RepoAgent: An LLM-Powered Open-Source Framework for Repository-level Code Documentation Generation

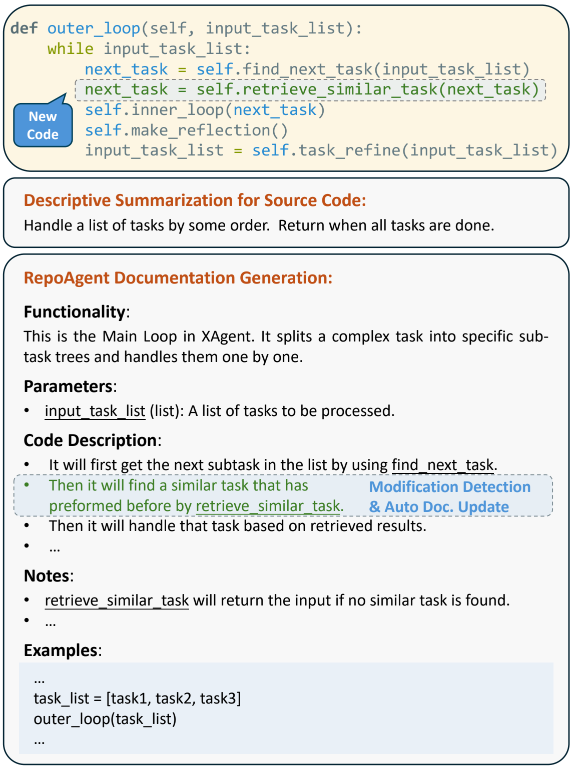
Qinyu Luo[1*], **Yining Ye**[1*], **Shihao Liang**[1], **Zhong Zhang**[1†],
Yujia Qin[1], **Yaxi Lu**[1], **Yesai Wu**[1], **Xin Cong**[1], **Yankai Lin**[2]
Yingli Zhang[3], **Xiaoyin Che**[3], **Zhiyuan Liu**[1†], **Maosong Sun**[1]
[1]Tsinghua University [2]Renmin University of China [3]Siemens AG
`qinyuluo123@gmail.com, yeyn2001@gmail.com`

Abstract

Generative models have demonstrated considerable potential in software engineering, particularly in tasks such as code generation and debugging. However, their utilization in the domain of code documentation generation remains underexplored. To this end, we introduce RepoAgent, a large language model powered open-source framework aimed at proactively generating, maintaining, and updating code documentation. Through both qualitative and quantitative evaluations, we have validated the effectiveness of our approach, showing that RepoAgent excels in generating high-quality repository-level documentation. The code and results are publicly accessible at `https://github.com/OpenBMB/RepoAgent`.

1 Introduction

Developers typically spend approximately 58% of their time on program comprehension, and high-quality code documentation plays a significant role in reducing this time (Xia et al., 2018; de Souza et al., 2005). High-quality documentation significantly lowers the learning curve for new project members, thereby accelerating their contributions and fostering a vibrant open-source community through enhanced participation and collaboration. However, maintaining code documentation also consumes a considerable amount of time, money, and human labor (Zhi et al., 2015), and not all projects have the resources or enthusiasm to prioritize documentation as their top concern.



```
def outer_loop(self, input_task_list):
    while input_task_list:
        next_task = self.find_next_task(input_task_list)
        next_task = self.retrieve_similar_task(next_task)
        self.inner_loop(next_task)
        self.make_reflection()
        input_task_list = self.task_refine(input_task_list)
```

Descriptive Summarization for Source Code:

Handle a list of tasks by some order. Return when all tasks are done.

RepoAgent Documentation Generation:

Functionality:

This is the Main Loop in XAgent. It splits a complex task into specific subtask trees and handles them one by one.

Parameters:

- input_task_list (list): A list of tasks to be processed.

Code Description:

- It will first get the next subtask in the list by using find_next_task.
- Then it will find a similar task that has preformed before by retrieve_similar_task.
- Then it will handle that task based on retrieved results.
- …

Notes:

- retrieve_similar_task will return the input if no similar task is found.
- …

Examples:

```
…
task_list = [task1, task2, task3]
outer_loop(task_list)
…
```

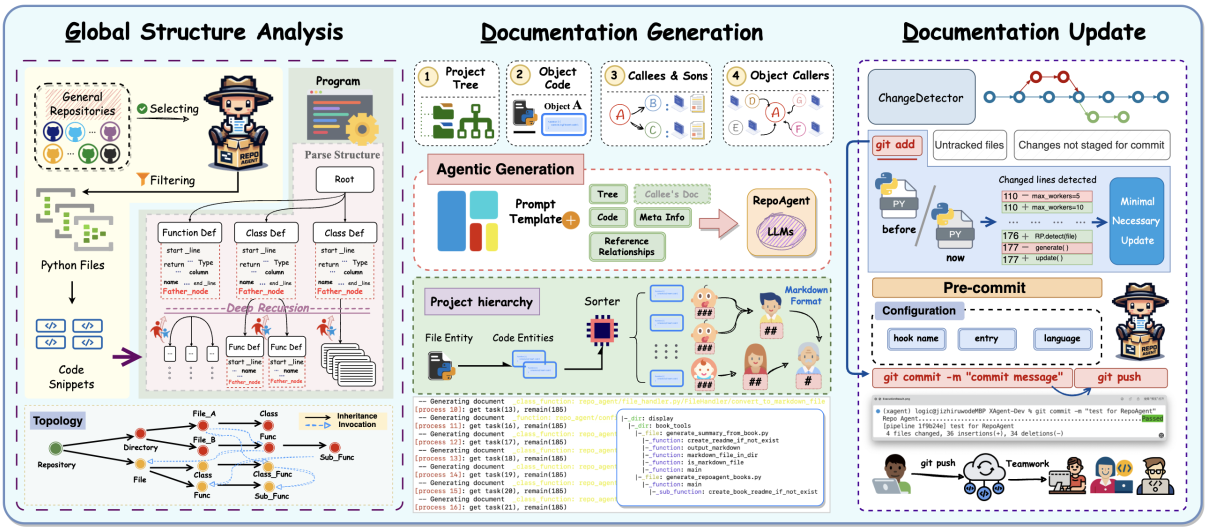
Figure 1: The comparison of code documentation generated by the plain summarization method and the newly proposed RepoAgent .

