Final Project

An Empirical Study of the Maintainability of ChatGPT's answers

Thomas Trépanier

Montréal, Canada
thomas.trépanier@polytml.ca

***Abstract*—**

***Index Terms*—llm, maintainability, ChatGPT, software repository mining**

I. Introcution

II. Background & Related Work

A. Automatic Code Smell Detection

As early as 2009, work was done to automatically detect code smells in Python projects. Chen *et al.* proposed $Pysmell$ [1], a program that could detect 11 Python smells with up to 97.7% detection accuracy. However, the tool is too outdated for the purpose of this research, as it supports Python 2.7 only and most of the code snippets in the dataset are written in Python 3, which makes the tool incompatible with the dataset.

In 2021, Wang *et al.* proposed $PyNose$ [2], a tool that could detect 17 Python smells with up to 94.0% detection accuracy. However, their tool is not publicly available, so we were not able to use it for this research.

A tool found on Github named $PyScent$ was published in 2019 by Whyjai17 [3]. It is a Python program that uses a existing libraries such as pylint [4], pyflakes [5], radon [6] and cohesion [7] to detect 11 code smells in a Python project. It then generates a report in PDF format that contains the results of the analysis. Even though its accuracy has not been studied, following the results of $Pysmell$ and $PyNose$, it is safe to assume that its results are accurate enough for our research.

B. Code Quality of Stack Overflow's answers

In 2020, Meldrum *et. al.* [8] studied the code quality of Stack Overflow's answers for Java problems. They found that Stack Overflow's answers are not always devoid of code quality problems, finding that there are about 5 violations of *programming rule* per code snippet, and 10.5 readability violations per. They proposed that developers should be cautious when using code snippets from Stack Overflow, and that they should be reviewed before being used. This study aims at answering similar questions for Python related questions on Stack Overflow.

C. Relation between Code Smell and Software Quality

The ultimate goal of detection code smells in a project is to evaluate the quality of the software being produced. To this end, several studies have been conducted to determine the relation between code smells and software quality.

In 2012, Yamashita and Moonen [9] studied if code smells reflect important maintainability aspects. They found that code smells do reflect important maintainability features, espacially at the file level. Since we are working with code snippets of similar size to a file, it could mean that code smells found in code snippets impact the maintainability of the software they are integrated in.

Furthermore, in 2018, Spadini *et al.* [10] studied the relation between test smells and software quality. They found that test smells lead to more defect-prone production code. Knowing that ChatGPT and other AI code assistants, such as Github Copilot [11], are often relied on when writing unit tests, it is important to keep this result in mind as the code smells they generate could lead to more defects in the production code.

Finally, in 2020, Kaur [12] found that code smells may have opposit effects on software quality, meaning that some code smell can improve software qualities. This can be explained by code smells such as *Code Repetition*, which negatively impacts the maintainability of the code base, but can also lead to faster ande more performant code. This result is important to keep in mind, as it could mean that some code smells found in ChatGPT's answers could have a positive impact on the software quality of the code base they are integrated in.

III. Methodology

The methodology to answer *RQ1* and *RQ2* is mostly the same and can be divided into 5 steps: (1) data collection, (2) code snippet extraction, (3) code snippets filtering, (4) code smell detection, and (5) code smell analysis. The main difference between the two research questions is the data collection step. For *RQ1*, the data used is the DevGPT dataset provided by the MSR2024 challenge [13]. For *RQ2*, a dataset was created using StackOverflow's API following the methodology described in section III-A.

All of the code and algorithms to extract and process data was written in Python. The code is available on GitHub [1].

A. Data Collection

1) DevGPT Dataset: All of the provided snapshots were used in the study, as well as all of the JSON files in each of the snapshots. The JSON files contain information about the

[1]https://github.com/ThomasTrepanier/log6307-final-project

conversations between the developers and the ChatGPT bot.

2) StackOverflow Dataset: In order to build a dataset of StackOverflow's code snippets, the following approach was used:

1) **Fetching Questions from StackOverflow's API:** The first step was to query StackOverflow's API to get a list of Python related questions asked between November 23rd 2018 and November 23rd 2023. These dates were chosen as they represent a recent period of time that would correspond to the kind of code snippets found in the DevGPT dataset.

