

Is ChatGPT a Good Software Librarian? An Exploratory Study on the Use of ChatGPT for Software Library Recommendations

JASMINE LATENDRESSE, Concordia University, Canada

SAYEDHASSAN KHATOONABADI, Concordia University, Canada

AHMAD ABDELLATIF, University of Calgary, Canada

EMAD SHIHAB, Concordia University, Canada

Software libraries play a critical role in the functionality, efficiency, and maintainability of software systems. As developers increasingly rely on Large Language Models (LLMs) to streamline their coding processes, the effectiveness of these models in recommending appropriate libraries becomes crucial yet remains largely unexplored. In this paper, we assess the effectiveness of ChatGPT as a software librarian and identify areas for improvement. We conducted an empirical study using GPT-3.5 Turbo to generate Python code for 10,000 Stack Overflow questions. Our findings show that ChatGPT uses third-party libraries nearly 10% more often than human developers, favoring widely adopted and well-established options. However, 14.2% of the recommended libraries had restrictive copyleft licenses, which were not explicitly communicated by ChatGPT. Additionally, 6.5% of the libraries did not work out of the box, leading to potential developer confusion and wasted time. While ChatGPT can be an effective software librarian, it should be improved by providing more explicit information on maintainability metrics and licensing. We recommend that developers implement rigorous dependency management practices and double-check library licenses before integrating LLM-generated code into their projects.

Additional Key Words and Phrases: Software engineering, software ecosystems, large language models.

ACM Reference Format:

Jasmine Latendresse, SayedHassan Khatoonabadi, Ahmad Abdellatif, and Emad Shihab. 2024. Is ChatGPT a Good Software Librarian? An Exploratory Study on the Use of ChatGPT for Software Library Recommendations. 1, 1 (August 2024), 20 pages. <https://doi.org/10.1145/nnnnnnn.nnnnnnn>

1 INTRODUCTION

Modern software development thrives on code reuse. Open source libraries provide pre-written code blocks, significantly reducing development time and effort [1–3]. While libraries enhance functionality, they also introduce dependencies — interconnections between code components — that can lead to increased complexity and dependency management challenges [4–6].

One critical aspect of dependency management is library selection [7]. Choosing the right library impacts factors like code maintainability, performance, and security. Previous studies have explored how developers select libraries, and highlighted primarily ad-hoc processes based on past experiences, expert advice, and online resources [8, 9].

Authors' addresses: Jasmine Latendresse, Concordia University, Montreal, Canada, jasmine.latendresse@mail.concordia.ca; SayedHassan Khatoonabadi, Concordia University, Montreal, Canada, sayedhassan.khatoonabadi@concordia.ca; Ahmad Abdellatif, University of Calgary, Calgary, Canada, ahmad.abdellatif@ucalgary.ca; Emad Shihab, Concordia University, Montreal, Canada, emad.shihab@concordia.ca.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM XXXX-XXXX/2024/8-ART

<https://doi.org/10.1145/nnnnnnn.nnnnnnn>

The current, often informal, process of selecting libraries presents an opportunity for new tools to assist developers. Large Language Models (LLMs) are emerging as powerful programming assistants, capable of generating code and streamlining development workflows [10]. They have shown potential in improving developer productivity through functionalities like code completion and search [11, 12], but their reliability as "software librarians" (recommending appropriate libraries) remains unexplored.

Our work aims to explore a specific aspect of LLM functionality: their potential as software librarians for recommending libraries in code generation tasks. In this paper, we investigate the effectiveness of LLMs in such context by conducting an empirical study using ChatGPT [13] (specifically GPT 3.5 Turbo) to generate Python code for a large set of coding problems derived from real-world Stack Overflow [14] questions. This allows us to assess LLM performance in a context that mimics actual developer usage. By analyzing the libraries recommended by ChatGPT, we aim to answer the following research questions:

RQ1 What are the characteristics of the software libraries recommended by ChatGPT?

Our results show that ChatGPT uses third-party libraries almost 10% more than human developers. It also tends to use stable and well-established libraries, with fewer dependencies. Moreover, the licenses of ChatGPT's are mostly permissive, but also include copyleft and public domain options. However, our analysis highlight potential challenges for developers, as 6.5% of ChatGPT library recommendations resulted in import or installation failures. To shed light on these challenges and understand the specific reasons behind these failures, we delve deeper in our second research question.

RQ2 What are the challenges encountered by developers when using ChatGPT for library recommendations? We find that 6.5% of the libraries used by ChatGPT lead to import or installation failures, mainly because of aliases (alternative names for existing libraries), implicit imports of modules (no mention of the explicit parent library), placeholder suggestions, and context-specific recommendations that might not generalize well.

Our findings shed light on how well LLMs recommend libraries and the factors that influence their selection. This will not only inform the development of more robust LLM software librarians but also provide actionable recommendations for practitioners to fully benefit from LLM-based tools in streamlining their workflows. We also identify areas where LLMs excel and areas where developers need to exercise caution. Our work makes the following contributions:

- Empirical evidence on the characteristics of libraries used in GPT-generated code and the potential challenges in using these libraries out of the box.
- A set of actionable recommendations for developers using LLMs to streamline their workflows, including considerations for potential issues with LLM-suggested libraries.
- A dataset of ChatGPT-generated Python code alongside human-written solutions for Stack Overflow prompts, enabling a direct comparison of library selection practices.¹

Paper Organization. The rest of the paper is organized as follows. Section 2 describes our dataset and methodology. Section 3 presents the findings of our two research questions. Section 4 discusses the implications of our findings along with our recommendations. Section 5 discusses related work. Section 6 outlines the threats to the validity of our study. Section 7 concludes this paper.

2 STUDY DESIGN

Figure 1 presents an overview of our approach to curate the dataset used in this study. In order to conduct our qualitative and quantitative analysis, we focus on the Python programming language

¹<https://zenodo.org/records/11506864>

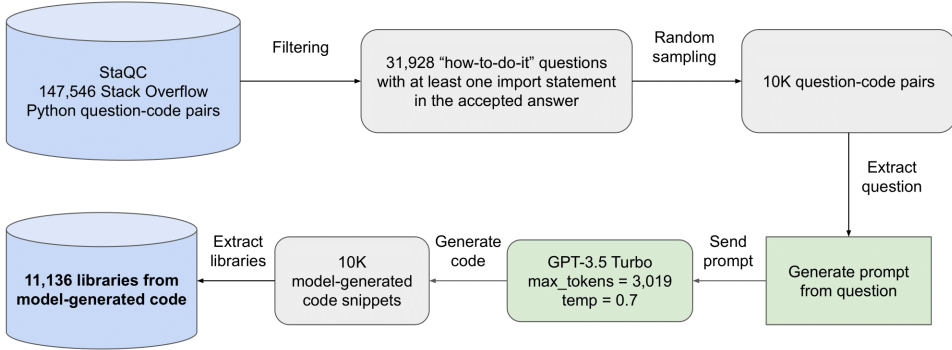


Fig. 1. Overview of our approach for collecting libraries ChatGPT-generated code snippets.

for its popularity, extensive library ecosystem, as well as the demonstrated capabilities of LLMs in generating and understanding Python code [15–18]. Below we describe the steps to curate the dataset used in our study and the experimental setup.