To alleviate the burden of maintaining code documentation, early attempts at automatic documentation generation aimed to provide descriptive summaries for source code (Sridhara et al., 2010; Rai et al., 2022; Khan and Uddin, 2022; Zhang et al., 2022), as illustrated in Figure 1. However, they still have significant limitations, particularly in the following aspects: **(1) Poor summarization**. Previous methods primarily focused on summarizing isolated code snippets, overlooking the dependencies of code within the broader repository-level context. The generated code summaries are overly abstract and fragmented, making it difficult to accurately convey the semantics of the code and compile the code summaries into documentation. **(2) Inadequate guidance**. Good documentation not only accurately describes the code's functionality, but also meticulously guides developers on the correct usage of the described code (Khan and Uddin, 2022; Wang et al., 2023). This includes, but is not limited to, clarifying functional boundaries, highlighting potential misuses, and presenting examples of inputs and outputs. Previous methods still fall short of offering such comprehensive guidance. **(3) Passive update**. Lehman's first law of software evolution states that a program in use will continuously evolve to meet new user needs (Lehman, 1980). Consequently, it is crucial for the documentation to be updated in a timely manner to align with code changes, which is the capability that previous methods overlook. Recently, Large Language Models (LLMs) have made significant progress (OpenAI, 2022, 2023), especially in the code

---

[*] Indicates equal contribution.
[†] Corresponding Author.
Demo Video: `https://youtu.be/YPPJBVOP71M`

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing
System Demonstrations, pages 436–464
November 12-16, 2024 ©2024 Association for Computational Linguistics



```
 -- Generating document  _class_function: repo_agent/file_handler.py/FileHandler/convert_to_markdown_file
[process 10]: get task(13), remain(185)
 -- Generating document  _function: repo_agent/confi
[process 11]: get task(16), remain(185)
 -- Generating document  _class_function: repo_agent
[process 12]: get task(17), remain(185)
 -- Generating document  _class_function: repo_agent
[process 13]: get task(18), remain(185)
 -- Generating document  _class_function: repo_agent
[process 14]: get task(19), remain(185)
 -- Generating document  _class_function: repo_agent
[process 15]: get task(20), remain(185)
 -- Generating document  _class_function: repo_agent
[process 16]: get task(21), remain(185)
```

```
|-_dir: display
  |-_dir: book_tools
    |-_file: generate_summary_from_book.py
      |-_function: create_readme_if_not_exist
      |-_function: output_markdown
      |-_function: markdown_file_in_dir
      |-_function: is_markdown_file
      |-_function: main
    |-_file: generate_repoagent_books.py
      |-_function: main
        |-_sub_function: create_book_readme_if_not_exist
```

```
110 − max_workers=5
110 + max_workers=10
…   …   …   …
176 + RP.detect(file)
177 − generate( )
177 + update( )
```

```
● (xagent) logic@jizhiruwodeMBP XAgent-Dev % git commit -m "test for RepoAgent"
  Repo Agent.................................................................Passed
  [pipeline 1f9b24e] test for RepoAgent
   4 files changed, 36 insertions(+), 34 deletions(-)
```

Figure 2: The RepoAgent method consists of Global Structure Analysis, Documentation Generation, and Documentation Update. Each component can be executed independently or packaged as a hook for tooling purposes. When operating as a whole, RepoAgent ensures the capability to construct and maintain documentation for a repository from scratch, elevating documentation to the same level of importance as code, facilitating synchronization and collaboration among teams.

understanding and generation realm (Nijkamp et al., 2023; Li et al., 2023; Chen et al., 2021; Rozière et al., 2023; Xu et al., 2024; Sun et al., 2023; Wang et al., 2023; Khan and Uddin, 2022). Given these advancements, it is natural to ask: Can LLMs be used to generate and maintain repository-level code documentation, addressing the aforementioned limitations?

In this study, we introduce REPOAGENT, the first framework powered by LLMs, designed to proactively generate and maintain comprehensive documentation for the entire repository. A running example is demonstrated in Figure 1. REPOAGENT offers the following features: **(1) Repository-level documentation:** REPOAGENT leverages the global context to deduce the functional semantics of target code objects within the entire repository, enabling the generation of accurate and semantically coherent structured documentation. **(2) Practical guidance:** REPOAGENT not only describes the functionality of the code but also provides practical guidance, including notes for code usage and examples of input and output, thereby facilitating developers' swift comprehension of the code repository. **(3) Maintenance automation:** REPOAGENT can seamlessly integrate into team software development workflows managed with Git and proactively take over documentation maintenance, ensuring that the code and documentation remain synchronized. The process is fully automated, leveraging advanced algorithms to ensure it accommodates all conceivable Git operations and change scenarios, thereby eradicating the necessity for manual updates and significantly diminishing the risk of human error.

We qualitatively showcased the code documentation generated by REPOAGENT for real Python repositories. The results reveal that REPOAGENT is adept at producing documentation of a quality comparable to that created by humans. Quantitatively, in two blind preference tests, the documentation generated by REPOAGENT was favored over human-authored documentation, achieving preference rates of 70% and 91.33% on the Transformers and LlamaIndex repositories, respectively. These evaluation results indicate the practicality of the proposed REPOAGENT in automatic code documentation generation.

2 RepoAgent

REPOAGENT consists of three key stages: **global structure analysis**, **documentation generation**, and **documentation update**. Figure 2 shows the overall design of REPOAGENT. The **global structure analysis** stage involves parsing necessary meta information and global contextual relationships from the source code, laying the foundation for REPOAGENT to infer the functional semantics of the target code. In the **documentation generation** stage, we have designed a sophisticated strategy that leverages the parsed meta information and global contextual relationships to prompt the LLM to generate fine-grained documentation that is of practical guidance. In the **documentation update** stage, REPOAGENT utilizes Git tools to track code changes and update the documentation accordingly, ensuring that the code and documentation remain synchronized throughout the entire project lifecycle.