   Furthermore, the questions had to have at least five answers with one of them being accepted, and having been viewed 1000 times. These criteria were chosen to ensure that the questions were relevant and that the code snippets found in the answers would be of relatively good quality.

2) **Retrieving the answers to the questions:** The second step was to fetch the answers to the questions retrived in step 1. To do so, the *question_id* of each question was used to query StackOverflow's API and retrieve the answers associated with the question. The answers were then stored in a JSON file.

B. Code Snippet Extraction

1) DevGPT Dataset: The DevGPT dataset contains conversations between developers and the ChatGPT bot. Each JSON file contains a data structure called `ChatgptSharing`, itself containing a list of `Conversation`. Each `Conversation` contains the prompt given by the developer, the answer given by ChatGPT, and a data structure called `ListOfCode` containing the code snippets included in ChatGPT's answer.

This `ListOfCode` itself provides the code snippet emitted by ChatGPT as well as the programming language used in the snippet.

In order to build a dataset of all of the code snippets generated by ChatGPT, a program was written to parse all of the JSON files of all of the snapshots and extract all the `ListOfCode` data structures.

2) StackOverflow Dataset: The answers retrieved from StackOverflow's API consisted of HTML bodies containing HTML elements to delimit the code snippet. Therefore, the HTML bodies were parsed using the following Regex to extract the code snippet itself: `<pre><code>(.*?)</code></pre>`.

The snippet was then parsed from its HTML format to a Unicode format and stored in a JSON file.

C. Code Snippets Filtering

Not all code snippets extracted were used for the code smell detection, since not all of them were suited for the analysis. The definition of a *suitable* code snippet is as follows:

1) The code snippet must be in Python. This is because the code smell detection tool used, *PyScent*, is a Python tool.

2) The code snippet must be at least 5 lines long, and contain either a function (defined with the `def` keyword) or a class (defined with the `class` keyword). This is to filter out short answers or code snippets that were not actually Python code, such as JSON objects or library import statements.

3) The code snippet must be written entirely in english. This is because non-unicode characters are not supported by *PyScent*.

4) The code snippet must not contain any syntax errors. This is because *PyScent* is not able to detect code smells in code that does not compile. To remove the snippets with syntax errors, a program was written to delete export each code snippet to a `.py` file and try to compile it. If a `SyntaxError` or `TypeError` was thrown by Python, the file was assumed to be erronous and deleted from the dataset.

D. Code Smell Detection

The last step was to detect code smells in the code snippets. To do so, the *PyScent* tool was used. *PyScent* is a Python tool that detects code smells in Python code and the source code can be found on Github [2].

PyScent detects the 11 following code smells:

- **Long Method:** A method with over 50 statements.
- **Long Parameter List:** A method that takes more than 5 arguments.
- **Long Branches:** An if statement with more than 12 branches.
- **Too Many Attributes:** A class that has more than 7 attributes.
- **Too Many Methods:** A class that has more than 20 public methods.
- **Useless Try/Except Clauses:** A try/except clause that catches too many exception (using the `Exception` or `Standard Error`), or whose catch clauses are empty.
- **Shotgun Surgery:** A Class that calls more than 5 methods from other classes.
- **Class Cohesion:** A class that has cohesion score lower 30%. The cohesion score is computed using the Cohesion project on Gitbut [3].

[2]https://github.com/whyjay17/Pyscent
[3]https://github.com/mschwager/cohesion

- **Code Complexity:** A block that has a cyclomatic complexity score lower than 'C'. The cyclomatic complexity score is computed using the Radon project [4].
- **Long Lambda:** A lambda function that has more than 60 characters.
- **Long List Comprehension:** A list comprehension that has more than 72 characters.

To analyse the code snippets, *PyScent* needs them to be in `.py` files under one directory. Therefore, a program was written to export all of the code snippets to `.py` files and run *PyScent* on them.

The results of the code smell detection are presented in section IV.

IV. Results

In this section, we first present the result of the code smell analysis performed by Pyscent in section IV-A, then, we compare the results between ChatGPT and Stack Overflow in IV-B.

A. Code Smell Results

Following the code snippet analysis by Pyscent, a report was generated detailling the number of smells detected for each Code Smell described in section III. The report also presents the number of possible occurences of each Code Smell, so for example, there were 1247 methods detected in ChatGPT's code snippets, and 9 of them were detected as Long Parameter.