2.1 Dataset

We prepared our dataset using the dataset of Stack Overflow question-code pairs curated by Yao et al. [19]. Their set contained 147,546 Python question-code pairs automatically mined from Stack Overflow using a bi-view hierarchical neural network. We selected this dataset because it provides human-written code, which is necessary for our comparative analysis with ChatGPT-generated code. Also, the dataset is comprised of "how-to-do-it" Stack Overflow questions (e.g., how to grab from JSON in selenium python²), which facilitates prompt crafting for our experiment. Finally, the dataset is labeled, allowing us to filter for specific types of answers, such as multi-code versus single-code answer. Thus, to facilitate code extraction, we select questions annotated single-code answer posts, which are accepted answers that contain only one code snippet.³ Finally, to ensure that the selected questions are specifically about library usage, we only keep the ones where the code snippet of the corresponding accepted answer includes at least one import statement. This process yields a total of 31,928 question-code pairs. Then, to ensure that our analysis is manageable, we select a random sample of 10,000 question-code pairs (confidence level of 99% with a ~1% confidence interval).

2.2 Experiment Setup

To create model-generated code for each Stack Overflow accepted answer, we take a two-step approach. The first step is to construct the prompts to specify the type of tasks we expect the model to complete. In our case, the model will be generating code to answer programming questions in Python. Therefore, the prompts are structured with the base prompt "Generate Python code for

²<https://stackoverflow.com/questions/26661808/how-to-grab-from-json-in-selenium-python>

³<https://stackoverflow.com/questions/36479006/how-to-post-data-and-binary-data-using-urllib2-in-python>

the following question" to which we append the Stack Overflow questions. Then, to facilitate the post-processing of responses, we ask the model to "only generate code, no explanations". Thus, the final prompt is "Generate Python code for the following question: <question>. Only generate code. No explanations."

Then, after generating the prompts, we feed them to the model to generate the corresponding code. To achieve this, we use the OpenAI API to query the language model GPT-3.5 Turbo. We selected this specific model due to its free availability, which lowers barriers to use, enhancing the inclusivity of our study and simplifying replication, as well as for its rapid response time. This corroborates with the work of Destefanis et al. [20] who have shown that GPT-3.5 Turbo outperforms other models in generating correct code, with a 90.6% correctness rate across various problem categories.

The parameters that we used for the model are shown in Figure 1. These parameters are essentially to guide the model's response generation. One such parameter is *temperature* (temp), which affects the creativity or randomness of the response. In our study, we use ChatGPT's default temperature which is of 0.7.⁴ Another important parameter is the *maximum number of tokens* (max_tokens), which determines the maximum length of the generated response. We set this to 3,019 tokens, which is the longest code snippet from the accepted answers on Stack Overflow. This value was chosen to ensure that the model-generated responses had sufficient scope to be complete while replicating the format and detail level found in the code snippets of our dataset.

Finally, because the focus of our research is on the use of libraries in large language models, we use regular expressions to extract the names of the used libraries from import statements in the model-generated code snippets. Additionally, we only extract the parent libraries and not the individual modules because our study's primary interest lies in understanding the usage of libraries. For example, from the import statement `from X import Y`, we extract X. This yields a total of 11,136 libraries (764 unique) across 10,000 question-code pairs.

3 RESULTS

In this section, we present the results of our three research questions. For each research question, we present our motivation, the approach to answer the question, and key findings.

RQ1: What are the characteristics of the software libraries recommended by ChatGPT?

Motivation. As developers increasingly turn to tools like ChatGPT and other similar tools for coding assistance [21], the libraries included in the recommended code become more important as the code's performance, security, and functionality are directly impacted by these libraries [4, 22, 23]. Thus, in this RQ, we aim to understand the characteristics of libraries used in model-generated code to understand how software development is changing with the inclusion of LLMs. This knowledge will help us gain insights into the model's underlying knowledge base and its alignment with current software development practices. This is particularly relevant for developers relying on ChatGPT's suggestions to expedite their coding process. For that purpose, we will investigate the types of libraries recommended by ChatGPT, their popularity and maintenance characteristics, as well as their licenses.

Approach. To answer this RQ, our approach involves a three-step analysis. First, we classify libraries into three categories: *standard* (included in the standard Python package), *third-party* (available on the PyPi repository), and *other* (neither standard nor third-party). Subsequently, we examine the characteristics of libraries in ChatGPT-generated code. Finally, we analyze the licenses of third-party libraries. We outline each step below.

⁴<https://gpt.space/blog/how-to-use-openai-model-temperature-for-better-ai-chat-responses>

First, To identify *standard* Python libraries, we use `stdlib-list`,⁵ a Python package that lists all the standard libraries included in each Python version since 2.6. Then, for each library, we check if it is in the list of standard Python libraries. If so, we classify the library as *standard*. For the libraries that were not classified as *standard*, we verify if they are third-party by utilizing the PyPi API.⁶ If the API response returned a status code of 200, indicating a successful retrieval, we classify the library as *third-party*.

Second, we analyze the characteristics of third-party libraries. The reason we focus on third-party libraries is that standard libraries are always available with a Python installation, which means that characteristics such as the number of dependents and the number of dependencies do not apply in this context. We select the characteristics based on the work of Jafari et al. [22] on dependency update strategies and library characteristics, and the work of Vargas et al. [24] on practitioners' perspective on selecting third-party libraries. To obtain such characteristics, we utilize the `libraries.io` API, a tool that tracks open source software libraries across many programming languages and platforms.

The third step is to identify the license category for each third-party library. For this, we manually labeled each one with the Software Package Data Exchange (SPDX) license identifier by examining the license information available on the PyPi page. For example, the library `numpy` has the SPDX identifier "BSD" which indicates that the license is a BSD style license. Then, we cross-referenced this information with the Open Source Initiative's (OSI)⁷ list of approved licenses to identify the license category. For example, the BSD style of license belongs to the "permissive" license category.

Finally, based on prior work on dependency analysis [22], we group the library characteristics into three main categories:

- **Popularity:** This category measures how widely adopted and trusted a library is within the Python ecosystem. We analyze three key characteristics to gauge popularity:
 - Forks: The number of GitHub repositories that share the code and visibility settings with the original upstream repository of the library.⁸
 - Stars: The number of GitHub users that have marked the library's repository as a favorite.
 - Dependent Count: Indicates how many libraries rely on a specific library.
- **Maintenance:** This category assesses the health, maturity, and activity level of the libraries based on the following characteristics:
 - Dependency Count: Represents how many other libraries a particular library depends on. It is based on the library's state just before September 2021, which marks the date of the last data included in GPT-3.5 Turbo's training.
 - Age: The age of the project in months; calculated from when the library was first released until the date of metadata collection (February 2024).
 - Version Frequency: The number of released versions divided by the age in months.
 - SourceRank: The SourceRank metric of a library provided by `libraries.io`; derived from the latest version of a library as of the date of the metadata collection (February 2024).
- **Licensing:** Licensing defines the terms under which the library can be used, modified, and distributed, which impacts the legal aspects of a software. In this category, we evaluate whether the recommended third-party libraries have an explicit usage license (indicated in the PyPi page of the library) and, if so, the category of the license (i.e., permissive, copyleft, weak

⁵<https://pypi.org/project/stdlib-list/>

⁶<https://warehouse.pypa.io/api-reference/index.html>

⁷<https://opensource.org/licenses>

⁸<https://docs.github.com/en/pull-requests/collaborating-with-pull-requests/working-with-forks/about-forks>

copyleft, public domain). This categorization is essential to understanding the restrictions of using each library, which helps us determine their usage for different development scenarios.

Results. Table 1 shows the distribution of the types of libraries found in ChatGPT-generated code against human code. The first column (Source) represents the source of the library (i.e., ChatGPT or Human). The second and third columns show the number of occurrences (# Occ.) and the percentage of occurrences (% Occ.) for each type of library, respectively. Finally, the last two columns show the number (# Unique) and percentage (% Unique) of unique libraries in each type, respectively. The results indicate that in 48.3% of the cases, ChatGPT utilizes third-party libraries, compared to 45.2% where standard Python libraries are used. In comparison, human-generated code from accepted answer snippets use third-party libraries in 38.7% of the cases, highlighting a difference of 9.6%. This suggests that ChatGPT includes more third-party libraries, especially when compared to human-generated. In 6.5% of the cases, ChatGPT recommends libraries that fall into the *other* category, indicating these libraries are neither identified as *standard* nor *third-party*. We delve deeper into this category and its implications in RQ2.