2.1 Global Structure Analysis

An essential prerequisite for generating accurate and fine-grained code documentation is a comprehensive understanding of the code structure. To achieve this goal, we proposed a project tree, a data structure that maintains all code objects in the repository while preserving their semantic hierarchical relationships. Firstly, we filter out all non-Python files within the repository. For each Python file, we apply Abstract Syntax Tree

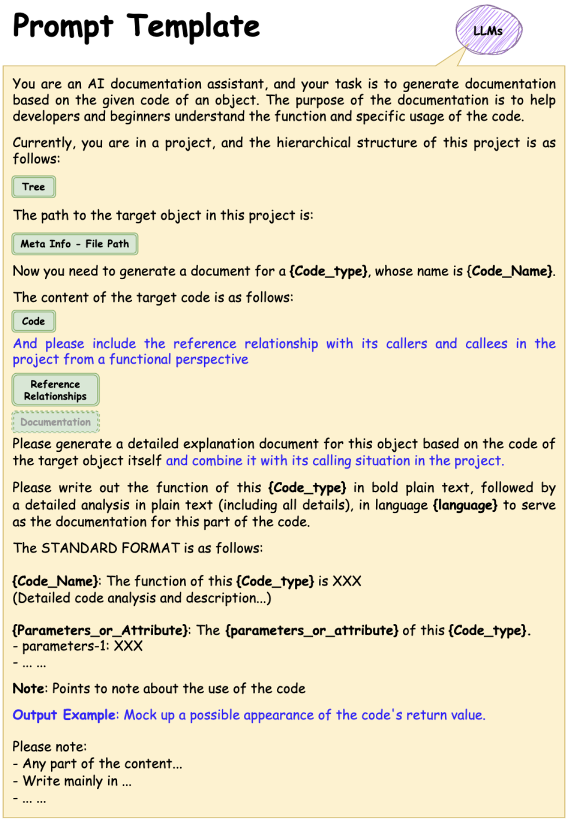
(AST) analysis (Zhang et al., 2019) to recursively parse the meta information of all Classes and Functions within the file, including their type, name, code snippets, etc. These Classes and Functions associated with their meta information are used as the atomic objects for documentation generation. It is worth noting that the file structures of most well-engineered repositories have reflected the functional semantics of code. Therefore, we first utilize it to initialize the project tree, whose root node represents the entire repository, middle nodes and leaf nodes represent directories and Python files, respectively. Then, we add the parsed Classes and Functions as new leaf nodes (or sub-trees) to the corresponding Python file nodes to form the final project tree.

Beyond the code structure, the reference relationships within the code, as a form of important global contextual information, can also assist the LLM in identifying the functional semantics of the code. Also, references to a target function can be considered natural in-context learning examples (Wei et al., 2022) to teach the LLM to use the target function, thereby helping generate documentation that is of practical guidance. We consider two types of reference relationships: `Caller` and `Callee`. We use the Jedi library[1] to extract all bi-directional reference relationships in the repository, and then ground them to the corresponding leaf nodes in the project tree. The project tree augmented with the reference relationships forms a Directed Acyclic Graph[2] (DAG).

2.2 Documentation Generation

RepoAgent aims to generate fine-grained documentation that is of practical guidance, which includes detailed **Functionality**, **Parameters**, **Code Description**, **Notes**, and **Examples**. A backend LLM leverages the parsed meta information and reference relationships from the previous stage to generate documentation with the required structure using a carefully designed prompt template. An illustrative prompt template is shown in Figure 3, and a complete real-world prompt example is given in Appendix C.1.

The prompt template mainly requires the following parameters: The **Project Tree** helps RepoAgent perceive the repository-level context. The **Code Snippet** serves as the main source of information for RepoAgent to generate the documentation. The **Reference Relationships** provide semantic invocation relationships between code objects and assist RepoAgent in generating guiding notes and examples. The **Meta Information** indicates the necessary information such as type, name, relative file path of the target object, and is used for post-processing of the documentation. Additionally, we can include the previously generated **Documentation** of a direct child node of an object as auxiliary information to help code understanding. This is optional, as omitting it can save costs significantly.



Prompt Template

LLMs

You are an AI documentation assistant, and your task is to generate documentation based on the given code of an object. The purpose of the documentation is to help developers and beginners understand the function and specific usage of the code.

Currently, you are in a project, and the hierarchical structure of this project is as follows:

Tree

The path to the target object in this project is:

Meta Info - File Path

Now you need to generate a document for a **{Code_type}**, whose name is **{Code_Name}**.

The content of the target code is as follows:

Code

And please include the reference relationship with its callers and callees in the project from a functional perspective

Reference Relationships

Documentation

Please generate a detailed explanation document for this object based on the code of the target object itself and combine it with its calling situation in the project.

Please write out the function of this **{Code_type}** in bold plain text, followed by a detailed analysis in plain text (including all details), in language **{language}** to serve as the documentation for this part of the code.

The STANDARD FORMAT is as follows:

{Code_Name}: The function of this **{Code_type}** is XXX
(Detailed code analysis and description...)

{Parameters_or_Attribute}: The **{parameters_or_attribute}** of this **{Code_type}**.
- parameters-1: XXX
- ... ...

Note: Points to note about the use of the code

Output Example: Mock up a possible appearance of the code's return value.

Please note:
- Any part of the content...
- Write mainly in ...
- ... ...

Figure 3: Prompt template used for documentation generation, some details are omitted. Variables within the braces are assigned according to different objects. The blue parts are dynamically filled based on the Meta Info of different objects, enriching the documentation content according to the object characteristics. The Documentation within the dashed boxes can be dynamically utilized according to the program settings. If the documentation information is not used, the program may not execute in topological order.

RepoAgent follows a bottom-to-top topological order to generate documentation for all code objects in the DAG, ensuring that the child nodes of each node, as well as the nodes it references, have their documentation generated before it. After the documentation is generated, RepoAgent compiles it into a human-friendly Markdown format. For example, objects of different levels are associated with different Markdown headings (e.g., ##, ###). Finally, RepoAgent utilizes GitBook[3] to render the Markdown formatted documentation into a convenient web graphical interface, which enables easy navigation and readability for documentation readers.

