

# A Study of LLMs’ Preferences for Libraries and Programming Languages

Lukas Twist

King’s College London  
London, UK  
lukas.twist@kcl.ac.uk

Don Syme

GitHub Next  
London, UK  
dsyme@github.com

Jie M. Zhang

King’s College London  
London, UK  
jie.zhang@kcl.ac.uk

Joost Noppen

AI Research Labs, BT plc  
Ipswich, UK  
johannes.noppen@bt.com

Mark Harman

University College London  
London, UK  
mark.harman@ucl.ac.uk

Helen Yannakoudakis

King’s College London  
London, UK  
helen.yannakoudakis@kcl.ac.uk

Detlef Nauck

AI Research Labs, BT plc  
Ipswich, UK  
detlef.nauck@bt.com

## Abstract

Large Language Models (LLMs) are increasingly used to generate code, influencing users’ choices of libraries and programming languages in critical real-world projects. However, little is known about their systematic biases or preferences toward certain libraries and programming languages, which can significantly impact software development practices. To fill this gap, we perform the first empirical study of LLMs’ preferences for libraries and programming languages when generating code, covering eight diverse LLMs. Our results reveal that LLMs exhibit a strong tendency to overuse widely adopted libraries such as NumPy; in up to 48% of cases, this usage is unnecessary and deviates from the ground-truth solutions. LLMs also exhibit a significant preference toward Python as their default language. For high-performance project initialisation tasks where Python is not the optimal language, it remains the dominant choice in 58% of cases, and Rust is not used a single time. These results indicate that LLMs may prioritise familiarity and popularity over suitability and task-specific optimality. This will introduce security vulnerabilities and technical debt, and limit exposure to newly developed, better-suited tools and languages. Understanding and addressing these biases is essential for the responsible integration of LLMs into software development workflows.

## 1 Introduction

The last few years have seen Large Language Models (LLMs) make huge advances [55], particularly excelling in code generation tasks [33]. Widely accessible tools—such as ChatGPT [60] and GitHub Copilot [9]—enable developers of all experience levels to use generative AI in their workflow, leading to a significant impact on newly produced code [70, 42]. Existing research has focused mainly on evaluating and improving LLM correctness, security, and efficiency in generating code [8]. Although LLMs demonstrate impressive performance, little is known about their preferences when selecting libraries and programming languages for various coding tasks.

This gap is particularly significant as LLMs are increasingly used as coding assistants by both professional developers and by end

users who often lack the expertise to assess whether a recommended library or programming language is appropriate [1, 38]. If LLMs exhibit systematic biases in libraries and programming languages, they can mislead users into making flawed decisions, compromising the reliability, security, and maintainability of software.

To fill this gap, this paper provides the first empirical investigation of the library and programming language preferences of LLMs when generating code. To ensure a broad understanding of LLM preferences, *eight* widely-used LLMs are chosen for the study, covering both open- and closed-source and a range of sizes: GPT-4o [60], GPT-3.5 [2], Claude3.5 (Haiku and Sonnet) [3], Llama3.2 [22], Qwen2.5 [30], DeepSeekLLM [15] and Mistral7b [32].

We begin with libraries—vital components of modern software development [68]—because developers often omit them from prompts [26] and may not know which dependencies a task requires [43], giving LLMs the opportunity to influence library adoption. We then consider the programming language preferences of LLMs. Although professional developers may specify a programming language in the prompt, omitting this information allows us to probe the inherent biases LLMs exhibit when left to make coding choices on their own. This analysis is particularly important as more end users, especially those with limited programming experience, increasingly rely on LLMs’ default choices, often lacking the expertise to evaluate which languages are most appropriate for a given task [57], which is a trend especially evident in “vibe-coding” scenarios [67].

For both libraries and programming languages, we examine LLM preferences in two key scenarios: writing code for practical tasks from widely studied benchmarks; and generating the initial code for new projects, a crucial phase where LLMs have been found out to be particularly helpful [64] and foundational technology choices are made. Additionally, we explore whether the coding technologies that LLMs provide as recommendations match those they actually use when generating code, and whether all studied LLMs share similar preferences.

Our experiments lead to the following primary observations:  
**(1)** All LLMs we study **heavily favour well established libraries** over high-quality alternatives. In particular, LLMs tend to overuse

widely adopted data science libraries. For example, NumPy is used unnecessarily in up to 48% of cases, diverging from the ground-truth solutions. (2) For programming languages, all LLMs we study exhibit a **significant preference toward using Python**. To our surprise, even for tasks where high performance and memory safety are critical—and therefore Python is suboptimal—Python remains the dominant choice as the most used language in 58% of cases, while Rust is not used a single time. (3) There is **extremely low consistency** between the libraries and programming languages recommended by LLMs and those they actually use.

LLMs may develop these strong biases due to various factors, such as the prevalence of certain libraries in open source repositories, the distribution of programming languages included in their training data [76], or post-training alignment and fine-tuning processes [89]. For example, if an LLM is trained on codebases where a particular library is favoured, it may exhibit a strong inclination toward generating code that uses the given library, even in cases where it might not be necessary.

Understanding and mitigating programming language and library bias in LLMs is *not* merely an academic concern. The systemic biases we observe have immediate and far-reaching practical implications: (1) Strong biases in default choices **introduce technical debt and security risks** because of suboptimal choices, this is especially problematic for safety-critical systems. (2) LLMs now frequently augment training data with self-generated data [87], where language and library favouritism will create a **dangerous feedback loop**, further reducing diversity in synthetic data. (3) Bias toward well established languages and libraries **actively blocks newer languages and open-source tools**. This not only creates a discoverability barrier for newer or community-driven tools but also inhibits innovation, particularly in open-source environments.

We call for contributions from the research and developer communities to better understand, evaluate, and mitigate these biases, ensuring that LLMs support a more diverse, responsible, and forward-looking software development landscape. The primary contributions of this paper are as follows.

- Presenting the first empirical study of LLM preferences for libraries and programming languages when generating code, revealing a clear systemic bias and lack diversity, with significant preferences toward established libraries and Python.
- Increasing awareness of the need to measure and understand the diversity and preferences within LLM-generated code, providing a discussion of how this could impact the coding landscape and open-source software ecosystem.
- Releasing our code, datasets and complete results publicly via our GitHub repository<sup>1</sup>, to encourage further investigation in this area.

The rest of the paper is organised as follows. Section 2 summarises work in related areas. Section 3 details the research questions, and the experiments conducted to answer them. Section 4 presents the results and answers to the research questions. Section 5 further discusses the findings and their broader impacts. Section 6 presents the threats to validity. Section 7 concludes the paper.

<sup>1</sup>Project repository: <https://github.com/itsluketwist/llm-code-bias>

## 2 Related Work

To our knowledge, this is the first study to examine LLM biases for libraries and programming languages when writing code. In the following, we cover related questions from prior work.

**Bias and Unfairness in LLMs:** Whether an LLM has preferences for specific programming languages and libraries is related to bias and whether an LLM provides an unfair representation of the world. Bias within an LLM’s natural language (NL) responses has been extensively studied [20], but bias within any generated code is widely under explored. Most works focus on the trustworthiness of the code [61, 78, 53], or how LLMs can propagate harmful social biases within their code [49, 29] - not whether the coding choices themselves include bias. A study has been conducted on the similarities between human and LLM preferences in task responses [45], but it does not go into depth on code generation preferences.

**LLM-based Code Generation:** Code-generation in LLMs has been extensively studied [33, 86], yet the best way to evaluate the performance of LLMs on code generation tasks is an open problem [62]. Existing approaches typically use benchmarks that contain a dataset of programming problems and tests to check the functionality of generated code [62], although early benchmarks have been shown to be limited in their scope [82]. Using LLMs to perform their own code evaluation has also been suggested [72], as well as the need to test LLMs in more realistic repository-level code-generation scenarios [34].

**Library-oriented Code:** Using libraries to enhance code is a key part of modern software development [68]. Multiple studies discuss the generation of library-oriented code [84, 48], including when private libraries must be used [85], and the API recommendation problem [50, 75] within LLMs. None of the above discuss the potential impact of bias in LLM library preferences.

**Software Engineering Decisions:** This paper argues that LLMs will affect how a developer makes decisions during software development. The literature discusses how developers select libraries [71, 43], and make implementation choices [47], without the use of LLMs. Using ChatGPT as a software librarian has been suggested to aid these decisions [44], as well as how users’ distrust towards LLMs may impact the choices they make when using them as coding assistants [11]. An study has also been done to analyse how an LLM can best benefit a software engineering project [64].

## 3 Experimental Design

This section presents the research questions (RQs) and describes the experiments designed to answer them. For transparency and reproducibility, we have released the full code for all experiments described here on our GitHub repository<sup>1</sup>.

### 3.1 Research Questions

RQ1: **Library Preference:** *What library preferences do LLMs exhibit in code generation?* To answer this RQ, we analyse the library preferences in LLMs using library-agnostic benchmark tasks (Section 3.3.1) and more realistic project initialisation tasks (Section 3.3.2).

RQ2: **Language Preference:** *What programming language preferences do LLMs exhibit in code generation?* To answer this RQ, we analyse the programming language preferences in LLMs

- using widely adopted benchmark tasks (Section 3.4.1) and more realistic project initialisation tasks (Section 3.4.2).
- RQ3: **Recommendation Consistency:** *Is there any consistency between the libraries and programming languages LLMs recommend for a task and those they actually use?* To answer this RQ, we ask LLMs to rank their recommended libraries (Section 3.3.2) and programming languages (Section 3.4.2) for project initialisation tasks; then evaluate the alignment between their recommendations and actual choices using Kendall's  $\tau$  coefficient.
- RQ4: **LLM Similarities:** *Do different LLMs have similar preferences when selecting libraries and programming languages?* To answer this RQ, we examine the libraries and programming languages used in all experiments (Sections 3.3.1, 3.3.2, 3.4.1, 3.4.2) and assess the similarity between LLMs using Kendall's  $\tau$  coefficient.

## 3.2 LLM Selection

We use a diverse set of LLMs in this study, to gain a broad understanding of LLM preferences. LLMs are chosen to vary in number of parameters, availability (open- or closed-source), and intended use case (general or code-specific). This setup enables us to explore whether the underlying architecture influences the results. *Eight* LLMs were chosen, with full details given in Table 1.

**LLM Usage:** To reflect the typical usage of LLMs by developers, which often overlooks the role of parameters [17], each LLM is prompted using the default values for the *temperature* and *top\_p* parameters. Closed-source LLMs are prompted using the default values of their corresponding APIs; open-source models are prompted using the default values given on each LLM's *Hugging Face*<sup>2</sup> repository. The parameter values are given in Table 1.

Furthermore, we conduct each LLM interaction in a fresh API session, to avoid bias from prompt caching or leakage [24]; and we do not use a system prompt to ensure that each LLM has its base functionality considered without external influence [54].