Table 2 presents a statistical summary of the characteristics across the 437 third-party libraries used by ChatGPT.

Popularity. In terms of popularity, the high standard deviation in the number of stars (10973.4), fork (2916.2), and dependents (5072.5) indicate a wide range in the popularity of the libraries used by ChatGPT. The maximum values observed (30,753 forks, 125,980 stars, and 58,040 dependents) suggest that some of these are highly popular, and are integral to a large number of projects (e.g., requests). In comparison, libraries found in human-generated code show lower mean values for forks (487.7) and stars (2,113.4), a similarly high maximum, indicating that while human developers also use popular libraries, ChatGPT does so more often. The median values further illustrate this, with ChatGPT having medians of 98 forks, 430 stars, and 98 dependents, compared to humans having medians of 7 forks, 23.5 stars, and 14 dependents. This suggest that for the same question and context, ChatGPT tends to favour more popular libraries than human developers.

Maintenance. Regarding maintenance, Table 2 shows that the maximum number of dependencies for a library is 59 (e.g., Zope), which shows that some libraries are dependent on a high number of other libraries. However, the median value of 0 for dependency count indicates that a significant number of libraries are standalone or have minimal dependencies (e.g., numpy with 0 dependency or scipy with one dependency). This shows that ChatGPT has a tendency towards using libraries that do not require additional installations or configurations, which can potentially simplify the development process for users since the dependency count can impact the ease of integration, maintenance, and potential issues related to version conflicts or transitive dependencies [4, 22]. In comparison, libraries found in human-generated code also show a median of 0 dependencies but with a slightly lower mean of 2.3. This shows a similar tendency towards human developers using libraries with fewer dependencies.

Moreover, our results show that the overall SourceRank of the libraries is relatively high. According to the libraries.io documentation,⁹ the maximum score for SourceRank is around 30 points, a difference of 8 points from the median SourceRank of 22 for the libraries used by ChatGPT. This suggests that more than half of the libraries are rated above the mid-range in terms of their overall quality and community trust and have a score close to the maximum possible score. Three libraries are found to be above this threshold: requests, pandas, and numpy, with SourceRank of 32 and 31, respectively. On the other hand, human developers use libraries with a median SourceRank of 12 and a mean of 13.3. This indicates that ChatGPT tends to use libraries that are recognized for their quality and maintenance more often than human developers.

⁹<https://docs.libraries.io/overview.html#sourcerank>

Table 1. Types of Libraries Used by ChatGPT.

	Source	# Occ.	% Occ.	# Unique	% Unique
Third-party	ChatGPT	5,377	48.3%	437	57.2%
	Human	5,891	38.7%	720	57.8%
Standard	ChatGPT	5,029	45.2%	145	19%
	Human	7,538	49.6%	159	12.8%
Other	ChatGPT	730	6.5%	182	23.8%
	Human	1,778	11.7%	366	29.4%
Total	ChatGPT	11,136	100%	764	100%
	Human	15,207	100%	1,245	100%

Finally, we consider the age and version frequency of a library as interrelated characteristics because they provide insights into its maturity and maintenance activity [22]. The median age of the libraries used by ChatGPT is 159 months (approximately 13 years), while for libraries in human-generated code, it is around 125.1 months (approximately 10.5 years). This indicates in both cases, half of the libraries have been around for over a decade, implying a high level of maturity. The version frequency further illustrates the maintenance activity, with ChatGPT showing a median of 0.24 and a mean of 0.94, compared to human developers with a median of 0.16 and a mean of 0.59. This suggests that libraries used by ChatGPT are updated slightly more frequently than those used by human developers.

Licensing. Table 3 shows the categories of licenses associated with the libraries used by ChatGPT compared to those used by human developers. Our results indicate a strong dominance of *permissive* licenses among libraries used both by ChatGPT and human developers. Specifically, 71.7% of the libraries used by ChatGPT fall under the permissive category, which includes licenses such as MIT, Apache 2.0, and BSD. This is only slightly higher than the 69.3% observed in libraries found in the human-generated code from our dataset. This indicates that the vast majority of the libraries in both ChatGPT and human-generate code impose minimal restrictions on how they can be used, modified, and distributed.

In contrast, *copyleft* licenses (e.g., GPL, LGPL, AGLP), which require derivative works to also be open source and distributed under the same license, are less prevalent. ChatGPT used libraries with copyleft licenses 14.2% of the time, followed closely by 13.9% in human-generated code. *Public domain* licenses (e.g., CC0, Unlicense), which place the library in the public domain and relinquish all rights, have a low percentage of usage among both sets of libraries. ChatGPT includes public domain licenses in 1.5% of cases, compared to 1.9% in human-generated code. *Weak copyleft* licenses (e.g., EPL, MPL), which are less restrictive than full copyleft but still require some openness for derivative works, represent 0.9% and 1.1% of the licenses in ChatGPT and human libraries, respectively. These results suggest that both ChatGPT and human developers might favor licenses that are either fully permissive or fully copyleft over licenses with more nuanced requirements.

Interestingly, we observe the presence of libraries with unspecified licenses, which account for 10.4% of the libraries used by ChatGPT and 12.9% of the libraries used by human developers. The slightly higher percentage of unspecified licenses in libraries used in human-generated code could indicate that developers tend to overlook license specifications more often than ChatGPT. It could also suggest that human developers tend to include more obscure libraries where licensing

Table 2. Characteristics of Libraries Used by ChatGPT.

Category	Characteristic	Source	Median	Mean	Std. Deviation	Min	Max
Popularity	No. of forks	ChatGPT	98	907.6	2916.2	0	30753
		Human	7	487.7	2063.4	0	30497
	Stars	ChatGPT	430	3866	10973.4	0	125980
		Human	23.5	2113.4	7796.5	0	127601
	No. dependents	ChatGPT	98	1328.5	5072.5	0	58040
		Human	14	840.8	4205.4	0	60232
Maintenance	No. dependencies	ChatGPT	0	3.2	7.5	0	59
		Human	0	2.3	6	0	59
	sourceRank	ChatGPT	22	15.9	7	1	32
		Human	12	13.3	7	0	32
	Age (months)	ChatGPT	159	116.9	59.7	2.6	259
		Human	125.1	117.2	58.6	4.3	261
	Version frequency	ChatGPT	0.24	0.94	2.4	0.01	21.2
		Human	0.16	0.59	1.6	0.01	15.5

Table 3. Library License Categories.

Category	Definition	ChatGPT		Human	
		# Occ.	% Occ.	# Occ.	% Occ.
Permissive	Allows broad reuse with minimal restrictions.	324	71.7%	500	69.3%
Copyleft	Requires derivative works to use the same license.	64	14.2%	100	13.9%
Public Domain	No exclusive intellectual property rights.	7	1.5%	14	1.9%
Weak Copyleft	Allows linking with other licenses	4	0.9%	8	1.1%
Other	Miscellaneous licenses not in the above categories such as "Proprietary", "Non-standard", and "OtherProprietary".	6	1.3%	5	0.7%
Unspecified	License not specified.	47	10.4%	93	12.9%
Total		452	100%	722	100%

information is not readily available (e.g., app2¹⁰). The presence of unspecified licenses can be a potential risk for developers since it can potentially lead to legal ambiguities [25].

Lastly, the "Other" category, which includes licenses that are less common or too obscure to classify (i.e., *OtherProprietary*, *Proprietary*, and *Non-standard*), shows a minimal difference between libraries used by ChatGPT and human developers, at 1.3% and 0.7% respectively. These results indicate that in both cases, less conventional licensing arrangements are uncommon.