2.3 Documentation Update

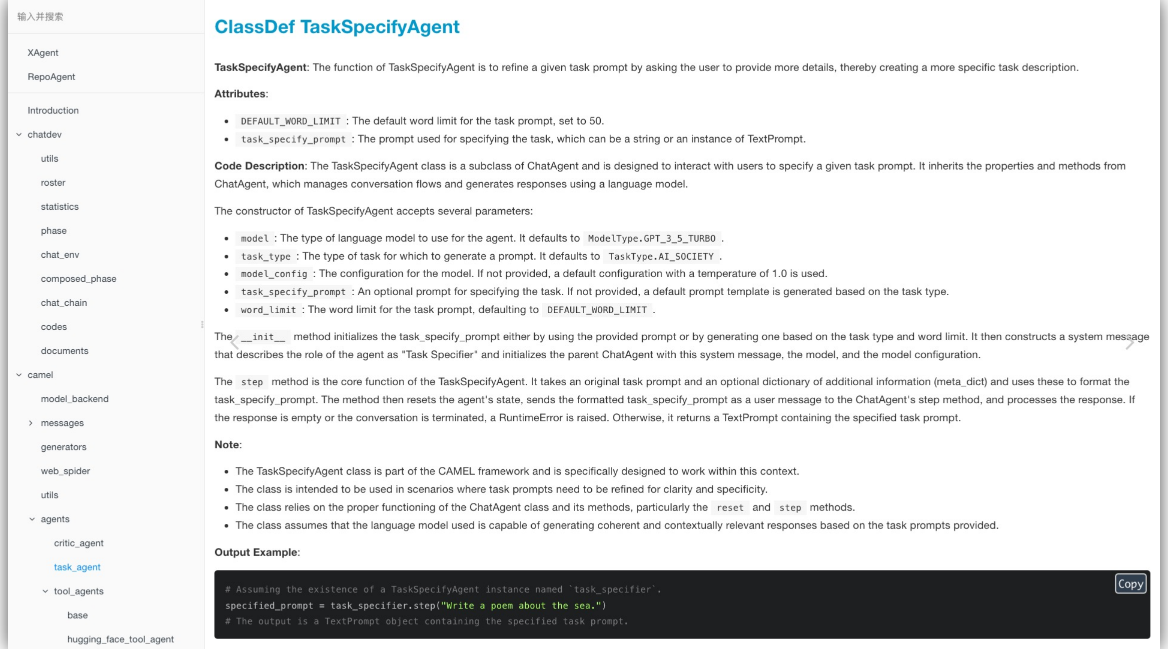
RepoAgent supports automatic tracking and updating of documentation through seamless collaboration with Git. The pre-commit hook of Git is utilized to enable RepoAgent to detect any code changes and perform documentation updates. After the update, the hook sub-

---

[1] `https://github.com/davidhalter/jedi` Extensible to programming languages other than Python by replacing code parsing tools.

[2] We simply ignored circular dependencies to avoid loops, as most of these situations may have bugs.

[3] `https://www.gitbook.com/`



ClassDef TaskSpecifyAgent

TaskSpecifyAgent: The function of TaskSpecifyAgent is to refine a given task prompt by asking the user to provide more details, thereby creating a more specific task description.

Attributes:

- `DEFAULT_WORD_LIMIT` : The default word limit for the task prompt, set to 50.
- `task_specify_prompt` : The prompt used for specifying the task, which can be a string or an instance of TextPrompt.

Code Description: The TaskSpecifyAgent class is a subclass of ChatAgent and is designed to interact with users to specify a given task prompt. It inherits the properties and methods from ChatAgent, which manages conversation flows and generates responses using a language model.

The constructor of TaskSpecifyAgent accepts several parameters:

- `model` : The type of language model to use for the agent. It defaults to `ModelType.GPT_3_5_TURBO` .
- `task_type` : The type of task for which to generate a prompt. It defaults to `TaskType.AI_SOCIETY` .
- `model_config` : The configuration for the model. If not provided, a default configuration with a temperature of 1.0 is used.
- `task_specify_prompt` : An optional prompt for specifying the task. If not provided, a default prompt template is generated based on the task type.
- `word_limit` : The word limit for the task prompt, defaulting to `DEFAULT_WORD_LIMIT` .

The `__init__` method initializes the task_specify_prompt either by using the provided prompt or by generating one based on the task type and word limit. It then constructs a system message that describes the role of the agent as "Task Specifier" and initializes the parent ChatAgent with this system message, the model, and the model configuration.

The `step` method is the core function of the TaskSpecifyAgent. It takes an original task prompt and an optional dictionary of additional information (meta_dict) and uses these to format the task_specify_prompt. The method then resets the agent's state, sends the formatted task_specify_prompt as a user message to the ChatAgent's step method, and processes the response. If the response is empty or the conversation is terminated, a RuntimeError is raised. Otherwise, it returns a TextPrompt containing the specified task prompt.

Note:

- The TaskSpecifyAgent class is part of the CAMEL framework and is specifically designed to work within this context.
- The class is intended to be used in scenarios where task prompts need to be refined for clarity and specificity.
- The class relies on the proper functioning of the ChatAgent class and its methods, particularly the `reset` and `step` methods.
- The class assumes that the language model used is capable of generating coherent and contextually relevant responses based on the task prompts provided.

Output Example:

```
# Assuming the existence of a TaskSpecifyAgent instance named `task_specifier`.
specified_prompt = task_specifier.step("Write a poem about the sea.")
# The output is a TextPrompt object containing the specified task prompt.
```

Figure 4: Demonstration of code documentation generated by REPOAGENT for the ChatDev repository.

mits both the code and documentation changes, ensuring that the code and documentation remain synchronized. This process is fully automated and does not require human intervention.