**Table 1:** Summary of LLMs used in this study.

Model	Version	Release	Knowledge cut-off	Open-source?	Code model?	Size	Parameters	
							temp	top_p
GPT-4o [60]	gpt-4o-mini-2024-07-18	July '24	Oct. '23	✗	✗	-	1.0	1.0
GPT-3.5 [2]	gpt-3.5-turbo-0125	Nov. '22	Sep. '21	✗	✗	-	1.0	1.0
Sonnet3.5 [3]	claude-3.5-sonnet-20241022	Oct. '24	July '24	✗	✗	-	1.0	1.0
Haiku3.5 [3]	claude-3.5-haiku-20241022	Oct. '24	July '24	✗	✗	-	1.0	1.0
Llama3.2 [22]	llama-3.2-3b-instruct-turbo	Sep. '24	Dec. '23	✓	✗	3B	0.6	0.9
Mistral7B [32]	mistral-7b-instruct-v0.3	May '24	May '24	✓	✗	7B	0.7	1.0
Qwen2.5 [30]	qwen2.5-coder-32b-instruct	Nov. '24	Mar. '24	✓	✓	32B	0.7	0.8
DeepSeekLLM [15]	deepseek-llm-67b-chat	Nov. '23	May '23	✓	✗	67B	0.7	0.95

## 3.3 Experimental Setup for Library Preferences

To investigate library preferences, we focus on a single programming language to enable a more in-depth analysis. Python is chosen for library-related experiments due to its vast collection of

open-source libraries<sup>3</sup>, its easy-to-parse import syntax (as discussed in 3.5.1), and its popularity in the open-source community [69].

Previous work shows that users often start tasks with NL descriptions without knowing which—or that any—libraries are required [40]; and when prompting LLMs, will initially omit details, iteratively refining their requirements over time [26]. Therefore, we investigate library preferences using tasks that omit the specific libraries to use, to mirror real-world prompting.

**3.3.1 Benchmark Tasks.** We begin by investigating the library preferences of LLMs when writing code for self-contained tasks, such as those commonly found in widely used benchmarks [77].

**Dataset Selection:** Typically, library-specific Python benchmarks (such as DS-1000 [41]) focus on evaluating LLM performance within narrow domains and are ill-suited for investigating library choice across diverse tasks. To ensure reliable results, it is also vital to avoid data contamination between the datasets that we choose and the LLMs that we test [10]. Therefore, we choose to adopt BigCodeBench [88]. BigCodeBench contains 1140 Python coding tasks across 7 domains (general, computation, visualisation, system, time, network and cryptography) - 813 of which include external libraries in their ground truth solution. Although seeded from the ODEX dataset [79], its task descriptions have been back-translated using the NL-Augmenter method [16], and the latest version's February 2025 release postdates our LLMs' knowledge cutoffs, both of which minimise bias from previous exposure.

**Dataset Processing:** We make use of the NL description for each BigCodeBench task. Some tasks implicitly reveal the libraries used in their ground-truth solutions, therefore, to eliminate bias, we applied a regex filter to remove any task whose description mentioned one of those libraries. The filtered dataset contains 525 tasks, with 34 distinct libraries in their ground truth solutions.

**Prompt Strategy:** We prompt each LLM to write code using Prompt 1. Our prompt is modelled on the prompt used by BigCodeBench [88], with the additional directive to use an external library. *Three* responses are generated per task to reduce the impact of the inherent randomness of the LLMs [66], ensuring a fair representation of the preferred libraries.

PROMPT 1. “<task description> You should write self-contained python code. Choose, import and use at least one external library.”

**3.3.2 Project Initialisation Tasks.** To examine library preferences in a realistic context, we ask LLMs to write the initial code for various project descriptions. LLMs have been shown to be particularly helpful in the early stages of software development projects [64]; and with the advent of “vibe-coding” [67]—where users provide only high-level descriptions, allowing the LLM to determine the implementation details—this is an increasingly common use case of LLMs, to speed up repetitive tasks when establishing projects. It is also a natural decision point for adopting a new library, where the LLM's preferences may influence users' choices.

**Task Selection:** The project descriptions need to be realistic for users to want to build, require an external library to support core functionality, and have many viable libraries that could be used. Therefore, the chosen project descriptions are inspired by

<sup>2</sup>Hugging Face: <https://huggingface.co/models>

<sup>3</sup>Python Package Index: <https://pypi.org/>

the groups of related libraries on *Awesome Python*<sup>4</sup>, ensuring that there are multiple possible core libraries for each task. The task descriptions used are listed below, along with a selection of the possible libraries for each.

- (1) **Database Object-Relational Mapping (ORM):** “*Write the initial code for a database project with an orm layer.*”  
Possible libraries: Django, peewee, Pony, pydAL, SQLAlchemy.
- (2) **Deep-learning:** “*Write the initial code for a deep learning project implementing a neural network.*”  
Possible libraries: Caffe, Keras, PyTorch, scikit-learn, TensorFlow.
- (3) **Distributed computing:** “*Write the initial code for a distributed computing project.*”  
Possible libraries: Celery, Dask, Luigi, PySpark, Ray.
- (4) **Web-scaper:** “*Write the initial code for a web scraping and analysis library.*”  
Possible libraries: BeautifulSoup, lxml, MechanicalSoup, Scrapy.
- (5) **Web-server:** “*Write the initial code for a backend API web server.*”  
Possible libraries: Django Rest, FastAPI, Flask, Pyramid, Starlette.

**Prompt Strategy:** We generate responses from the LLMs using Prompt 1. Project initialisation is inherently open-ended, with some projects likely requiring multiple libraries - our prompt allows for this, giving the LLMs freedom in how they write the code. Additionally, to account for LLM nondeterminism and to accurately capture each LLM’s library preferences, we sample responses to each prompt 100 times per LLM and analyse the resulting distribution of libraries used, following recommended best practice for handling evaluation variance [51].

For RQ3: *Recommendation Consistency*, we examine whether LLMs are consistent between their library recommendations for a task and what they actually use. To assess this, each model is asked to provide an ordered list of recommended Python libraries for each task using Prompt 2. This process is repeated three times to mitigate randomness, with the arithmetic mean of the ranks calculated to determine the final rank per LLM.

PROMPT 2. “*List, in order, the best (python libraries / languages) for the following task: <task description>*”

### 3.4 Experimental Setup for Language Preferences

Next, we investigate LLM preferences for the programming languages they use when not given a specific directive. Although this is less common among professional developers—mostly seen from novice programmers who struggle to write detailed prompts [57], and “vibe-coders” [67], that leave more of the technical decisions to the LLM—it will allow us to further study the inherent biases LLMs exhibit when allowed to make coding decisions.

**3.4.1 Benchmark Tasks.** We begin by investigating the programming language preferences of LLMs when solving language-agnostic coding tasks from datasets that are widely adopted to benchmark model performance.

**Dataset Selection:** We require the benchmark tasks to be described in NL only, as any code present is likely to influence the LLMs’ choice of programming language. Furthermore, to mitigate

the risk of data leakage [81], we prioritise benchmarks with ground-truth solutions in multiple programming languages, or those likely to have had solutions published online in multiple languages. This approach ensures that any prior exposure to solutions is distributed across programming languages, removing the likelihood that the observed preference is merely an artifact of contamination. In total, we selected six datasets, organised into *three* categories.

- (1) **Basic:** Short coding problems that an entry-level programmer could solve.
  - **Multi-HumanEval** [4]: A multi-language version of the popular HumanEval [9] dataset originally published by OpenAI in 2021, to test code generation from docstrings.  
Tasks: 160. Solution languages: C#, Go, Java, Javascript, Kotlin, Perl, PHP, Python, Ruby, Scala, Swift, Typescript.
  - **MBXP** [4]: A multi-language version of the popular MBPP [5] dataset originally published by Google in 2021, containing simple problems that require short solutions.  
Tasks: 968. Solution languages: C#, Go, Java, Javascript, Kotlin, Perl, PHP, Python, Ruby, Scala, Swift, Typescript.
- (2) **Real-world:** Code generation tasks constructed from realistic coding situations that developers have faced. Because the problems are realistic, there are likely to be solutions available online in many languages.
  - **AixBench** [27]: A method-level code generation dataset built from comments found in public GitHub Java methods.  
Tasks: 161 total, 129 valid. Solution language: Java.
  - **CoNaLa** [83]: A code generation from NL dataset built from coding problems found on Stack Overflow.  
Tasks: 2879 total, 2659 valid. Solution language: Python.
- (3) **Coding challenge:** In-depth problems that typically require more thinking in order to solve. Sourced from online coding competition websites, they are language-agnostic by nature. Users can solve the problems in any language they choose, so there are solutions published online in many languages [19].
  - **APPS** [28]: A large collection of problems from various open-access coding websites, covering various difficulty levels.  
Tasks: 10000 total, 8623 valid. Solution language: Python.
  - **CodeContests** [46]: A competitive coding dataset from 2024, initially used to train AlphaCode [13].  
Tasks: 13610 total, 12830 valid. Solution languages: C++, Java, Python.

**Dataset Processing:** We found that some task descriptions include programming language names. Therefore, to eliminate bias, we performed a regex string search on all task descriptions for any programming language in the TIOBE Index [31] top 50, discarding those tasks. Preliminary experiments showed that a sample of 200 tasks gave a good representation of the results for the larger datasets, while being similar in size to the smaller datasets, allowing results to be more fairly aggregated. Therefore, after filtering, we randomly select 200 problems to use from each dataset.

**Prompt Strategy:** We prompt each LLM to generate code for each benchmark task, using Prompt 3. We use a prompt that is compatible with the varying task formats in the datasets, requesting an explanation to encourage the LLMs to respond with NL and code in Markdown format. As in Section 3.3.1, *three* responses are generated per task to reduce the effect of LLM randomness [66], getting a more fair representation of their language preferences.

<sup>4</sup>Awesome Python: <https://www.awesomepython.org/>

PROMPT 3. “Generate a code-based solution, with an explanation, for the following task or described function: <task description>”

**3.4.2 Project Initialisation Tasks.** For the same reasons discussed in Section 3.3.2, LLMs are asked to write the initial code for various language-agnostic project descriptions.

**Task Selection:** Our initial results imply that LLMs have a general preference towards using Python. Therefore, to challenge the LLMs’ default programming language preference, we choose project descriptions with non-functional requirements such that Python is regarded as a suboptimal choice. Python is widely appreciated for its developer-friendly syntax and rapid prototyping capabilities, but is not considered a high-performance language in all situations due to its interpretive nature and inefficient concurrency implementations [21]. For computationally intensive tasks, compiled languages like C, C++, and Rust are generally preferred; they provide greater memory control, lower execution overhead, and superior parallel processing [14]. The chosen task descriptions are listed below.