¹⁰<https://pypi.org/project/app2/>

Our findings suggest that ChatGPT tends to use libraries that are well-established and stable as opposed to popular according to their SourceRank, Age, and Version Frequency. We also found that ChatGPT tends to use libraries with fewer dependencies more frequently than those with a high dependency count, implying a preference for more standalone libraries. Moreover, similar to human developers ChatGPT tends to use libraries that have permissive licenses, with a smaller but still significant amount of libraries with copyleft and public domain licenses.

RQ2: What are the challenges encountered by developers when using ChatGPT for library recommendations?

Motivation. In RQ1, we find that certain libraries recommended by ChatGPT are neither part of the standard Python libraries, nor published on the PyPi repository. Such libraries are potentially problematic because it means that they do not work out of the box, which can confuse developers and hinder their workflow. Thus, identifying and understanding challenges such as this is crucial for improving the practical utility of LLMs in real-world settings. When ChatGPT uses libraries that do not work out of the box (i.e., they do not exist on the current PyPi repository and are not part of the default Python installation), it not only impacts the productivity of developers but also raises concerns about the trustworthiness and validity of the generated code [17]. By studying the challenges developers encounter in the context of LLMs as software librarians, we can develop actionable strategies to mitigate them. Thus, this RQ aims to pinpoint the specific difficulties developers face when using ChatGPT for library recommendations.

Approach. To understand the underlying causes behind the presence of problematic libraries in ChatGPT-generated code, we first determine whether these libraries were explicitly mentioned in the question. If so, we categorize the library as *hard-coded*, indicating that ChatGPT was directed to use a library not found in standard Python libraries or the PyPi [26] repository.

In cases where the unclassified library is not hard-coded, we verify whether it is deprecated (as of March 2024). We devised a two-step approach to determine the deprecation status of a library. The first step targets libraries that while not found on the PyPi registry, might have been previously part of the standard Python installation, and involves querying the Python documentation website [27] for Python 3 and Python 2 versions. In this case, we consider a library deprecated if the request for Python 3 is successful (i.e., status code is 200) but with a changed library name in the response URL (e.g., `urllib2` becomes `urllib.request`). Alternatively, if the request for Python 3.x fails (e.g., status code is 404), we request the documentation website for Python 2.x; a successful request leads to the library being labeled as deprecated.

The second step covers the libraries that might have been on the PyPi registry, but have been removed since then. This step involves using virtual environments and the `subprocess` module to automate the process. For each library, we create a unique virtual environment and install the library. During both the installation and import stages, we check for deprecation warnings. This is done by capturing and analyzing the output from the `pip install` command and a Python script that imports the library.

Next, once we have ruled out the possibility of a problematic library being deprecated or hard-coded, we look at other possibilities. For this, we take the remaining set of recommended libraries and perform open coding. This process involves analyzing the Stack Overflow question used to generate the prompt, including the question body, any provided code snippets, and the accepted answer's body. Through this analysis, we developed a set of nine mutually exclusive labels to categorize the libraries. The open coding was performed by one author, with the labeling scheme

having been collaboratively developed by both authors. After the initial labeling, the results were validated through discussions with the second author. The reason behind our decision to involve only one author in the labeling process is that the nature of the labeling in this context is largely factual rather than subjective. For example, `cv2` is an alias for `opencv-python`, is a factual classification based on established naming conventions and not open to interpretation. Thus, the labels in our scheme were defined based on concrete characteristics of the libraries and their usage.

Results. Table 4 shows the number and ratio of the problematic library recommendations made by ChatGPT. Below we outline each of the categories that resulted in an import or installation failure.

Hard-coded Libraries (32.9%). A significant portion (one-third) of problematic recommendations are hard-coded libraries. This implies that ChatGPT generates responses based on contextually mentioned libraries, regardless of their existence or current support status. Such behavior can mislead developers into believing that the requested library is real, even if it doesn't exist, undermining the trustworthiness of LLM-generated code.

Alias (30.3%). Another prevalent challenge arises from ChatGPT's use of aliases in library recommendations. The Python ecosystem lacks a standardized naming convention for distributions, leading to discrepancies between library import names and installation names [28]. For example, the library `opencv` has the distribution name `opencv` but is imported as `cv2`. While aliases themselves aren't inherently problematic, they introduce a layer of complexity that developers need to navigate. This highlights the need for LLMs to be aware of these naming conventions and favor explicit imports (e.g., `from numpy import array`) to enhance code clarity and avoid potential errors.

Module (16.2%). Our results reveal challenges related to how ChatGPT distinguishes between modules and libraries. In Python, a library is a collection of modules, while a module is a single Python file. Python provides several ways to import modules and libraries, which we categorize into "implicit" and "explicit" imports. LLMs should understand these nuances and adapt their recommendations accordingly. Inappropriate import statements (e.g., importing `array` without specifying the need for `numpy`) can lead to confusion and installation failures if developers attempt to install the module as a standalone entity.

Placeholder (14.7%). Placeholders appear as generic or illustrative names that possibly intend to guide developers, and are often used when the question is more vague (e.g., Pythonic way to write long import statements¹¹). Placeholders can become problematic when what is intended to be a generic example ends up being interpreted as an actual library. For instance, in our dataset, there was a case where the import statement was `import ParentModel`, which was intended as a placeholder. This ambiguity can cause confusion, where it would be more beneficial for ChatGPT to provide actionable examples as oppose to illustrative examples.

Deprecated (1.4%). A smaller but still significant portion of problematic libraries are tied to deprecated libraries. Such libraries may have outdated functionality, or are no longer supported or available through PyPi. A particular case that may generate confusion is when the syntax of the library is outdated. For instance, the `queue` library is imported as `Queue` in Python 2, but is imported as `queue` in Python 3. This suggests that models should have updated, timely knowledge of library statuses.

Mistake (1%). The *Mistake* category represents cases where ChatGPT appears to have made genuine errors in suggesting libraries. Table 5 shows specific instances where ChatGPT recommended non-existent libraries or inappropriate import statements in response to various StackOverflow questions from our dataset. For example, the suggestion `fuzzyhashmap` in response to the question about fuzzy hash tables in Python is not a legitimate library and indicates that ChatGPT might have fabricated the name based on related terminology in the question.

¹¹<https://stackoverflow.com/questions/26253545/pythonic-way-to-write-long-import-statements>

Table 4. Causes for Import or Installation Failures of ChatGPT Library Recommendations.

Cause	# Occ.	% Occ.
Hard-coded	241	33%
Alias	223	30.5%
Module	119	16.3%
Placeholder	108	14.8%
Deprecated	10	1.4%
Mistake	7	1%
Other	22	3%
Total	730	100%

Table 5. Incorrect library imports by ChatGPT based on Stack Overflow questions.

StackOverflow Question	Import Statement
"Fuzzy" hashtables in Python to replace elif chains?	import fuzzyhashmap
Can I work with a socket I created in another file in python?	import create_socket
How can I use ptiPython with Flask-Script together?	import flask_ptipython
Working with Import Statements in OSX Mavericks.	import opencv,import scikit, import Queue
Java and Python implementation of Blowfish produce-different results.	import javax

Other (1.2%). The *Other* category groups recommendations for custom libraries (libraries not widely available, e.g., `progress_statusbar` refers to a custom module from a Python development cookbook [29]), external applications that can be extended with Python (but not installed via PyPi), and libraries dependent on specific environments (e.g., Android Python interpreter).