Local code changes generally do not affect other code due to the low coupling principle, it is not necessary to regenerate the entire documentation with each minor code update. REPOAGENT only updates the documentation of affected objects. The updates are triggered when (1) an object's source code is modified; (2) an object's referrers no longer reference it; or (3) an object gets new references. It is worth noting that the update is not triggered when an object's reference objects change, because we adhere to the dependency inversion principle (Martin, 1996), which states that high-level modules should not depend on the implementations of low-level modules.

3 Experiments

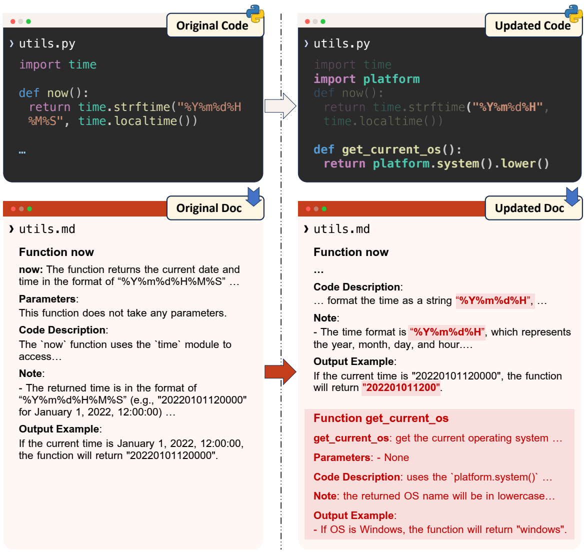
3.1 Experimental Settings

For the purpose of generating documentation, we selected 9 Python repositories, spanning a wide range of scales from less than 1,000 to over 10,000 lines of code. This selection encompasses both well-established projects with significant followings and newly emerged ones that have quickly gained recognition on GitHub for their quality. Distinguished by their high-quality code and considerable complexity, these repositories are meticulously characterized by various metrics, including the number of lines of code, classes, and functions. The detailed statistics of the repositories are provided in Appendix A.1. We adopted the API-based LLMs `gpt-3.5-turbo` (OpenAI, 2022) and `gpt-4-0125` (OpenAI, 2023), along with the open-source LLMs `Llama-2-7b` and `Llama-2-70b` (Touvron et al., 2023) as backend models for REPOAGENT .

3.2 Case Study

We use the ChatDev repository (Qian et al., 2023) and the `gpt-4-0125` backend for a case study. The generated documentation is illustrated in Figure 4. Documentation generated by REPOAGENT is structured into several parts, starting with a clear, concise sentence that articulates the object's functionality. Following this, the parameters section enumerates all relevant parameters along with their descriptions, aiding developers in understanding how to leverage the provided code. Moreover, the code description section comprehensively elaborates on all aspects of the code, implicitly or explicitly demonstrating the object's role and its associations with other code within the global context. In addition, the notes section further enriches these descriptions by covering usage considerations for the object at hand. Notably, it highlights any logical errors or potential optimization within the code, thereby prompting advanced developers to make modifications. Lastly, if the current object yields a return value, the model will generate an examples section, filled with simulated content to clearly demonstrate the expected output. This is highly advantageous for developers, facilitating efficient code reuse and unit test construction.

Once the code is changed, the documentation update will be triggered, as illustrated in Figure 5. Upon code changes in the staging area, REPOAGENT identifies affected objects and their bidirectional references, up-



```
› utils.py
  import time

  def now():
    return time.strftime("%Y%m%d%H
    %M%S", time.localtime())

  …
```

```
› utils.py
  import time
  import platform
  def now():
    return time.strftime("%Y%m%d%H",
    time.localtime())

  def get_current_os():
    return platform.system().lower()
```

› utils.md

Function now

now: The function returns the current date and time in the format of “%Y%m%d%H%M%S” …

Parameters:
This function does not take any parameters.

Code Description:
The `now` function uses the `time` module to access…

Note:
- The returned time is in the format of “%Y%m%d%H%M%S” (e.g., "20220101120000" for January 1, 2022, 12:00:00) …

Output Example:
If the current time is January 1, 2022, 12:00:00, the function will return "20220101120000".

› utils.md

Function now

…

Code Description:
… format the time as a string **“%Y%m%d%H”**, …

Note:
- The time format is **“%Y%m%d%H”**, which represents the year, month, day, and hour.…

Output Example:
If the current time is "20220101120000", the function will return **"202201011200"**.

Function get_current_os

get_current_os: get the current operating system …

Parameters: - None

Code Description: uses the `platform.system()` …

Note: the returned OS name will be in lowercase…

Output Example:
- If OS is Windows, the function will return "windows".

Figure 5: Documentation update for functions of ChatDev.

dates documentation for the minimally impacted scope, and integrates these updates into a new Markdown file, which includes additions or global removals of objects' documentation. This automation extends to integrating the pre-commit hook of Git to detect code changes and update documentation, thus seamlessly maintaining documentation alongside project development. Specifically, when code updates are staged and committed, REPOAGENT is triggered, automatically refreshing the documentation and staging it for the commit. It confirms the process with a "Passed" indicator, without requiring extra commands or manual intervention, preserving developers' usual workflows.

3.3 Human Evaluation

Given the lack of reliable automatic evaluation methods for capturing the nuances of code documentation, we chose human evaluation to assess our method's documentation quality. A preference test was designed to compare human-authored documentation directly against that produced by REPOAGENT . For this purpose, 150 pieces of documentation content were randomly sampled, including 100 class objects and 50 function-level objects from both the Transformers and LlamaIndex repositories. Three independent evaluators were then enlisted to impartially assess the documentation quality, following a protocol detailed in Appendix A.2.2. The findings from this rigorous comparison are summarized in Table 1, underscore RepoAgent's notable effectiveness in producing documentation that surpasses human-authored content, achieving win rates of 0.70 and 0.91, respectively.

	Total	Human	Model	Win Rate
Transformers	150	45	105	**0.70**
LlamaIndex	150	13	137	**0.91**

Table 1: Results of human preference test on human-authored and model-generated code documentation.