Furthermore, to be able to compare the results between ChatGPT and Stack Overflow, the percentage of each Code Smell was calculated. This percentage is the number of smells detected divided by the number of possible occurences of the smell.

The results of the analysis are presented in tables I and II.

TABLE I
ChatGPT's Code Smells

Smell Name	Smell Count	# of poss. occ.	%
Long Methods	0	1247	0.000%
Long Parameter	9	1247	0.722%
Long Branches	1	N/A	N/A
Too Many Attribute	11	250	4.400%
Too Many Methods	0	250	0.000%
Useless try/catch	1	101	0.990%
Shotgun Surgery	175	250	70.000%
Low Class Cohesion	0	250	0.000%
Code Complexity (< C)	17	1497	1.136%
Long Lambda	0	43	0.000%
Long List Comprehension	16	59	27.119%

For ChatGPT, we can see that the most common code smell is Shotgun Surgery, with a 70% occurence rate. This is not surprising, since ChatGPT is not aware of the code context of the developer. Therefore, it will often generate code snippets that are not related to the context.

We can also observe that ChatGPT tends to write long list comprehension. This might be because the training data it was fed contained a lot of long list comprehension. However, it does not seem to generate long lambdas, which might mean that it is aware of this code smell.

Finally, we can see that ChatGPT does not generate long methods, which could mean it is also aware of this code smell, but it could also be because ChatGPT does not tend to write long code snippets in general, so the code smell cannot appear.

TABLE II
Stack Overflow's Code Smells

Smell Name	Smell Count	# of poss. occ.	%
Long Methods	5	3438	0.145%
Long Parameter	18	3438	0.524%
Long Branches	14	N/A	N/A
Too Many Attribute	4	592	0.676%
Too Many Methods	0	592	0.000%
Useless try/catch	19	204	9.314%
Shotgun Surgery	376	592	63.514%
Low Class Cohesion	14	592	2.365%
Code Complexity (< C)	41	4030	1.017%
Long Lambda	19	146	13.014%
Long List Comprehension	56	439	12.756%

From Table II, we can see that Stack Overflow's code snippets also tend to have a lot of Shotgun Surgery and long list comprehension, but not as much as ChatGPT. This is probably because Stack Overflow's code snippets are written by humans, who are more aware of those Code Smells.

However, Stack Overflow's code snippet present a higher occurence of useless try/catch and long lambdas compared to ChatGPT's. The first one might be because the developer does not put a lot of thought into the type of error handled by the try/catch to save time, and the second one could be because they want to impress the reader with a complex lambda.

Finally, it might be surprising to see that Stack Overflow's code snippets present Code Smells that are not present in any of ChatGPT's, such as long methods, low class cohesion and long lambdas, since we would expect developers to be more aware of Code Smells than ChatGPT.

B. Code Smell Comparison

Table III presents the comparison between ChatGPT and Stack Overflow's code smell percentages where the difference is significative ($> 0.5\%$). The percentage difference is calculated by substracting the percentage of Stack Overflow's code smells from the percentage of ChatGPT's code smells.

Therefore, a positive value means that there are more code smells of this type in Stack Overflow's code snippets, whereas a negative value means that ChatGPT uses this code smell more often.

Finally, an overall percentage difference is computed using the sum of the percentage difference for all of the code smells.

From Table III, we can see that ChatGPT's code snippets present more code smells than Stack Overflow's, with an overall difference of -1.04%. This means that the code quality of ChatGPT is more or less similar to that of Stack Overflow.

[4]https://pypi.org/project/radon/

TABLE III
SO VS. CHATGPT CODE SMELLS

Smell Name	SO (%)	ChatGPT %	difference (%)
Too Many Attribute	0.68%	4.40%	-3.72%
Useless try/catch	9.31%	0.99%	8.32%
Shotgun Surgery	63.51%	70.00%	-6.49%
Low Class Cohesion	2.37%	0%	2.37%
Long Lambda	13.01%	0%	13.01%
Long List Comprehension	12.76%	27.12%	-14.36%
Overall			**-1.04%**

V. DISCUSSION

A. RQ#1: Code Smells in ChatGPT's answers

It is not surprising to find that ChatGPT produces code smells, as any code base is bound to present some at some point. However, it is interesting to note that it does seem to have an understanding of some code smells and tries to avoid them, such as long methods, long lambdas and Too Many Methods. As discussed before, long methods may be absent because of the short nature of ChatGPT's answer, but as we saw with the Stack Overflow analysis, long methods can still appear in code snippets.