- (1) **Concurrency:** “Write the initial code for a high-performance web server to handle a large number of concurrent requests.” High performance web servers require efficient thread management and concurrency control, areas where Rust (with full async capabilities) or C++ (with multithreading) outperform Python [7].
- (2) **Graphical interface:** “Write the initial code for a modern cross-platform application with a graphical user interface.” GUIs benefit from native performance optimisations and low-latency rendering, Python is inefficient with native libraries, making low-level languages like C and C++ more suitable [36].
- (3) **Low-latency:** “Write the initial code for a low-latency trading platform that will allow scaling in the future.” Financial trading systems demand minimal execution delays, and benefit from precise control over hardware and low-level memory management, areas where C++ and Go are better than Python [58].
- (4) **Parallel processing:** “Write the initial code for a high performance parallel task processing library.” Python is suboptimal here because it does not have true parallelism due to the global interpreter lock [23], compiled languages with true built-in parallelism like C, C++, and Rust are more suitable.
- (5) **System-level programming:** “Write the initial code for a command line application to perform system-level programming.” System-level programming inherently relies on direct hardware interaction and efficient memory usage, making C, C++, and Rust better choices than Python [14].

**Prompt Strategy:** We generate responses using only the task descriptions as prompts. This gives the LLMs freedom in how they approach writing the code for these more open-ended tasks, where it is feasible that multiple languages could be used in the response, and will allow us to capture their true preferences. We again follow best practices for reducing evaluation variation; employing the same strategy from Section 3.3.2, we sample responses to each prompt 100 times per LLM before analysing the resulting distribution of languages used.

For RQ3 Internal Consistency, the same process as in Section 3.3.2 was followed to explore the consistency between the LLMs’ programming language recommendations and the languages they actually use. LLMs were asked for an ordered list of recommendations using Prompt 2, repeated three times to mitigate randomness, with

the arithmetic mean of the ranks calculated to determine the final rank per LLM.

### 3.5 Result Analysis

**3.5.1 Response Data Extraction.** The LLMs used in the study typically format their responses as Markdown when they provide both code and NL, allowing us to perform an automatic process to extract the Python libraries or programming languages used. Although typical, it is not guaranteed that LLMs will respond this way; therefore, any responses from which code cannot be detected are saved separately and subjected to manual analysis for inclusion in our results.

In Markdown, code blocks are denoted by a triple backtick followed by the programming language name [18], therefore we can use regex matching to extract code blocks and programming languages. For library experiments, if a Python code block has been extracted, further regex matching is used to extract the imported libraries. There are two correct syntaxes for absolute imports in Python [59]: “import <library name>”; and “from <library name> import <object name>”. To verify correctness, the automatic extraction process was manually validated on 100 random samples from both the library and the language experiments.

**3.5.2 Rank Correlation Analysis.** We used Kendall’s  $\tau$  coefficient to measure rank correlation for RQ3: Recommendation Consistency and RQ4: LLM Similarities. Kendall’s  $\tau$  coefficient is a non-parametric measure of rank correlation that quantifies the ordinal association between two variables by comparing the number of concordant and discordant pairs in their rankings [37]. We use the Kendall’s  $\tau_b$  coefficient variant that allows ties in both variables.

For RQ3: Recommendation Consistency, we use the results of the experiments investigating libraries and languages chosen for project initialisation tasks. We compare the LLMs’ rankings when asked for recommendations, with the ranking provided by the percentage of tasks where each language or library is used.

For RQ4: LLM Similarities, we compare LLMs using results from all experiments. For each pair of LLMs, we rank the libraries or languages by the percentage of responses in which they are used, and calculate Kendall’s  $\tau$  coefficient between these rankings.

## 4 Results

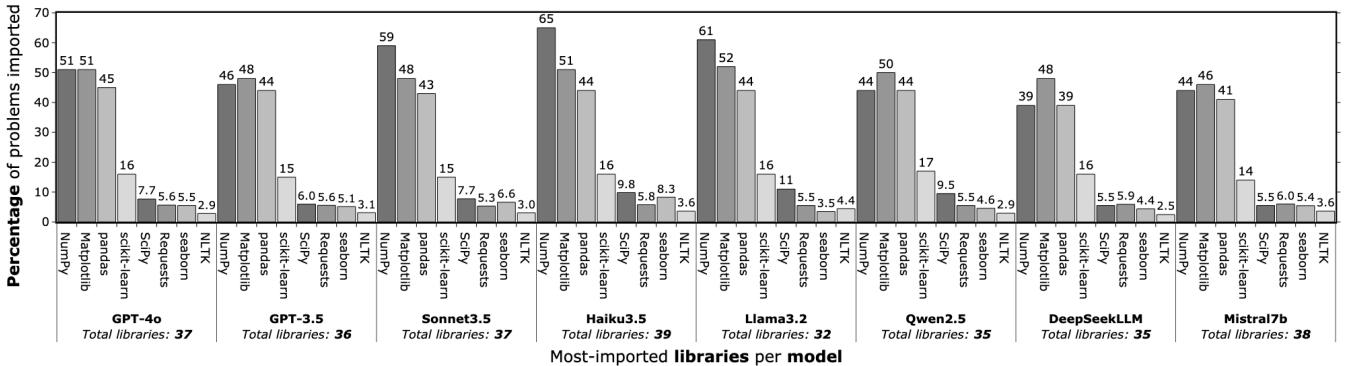
This section presents the results. Note that it is expected that the percentages in the results do not add up to 100% because LLMs were prompted multiple times per task, and could respond with different libraries or languages, or sometimes without code.

### 4.1 RQ1: Library Preference

This RQ aims to assess LLMs’ preferences when choosing Python libraries for code generation tasks.

**4.1.1 Preferences for benchmark tasks.** Figure 1 shows the libraries used by the LLMs when writing code for library-agnostic tasks from BigCodeBench. All LLMs show similar preferences across the problems, with a similar distribution in the most-used libraries.

In each case, the top 3 libraries are identical (NumPy, pandas and Matplotlib) and heavily preferred, with a gap of at least 22% between themselves and the fourth most-used. It is unsurprising



**Figure 1: Library Preferences, Benchmark Tasks (RQ1).** Libraries used by LLMs when generating code for tasks from the BigCodeBench dataset. For each LLM, the most-used libraries are given with the percentage of problems they were imported for, along with total unique libraries used. All libraries not shown are imported for less than 2% of problems.

**Table 2: Library Preferences, Project Initialisation Tasks (RQ1, RQ3).** Libraries used by LLMs when writing initial project code. For each LLM and project, the preferred libraries ( $l$ ) are given, along with the percentage ( $p$ ) of responses that used the library and the rank ( $r$ ) assigned to that library from the LLM. Libraries considered to provide core functionality are marked with a \* and are listed first.

Library Task	GPT-4o			GPT-3.5			Sonnet3.5			Haiku3.5			Llama3.2			Qwen2.5			DeepSeek			Mistral7b			
	$l$	$p$	$r$	$l$	$p$	$r$	$l$	$p$	$r$	$l$	$p$	$r$	$l$	$p$	$r$	$l$	$p$	$r$	$l$	$p$	$r$	$l$	$p$	$r$	
Database ORM	SQLAlchemy*	100%	#1	SQLAlchemy*	100%	#1	SQLAlchemy*	100%	#1	SQLAlchemy*	100%	#1	SQLAlchemy*	100%	#1	SQLAlchemy*	99%	#1	SQLAlchemy*	100%	#1	SQLAlchemy*	100%	#1	
	models	11%	-	-	-	-	-	-	-	Pydantic	3%	-	models	96%	-	TensorFlow	1%	-	-	-	-	-	-	-	
	database	3%	-	-	-	-	-	-	-	dotenv	3%	-	db.py	1%	-	-	-	-	-	-	-	-	-	-	
Deep-learning	Total used	-	-	-	-	-	-	-	-	Total used	5	-	-	-	-	-	-	-	-	-	-	-	-	-	
	TensorFlow*	97%	#1	TensorFlow*	85%	#1	PyTorch*	100%	#2	scikit-learn*	82%	#3	TensorFlow*	100%	#3	TensorFlow*	100%	#1	TensorFlow*	100%	#1	Keras*	100%	#3	
	scikit-learn*	4%	#10	PyTorch*	7%	#2	TorchVision*	93%	#2	scikit-learn*	5%	#5	PyTorch*	25%	#2	NumPy	100%	#1	scikit-learn*	91%	#7	scikit-learn*	44%	#4	
	PyTorch*	3%	#2	TorchVision*	6%	#3	Celas*	1%	#4	TorchVision*	96%	#8	Matplotlib	91%	#8	Matplotlib	91%	#8	TensorFlow	1%	#1	NumPy	100%	-	
	TorchVision*	2%	-	scikit-learn*	2%	#4	Matplotlib	96%	#8	TorchVision*	14%	-	NumPy	11%	#4	NumPy	86%	-	NumPy	100%	#1	-	-	-	
	NumPy	76%	-	TorchVision*	1%	-	NumPy	11%	#4	Matplotlib	26%	-	Matplotlib	-	-	Matplotlib	-	-	Matplotlib	-	-	-	-	-	
	Matplotlib	64%	-	NumPy	47%	#5	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
Distributed computing	Dask*	73%	#1	MP4ipy*	33%	#5	Dask*	24%	#2	Dask*	89%	#1	Dask*	100%	#1	Ray*	100%	#2	-	-	-	Dask*	100%	#1	
	MP4ipy*	5%	#8	Dask*	31%	#1	Ray*	20%	#1	Ray*	4%	#2	NumPy	55%	#9	-	-	-	-	-	-	-	NumPy	100%	-
	Ray*	2%	#2	NumPy	6%	-	Celery*	1%	#4	Celery*	2%	#4	-	-	-	-	-	-	-	-	-	-	-	-	
	NumPy	28%	-	Joblib	1%	-	Redis	49%	-	Redis	71%	-	-	-	-	-	-	-	-	-	-	-	-	-	
	Total used	8	-	-	-	-	Total used	8	-	Total used	6	-	-	-	-	-	-	-	-	-	-	-	-	-	
Web-scraping	BS4*	100%	#2	BS4*	100%	#1	BS4*	100%	#2	BS4*	100%	#3	BS4*	100%	#1	BS4*	100%	#1	BS4*	100%	#2	BS4*	100%	#2	
	Requests*	100%	#1	Requests*	100%	#4	Requests*	100%	#1	Requests*	100%	#3	Requests*	100%	#1	pandas	100%	#8	Requests*	100%	#2	Requests*	100%	#1	
	pandas	27%	#5	pandas	13%	#3	pandas	100%	#5	pandas	100%	#6	pandas	100%	#5	-	-	-	pandas	100%	#3	-	-		
	-	-	-	-	-	-	-	-	-	Total used	6	-	-	-	-	-	-	-	-	-	-	-	-		
Web-server	Flask*	98%	#1	Flask*	100%	#1	FastAPI*	48%	#1	Flask*	69%	#2	Flask*	100%	#1	Flask*	100%	#1	Flask*	100%	#1	Flask*	100%	#1	
	FastAPI*	2%	#2	-	-	-	Flask*	39%	#2	FastAPI*	23%	#1	Flask-SQLA	100%	-	-	-	-	-	-	-	-	-	-	-
	Flask-REST	19%	-	-	-	-	Pydantic	48%	#12	Flask-Cors	52%	-	-	-	-	-	-	-	-	-	-	-	-	-	
	Flask-Cors	15%	-	-	-	-	Uvicorn	48%	#14	Flask-SQLA	31%	-	-	-	-	-	-	-	-	-	-	-	-	-	
	Total used	7	-	-	-	-	Flask-Cors	14%	-	Total used	12	-	-	-	-	-	-	-	-	-	-	-	-	-	

Note some library names have been shortened in the table: Flask-REST is Flask-RESTful, Flask-SQLA is Flask-SQLAlchemy, BS4 is BeautifulSoup.

that these libraries are the most-preferred due to a large number of the contained problems being in the computation and visualisation domains, but they are often imported when they are not necessary. For example, the most used library, NumPy, is imported in responses to 192/301 tasks where it is not a part of the ground truth solution - showing a preference to use it over alternative options for any task.