3.4 Quantitative Analysis

Reference Recall. We evaluated the models' perception of global context by calculating the recall for identifying reference relationships of code objects. We sampled 20 objects from each of 9 repositories and compared 3 documentation generation methods for their recall in global caller and callee identification. The comparison methods included a machine learning based method that uses LSTM for comment generation (Iyer et al., 2016), long context concatenation leveraging LLMs with up to 128k context lengths to process entire project codes for identifying calling relationships, single-object generation method that only provides code snippets to LLMs.

Figure 6 demonstrates the recall for identifying reference relationships. The machine learning based method is unable to identify reference relationships, whereas the Single-object method partially identifies callees but not callers. The Long Context method, despite offering extensive code content, achieves only partial and non-comprehensive recognition of references, with recall declining as context increases. In contrast, our approach utilizes deterministic tools Jedi and bi-directional parsing to accurately convey global reference relationships, effectively overcoming the scope limitations that other methods encounter in generating repository-level code documentation.

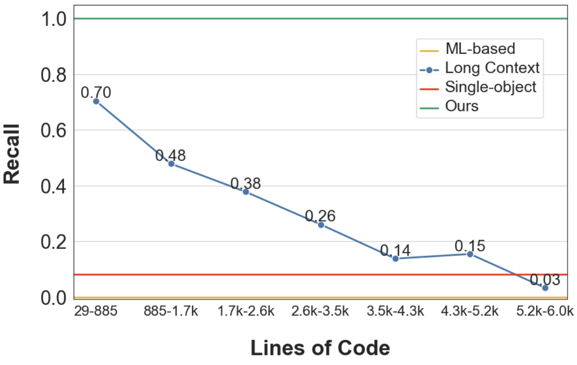


Figure 6: Recall for identifying reference relationships.

Format Alignment. Adherence to the specified format is critical in documentation generation. The generated documentation should consist of 5 basic parts, where the *Examples* is dynamic, depending on whether the code object has a return value or not. We evaluated the ability of LLMs to adhere to the format using all 9 repositories, the results are shown in Figure 7. Large models like GPT series and `Llama-2-70b` perform very well in format alignment, while the smaller model `Llama-2-7b` performs poorly, especially in terms of the examples.

Repository	`Llama-2-7b`	`Llama-2-70b`	`gpt-3.5-turbo`	`gpt-4-0125`
unoconv	0.0000	0.5000	**1.0000**	**1.0000**
simdjson	0.4298	0.6336	**1.0000**	0.9644
greenlet	0.5000	0.7482	0.9252	**0.9615**
code2flow	0.5145	0.6171	0.9735	**0.9803**
AutoGen	0.3049	0.5157	0.8633	**0.9545**
AutoGPT	0.4243	0.5611	0.8918	**0.9527**
ChatDev	0.5387	0.6980	0.9164	**0.9695**
MemGPT	0.4582	0.5729	0.9285	**0.9911**
MetaGPT	0.3920	0.5819	0.9066	**0.9708**

Table 2: Accuracy of identifying function parameters with different LLMs as backends.

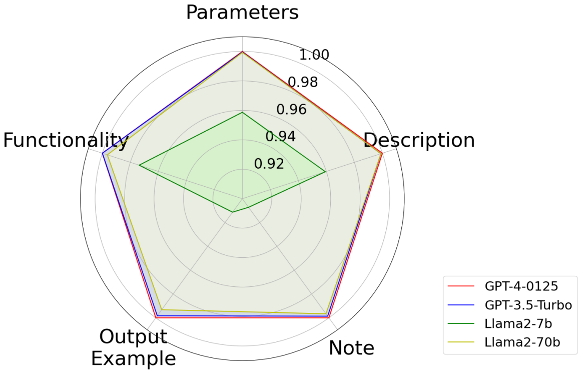


Figure 7: Format alignment accuracy of different LLMs.

Parameter Identification. We further evaluated the models' capability to identify parameters on all 9 repositories, the results are shown in Table 2. It is worth noting that we report the accuracy instead of recall, because models may hallucinate non-existent parameters, which should be taken into account. As seen in the table, the GPT series significantly outperforms the LLaMA series in parameter identification, and `gpt-4-0125` performs the best.

4 Related Work

Code Summarization. The field focuses on generating succinct, human-readable code summaries. Early methods were rule-based or template-driven (Haiduc et al., 2010; Sridhara et al., 2010; Moreno et al., 2013; Rodeghero et al., 2014). With advancements in machine learning, learning-based approaches like CODE-NN, which utilize LSTM units, emerged for summary creation (Iyer et al., 2016). The field further evolved with attention mechanisms and transformer models, significantly enhancing the ability to model long-range dependencies (Allamanis et al., 2016; Vaswani et al., 2017), indicating a shift towards more context-aware and flexible summarization techniques.

LLM Development. The development and application of LLMs have revolutionized both NLP and software engineering fields. Initially, the field was transformed by masked language models like BERT (Devlin et al., 2019), followed by advancements in encoder-decoder models, such as the T5 series (Raffel et al., 2020), and auto-regressive models like the GPT series (Radford et al., 2018). Auto-regressive models, notable for their sequence generation capabilities, have been effectively applied in code generation (Nijkamp et al., 2023; Li et al., 2023; Chen et al., 2021; Rozière et al., 2023; Xu et al., 2024), code summarization (Sun et al., 2023), and documentation generation (Wang et al., 2023; Khan and Uddin, 2022), highlighting their versatility in programming and documentation tasks. Concurrently, LLM-based agents have become ubiquitous (XAgent, 2023; Qin et al., 2024; Lyu et al., 2023; Ye et al., 2023; Qin et al., 2023), especially in software engineering (Chen et al., 2024; Qian et al., 2023; Hong et al., 2024), facilitating development through role-play and the automatic generation of agents (Wu et al., 2023), thereby enhancing repository-level code understanding, generation and even debugging (Tian et al., 2024). With the development of LLM-based agents, repository-level documentation generation become solvable as an agent task.

