/blog/code-llama-large-language-model-coding/>.
- [64] C. Wang, J. Zhang, Y. Feng, T. Li, W. Sun, Y. Liu, and X. Peng, “Teaching Code LLMs to Use Autocompletion Tools in Repository-Level Code Generation,” *arXiv preprint arXiv:2401.06391*, 2024. Available: <https://arxiv.org/abs/2401.06391>.
- [65] S. Jain, A. Dora, K. S. Sam, and P. Singh, “LLM Agents Improve Semantic Code Search,” *arXiv preprint arXiv:2408.11058*, 2024. Available: <https://arxiv.org/abs/2408.11058>.
- [66] Z. Feng, D. Guo, D. Tang, N. Duan, X. Feng, M. Gong, L. Shou, B. Qin, T. Liu, D. Jiang, and M. Zhou, “CodeBERT: A Pre-Trained Model for Programming and Natural Languages,” *arXiv preprint arXiv:2002.08155*, 2020. Available: <https://arxiv.org/pdf/2002.08155.pdf>.
- [67] Y. Li, D. Choi, J. Chung, N. Kushman, J. Schrittwieser, R. Leblond, T. Eccles, J. Keeling, F. Gimeno, A. Dal Lago, T. Hubert, P. Choy, C. de Masson d’Autume, I. Babuschkin, X. Chen, P.-S. Huang, J. Welbl, S. Gowal, A. Cherepanov, J. Molloy, D. J. Mankowitz, E. S. Robson, P. Kohli, N. de Freitas, K. Kavukcuoglu, and O. Vinyals, “Competition-Level Code Generation with AlphaCode,” *DeepMind*, 2022. Available: <https://arxiv.org/pdf/2203.07814.pdf>.
- [68] W. Hou and Z. Ji, “Comparing large language models and human programmers for generating programming code,” *arXiv preprint arXiv:2403.00894*, 2024. Available: <https://arxiv.org/abs/2403.00894>.
- [69] J. A. Prenner, H. Babii, and R. Robbes, “Can OpenAI’s Codex Fix Bugs?” *IEEE*, 2022. Available: <https://ieeexplore.ieee.org.ezproxy.lib.ou.edu/stamp/stamp.jsp?tp=&arnumber=9809175>.
- [70] H. F. Eniser, H. Zhang, C. David, M. Wang, M. Christakis, B. Paulsen, J. Dodds, and D. Kroening, “Towards Translating Real-World Code with LLMs: A Study of Translating to Rust,” *arXiv preprint arXiv:2405.11514*, 2024. Available: <https://arxiv.org/abs/2405.11514>.