The range of Python libraries that are available for download is expansive - in January 2025, over 7000 different open-source libraries surpassed 100,000 downloads [73] (this is still just a fraction of those available). However, all models use similarly low numbers of distinct libraries, ranging from 32 (Llama3.2) to 39 (Haiku3.5). Although not all of the available libraries are suitable, using less than 40 over a wide range of tasks shows a severe lack of diversity.

**Case Analysis:** It is clear from the results that LLMs show a preference for established libraries. To understand the extent of this bias, we perform a case analysis on libraries used in different domains, to see if there are any high-quality libraries neglected by LLMs. In particular, we use GitHub star growth over time as an objective heuristic for quality, and look for alternative libraries that show signs of higher quality but have lower usage rates.

Figure 2 shows the GitHub star growth rates for competing libraries. We observe that high-quality alternatives exist in the computation and visualisation domains. For computation problems, pandas (2010) is heavily favoured by LLMs, being used for 58% of tasks. Polars (2020) is a newer library with similar functionality (including usage of DataFrame objects) and twice as fast GitHub star growth, but is not used by LLMs at all. Similarly, Matplotlib (2011) and seaborn (2012) are heavily used by LLMs for visualisation problems, for 57% and 17% of tasks respectively. Plotly (2013) has additional functionality and signs of higher potential, but is used for only one problem and by three LLMs.

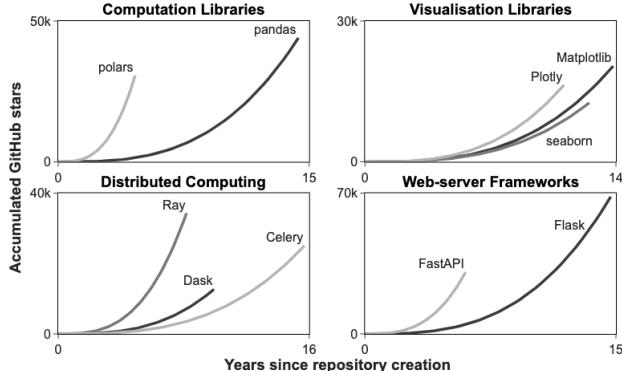
**4.1.2 Preferences for project initialisation tasks.** Table 2 shows the libraries imported when LLMs wrote the initial code for various Python project descriptions.

We observe that DeepSeek and Mistral7b have the least diversity, using the same libraries in every response. Llama3.2 and Qwen2.5 are only slightly better, using the same core libraries in every response and only varying the supporting libraries. The closed-source GPT and Claude LLMs showed much more varied preferences, with

**Table 3: Language Preferences, Benchmark Tasks (RQ2).** Programming languages used by LLMs when writing code for benchmark tasks. Column “Dataset” shows the name of the benchmark, as well as the programming languages for which the ground truth solutions are provided in each benchmark. For each LLM and dataset, the preferred languages ( $l$ ) are given, along with the percentage ( $p$ ) of dataset problems that have a response using that language, and the total count of different languages used (when necessary). The most-used language in each case is in bold.

Dataset	GPT-4o		GPT-3.5		Sonnet3.5		Haiku3.5		Llama3.2		Qwen2.5		DeepSeek		Mistral7b	
	$l$	$p$	$l$	$p$												
Multi-HumanEval (10 languages*)	<b>Python</b>	100.0%	<b>Python</b>	100.0%	<b>Python</b>	98.8%	<b>Python</b>	99.4%	<b>Python</b>	100.0%	<b>Python</b>	100.0%	<b>Python</b>	100.0%	<b>Python</b>	100.0%
-	-	-	-	-	JavaScript	1.2%	Total used	4	-	-	-	-	-	-	Total used	3
MBXP (10 languages*)	<b>Python</b>	100.0%	<b>Python</b>	100.0%	<b>Python</b>	97.0%	<b>Python</b>	99.5%	<b>Python</b>	100.0%	<b>Python</b>	100.0%	<b>Python</b>	98.5%	<b>Python</b>	99.0%
-	-	-	JavaScript	1.0%	JavaScript	7.5%	Java	2.5%	-	-	-	-	-	-	JavaScript	0.5%
-	-	-	-	-	Java	1.0%	C++	2.0%	-	-	-	-	-	-	-	-
-	-	-	-	-	C	0.5%	JavaScript	2.0%	-	-	-	-	-	-	-	-
AixBench (Java)	<b>Python</b>	75.2%	<b>Python</b>	64.3%	<b>Java</b>	53.5%	<b>Java</b>	59.7%	<b>Python</b>	75.2%	<b>Python</b>	68.2%	<b>Python</b>	78.3%	<b>Python</b>	62.0%
Java	27.1%	Java	38.8%	Python	38.8%	Python	55.0%	Java	18.6%	Java	26.4%	Java	18.6%	Java	24.8%	
JavaScript	10.9%	JavaScript	10.1%	JavaScript	18.6%	JavaScript	20.2%	C#	3.1%	JavaScript	3.1%	JavaScript	2.3%	JavaScript	14.0%	
Total used	7	Total used	7	Total used	9	Total used	12	Total used	8	Total used	5	C#	0.8%	Total used	9	
CoNaLa (Python)	<b>Python</b>	99.0%	<b>Python</b>	98.5%	<b>Python</b>	98.0%	<b>Python</b>	99.0%	<b>Python</b>	99.0%	<b>Python</b>	98.0%	<b>Python</b>	99.0%	<b>Python</b>	97.5%
JavaScript	6.0%	JavaScript	2.5%	JavaScript	12.5%	JavaScript	15.5%	JavaScript	1.0%	JavaScript	3.0%	JavaScript	1.5%	JavaScript	4.0%	
Java	4.5%	Java	1.5%	Java	3.5%	Java	13.0%	-	-	Java	2.0%	Java	1.0%	Java	2.5%	
Total used	7	Total used	9	Total used	11	Total used	11	-	-	Total used	5	C++	0.5%	Total used	5	
APPS (Python)	<b>Python</b>	99.5%	<b>Python</b>	99.5%	<b>Python</b>	93.5%	<b>Python</b>	93.5%	<b>Python</b>	98.0%	<b>Python</b>	98.0%	<b>Python</b>	98.5%	<b>Python</b>	98.5%
JavaScript	1.0%	JavaScript	1.5%	C++	10.0%	JavaScript	7.5%	C++	1.5%	Ruby	0.5%	Ruby	0.5%	Ruby	0.5%	
Ruby	0.5%	Ruby	0.5%	JavaScript	5.5%	JavaScript	4.5%	JavaScript	0.5%	Java	0.5%	Java	0.5%	Java	0.5%	
R	0.5%	Java	0.5%	Total used	5	Total used	6	C	0.5%	R	0.5%	-	-	R	0.5%	
CodeContests (C++, Java, Python)	<b>Python</b>	100.0%	<b>Python</b>	98.5%	<b>Python</b>	96.5%	<b>Python</b>	94.0%	<b>Python</b>	97.5%	<b>Python</b>	97.0%	<b>Python</b>	96.5%	<b>Python</b>	97.5%
-	-	C++	2.0%	C++	19.5%	C++	15.0%	C++	2.0%	C++	6.0%	-	-	C++	2.0%	
-	-	-	-	Java	0.5%	-	-	JavaScript	0.5%	-	-	-	-	JavaScript	0.5%	
All datasets	<b>Python</b>	96.8%	<b>Python</b>	95.1%	<b>Python</b>	89.8%	<b>Python</b>	92.0%	<b>Python</b>	96.1%	<b>Python</b>	94.9%	<b>Python</b>	96.1%	<b>Python</b>	94.1%
Java	4.0%	Java	5.0%	Java	7.3%	Java	10.3%	Java	2.3%	Java	3.5%	Java	2.4%	Java	3.5%	
JavaScript	2.6%	JavaScript	2.1%	JavaScript	7.1%	C++	7.8%	C++	0.6%	JavaScript	1.7%	JavaScript	0.6%	JavaScript	2.6%	
C#	0.5%	C++	0.6%	C++	6.2%	JavaScript	6.6%	JavaScript	0.5%	JavaScript	0.9%	C#	0.1%	C#	0.6%	
Total used	11	Total used	8	Total used	13	Total used	14	Total used	9	Total used	8	Total used	6	Total used	13	

\* Multi-HumanEval and MBXP benchmarks contain solutions in the following languages: C#, Go, Java, JavaScript, Kotlin, Perl, PHP, Python, Ruby, Scala, Swift, TypeScript.



**Figure 2: Case Analysis.** GitHub growth statistics for libraries studied in case analysis. For each library, the accumulated GitHub stars are plotted over its lifetime (data from GitHub, December 2024). The top/bottom two graphs show case analysis for benchmark/project initialisation tasks.

different core libraries in 3/5 tasks (“Deep learning”, “Distributed computing” and “Web-server”).

Notably, there are multiple instances where a library is imported despite being unnecessary for the given task: Qwen2.5 imported TensorFlow for “Database ORM”; and multiple models imported NumPy and pandas for “Distributed computing”. This overuse of data-science libraries is consistent with the results for benchmark tasks and reinforces the bias towards their usage. These preferences could be correlated with the recent growth in Python usage due to the growth in AI and data-science-based code [69].

**Case Analysis:** The results once again show a preference for more established libraries. We perform a further case analysis (as described in Section 4.1.1), to show how LLMs rarely use high-quality alternatives to their preferred libraries. Figure The results are shown in Figure 2, where we observe that there are high-quality alternatives for the tasks “Web-server” and “Distributed computing”. For the “Web-server” task the potential Python frameworks are plentiful, Flask (2010) is a traditional option and used by the LLMs in 88% of responses. FastAPI (2018) is much newer and has grown its GitHub stars almost three times as quickly - yet it was used in only 9% of responses, and by only three LLMs. For the “Distributed computing” task, Dask (2015) is the library preferred by most LLMs, used in 52% of responses. Ray (2016) and Celery (2009) both have similar functionality along with more stars and faster growth, but they were used minimally (in 16% and 0.4% of responses, respectively), and by only four LLMs.

**Answer to RQ1:** LLMs strongly favour older, well-established libraries; often overlooking newer, highly-rated alternatives, as shown in our four case studies, where usage rate differences range from 36% to 79%. LLMs also heavily favour data-science libraries, even when they are not necessary: for 48% of the benchmark tasks where NumPy was imported, its usage differs from the ground-truth solutions.