5 Conclusion and Discussion

In this paper, we introduce REPOAGENT, an open source framework designed to generate fine-grained repository-level code documentation, facilitating improved team collaboration. The experimental results suggest that REPOAGENT is capable of generating and proactively maintaining high-quality documentation for the entire project. REPOAGENT is expected to free developers from this tedious task, thereby improving their productivity and innovation potential.

In future work, we consider how to effectively utilize this tool and explore ways to apply REPOAGENT to a broader range of downstream applications in the future. To this end, we believe that chatting can serve as a natural tool for establishing a communication bridge between code and humans. Currently, by employing our approach with retrieval-augmented generation, which combines code, documentation, and reference relationships, we have achieved preliminary results in what we called "Chat With Repo", which marks the advent of a novel coding paradigm.

Limitations

Programming Language Limitations. RepoAgent currently relies on the Jedi reference recognition tool, limiting its applicability exclusively to Python projects. A more versatile, open-source tool that can adapt to multiple programming languages would enable broader adoption across various codebases, which will be addressed in future iterations.

Requirement for Human Oversight. AI-generated documentation may still require human review and modification to ensure its accuracy and completeness. Technical intricacies, project-specific conventions, and domain-specific terminology may necessitate manual intervention to enhance the quality of generated documentation.

Dependency on Language Model Capabilities. The performance of RepoAgent significantly depends on the backend LLMs and associated technologies. Although current results have shown promising progress with API-based LLMs like GPT series, the long-term stability and sustainability of using open-source models still require further validation and research.

Lack of Standards for Evaluation. It is difficult to establish a unified quantitative evaluation method for the professionalism, accuracy, and standardization of generated documentation. Furthermore, it is worth noting that the academic community currently lacks benchmarks and datasets of exemplary human documentation. Additionally, the subjective nature of documentation further limits current methods in terms of quality assessment.

Broader Impact

Enhancing Productivity and Innovation. RepoAgent automates the generation, update and maintenance of code documentation, which is traditionally a time-consuming task for developers. By freeing developers from this burden, our tool not only enhances productivity but also allows more time for creative and innovative work in software development.

Improving Software Quality and Collaboration. High-quality documentation is crucial for understanding, using, and contributing to software projects, facilitating developers' swift comprehension of projects. RepoAgent 's ability ensures long-term high consistency in code documentation. We posit that integrating RepoAgent closely with the project development process can introduce a new paradigm for standardizing and making repositories more readable. This, in turn, is expected to stimulate active community contributions and rapid development with higher overall quality of software projects.

Educational Benefits. RepoAgent can serve as an educational tool by providing clear and consistent documentation for codebases, making it easier for students and novice programmers to learn software development practices and understand complex codebases.

Bias and Inaccuracy. While RepoAgent aims to generate high-quality documentation, there's a potential risk of generating biased or inaccurate content due to model hallucination.

Security and Privacy Concerns. Currently, RepoAgent mainly relies on remote API-based LLMs, which will have the opportunity to access users' code data. This may raise security and privacy concerns, especially for proprietary software. Ensuring data protection and secure handling of the code is crucial.

Acknowledgments

The work was supported by the National Key R&D Program of China No.2022ZD0116312),the Postdoctoral Fellowship Program of CPSF under Grant Number GZC20240831 and Institute Guo Qiang at Tsinghua University.

We appreciate the suggestions and assistance from all the fellow students and friends in the community, including Arno (Bangsheng Feng), Guo Zhang, Qiang Guo, Yang Li, Yang Jiao, and others.

References

Miltiadis Allamanis, Hao Peng, and Charles Sutton. 2016. A convolutional attention network for extreme summarization of source code. In *Proceedings of the 33nd International Conference on Machine Learning*, volume 48, pages 2091–2100, New York City, NY, USA.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Pondé de Oliveira Pinto, Jared Kaplan, Harrison Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Joshua Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating large language models trained on code. *Computing Research Repository*, arXiv:2107.03374.

Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, Yujia Qin, Xin Cong, Ruobing Xie, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2024. AgentVerse: Facilitating multi-agent collaboration and exploring emergent behaviors. In *Proceedings of the the 12th International Conference on Learning Representations*, Vienna, Austria.

Sergio Cozzetti B. de Souza, Nicolas Anquetil, and Káthia Marçal de Oliveira. 2005. A study of the documentation essential to software maintenance. In *Proceedings of the 23rd Annual International Conference on Design of Communication: documenting & Designing for Pervasive Information*, pages 68–75, Coventry, UK.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Sonia Haiduc, Jairo Aponte, Laura Moreno, and Andrian Marcus. 2010. On the use of automated text summarization techniques for summarizing source code. In *Proceedings of the 17th Working Conference on Reverse Engineering*, pages 35–44, Beverly, MA, USA.

Sirui Hong, Xiawu Zheng, Jonathan Chen, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, et al. 2024. MetaGPT: Meta programming for multi-agent collaborative framework. In *Proceedings of the the 12th International Conference on Learning Representations*, Vienna, Austria.

Srinivasan Iyer, Ioannis Konstas, Alvin Cheung, and Luke Zettlemoyer. 2016. Summarizing source code using a neural attention model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2073–2083, Berlin, Germany. Association for Computational Linguistics.

Junaed Younus Khan and Gias Uddin. 2022. Automatic code documentation generation using GPT-3. In *Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering*, pages 174:1–174:6, Rochester, MI, USA.

M.M. Lehman. 1980. Programs, life cycles, and laws of software evolution. *Proceedings of the IEEE*, 68(9):1060–1076.