Our findings indicate that two thirds of problematic libraries are hard-coded or alias-driven, with 32.9% and 30.3%, respectively. Our analysis also reveals a mix-up between modules and libraries, the introduction of placeholders, as well as context-specific recommendations. Such library recommendations can lead developers into confusion and waste of time by trying to unsuccessfully import or install the recommended library.

4 DISCUSSION

In this section, we discuss the implications of our work and propose recommendations for practitioners and researchers in the context of LLMs as software librarians.

4.1 Are LLMs Good Software Librarians After All?

Our results indicate that ChatGPT generated only five hallucinations out of 10,000 questions, and only 6.5% of the libraries were potentially problematic. Based on these findings, we argue that LLMs like ChatGPT have the potential to be good software librarians, though several important considerations must be addressed, which we discuss below.

Licensing Awareness. One significant area of improvement for ChatGPT as a software librarian is its lack of communication of software licenses. While most of the recommended libraries had permissive licenses, 14.2% were copyleft licensed- a more restrictive type of license- and 10.4% had no license specified. This was not explicitly communicated by the model, which can lead to potential legal and compliance issues for developers. Permissive licenses, such as the MIT or Apache licenses, allow developers to use, modify, and distribute the software with minimal restrictions, providing flexibility and reducing legal risks. On the other hand, copyleft licenses, such as the GNU General Public License (GPL), require that any derivative works also be distributed under the same license. This implies that if developers use a GPL-licenses library in their project, they may be obligated to release their own code under the GPL as well. This can be problematic for proprietary software developers who do not wish to open-source their code. To better understand this, we can consider a scenario where a developer is working on a proprietary project and uses ChatGPT to generate code. The model recommends a library without specifying its license, for example PyQt¹², one of the most popular cross-platform GUI libraries for Python. The developer, assuming it is safe to use, integrates the library into their codebase, but later discovers that it is GPL-licensed. This forces the developer to either comply with GPL requirements, which may not be feasible or desirable, or replace the library, which involves additional effort to refactor the code and ensure compatibility.

To avoid such problematic scenarios, developers can enhance the model's responses by including safeguards such as explicit license information for each recommended library. For example, when suggesting a library, the model should include a note such as "This library is licensed under the GNU GPL v3.0". Moreover, if the generated output contains a library, the response should clearly state the license type and restrictions, e.g., "The library PyQt5 is licensed under the GNU GPL v3.0, a copyleft license that may require your code to also be open-sourced under the same license."

Also, LLM (including GPT) developers should train models to detect and flag libraries with restrictive licenses, and provide alternatives with permissive licenses where possible. For instance, if the model detects a GPL-licensed library, it should offer an alternative like: "While PyQt5 is suitable, it is GPL-licensed. You might consider Tkinter, which has a more permissive license (PSF License Agreement)."

ChatGPT Unconditionally Recommends Libraries. One notable finding from RQ2 is that 33% of the libraries that did not work out of the box in ChatGPT code were actually explicitly mentioned in the user question. This highlights a critical issue: if a user explicitly instructs the model to use libraries that are not standard, recognized, or even existent, ChatGPT will still attempt to generate code utilizing that library without any warning. This can lead to several significant problems.

First, ChatGPT may generate code for libraries that are non-existent or no longer supported. This can lead to confusion and wasted time as developers attempt to install and use a library that simply isn't there. For example, a user asked "how to post data and binary data using urllib2 in Python?".¹³ In this case, urllib2 is a library that is no longer supported and no longer available on PyPi. Despite this, ChatGPT generated code using urllib2 as shown in Listing 1. While the code itself is correct, installing urllib2 is not possible with Python 3. Although the Stack Overflow question from which the prompt was generated is eight years old, tools like ChatGPT should have up-to-date knowledge of libraries to ensure that developers use current and supported libraries in their code. From the developer's side, fixing such issues requires the developer to recognize that the library does not exist or is deprecated, which may not be immediately obvious, especially to less experienced developers. Thus, the reliance on user knowledge to identify and correct library recommendations undermines the efficacy and reliability of LLMs as software librarians. As such,

¹²<https://pypi.org/project/PyQt5/>

¹³<https://stackoverflow.com/questions/36479006/how-to-post-data-and-binary-data-using-urllib2-in-python>

we propose that LLM developers enhance models to include warnings when a user requests a non-existent or deprecated library along with the generated code. For example, if a user asks for a deprecated library like `urllib2`, the model should respond with, "The library `urllib2` is no longer supported. Consider using requests instead." This will help developers avoid wasting time on outdated libraries and ensure they use current, supported ones.

Furthermore, there is a critical security implication arises when ChatGPT unconditionally generates code for libraries that are explicit in the request. For example, if a user mistakenly asks for `padnas` instead of `pandas`, ChatGPT might generate code for the misspelled library name. If the user attempts to install `padnas`, they might inadvertently install a malicious library that takes advantage of typosquatting- a common tactic where attackers upload malicious packages with names similar to popular libraries to exploit such mistakes. In this specific example, `padnas`¹⁴ is a harmless library designed to resemble `pandas` to prevent such attacks, which highlights that these scenarios are not uncommon.

To mitigate these security risks, we recommend that both LLM developers and practitioners take proactive measures. For example, LLM developers should enhance models to verify and correct library names before generating code. If the model detects a likely typo, it should suggest the correct library name, such as, "It looks like you meant `pandas`. Generating code for `pandas` instead of `padnas`". Practitioners should also adopt best practices when using LLMs such as ChatGPT to assist in programming tasks [30]. Notably, they should cross-check library names and ensure they are correctly spelled and widely recognized before attempting to install them. It is critical that practitioners do not blindly adopt generated code without first verifying the overall health of the libraries, as the health of libraries is constantly evolving, and new vulnerabilities may have emerged since ChatGPT was last trained. For this, practitioners can use tools such as the `sourceRank` provided by `libraries.io`, or `Snyk`,¹⁵ a security platform for securing code, open source dependencies, and cloud platforms.

```

1  import urllib2
2  import json
3
4  url = 'http://url.com?u=user&p=pass'
5  data = json.dumps({'config': 'configData'})
6  cont_len = len(data)
7  req = urllib2.Request(url, data,
8      {'Content-Type': 'application/json', 'Content-Length': cont_len})
9
10 f = urllib2.urlopen(req)
11 response = f.read()
12 f.close()

```

Listing 1. ChatGPT-generated code for deprecated `urllib2`.

We conclude that while LLMs like ChatGPT can be good software librarians, they can be improved to better assist developers in streamlining their work. This can be achieved by implementing warning mechanisms to alert users when they request libraries that are deprecated or potentially non-existent, integrating verification tools that cross-reference libraries against known repositories like PyPi, and incorporating alerts for potential security risks like typosquatting. Additionally, providing more explicit information and context for recommended libraries, such as licensing, will help users mitigate potential conflicts and ensure easier integration into their codebase.

¹⁴<https://pypi.org/project/padnas/>

¹⁵<https://snyk.io/>

4.2 The Impact on Dependency Management

Our study highlights several implications for practitioners who are increasingly relying on LLMs as programming assistants. Here, we discuss the potential impact on dependency management and development workflow.

Increased Use of Third-Party Libraries. Our results show that ChatGPT tends to use third-party libraries more frequently than human developers, favoring well-established and widely adopted libraries. The use of third-party libraries can significantly impact dependency management as it introduces additional complexity and risks into software projects [6, 31]. Each new library brings its own set of dependencies, the majority of which are transitive (dependencies induced by other dependencies) [4]. This can lead to *dependency hell*, a term used to describe when there are so many dependencies and potential version conflicts that resolving them becomes a significant challenge and developers spend more time managing dependencies than writing actual code [5, 32].