## 4.2 RQ2: Language Preference

This RQ aims to assess LLMs' preferences when choosing programming languages for code-generation tasks.

**Table 4: Language Preferences, Project Initialisation Tasks (RQ2, RQ3).** Programming languages chosen by LLMs for initial project code in scenarios where Python is not the most suitable option. For each LLM and project, each language (*l*) used by the LLM is given, along with the percentage (*p*) of responses that used the language and the rank (*r*) assigned to that language from the LLM. The most-used language in each case is in bold.

Language Task	GPT-4o			GPT-3.5			Sonnet3.5			Haiku3.5			Llama3.2			Qwen2.5			DeepSeek			Mistral7b		
	<i>l</i>	<i>p</i>	<i>r</i>	<i>l</i>	<i>p</i>	<i>r</i>	<i>l</i>	<i>p</i>	<i>r</i>	<i>l</i>	<i>p</i>	<i>r</i>	<i>l</i>	<i>p</i>	<i>r</i>	<i>l</i>	<i>p</i>	<i>r</i>	<i>l</i>	<i>p</i>	<i>r</i>	<i>l</i>	<i>p</i>	<i>r</i>
Concurrency	JavaScript	73% #3		JavaScript	99% #5		JavaScript	100%	#5	Python	100%	-	Go	100%	#2	Python	100%	#12	JavaScript	94% #5		JavaScript	100%	-
	Python	28% #8		Python	1% #6	-	-	-	-	-	-	-	-	-	-	Python	6% #4	-	-	-	-	-	-	-
	Go	2% #1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
Graphical interface	JavaScript	72% #1		JavaScript	64% #1		Python	99% #3		Python	81% #3		Python	100% #5		JavaScript	100%	-	Python	100% #2		Dart	81% #2	
	Dart	38% #5		C++	29% -		Dart	2% #1		JavaScript	19% #2	-	-	-	-	Dart	100% #6	-	-	-	-	JavaScript	48% #1	
	Python	14% #3		Python	11% #2	-	-	-	-	TypeScript	2%	-	-	-	-	-	-	-	-	-	-	-	-	-
Low-latency	C++	3% #12	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	Python	100% #5		Python	96% #3		Java	63% #2		Python	100% #5		Python	100% #5		Python	100% #5		Python	100% #3		Python	53% #3	
	-	-		Java	2% #2		C++	30% #1		-	-	-	-	-	-	-	-	-	-	-	none	47%	-	-
Parallel processing	-	-		-	JavaScript	2% -	Python	7% -	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	-	-		-	C++	1% #1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	Python	99% #5		Python	80% #5		Python	96% -		Python	100%	-	Python	100% #6		C++	100% #1		Python	100% #3		Python	100% #6	
System-level programming	C++	1% #1		C++	11% #1		C++	4% #2		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	-	-		Java	9% #4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
System-level programming	Python	81% #5		C	63% #1		C	50% #1		Python	100%	-	Python	100% #6		C	100% #1		C++	100% #2		C++	51% #1	
	C	23% #1		Python	37% #5		Python	50% #6	-	-	-	-	-	-	-	-	-	-	-	-	-	C	49% #1	

**4.2.1 Preferences for benchmark tasks.** Table 3 shows the programming languages used by LLMs when writing code for language-agnostic tasks from benchmarks datasets. We observe a clear preference towards using Python across all LLMs. For each dataset apart from AixBench, LLMs used Python for at least 93.5% of tasks, the highest percentage of responses in a non-Python language was considerably lower at 19.5%. On benchmarks intentionally chosen for their multilingual solution ecosystems (e.g., CodeContests: C++, Java, Python), the models disproportionately default to Python, suggesting the bias is not merely an artifact of single-language contamination.

The “Real-world” datasets (AixBench and CoNaLa) had up to 12 different languages used in the responses, significantly more than the other datasets - which had at most 5 different languages used, showing that the LLMs opted for more diverse languages in real-world situations. In general, the range of programming languages that LLMs choose to use is very limited.

**4.2.2 Preferences for project initialisation tasks.** Table 4 shows the languages used when LLMs wrote the initial code for various project descriptions where highly-performant languages—such as C, C++ and Rust [14]—are considered optimal.

Even for project initialisation tasks where Python is the sub-optimal choice, there is still a clear preference of LLMs for using it. Python is the preferred language in 23/40 cases even though it was given a low rank each time. Haiku3.5, Llama3.2 and DeepSeek were the worst offenders here, heavily using Python for 4/5 of the tasks. Javascript is the next most used, but still only the preferred choice in 8/40 instances. This result is consistent with the preference towards Python when solving benchmark tasks. We suspect that the trend towards favouring Python and Javascript stems from the LLMs being trained to prefer more user-friendly languages.

LLama3.2 and Qwen2.5 are the LLMs with the least diversity in their preferences, using the same programming languages in every response. Haiku3.5 and DeepSeek only show slightly more variability, having multiple options for a single project description. The GPT models have the most variability, using a range of languages across their responses to each task, but they still only use at most four different languages for a single project.

**4.2.3 Why is Python so dominant?** The bias in an LLM is usually attributed to the training / pre-training stages, where the knowledge of the LLM begins to mirror the underlying data corpora [25]. Therefore, we might expect the Python bias to arise from an abundance of Python in the training data. The precise code corpora used to train LLMs typically remain undisclosed [6], but there is strong evidence that they are heavily based on large-scale GitHub archives [52].

Although Python has seen a sudden surge in popularity, it has not historically been the most used programming language; only in 2024 did it record the highest number of GitHub contributors [69], and begin to lead the TIOBE Index [31]. This would suggest that an even sampling of GitHub would yield a more balanced language distribution than the heavy Python preference we observe. Therefore, we posit that it is likely that LLM creators are favouring Python when constructing their training datasets—or encouraging Python in the post-training / fine-tuning stages—to optimise performance on popular benchmarks, most of which are Python-centric.

**Answer to RQ2:** All LLMs demonstrate a strong preference for Python as their default programming language. Specifically, for benchmark tasks, LLMs use Python for **90–97%** of all problems. When generating initial project code where Python is a suboptimal programming language choice, Python usage remains dominant, appearing as the most-used language in **58%** of cases.

### 4.3 RQ3: Recommendation Consistency

This RQ aims to assess whether the programming languages or libraries LLMs recommend for a task are consistent with what they actually use. To answer this RQ, we ask the LLMs to make recommendations for each project initialisation task, and use Kendall’s  $\tau$  coefficient to compare this ranking to what they actually use. The results are shown in Table 5.

From the table, we observe that there is extremely low consistency between the technologies recommended by LLMs and the LLMs’ final choices when writing code. For library tasks, only “Distributed computing” and “Web-scraper” tasks have statistically significant correlations for multiple LLMs, probably due to the fact

**Table 5: Recommendation Consistency, Project Initialisation Tasks (RQ3).** Kendall's  $\tau$  correlation between LLMs' recommendations and their actual choices. Only statistically significant correlations are given ( $p$ -values  $< 0.05$ ) - there were none for language tasks. Values near 1.0/-1.0 indicate strong agreement/disagreement.

	GPT-4o	GPT-3.5	Sonnet3.5	Haiku3.5	Llama3.2	Qwen2.5	DeepSeek	Mistral7b
Library Tasks	-	-	-	-	-	-	-	-
Database ORM	-	-	-	-	-	-	-	-
Deep-learning	-	0.57	-	-	-	-	-	-
Distributed computing	0.41	-	0.60	0.60	-	-	-	-
Web-scrapers	0.49	-	0.42	-	-	0.55	0.60	0.73
Web-server	-	-	-	-	-	-	-	-

that there are more widely accepted and commonly used libraries in these domains. For language tasks, none of the correlations show statistical significance.

This inconsistency suggests that LLMs lack a universal knowledge representation of “which language best fits these requirements”, which they can apply to both code and NL answers. Instead, for both tasks, they generate text token by token, predicting the most likely next word based on patterns in their training data. This means that when actually generating code, LLMs can default to the programming languages most frequently seen in their training data [35], regardless of suitability. The inconsistency we observe highlights potential issues in the decision-making process of LLMs, raising concerns about their reliability in guiding developers toward optimal library and language choices. In Section 5.2.2, we discuss whether the use of prompt engineering strategies to enhance reasoning can mitigate this inconsistency.

**Answer to RQ3:** There is low consistency between the libraries and programming languages LLMs recommend and those that they actually adopt, with the majority of correlations not statistically significant. For project initialisation tasks, LLMs contradict their own language recommendations 83% of the time.

#### 4.4 RQ4: LLM Similarities

This RQ looks at whether coding preferences are similar across different LLMs. To answer this RQ, Kendall's  $\tau$  coefficient is calculated for each previous experiment, correlating the preferences of each pair of LLMs. Table 6 shows the results for the benchmark tasks.

From the table, we can observe that there is a median coefficient of 0.54 (range 0.40–0.67) across all LLM pairs for languages used for benchmark tasks, with only three results not statistically significant. For libraries, coefficients for all LLM pairs are statistically significant, with a median coefficient of 0.53 (range 0.40–0.65). This indicates that all LLMs have similar preferences when solving benchmark tasks. For project initialisation tasks, the correlation between preferences is much less clear. Only 16% of coefficients between LLMs have statistical significance, with 13% undefined due to an exact match of a single choice of technology.

We assume that the similarity across models reflects shared training data sources—such as public GitHub repositories [52]—while differences arise from their specific data and training variations.

These differences appear to be more pronounced for the open-ended project initialisation tasks, hence the lack of significance in those coefficients.

**Answer to RQ4:** Different LLMs have a notable alignment on their preferences for programming languages and libraries when solving benchmark tasks, with an overall median  $\tau$  coefficient of **0.53** (range 0.40–0.67). There is no clear evidence of alignment on language and library preferences for project initialisation tasks – only **16%** of  $\tau$  coefficients have statistical significance.

**Table 6: LLM Similarities, Benchmark Tasks (RQ4).** Kendall's  $\tau$  correlation between different LLMs', comparing the languages and libraries they use for benchmark tasks. Only statistically significant correlations are given ( $p$ -values  $< 0.05$ ). Values near 1.0/-1.0 indicate strong agreement/disagreement.

	GPT-4o	GPT-3.5	Sonnet3.5	Haiku3.5	Llama3.2	Qwen2.5	DeepSeek	Mistral7b	Language Tasks
Library Tasks	-	0.66	0.58	0.49	0.44	-	0.45	0.52	
GPT-4o	-	0.58	-	0.58	0.55	0.57	-	0.62	0.67
GPT-3.5	0.58	-	0.49	-	0.50	0.60	-	0.47	0.51
Sonnet3.5	0.49	0.62	-	0.50	0.60	-	0.47	-	
Haiku3.5	0.50	0.65	0.60	-	0.56	0.40	0.57	0.43	
Llama3.2	0.65	0.57	0.50	0.55	-	0.51	0.53	0.65	
Qwen2.5	0.57	0.62	0.54	0.61	0.51	-	-	0.44	
DeepSeek	0.51	0.62	0.44	0.52	0.55	0.47	-	0.65	
Mistral7b	0.44	0.55	0.40	0.44	0.49	0.43	0.50	-	

## 5 Discussion

This section further discusses our findings and provides extended analysis on techniques that may impact LLM preferences.