The absence of the Too Many Methods code smell could also be explained by the nature of ChatGPT's answers, as well as the nature of Python. ChatGPT's answers are short, and Python is not an object-oriented language. Therefore, the number of methods in ChatGPT's answers is bound to be low. It is interesting to note however that $Pyscent$ found 11 occurences of the Too Many Attributes code smell, meaning that code smells related to OOP can still be found in ChatGPT's answers.

As for the absence of long lambdas, it is possible that the relatively low number of lambda functions in the dataset led to its absence. However, it is also possible that ChatGPT has a better understanding of the long lambda code smell than it does of other code smells, and therefore tries to avoid it.

It would be difficult nonetheless to identify if ChatGPT has a real understanding of Code Smells as it learned the concept on the internet, or if because the way it was trained to generate code did not include such code smells.

B. RQ#2: SO vs. ChatGPT Code Smells

What could be surprising however is the finding that ChatGPT does not seem to produce a significantly higher number of code smells than Stack Overflow.

In fact, ChatGPT's answer seem to be characterized by a higher amount of the Too Many Attribute, Shotgun Surgery and Long List Comprehension code smells, whereas Stack Overflow's answers are characterized by the presence of useless try/catch, low class cohesion and long lambdas.

What this finding really reminds us however, is that developers should always be careful when integrating exterior code in their projects and should adapt it to meet their standards and practices.

VI. THREATS TO VALIDITY

VII. CONCLUSION

REFERENCES

[1] Z. Chen, L. Chen, W. Ma and B. Xu, "Detecting Code Smells in Python Programs," 2016 International Conference on Software Analysis, Testing and Evolution (SATE), Kunming, China, 2016, pp. 18-23, doi: 10.1109/SATE.2016.10.

[2] T. Wang, Y. Golubev, O. Smirnov, J. Li, T. Bryksin and I. Ahmed, "PyNose: A Test Smell Detector For Python," 2021 36th IEEE/ACM International Conference on Automated Software Engineering (ASE), Melbourne, Australia, 2021, pp. 593-605, doi: 10.1109/ASE51524.2021.9678615.

[3] Whyjai17, " Pyscent". 2019. https://github.com/whyjay17/Pyscent

[4] Pylint 2.3.1, https://pypi.org/project/pylint/2.3.1/

[5] Pyflakes 2.1.1, https://pypi.org/project/pyflakes/2.1.1/

[6] Radon 3.0.1, https://pypi.org/project/radon/3.0.1/

[7] Cohesion 0.9.1, https://pypi.org/project/cohesion/0.9.1/

[8] S. Meldrum, S. A. Licorish, C. A. Owen, B. T. Roy Savarimuthu,, "Understanding stack overflow code quality: A recommendation of caution", Science of Computer Programming, Volume 199, 2020, 102516, doi: https://doi.org/10.1016/j.scico.2020.102516.

[9] A. Yamashita and L. Moonen, "Do code smells reflect important maintainability aspects?," 2012 28th IEEE International Conference on Software Maintenance (ICSM), Trento, Italy, 2012, pp. 306-315, doi: 10.1109/ICSM.2012.6405287.

[10] D. Spadini, F. Palomba, A. Zaidman, M. Bruntink and A. Bacchelli, "On the Relation of Test Smells to Software Code Quality," 2018 IEEE International Conference on Software Maintenance and Evolution (ICSME), Madrid, Spain, 2018, pp. 1-12, doi: 10.1109/ICSME.2018.00010.

[11] Github Copilot, https://github.com/features/copilot

[12] Kaur, A. "A Systematic Literature Review on Empirical Analysis of the Relationship Between Code Smells and Software Quality Attributes," Arch Computat Methods Eng 27, 1267-1296 (2020). https://doi.org/10.1007/s11831-019-09348-6

[13] NAIST-SE. "DevGPT: Studying Developer-ChatGPT Conversations". MSR-2024 Challenge. https://github.com/NAIST-SE/DevGPT