To address these challenges, LLM developers should enhance models to provide not only code with appropriate libraries but also information on the dependency footprint of each library. For example, when suggesting a library, the model could include a note such as "This library has 5 direct dependencies and 20 transitive dependencies." Moreover, LLMs could be trained to recommend libraries with fewer dependencies when multiple options are available to minimize the risk of dependency hell. LLM developers should also incorporate security and stability checks into the models by leveraging databases like the National Vulnerability Database¹⁶ (NVD) or GitHub Security Advisories¹⁷ (GSA).

In addition to the steps for LLM developers, we recommend that practitioners thoroughly evaluate the maintenance metrics of libraries recommended by ChatGPT before integrating them into their codebase. Firstly, practitioners should consider the number of dependencies associated with each library. A high number of dependencies increases the risk of conflicts and broadens the attack surface of the codebase [4, 6]. For instance, using tools like `npm ls` for JavaScript or `pipdeptree` for Python can help visualize and assess the dependency tree of a library. Practitioners should also consider the version frequency of a library; while frequent updates may indicate active maintenance, they can also be a sign of instability, and managing constant updates can be challenging [22]. Tools such as Dependabot¹⁸ or Renovate¹⁹ can automate the process of dependency updates and help keep track of changes without overwhelming the development process.

Libraries Recommended by ChatGPT are Frozen in Time. The knowledge of an LLM is essentially "frozen in time" at the point when it was last trained. This arises from the nature of training LLMs, which is an expensive and infrequent process coupled with the dynamic nature of libraries, which continually evolve with updates with add new features, fix bugs, and address security vulnerabilities. As a result, the libraries recommended by ChatGPT may not reflect the most current versions. Thus, practitioners might find that the functions or methods found in ChatGPT-generated code are deprecated or have been replaced by more efficient alternatives in the latest library versions. New versions of libraries can include breaking changes that are not compatible with previous versions, which can cause runtime errors and necessitate additional effort to refactor the code with work with the updated version. Moreover, libraries are frequently updated to address security vulnerabilities [4, 22], and using an outdated snapshot of a library version as recommended by ChatGPT to avoid compatibility issues can expose projects to security risks.

¹⁶<https://nvd.nist.gov/>

¹⁷<https://github.com/advisories>

¹⁸<https://github.com/dependabot>

¹⁹<https://github.com/renovatebot/renovate>

To mitigate these issues, we recommend that practitioners integrate Continuous Integration (CI) tools into their workflow to automate the testing of code against the latest versions of dependencies. CI tools can help identify compatibility issues early, allowing developers to address them before they become problematic [33]. Furthermore, LLM developers should enhance models by incorporating mechanisms to check the currency of recommended libraries. For instance, the LLM could provide a disclaimer such as, "This library is based on a version as of [last training date]. Please verify with the latest documentation." This would prompt practitioners to validate the suggested library version. LLM developers can also train models to prioritize stable libraries that have low frequency of breaking changes. For example, if multiple libraries provide similar functionalities, the model should recommend the one with a proven long-term stability, not necessarily the most popular one as our results of RQ1 indicate. This could be done by integrating data from sources like GitHub releases or library changelogs.

Mind Your Prompts. In light of the above-mentioned implications, one might consider asking ChatGPT for alternative standard libraries equivalent to third-party libraries. Standard libraries are generally more stable, better documented, and maintained as part of the core language distribution, which reduces the dependency management burden and minimizes the risks as mentioned above [34]. However, when asking ChatGPT for alternative libraries, practitioners should be mindful of the potential for the model to "hallucinate". An article by Vulcan Cyber states that between 25% and 40% of the libraries generated by ChatGPT are hallucinations [35]. In this article, the author asked ChatGPT to generate code to integrate a library in Node.js and then repeatedly requested alternative methods. As a result, ChatGPT began hallucinating and producing non-existent libraries. While this has significant security implications, it does not reflect a typical user interaction with ChatGPT, which we aimed to reflect in this study.

To mitigate these risks, we recommend that practitioners craft solid prompts from the start, providing enough context surrounding the type of library they are seeking so that the model understands what is expected early on. For example, instead of asking, "Can you suggest a library for HTTP requests in Python?", a more detailed prompt would be, "Can you suggest a standard Python library or widely-used third-party library for making HTTP requests, and provide a brief explanation of its advantages?" LLM developers can also help in addressing these issues by enabling the models to recognize when it is best to use a standard library versus a third-party library. Implementing a confidence scoring system for library recommendations, where the model indicates the certainty of its suggestions, can guide LLM users in evaluating the reliability of the recommendations. This can be integrated in the models' responses such as, "This library has a high confidence level based on currently training data."

Users as Active Evaluators. The results of RQ2 suggest that the context in which a library is recommended is crucial in determining its relevance and utility. Thus, problematic library usage is not inherently tied to ChatGPT making mistakes, but to the very context-dependent nature of the PyPi ecosystem. This means that developers using LLMs as programming assistants are not just consumers of the library recommendations but should be active evaluators to ensure that the libraries they integrate into their own code are compatible with their own context. For the LLM development community, this means improving the models' ability to recognize, understand, and prioritize context-specific information when generating code with software libraries. For instance, the model should consider a project's framework, the programming language version, and the specific functionality required.

5 RELATED WORKS

In this section, we discuss the related literature divided into two aspects. First, we discuss the works that have focused on dependency management and related challenges. Second, we discuss the works that report on the use of large language models as programming assistants.

5.1 Dependency Management Challenges

While open source software libraries significantly reduce development time and costs [1–3], depending on numerous libraries introduces complexity and potential dependency management challenges [4–6]. One such challenge is highlighted by Mujahid et al. [7] who identify library selection as a critical aspect of dependency management. In their work, they surveyed developers from the npm ecosystem to qualitatively understand the characteristics of highly-selected libraries. Their results show that JavaScript developers believe that such libraries are well-documented, popular, and free of vulnerabilities. Building upon this work, our study leverages those same characteristics to categorize libraries used by ChatGPT. Hauge et al. [8] show that organizational library selection is an ad-hoc process that often relies on a combination of past experiences, expert advice, and online resources. This is further discussed in the work of Haenni et al. [9] where the authors surveyed developers about their decision-making when selecting a library to integrate into their application. Their findings show that, in general, developers do not apply rationale when selecting libraries. Alternatively, developers opted for libraries that fulfilled the immediate task requirements. Given this lack of formal selection processes and the increasing popularity of LLMs as programming assistants, this motivated us to investigate the role of LLMs as software librarians.

Dependency hell is a concept discussed in several studies [36–38] and refers to when a project has an excessive number of dependencies, and managing these dependencies becomes difficult and error-prone. Chen et al. [32] discuss the impact of "trivial packages," referring to libraries implementing simple functionalities, on the npm ecosystem. Their survey highlights that developers struggle with the multiple dependencies introduced by these libraries, contributing to dependency hell. For instance, a developer reported the cascading effect of patching a deeply nested dependency, requiring updates throughout the dependency tree. Jafari et al. [22] investigate the relationship between npm library characteristics and the dependency update strategy opted by its dependents. The authors report that the release status, the number of dependents, and the age of a library are the most important indicators of the dependency update strategy. Raemaekers et al. [31] discuss the risks associated with the usage of third-party libraries. They identify key library attributes that could serve as risk indicators. Notably, they report that more popular libraries may be updated more frequently, which increases the chance for new bugs to get introduced into the codebase. These findings support our claim that LLMs can be improved software librarians by providing critical information about recommended libraries, such as maintenance metrics (e.g., number of dependencies, version update frequency, age). Such transparency can help developers anticipate and manage potential maintenance challenges associated with the increased complexity of dependencies.