### 5.1 Advantages & Disadvantages of LLMs' Code Preferences

Our results show that LLMs exhibit strong preferences—and thus a lack of diversity—in their choice of programming languages and libraries. Here we consider both the advantages and disadvantages of such preferences in cases where multiple choices, including the ones favoured by LLMs, are equally acceptable.

**Advantages:** LLMs often favor widely adopted technologies, which are typically mature, well-documented, and supported by large communities. This can enhance the user experience by making the generated code easier to understand, extend, and integrate. The use of established languages and libraries also increases the likelihood of generating robust code, as these technologies have usually been thoroughly tested in real-world applications.

**Disadvantages:** However, these biases can lead to code homogeneity, limiting creative solutions, and preventing LLMs from using more specialised or optimal tools. This can be seen multiple times in this study, notably when Rust was not used a single time by any LLM for five different high-performance tasks (Table 4).

Bias toward mainstream technologies can also result in unnecessary or suboptimal dependencies, potentially causing performance issues. This is evidenced by the willingness of all models in this

study to import and use data science libraries in a variety of tasks where they are not strictly necessary (Table 2).

Furthermore, the lack of diversity in LLM preferences will likely also lead to inadequate discoverability of open source software. In Section 4.1 we present *four* use-cases where LLMs show a heavy preference towards older libraries over newer high-quality alternatives. If these biases persist, they risk reinforcing the dominance of a limited set of tools, stifling competition and innovation within the open-source ecosystem.

In part, this will severely limit the motivation to develop new open-source software (OSS). OSS is indispensable—evidenced by the reliance on Kubernetes and Docker for modern software development [56] and how 60% of websites are hosted using open-source web servers [74]—therefore, it is crucial for LLMs to contribute to, rather than hinder, OSS growth.

Additionally, whilst Python is not an inherently bad default programming language for LLMs, any preference as strong as the one observed for Python will be harmful. Such strong preferences in LLMs will encourage homogeneity across the coding landscape and discourage the learning of specialised programming languages that are vital to many industries.

## 5.2 Extended Analysis

**5.2.1 Varying Temperature.** Temperature is considered the creativity parameter [63] of LLMs, altering the variability and randomness in the responses. We conduct an initial investigation into how adjusting the temperature may allow LLMs to diversify their choice of programming language during project initialisation tasks. We repeat the experiment in Section 3.4.2 for GPT-4o, with varying temperatures. The OpenAI API accepts temperatures of 0.0-2.0, but larger temperatures can lead to unreliable response parsing [65]. Therefore, we use the following temperatures: 0.0, 0.5, 1.0 and 1.5.

The results are shown in Table 7, altering temperature has a clear but minimal impact. In each instance, a higher temperature led to the most-used language being used in fewer responses, along with a wider variety of languages. On average, the usage frequency of the most-used language will drop by 13.46% when increasing the temperature from 1.0 to 1.5. Interestingly, a temperature of 0.0 does not guarantee that the most used language will remain consistent.

**Table 7: Varying Temperature.** Languages used by GPT-4o for project initialisation tasks, when varying temperature. For each temperature ( $t$ ) and project, the preferred languages ( $l$ ) are given, along with the percentage ( $p$ ) of responses that used that language, and a count of the total used (if necessary). The most-used language in each case is in bold.

Language Task	$t = 0.0$	$t = 0.5$	$t = 1.0$	$t = 1.5$
	$l$	$l$	$l$	$l$
	$p$	$p$	$p$	$p$
Concurrency	<b>JavaScript</b> 73%	<b>JavaScript</b> 63%	<b>JavaScript</b> 68%	<b>JavaScript</b> 58%
	Python 27%	Python 36%	Python 32%	Python 39%
	Go 1%	Go 1%	Go 3%	Go 6
Graphical interface	<b>JavaScript</b> 84%	<b>JavaScript</b> 94%	<b>JavaScript</b> 76%	<b>JavaScript</b> 62%
	Dart 41%	Dart 50%	Dart 48%	Dart 30%
	<i>Total used</i> 4	<i>Total used</i> 4	<i>Total used</i> 5	<i>Total used</i> 6
	-	-	-	-
Low-latency	<b>Python</b> 100%	<b>Python</b> 100%	<b>Python</b> 100%	<b>Python</b> 65%
	C++ 1%	-	-	JavaScript 1%
Parallel processing	<b>Python</b> 99%	<b>Python</b> 100%	<b>Python</b> 99%	<b>Python</b> 97%
	C++ 1%	-	C++ 1%	C++ 4%
System-level programming	<b>Python</b> 87%	<b>Python</b> 93%	<b>Python</b> 89%	<b>Python</b> 83%
	C 16%	C 9%	C 17%	C 17%
	-	-	-	Go 1%

**5.2.2 Reasoning via Prompt Engineering.** Chain-of-thought (CoT) prompting has been shown to elicit reasoning in LLMs [80]; it may enable more optimal programming language choices when coding. Therefore, we conduct an initial investigation into whether these strategies can improve the consistency between an LLM’s programming language recommendations and usage. We repeat the experiment in Section 3.4.2 for GPT-4o, asking for project initialisation code without specifying the programming language. We append the original prompts with one of the following, either text that has been shown to elicit zero-shot reasoning behaviour, or with the directive to first rank the languages:

- (1) **Step-by-step:** “Think step by step about which coding language you should use and why.” [39]
- (2) **Double-check:** “Double check the reasoning for your coding language choice before writing code.” [12]
- (3) **First-list:** “First, list in order, the best coding languages for the task, then use this list to inform your language choice.”

The results are shown in Table 8. The consistency of the recommendations has improved, with the top-ranked programming language now being the most used in 11/15 instances. Asking the LLM to first rank suitable languages in context shows the best alignment between its original ranking and the languages used, although it was still not perfect, showing potential for variability in the recommendations. Recommendations for the “Parallel processing” task are still inconsistent across all reasoning prompts, and the LLM now strongly favours Rust, even though it ranked C++ top.

We also tried to increase the diversity of programming languages by asking the LLMs to “provide another option” after their initial response. Typically, they would only vary the language for project initialisation tasks and not for benchmark tasks. The results can be found in our GitHub repository.

**Table 8: Reasoning via Prompt Engineering.** Languages used for project initialisation tasks by GPT-4o when inducing reasoning via prompt engineering. The languages used are given, with the rank assigned to the language by the LLM, and the percentage of responses that used the language for each prompt style. The most-used language in each case has its percentage is in bold.

Language Tasks	Language	Rank ↓	Step-by-step	Double-check	First-list
Concurrency	Go	#1	77%	47%	<b>98%</b>
	Rust	#2	0%	0%	2%
	JavaScript	#3	23%	53%	0%
Graphical interface	JavaScript	#1	84%	50%	<b>96%</b>
	Python	#3	9%	33%	1%
	Dart	#5	8%	21%	3%
Low-latency	C++	#1	<b>88%</b>	66%	<b>100%</b>
	Rust	#3	2%	12%	0%
	Go	#4	5%	15%	0%
	Python	#5	9%	7%	0%
	-	-	-	-	-
Parallel processing	C++	#1	20%	8%	37%
	Rust	#2	54%	<b>61%</b>	<b>61%</b>
	Go	#3	20%	17%	1%
	Python	#5	6%	15%	1%
	-	-	-	-	-
System-level	C	#1	92%	96%	<b>100%</b>
	Rust	#3	8%	5%	0%
	Python	#5	4%	1%	0%

## 6 Threats to Validity

**Internal validity:** Prompt realism and analysis. The threats to internal validity lie in our automatic data extraction and prompt design. To alleviate the first threat, we have both unit tested the code responsible, and manually verified the extraction process on 100 random samples from language and library experiments. To reduce

the second threat, we carefully designed our prompts based on existing studies that generate code with LLMs [88], and investigate how users interact with LLMs [57].

**External validity:** Data and LLMs. The threats to external validity lie in contamination of the datasets used and in the LLMs. We reduce the first threat by carefully selecting benchmark datasets to minimise data contamination, opting for datasets released after the LLMs' knowledge cutoff, or datasets with a variety of programming languages to rule out the possibility that the observed preference is merely caused by contamination. Another threat is the non-deterministic nature of LLMs and their opaque updates [66], both of which can introduce variability between runs. We reduce this threat by repeating the experiments multiple times and specifying the exact LLM version to use.

## 7 Conclusion

This paper provides the first empirical study on LLM preferences for libraries and programming languages when generating code. Our findings demonstrate that LLMs consistently favour well-established libraries over high-quality alternatives; and exhibit a significant preference towards Python, even for tasks where there are more suitable choices. This lack of diversity risks homogenising the coding landscape and undermining open-source discoverability.

We call for targeted efforts to improve the diversity of LLM-generated code, via techniques such as prompt engineering, adaptive decoding, reinforcement learning, and retrieval-augmented generation; and investigation into other potential coding biases, such as paradigms, architectures, data structures, and typing conventions. By tackling these challenges, we can ensure that future LLMs enrich, rather than narrow, the software ecosystem.

## References

- [1] Mehmet Akhoroz and Caglar Yildirim. 2025. Conversational AI as a Coding Assistant: Understanding Programmers' Interactions with and Expectations from Large Language Models for Coding. (Mar. 14, 2025). arXiv: 2503.16508 [cs]. Retrieved July 18, 2025 from <http://arxiv.org/abs/2503.16508>. Pre-published.
- [2] Andrew Peng et al. 2023. GPT-3.5 Turbo fine-tuning and API updates. (Aug. 22, 2023). Retrieved Dec. 17, 2024 from <https://openai.com/index/gpt-3-5-turbo-fine-tuning-and-api-updates/>.
- [3] Anthropic. 2024. Claude 3 Model Card. (Oct. 22, 2024). Retrieved Jan. 22, 2025 from <https://assets.anthropic.com/m/61e7d27f8c8f5919/original/Claude-3-Model-Card.pdf>.
- [4] Ben Athiwaratkun et al. 2023. Multi-lingual evaluation of code generation models. In *Proc. ICLR*. arXiv: 2210.14868. doi: 10.48550/arXiv.2210.14868.
- [5] Jacob Austin et al. 2021. Program Synthesis with Large Language Models. (Aug. 16, 2021). arXiv: 2108.07732. Retrieved Oct. 18, 2024 from <http://arxiv.org/abs/2108.07732>. Pre-published.
- [6] Rishi Bommasani et al. 2023. The Foundation Model Transparency Index. (Oct. 19, 2023). arXiv: 2310.12941 [cs]. Retrieved May 29, 2025 from <http://arxiv.org/abs/2310.12941>. Pre-published.
- [7] William Bugden and Ayman Alahmar. 2022. The Safety and Performance of Prominent Programming Languages. *International Journal of Software Engineering and Knowledge Engineering*, 32, 05, (May 2022), 713–744. doi: 10.1142/S0218194022500231.
- [8] Liguang Chen et al. 2024. A Survey on Evaluating Large Language Models in Code Generation Tasks. Version 1. *Journal of computer science and technology*. doi: 10.48550/ARXIV.2408.16498.
- [9] Mark Chen et al. 2021. Evaluating Large Language Models Trained on Code. (July 14, 2021). arXiv: 2107.03374. Retrieved Nov. 18, 2024 from <http://arxiv.org/abs/2107.03374>. Pre-published.
- [10] Yuxing Cheng et al. 2025. A Survey on Data Contamination for Large Language Models. (June 5, 2025). arXiv: 2502.14425 [cs]. Retrieved July 7, 2025 from <http://arxiv.org/abs/2502.14425>. Pre-published.
- [11] Rudrajit Choudhuri et al. 2024. What Guides Our Choices? Modeling Developers' Trust and Behavioral Intentions Towards GenAI. (Dec. 2, 2024). arXiv: 2409.04099 [cs]. Retrieved Dec. 17, 2024 from <http://arxiv.org/abs/2409.04099>. Pre-published.
- [12] Jishnu Ray Chowdhury and Cornelia Caragea. 2025. Zero-Shot Verification-guided Chain of Thoughts. (Jan. 21, 2025). arXiv: 2501.13122 [cs]. Retrieved July 15, 2025 from <http://arxiv.org/abs/2501.13122>. Pre-published.
- [13] 2024. Competitive programming with AlphaCode. Google DeepMind. (Dec. 17, 2024). Retrieved Dec. 18, 2024 from <https://deepmind.google/discover/blog/competitive-programming-with-alphacode/>.
- [14] Manuel Costanzo et al. 2021. Performance vs programming effort between Rust and C on multicore architectures: case study in n-body. In *Proc. CLEI*, 1–10. doi: 10.1109/CLEI53233.2021.9640225.
- [15] DeepSeek-AI et al. 2024. DeepSeek LLM: Scaling Open-Source Language Models with Longtermism. (Jan. 5, 2024). arXiv: 2401.02954 [cs]. Retrieved Dec. 17, 2024 from <http://arxiv.org/abs/2401.02954>. Pre-published.
- [16] Kaustubh Dhole et al. 2023. NL-Augmenter: A Framework for Task-Sensitive Natural Language Augmentation. *Northern European Journal of Language Technology*, 9, Leon Derczynski, (Ed). doi: 10.3384/nejl.2000-1533.2023.4725.
- [17] Benedetta Donato et al. 2025. Studying How Configurations Impact Code Generation in LLMs: the Case of ChatGPT. In *The Proceedings of the 33rd IEEE/ACM International Conference on Program Comprehension*. arXiv, (Feb. 7, 2025). arXiv: 2502.17450 [cs]. doi: 10.48550/arXiv.2502.17450.
- [18] 2025. Extended Syntax | Markdown Guide. Retrieved Feb. 7, 2025 from <https://www.markdownguide.org/extended-syntax/>.
- [19] Carlo A. Furia et al. 2024. Towards Causal Analysis of Empirical Software Engineering Data: The Impact of Programming Languages on Coding Competitions. *ACM Transactions on Software Engineering and Methodology*, 33, 1, (Jan. 31, 2024), 1–35. arXiv: 2301.07524 [cs]. doi: 10.1145/3611667.
- [20] Isabel O. Gallegos et al. 2024. Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics*, 50, 3, (Sept. 1, 2024), 1097–1179. doi: 10.1162/coli\_a\_00524.
- [21] Yulia Gavrilova. 2023. Pros and Cons of Python. Pros and Cons of Python. (Oct. 31, 2023). Retrieved Dec. 20, 2024 from <https://serokell.io/blog/python-pros-and-cons>.
- [22] Aaron Grattafiori et al. 2024. The Llama 3 Herd of Models. (Nov. 23, 2024). arXiv: 2407.21783 [cs]. Retrieved Dec. 17, 2024 from <http://arxiv.org/abs/2407.21783>. Pre-published.
- [23] Sam Gross. [n. d.] PEP 703 – Making the Global Interpreter Lock Optional in CPython | peps.python.org. Python Enhancement Proposals (PEPs). Retrieved May 29, 2025 from <https://peps.python.org/pep-0703/>.
- [24] Chenchen Gu et al. 2025. Auditing Prompt Caching in Language Model APIs. (Feb. 11, 2025). arXiv: 2502.07776 [cs]. Retrieved May 18, 2025 from <http://arxiv.org/abs/2502.07776>. Pre-published.
- [25] Yufei Guo et al. 2024. Bias in Large Language Models: Origin, Evaluation, and Mitigation. Version 1. (Nov. 16, 2024). arXiv: 2411.10915 [cs]. Retrieved July 9, 2025 from <http://arxiv.org/abs/2411.10915>. Pre-published.
- [26] Huizi Hao et al. 2024. An Empirical Study on Developers Shared Conversations with ChatGPT in GitHub Pull Requests and Issues. Version 1. (Mar. 15, 2024). arXiv: 2403.10468 [cs]. Retrieved May 29, 2025 from <http://arxiv.org/abs/2403.10468>. Pre-published.
- [27] Yiyang Hao et al. 2022. AixBench: A Code Generation Benchmark Dataset. (July 21, 2022). arXiv: 2206.13179. Retrieved Oct. 29, 2024 from <http://arxiv.org/abs/2206.13179>. Pre-published.
- [28] Dan Hendrycks et al. 2021. Measuring coding challenge competence with APPS. In *Proc. NeurIPS Datasets and Benchmarks*. arXiv: 2105.09938. doi: 10.48550/arXiv.2105.09938.
- [29] Dong Huang et al. 2023. Bias Testing and Mitigation in LLM-based Code Generation. arXiv.org. (Sept. 3, 2023). Retrieved Oct. 8, 2024 from <https://arxiv.org/abs/2309.14345v3>.
- [30] Binyuan Hui et al. 2024. Qwen2.5-Coder Technical Report. (Nov. 12, 2024). arXiv: 2409.12186 [cs]. Retrieved Dec. 17, 2024 from <http://arxiv.org/abs/2409.12186>. Pre-published.
- [31] Paul Jansen. 2025. TIOBE Index. TIOBE. Retrieved July 15, 2025 from <https://www.tiobe.com/tiobe-index/>.
- [32] Albert Q. Jiang et al. 2023. Mistral 7B. (Oct. 10, 2023). arXiv: 2310.06825 [cs]. Retrieved Dec. 17, 2024 from <http://arxiv.org/abs/2310.06825>. Pre-published.
- [33] Juyong Jiang et al. 2024. A Survey on Large Language Models for Code Generation. (June 1, 2024). arXiv: 2406.00515. Retrieved Oct. 10, 2024 from <http://arxiv.org/abs/2406.00515>. Pre-published.
- [34] Carlos E. Jimenez et al. 2024. SWE-bench: can language models resolve real-world github issues? In *Proc. ICLR*. arXiv: 2310.06770. doi: 10.48550/arXiv.2310.06770.
- [35] Erik Jones and Jacob Steinhardt. 2022. Capturing failures of large language models via human cognitive biases. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NIPS '22)*. Curran Associates Inc., Red Hook, NY, USA, (Nov. 28, 2022), 11785–11799. ISBN: 978-1-71387-108-8.