5.2 LLMs in Software Development Workflows

LLMs are increasingly used by software engineering practitioners to perform various tasks, such as code generation, and have shown the potential to improve developer productivity [11, 12, 39]. However, some studies have raised concerns about the reliability of LLM-generated code. For instance, Zhong et al. [40] report common API misuse patterns found in popular LLMs. The study reveals that in the case of GPT-4, 62% of the generated code contained API misuses. The authors argue that this is particularly problematic given that users of LLM code generation are generally not familiar with the APIs that LLMs generate code for, and cannot tell whether the provided code

is correct or not. Similarly, a Vulcan Cyber article claims that ChatGPT hallucinated in almost 40% of the programming questions it was asked [35]. Our findings diverge from those reported in the aforementioned studies. This discrepancy might be attributed to differences in experimental design. Zhong et al. employed "one-shot" or "few-shot" approaches, providing either irrelevant or relevant examples alongside the prompt. The Vulcan Cyber article describes a scenario where ChatGPT was instructed to generate code for a specific library and then repeatedly prompted for alternatives, potentially leading to hallucinated libraries. In contrast, our study aimed to simulate a more realistic developer-LLM interaction by prompting ChatGPT only once. This approach aligns with findings from Jin et al. [41] whose empirical study reveals that developers only request code regeneration from ChatGPT in 3% of conversations.

While LLMs are making significant advancements in software development, concerns regarding reliability and the potential for hallucinations remain. However, studies also highlight the potential of LLMs to improve developer productivity through functionalities like code completion and search. Ross et al. [11] developed an LLM-based programmer's assistant and evaluated their system on 42 participants. Their results reveal that participants were in majority positive in the assistant's potential for improving their productivity. Heitz et al. [12] evaluate and compare the performance of OpenAI's ChatGPT and Google's Gemini in programming code. They find that while the premium version of the models offers enhanced performance, their free counterparts remain highly relevant for a wide range of users in the context of software development. Also, the authors report that these models can significantly accelerate coding tasks and improve productivity, but necessitates rigorous, especially if the generated code is used in critical areas. Despite this, there is a gap in understanding how LLMs perform as software librarians, a critical role in the development process that impacts project maintainability, security, and compatibility.

Our study directly addresses this gap by investigating the effectiveness of LLMs in recommending libraries. We analyze libraries suggested by ChatGPT in response to real-world developer prompts from Stack Overflow questions. This allows us to assess LLM performance in a context that mimics actual developer usage and avoids potential biases introduced by experimental setups. By evaluating factors like library popularity and maintenance, licensing, and potential dependency challenges, our work aims to inform the development of more robust LLM software librarians, and provide developers with recommendations and considerations when using such tools to streamline their workflows.

6 THREATS TO VALIDITY

Internal validity considers the experimenter's bias and errors. Thus, a potential threat to the validity of our study is the manual categorization process of instances where a library import or installation resulted in a failure. To mitigate this, each case was discussed and reviewed by other authors and any inconsistencies were carefully addressed. Additionally, the parameters used when prompting the model can impact the generated responses. We addressed this by using ChatGPT's default parameters, mimicking a typical user interaction. Also, a time gap between the Stack Overflow questions and the generated code could introduce inconsistencies due to library changes over time. To mitigate this, we used the latest version of the libraries (as of February 2024) as a baseline for our analysis. Moreover, while the open coding process for categorizing the libraries was conducted by a single author, the labeling scheme was developed collaboratively and the results were validated through discussions with other authors. We opted not to use multiple coders or statistical measures like Cohen's kappa due to the objective nature of the labeling, which focused on factual classifications (library characteristics and usage) rather than subjective interpretations. Nevertheless, this decision may introduce a potential bias, although it was minimized through careful validation and clear, objective criteria. Finally, the stochastic nature of LLMs might lead

to variations in generated code for the same prompt. To enable the replicability of this study, we provide our dataset of generated code snippets.

External validity considers the generalizability of the findings. First, it is possible that the Stack Overflow code was part of ChatGPT's training data, which might influence its library recommendations by favoring libraries commonly used on that platform. To mitigate this, we compared the libraries used by ChatGPT to those used by human developers, allowing us to identify potential biases towards Stack Overflow-popular libraries. Next, our analysis is based on a sample of 10,000 Stack Overflow questions, which may not fully capture the diversity of real-world programming tasks and library usage patterns. To address this, we referred to the literature to identify typical developer interaction with ChatGPT to design our experimentation setup. Moreover, our focus on Python code might not translate to other programming languages. Despite the popularity of the Python language in software development, especially on Stack Overflow [42], investigating additional languages would be valuable to understand if these models exhibit similar or different behaviors. Finally, we only investigated the performance of one specific LLM (GPT 3.5 Turbo). Including a wider range of LLMs in future studies would provide a more comprehensive picture of how these models handle library recommendations.

7 CONCLUSION AND FUTURE WORK

This study investigated the effectiveness of LLMs as "software librarians" by analyzing libraries suggested by ChatGPT (specifically GPT-3.5 Turbo) for real-world coding problems derived from Stack Overflow questions. Our findings shed light on several key aspects of LLM library recommendations. First, ChatGPT uses third-party libraries 10% more often than human developers, and prioritized stable and well-established options. Additionally, ChatGPT tends to recommend libraries with few dependencies, which suggests a focus on more self-contained solutions. Interestingly, the libraries used by ChatGPT have a majority of permissive licenses, but a concerning amount of copyleft licenses, indicating that developers need to exercise extra caution in the way they use and distribute the generated code. As such, we recommend that tools like ChatGPT provide explicit information about licensing, providing alternatives to libraries with more restrictive usage.

Our analysis also revealed areas for improvement. We found that 6.5% of the libraries in ChatGPT-generated code led to import or installation failures and stemmed from aliases and implicit imports of modules. Moreover, two thirds of the problematic libraries were explicitly mentioned in the user question, indicating that ChatGPT generates code for libraries that may not be recognized or even existent. These issues can lead to developer confusion and frustration when encountering import or installation failures. We recommend that developers practice caution before integrating libraries recommended by ChatGPT into their codebase by cross-referencing the libraries with up-to-date resources such as official documentation. We also recommend that LLM programming assistants provide explicit information maintainability metrics of recommended libraries to guide developers in selecting and integrating appropriate libraries into their codebase. In the future, we plan to implement a tool that integrates with LLMs to perform checks on libraries and provide essential metadata (e.g., the number of dependencies and the version frequency) and license information. Additionally, we plan to replicate our study to include more models such as CodeT5 and Llama3.

We conclude this work by answering the question: Is ChatGPT a good software librarian? Our answer is a cautious yes.

REFERENCES

- [1] V.R. Basili, L. C. Briand, and W. L. Melo, "How reuse influences productivity in object-oriented systems," *Communications of the ACM*, vol. 39, no. 10, pp. 104–116, 1996.