- [36] Dawid Karczewski. 2021. Python vs C++: Technology Comparison. Retrieved Feb. 18, 2025 from <https://www.ideamotive.co/blog/python-vs-cpp-technology-comparison>.
- [37] Maurice G. Kendall and Jean Dickinson Gibbons. 1990. Rank Correlation Methods. (5th ed ed.). 1 online resource (vii, 260 pages) vols. Oxford University Press, New York, NY. <https://archive.org/details/rankcorrelationm000kend>.
- [38] Anjali Khurana et al. 2024. Why and When LLM-Based Assistants Can Go Wrong: Investigating the Effectiveness of Prompt-Based Interactions for Software Help-Seeking. (Mar. 18, 2024). arXiv: 2402.08030 [cs]. Retrieved July 18, 2025 from <http://arxiv.org/abs/2402.08030>.
- [39] Takeshi Kojima et al. 2022. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NIPS '22)*. Curran Associates Inc., Red Hook, NY, USA, (Nov 28, 2022), 22199–22213. ISBN: 978-1-71387-108-8. Retrieved Mar. 6, 2025 from.
- [40] Adrian Kuhn and Robert DeLine. 2012. On Designing Better Tools for Learning APIs. (June 2012). arXiv: 1402.1188 [cs]. Retrieved July 9, 2025 from <http://arxiv.org/abs/1402.1188>.
- [41] Yuhang Lai et al. 2022. DS-1000: a natural and reliable benchmark for data science code generation. In *Proc. ICML*. arXiv: 2211.11501. doi: 10.48550/arXiv.2211.11501.
- [42] Decrypt / Jose Antonio Lanz. 2023. Stability AI CEO: There Will Be No (Human) Programmers in Five Years. Decrypt. (July 3, 2023). Retrieved Nov. 13, 2024 from <https://decrypt.co/147191/no-human-programmers-five-years-ai-stability-ceo>.
- [43] Enrique Larios-Vargas et al. 2020. Selecting third-party libraries: the practitioners' perspective. In *Proc. ESEC/FSE*, 245–256. doi: 10.1145/3368089.3409711.
- [44] Jasmijn Latendresse et al. 2024. Is ChatGPT a Good Software Librarian? An Exploratory Study on the Use of ChatGPT for Software Library Recommendations. (Aug. 9, 2024). arXiv: 2408.05128 [cs]. Retrieved Dec. 19, 2024 from <http://arxiv.org/abs/2408.05128>. Pre-published.
- [45] Junlong Li et al. 2024. Dissecting Human and LLM Preferences. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Lun-Wei Ku et al., (Eds.) Association for Computational Linguistics, Bangkok, Thailand, (Aug. 2024), 1790–1811. doi: 10.18653/v1/2024.acl-long.99.
- [46] Yujia Li et al. 2022. Competition-level code generation with AlphaCode. *Science*, 378, 6624, 1092–1097. eprint: <https://www.science.org/doi/pdf/10.1126/science.abq1158>. doi: 10.1126/science.abq1158.
- [47] Jenny T. Liang et al. 2023. A Qualitative Study on the Implementation Design Decisions of Developers. In *45th [IEEE/ACM] International Conference on Software Engineering, [ICSE] 2023, Melbourne, Australia, May 14–20, 2023*. arXiv, (Jan. 24, 2023). arXiv: 2301.09789 [cs]. doi: 10.48550/arXiv.2301.09789.
- [48] Mingwei Liu et al. 2023. CodeGen4Libs: A Two-Stage Approach for Library-Oriented Code Generation. In *2023 38th IEEE/ACM International Conference on Automated Software Engineering (ASE)*. IEEE, Luxembourg, Luxembourg, (Sept. 11, 2023), 434–445. ISBN: 9798350329964. doi: 10.1109/ASE56229.2023.00159.
- [49] Yan Liu et al. 2023. Uncovering and Quantifying Social Biases in Code Generation. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*. arXiv, (May 24, 2023). arXiv: 2305.15377. doi: 10.48550/arXiv.2305.15377.
- [50] Zexiong Ma et al. 2024. Compositional API Recommendation for Library-Oriented Code Generation. In *Proceedings of the 32nd IEEE/ACM International Conference on Program Comprehension (ICPC '24)*. Association for Computing Machinery, New York, NY, USA, (June 13, 2024), 87–98. ISBN: 9798400705861. doi: 10.1145/3643916.3644403.
- [51] Lovish Madaan et al. 2024. Quantifying Variance in Evaluation Benchmarks. (June 14, 2024). arXiv: 2406.10229 [cs]. Retrieved July 7, 2025 from <http://arxiv.org/abs/2406.10229>. Pre-published.
- [52] Vahid Majdinasab et al. 2025. Trained Without My Consent: Detecting Code Inclusion in Language Models Trained on Code. *ACM Transactions on Software Engineering and Methodology*, 34, 4, (May 31, 2025), 1–46. arXiv: 2402.09299 [cs]. doi: 10.1145/3702980.
- [53] Ahmad Mohsin et al. 2024. Can We Trust Large Language Models Generated Code? A Framework for In-Context Learning, Security Patterns, and Code Evaluations Across Diverse LLMs. (June 18, 2024). arXiv: 2406.12513. Retrieved Oct. 24, 2024 from <http://arxiv.org/abs/2406.12513>. Pre-published.
- [54] Norman Mu et al. 2025. A Closer Look at System Prompt Robustness. (Feb. 15, 2025). arXiv: 2502.12197 [cs]. Retrieved July 6, 2025 from <http://arxiv.org/abs/2502.12197>. Pre-published.
- [55] Humza Naveed et al. 2024. A Comprehensive Overview of Large Language Models. (Oct. 17, 2024). arXiv: 2307.06435. Retrieved Nov. 18, 2024 from <http://arxiv.org/abs/2307.06435>. Pre-published.
- [56] 2024. Need of Dockers and Kubernetes in Modern Software Development - GeekMinds. (May 15, 2024). Retrieved Nov. 18, 2024 from <https://geekminds.com/need-of-dockers-and-kubernetes-in-modern-software-development/>.
- [57] Sydney Nguyen et al. 2024. How Beginning Programmers and Code LLMs (Mis)read Each Other. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. (May 11, 2024), 1–26. arXiv: 2401.15232 [cs]. doi: 10.1145/3613904.3642706.
- [58] Muhammed Nihal. 2024. The Race to Zero Latency: How to Optimize Code for High-Frequency Trading Quant Firms. Medium. (Aug. 13, 2024). Retrieved May 29, 2025 from <https://medium.com/@nihal.143/the-race-to-zero-latency-how-to-optimize-code-for-high-frequency-trading-quant-firms-362f828f9c16>.
- [59] Mbithie Nzomo. 2025. Absolute vs Relative Imports in Python – Real Python. Retrieved Feb. 5, 2025 from <https://realpython.com/absolute-vs-relative-python-n-imports/>.
- [60] OpenAI et al. 2024. GPT-4 Technical Report. (Mar. 4, 2024). arXiv: 2303.08774 [cs]. Retrieved Apr. 9, 2024 from <http://arxiv.org/abs/2303.08774>. Pre-published.
- [61] Arkil Patel et al. 2024. Evaluating In-Context Learning of Libraries for Code Generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. [NAACL] 2024, Mexico City, Mexico, June 16–21, 2024. arXiv, (Apr. 4, 2024). arXiv: 2311.09635. doi: 10.48550/arXiv.2311.09635.
- [62] Debalina Ghosh Paul et al. 2024. Benchmarks and Metrics for Evaluations of Code Generation: A Critical Review. In *2024 IEEE International Conference on Artificial Intelligence Testing (AITest)*. IEEE Computer Society, (July 1, 2024), 87–94. ISBN: 9798350365054. doi: 10.1109/AITest62860.2024.00019.
- [63] Max Peperkorn et al. 2024. Is Temperature the Creativity Parameter of Large Language Models? In *International Conference on Computational Creativity*. arXiv, (May 1, 2024). arXiv: 2405.00492 [cs]. doi: 10.48550/arXiv.2405.00492.
- [64] Sanka Rasnayaka et al. 2024. An empirical study on usage and perceptions of LLMs in a software engineering project. In *Proc. LLM4Code*, 111–118. doi: 10.1145/3643795.3648379.
- [65] Matthew Renze and Erhan Guven. 2024. The effect of sampling temperature on problem solving in large language models. In *Findings of EMNLP*. arXiv: 2402.05201. doi: 10.48550/arXiv.2402.05201.
- [66] June Sallou et al. 2024. Breaking the silence: the threats of using LLMs in software engineering. In *Proc. ICSE-NIER*, 102–106. doi: 10.1145/3639476.3639764.
- [67] Matthew Smith. 2025. AI Vibe Coding: Engineers' Secret to Fast Development - IEEE Spectrum. IEEE Spectrum. Retrieved May 29, 2025 from <https://spectrum.ieee.org/vibe-coding>.
- [68] Ian Somerville. 2016. *Software Engineering, Global Edition*. Pearson Education. ISBN: 978-1-292-09614-8.
- [69] GitHub Staff. 2024. Octoverse: AI leads Python to top language as the number of global developers surges. The GitHub Blog. (Oct. 29, 2024). Retrieved Feb. 10, 2025 from <https://github.blog/news-insights/octoverse-octoverse-2024/>.
- [70] Kyle Daigle Staff GitHub. 2024. Survey: The AI wave continues to grow on software development teams. The GitHub Blog. (Aug. 20, 2024). Retrieved Nov. 18, 2024 from <https://github.blog/news-insights/research/survey-ai-wave-grows/>.
- [71] Minao Hossain Tanzil et al. 2024. "How do people decide?": A Model for Software Library Selection. In *Proceedings of the 2024 IEEE/ACM 17th International Conference on Cooperative and Human Aspects of Software Engineering*. (Apr. 14, 2024), 1–12. arXiv: 2403.16245 [cs]. doi: 10.1145/3641822.3641865.
- [72] Weixi Tong and Tianyi Zhang. 2024. CodeJudge: evaluating code generation with large language models. In *Proc. EMNLP*, 20032–20051. doi: 10.18653/v1/2024.emnlp-main.1118.
- [73] 2025. Top PyPI Packages. Retrieved Feb. 8, 2025 from <https://hugovk.github.io/top-pypi-packages/>.
- [74] 2024. Usage Statistics and Market Share of Web Servers, November 2024. Retrieved Nov. 18, 2024 from [https://w3techs.com/technologies/overview/web\\_server](https://w3techs.com/technologies/overview/web_server).
- [75] Chaozheng Wang et al. 2024. A systematic evaluation of large code models in api suggestion: when, which, and how. In *Proc. ASE*. arXiv: 2409.13178. doi: 10.48550/arXiv.2409.13178.
- [76] Chaozheng Wang et al. 2024. Exploring Multi-Lingual Bias of Large Code Models in Code Generation. (Apr. 30, 2024). arXiv: 2404.19368 [cs]. Retrieved Feb. 17, 2025 from <http://arxiv.org/abs/2404.19368>. Pre-published.
- [77] Kaixin Wang et al. 2025. Software Development Life Cycle Perspective: A Survey of Benchmarks for Code Large Language Models and Agents. Version 2. (May 9, 2025). arXiv: 2505.05283 [cs]. Retrieved July 7, 2025 from <http://arxiv.org/abs/2505.05283>. Pre-published.
- [78] Ruotong Wang et al. 2024. Investigating and designing for trust in ai-powered code generation tools. In *Proc. ACM FAccT*. arXiv: 2305.11248. doi: 10.48550/arXiv.2305.11248.
- [79] Zhiruo Wang et al. 2023. Execution-based evaluation for open-domain code generation. In *Findings of EMNLP*. arXiv: 2212.10481. doi: 10.48550/arXiv.2212.10481.
- [80] Jason Wei et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*.

- Information Processing Systems* (NIPS '22). Curran Associates Inc., Red Hook, NY, USA, (Nov. 28, 2022), 24824–24837. ISBN: 978-1-71387-108-8. Retrieved Mar. 6, 2025 from.
- [81] Cheng Xu et al. 2024. Benchmark Data Contamination of Large Language Models: A Survey. (June 6, 2024). arXiv: 2406.04244 [cs]. Retrieved July 7, 2025 from <http://arxiv.org/abs/2406.04244>. Pre-published.
- [82] Ankit Yadav et al. 2024. PythonSaga: Redefining the Benchmark to Evaluate Code Generating LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. Yaser Al-Onaizan et al., (Eds.) Association for Computational Linguistics, Miami, Florida, USA, (Nov. 2024), 17113–17126. doi: 10.18653/v1/2024.findings-emnlp.996.
- [83] Pengcheng Yin et al. 2018. Learning to mine aligned code and natural language pairs from stack overflow. In *Proc. MSR*. arXiv: 1805.08949. doi: 10.48550/arXiv.1805.08949.
- [84] Daoguang Zan et al. 2022. CERT: continual pre-training on sketches for library-oriented code generation. In *Proc. IJCAI*, 2369–2375. doi: 10.24963/ijcai.2022/329.
- [85] Daoguang Zan et al. 2023. Private-Library-Oriented Code Generation with Large Language Models. (July 28, 2023). arXiv: 2307.15370. Retrieved Oct. 29, 2024 from <http://arxiv.org/abs/2307.15370>. Pre-published.
- [86] Ziyin Zhang et al. 2023. Unifying the Perspectives of NLP and Software Engineering: A Survey on Language Models for Code. Version 3. *Transactions on Machine Learning Research*, (Dec. 5, 2023). arXiv: 2311.07989 [cs]. doi: 10.48550/arXiv.2311.07989.
- [87] Xuekai Zhu et al. 2025. How to Synthesize Text Data without Model Collapse? (May 28, 2025). arXiv: 2412.14689 [cs]. Retrieved May 30, 2025 from <http://arxiv.org/abs/2412.14689>. Pre-published.
- [88] Terry Yu Zhuo et al. 2024. BigCodeBench: Benchmarking Code Generation with Diverse Function Calls and Complex Instructions. In *13th International Conference on Learning Representations (ICLR25)*. arXiv, (Oct. 7, 2024). arXiv: 2406.15877. doi: 10.48550/arXiv.2406.15877.
- [89] Daniel M. Ziegler et al. 2020. Fine-Tuning Language Models from Human Preferences. (Jan. 8, 2020). arXiv: 1909.08593 [cs]. Retrieved Feb. 18, 2025 from <http://arxiv.org/abs/1909.08593>. Pre-published.