- [2] A. Decan, T. Mens, and P. Grosjean, "An empirical comparison of dependency network evolution in seven software packaging ecosystems," *Empirical Software Engineering*, vol. 24, no. 1, pp. 381–416, 2019.
- [3] E. Murphy-Hill, C. Jaspán, C. Sadowski, D. Shepherd, M. Phillips, C. Winter, A. Knight, E. Smith, and M. Jorde, "What predicts software developers' productivity?" *IEEE Transactions on Software Engineering*, vol. 47, no. 3, pp. 582–594, 2019.
- [4] J. Latendresse, S. Mujahid, D. E. Costa, and E. Shihab, "Not All Dependencies are Equal: An Empirical Study on Production Dependencies in NPM," in *Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering (ASE'22)*, 2022.
- [5] S. Mujahid, "Effective dependency management for the javascript software ecosystem," Ph.D. dissertation, Concordia University, 2021.
- [6] Y. Wang, B. Chen, K. Huang, B. Shi, C. Xu, X. Peng, Y. Wu, and Y. Liu, "An empirical study of usages, updates and risks of third-party libraries in java projects," in *2020 IEEE International Conference on Software Maintenance and Evolution (ICSME)*. IEEE, 2020, pp. 35–45.
- [7] S. Mujahid, R. Abdalkareem, and E. Shihab, "What are the characteristics of highly-selected packages? a case study on the npm ecosystem," *Journal of Systems and Software*, vol. 198, p. 111588, 2023.
- [8] O. Hauge, T. Osterlie, C.-F. Sorensen, and M. Gereá, "An empirical study on selection of open source software-preliminary results," in *2009 ICSE Workshop on Emerging Trends in Free/Libre/Open Source Software Research and Development*. IEEE, 2009, pp. 42–47.
- [9] N. Haenni, M. Lungu, N. Schwarz, and O. Nierstrasz, "Categorizing developer information needs in software ecosystems," in *Proceedings of the 2013 international workshop on ecosystem architectures*, 2013, pp. 1–5.
- [10] J. T. Liang, C. Yang, and B. A. Myers, "A large-scale survey on the usability of ai programming assistants: Successes and challenges," in *Proceedings of the 46th IEEE/ACM International Conference on Software Engineering*, 2024, pp. 1–13.
- [11] S. I. Ross, F. Martinez, S. Houde, M. Muller, and J. D. Weisz, "The programmer's assistant: Conversational interaction with a large language model for software development," in *Proceedings of the 28th International Conference on Intelligent User Interfaces*, 2023, pp. 491–514.
- [12] L. B. Heitz, J. Chamas, and C. Scherb, "Evaluation of the programming skills of large language models," *arXiv preprint arXiv:2405.14388*, 2024.
- [13] "OpenAI," 2020. [Online]. Available: <https://openai.com/chatgpt/>
- [14] "Stack Overflow - where developers learn, share, & build careers." [Online]. Available: <https://stackoverflow.com/>
- [15] A. S. Saabith, T. Vinothraj, and M. Fareez, "Popular python libraries and their application domains," *International Journal of Advance Engineering and Research Development*, vol. 7, no. 11, 2020.
- [16] K. Srinath, "Python—the fastest growing programming language," *International Research Journal of Engineering and Technology*, vol. 4, no. 12, pp. 354–357, 2017.
- [17] V. Adamson and J. Bägerfeldt, "Assessing the effectiveness of chatgpt in generating python code," 2023.
- [18] M. E. Ellis, K. M. Casey, and G. Hill, "Chatgpt and python programming homework," *Decision Sciences Journal of Innovative Education*, vol. 22, no. 2, pp. 74–87, 2024.
- [19] Z. Yao, D. S. Weld, W.-P. Chen, and H. Sun, "Staqc: A systematically mined question-code dataset from stack overflow," in *Proceedings of the 2018 World Wide Web Conference*, 2018, pp. 1693–1703.
- [20] G. Destefanis, S. Bartolucci, and M. Ortu, "A preliminary analysis on the code generation capabilities of gpt-3.5 and bard ai models for java functions," *arXiv preprint arXiv:2305.09402*, 2023.
- [21] H. Hassani and E. S. Silva, "The role of chatgpt in data science: how ai-assisted conversational interfaces are revolutionizing the field," *Big data and cognitive computing*, vol. 7, no. 2, p. 62, 2023.
- [22] A. Javan Jafari, D. E. Costa, E. Shihab, and R. Abdalkareem, "Dependency update strategies and package characteristics," *ACM Trans. Softw. Eng. Methodol.*, vol. 32, no. 6, sep 2023. [Online]. Available: <https://doi.org/10.1145/3603110>
- [23] R. Cox, "Surviving software dependencies," *Communications of the ACM*, vol. 62, no. 9, pp. 36–43, 2019.
- [24] E. Larios Vargas, M. Aniche, C. Treude, M. Bruntink, and G. Gousios, "Selecting third-party libraries: the practitioners' perspective," in *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, ser. ESEC/FSE 2020. New York, NY, USA: Association for Computing Machinery, 2020, p. 245–256. [Online]. Available: <https://doi.org/10.1145/3368089.3409711>
- [25] J. Wu, L. Bao, X. Yang, X. Xia, and X. Hu, "A large-scale empirical study of open source license usage: Practices and challenges," 2024.
- [26] "PYPI · The Python Package Index." [Online]. Available: <https://pypi.org/>
- [27] "3.12.3 documentation." [Online]. Available: <https://docs.python.org/>
- [28] "Revisiting distribution name normalization," 12 2021. [Online]. Available: <https://discuss.python.org/t/revisiting-distribution-name-normalization/12348/3?u=abdur-rahmaanaj>
- [29] C. Precord, *WxPython application development cookbook*. Packt Publishing Ltd, 2015.

- [30] G. Marvin, N. Hellen, D. Jjing, and J. Nakatumba-Nabende, "Prompt engineering in large language models," in *International conference on data intelligence and cognitive informatics*. Springer, 2023, pp. 387–402.
- [31] S. Raemaekers, A. van Deursen, and J. Visser, "Exploring risks in the usage of third-party libraries," in *of the Belgian-Netherlands software eVOLUTION seminar*, vol. 31, 2011.
- [32] X. Chen, R. Abdalkareem, S. Mujahid, E. Shihab, and X. Xia, "Helping or not helping? why and how trivial packages impact the npm ecosystem," *Empirical Software Engineering*, vol. 26, pp. 1–24, 2021.
- [33] M. Hilton, T. Tunnell, K. Huang, D. Marinov, and D. Dig, "Usage, costs, and benefits of continuous integration in open-source projects," in *Proceedings of the 31st IEEE/ACM international conference on automated software engineering*, 2016, pp. 426–437.
- [34] J. Edmunds, PhD, "The standard library is all you need - Jordan Edmunds, PhD - Medium," 12 2023. [Online]. Available: <https://medium.com/@jordan.l.edmunds/the-standard-library-is-all-you-need-d5908628d001>
- [35] B. Lanyado, "Can you trust ChatGPT's package recommendations?" 3 2023. [Online]. Available: <https://vulcan.io/blog/ai-hallucinations-package-risk>
- [36] G. Fan, C. Wang, R. Wu, X. Xiao, Q. Shi, and C. Zhang, "Escaping dependency hell: finding build dependency errors with the unified dependency graph," in *Proceedings of the 29th ACM SIGSOFT International Symposium on Software Testing and Analysis*, 2020, pp. 463–474.
- [37] Y. Tanabe, T. Aotani, and H. Masuhara, "A context-oriented programming approach to dependency hell," in *Proceedings of the 10th ACM International Workshop on Context-Oriented Programming: Advanced Modularity for Run-time Composition*, 2018, pp. 8–14.
- [38] R. Abdalkareem, V. Oda, S. Mujahid, and E. Shihab, "On the impact of using trivial packages: An empirical case study on npm and pypi," *Empirical Software Engineering*, vol. 25, pp. 1168–1204, 2020.
- [39] T. Teubner, C. M. Flath, C. Weinhardt, W. van der Aalst, and O. Hinz, "Welcome to the era of chatgpt et al. the prospects of large language models," *Business & Information Systems Engineering*, vol. 65, no. 2, pp. 95–101, 2023.
- [40] L. Zhong and Z. Wang, "Can chatgpt replace stackoverflow? a study on robustness and reliability of large language model code generation," 2024.
- [41] K. Jin, C.-Y. Wang, H. V. Pham, and H. Hemmati, "Can chatgpt support developers? an empirical evaluation of large language models for code generation," *arXiv preprint arXiv:2402.11702*, 2024.
- [42] "Comparing tag trends with our Most Loved programming languages - Stack Overflow," 1 2023. [Online]. Available: <https://stackoverflow.blog/2023/01/26/comparing-tag-trends-with-our-most-loved-programming-languages/>