Raymond Li, Loubna Ben Allal, Yangtian Zi, Niklas Muennighoff, Denis Kocetkov, Chenghao Mou, Marc Marone, Christopher Akiki, Jia Li, Jenny Chim, Qian Liu, Evgenii Zheltonozhskii, Terry Yue Zhuo, Thomas Wang, Olivier Dehaene, Mishig Davaadorj, Joel Lamy-Poirier, João Monteiro, Oleh Shliazhko, Nicolas Gontier, Nicholas Meade, Armel Zebaze, Ming-Ho Yee, Logesh Kumar Umapathi, Jian Zhu, Benjamin Lipkin, Muhtasham Oblokulov, Zhiruo Wang, Rudra Murthy V, Jason Stillerman, Siva Sankalp Patel, Dmitry Abulkhanov, Marco Zocca, Manan Dey, Zhihan Zhang, Nour Moustafa-Fahmy, Urvashi Bhattacharyya, Wenhao Yu, Swayam Singh, Sasha Luccioni, Paulo Villegas, Maxim Kunakov, Fedor Zhdanov, Manuel Romero, Tony Lee, Nadav Timor, Jennifer Ding, Claire Schlesinger, Hailey Schoelkopf, Jan Ebert, Tri Dao, Mayank Mishra, Alex Gu, Jennifer Robinson, Carolyn Jane Anderson, Brendan Dolan-Gavitt, Danish Contractor, Siva Reddy, Daniel Fried, Dzmitry Bahdanau, Yacine Jernite, Carlos Muñoz Ferrandis, Sean Hughes, Thomas Wolf, Arjun Guha, Leandro von Werra, and Harm de Vries. 2023. StarCoder: may the source be with you! *Computing Research Repository*, arXiv:2305.06161.

Bohan Lyu, Xin Cong, Heyang Yu, Pan Yang, Yujia Qin, Yining Ye, Yaxi Lu, Zhong Zhang, Yukun Yan, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2023. Gitagent: Facilitating autonomous agent with github by tool extension. *Computing Research Repository*, arXiv:2312.17294.

Robert C Martin. 1996. The dependency inversion principle. *C++ Report*, 8(6):61–66.

Laura Moreno, Jairo Aponte, Giriprasad Sridhara, Andrian Marcus, Lori L. Pollock, and K. Vijay-Shanker. 2013. Automatic generation of natural language summaries for Java classes. In *Proceedings of the IEEE 21st International Conference on Program Comprehension*, pages 23–32, San Francisco, CA, USA.

Erik Nijkamp, Bo Pang, Hiroaki Hayashi, Lifu Tu, Huan Wang, Yingbo Zhou, Silvio Savarese, and Caiming Xiong. 2023. CodeGen: An open large language model for code with multi-turn program synthesis. In *Proceedings of the 11th International Conference on Learning Representations*, Kigali, Rwanda.

OpenAI. 2022. OpenAI: Introducing ChatGPT.

OpenAI. 2023. GPT-4 technical report. *Computing Research Repository*, arXiv:2303.08774.

Chen Qian, Xin Cong, Cheng Yang, Weize Chen, Yusheng Su, Juyuan Xu, Zhiyuan Liu, and Maosong Sun. 2023. Communicative agents for software development. *Computing Research Repository,*, arXiv:2307.07924.

Yujia Qin, Shengding Hu, Yankai Lin, Weize Chen, Ning Ding, Ganqu Cui, Zheni Zeng, Yufei Huang, Chaojun Xiao, Chi Han, Yi Ren Fung, Yusheng Su, Huadong Wang, Cheng Qian, Runchu Tian, Kunlun Zhu, Shihao Liang, Xingyu Shen, Bokai Xu, Zhen Zhang, Yining Ye, Bowen Li, Ziwei Tang, Jing Yi, Yuzhang Zhu, Zhenning Dai, Lan Yan, Xin Cong, Yaxi Lu, Weilin Zhao, Yuxiang Huang, Junxi Yan, Xu Han, Xian Sun, Dahai Li, Jason Phang, Cheng Yang, Tongshuang Wu, Heng Ji, Zhiyuan Liu, and Maosong Sun. 2023. Tool learning with foundation models. *Computing Research Repository*, arXiv:2304.08354.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Lauren Hong, Runchu Tian, Ruobing Xie, Jie Zhou, Mark Gerstein, dahai li, Zhiyuan Liu, and Maosong Sun. 2024. ToolLLM: Facilitating large language models to master 16000+ real-world APIs. In *The Twelfth International Conference on Learning Representations*, Vienna, Austria.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *Preprint*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Sawan Rai, Ramesh Chandra Belwal, and Atul Gupta. 2022. A review on source code documentation. *ACM Transactions on Intelligent Systems and Technology*, 13(5):1 – 44.

Paige Rodeghero, Collin McMillan, Paul W. McBurney, Nigel Bosch, and Sidney K. D'Mello. 2014. Improving automated source code summarization via an eye-tracking study of programmers. In *Proceedings of the 36th International Conference on Software Engineering*, pages 390–401, Hyderabad, India.

Baptiste Rozière, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Tal Remez, Jérémy Rapin, Artyom Kozhevnikov, Ivan Evtimov, Joanna Bitton, Manish Bhatt, Cristian Canton-Ferrer, Aaron Grattafiori, Wenhan Xiong, Alexandre Défossez, Jade Copet, Faisal Azhar, Hugo Touvron, Louis Martin, Nicolas Usunier, Thomas Scialom, and Gabriel Synnaeve. 2023. Code Llama: Open foundation models for code. *Computing Research Repository,*, arXiv:2308.12950.

Giriprasad Sridhara, Emily Hill, Divya Muppaneni, Lori L. Pollock, and K. Vijay-Shanker. 2010. Towards automatically generating summary comments for java methods. In *Proceedings of the 25th IEEE/ACM international conference on Automated software engineering*, pages 43–52, Antwerp, Belgium.

Weisong Sun, Chunrong Fang, Yudu You, Yuchen Chen, Yi Liu, Chong Wang, Jian Zhang, Quanjun Zhang, Hanwei Qian, Wei Zhao, et al. 2023. A prompt learning framework for source code summarization. *Computing Research Repository*, arXiv:2312.16066.

Runchu Tian, Yining Ye, Yujia Qin, Xin Cong, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2024. DebugBench: Evaluating debugging capability of large language models. *Computing Research Repository*, arXiv:2401.04621.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Computing Research Repository*, arXiv:2307.09288.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, pages 5998–6008, Long Beach, CA, USA.

Shujun Wang, Yongqiang Tian, and Dengcheng He. 2023. gDoc: Automatic generation of structured API documentation. In *Companion Proceedings of the ACM Web Conference 2023*, pages 53–56, Austin, TX, USA.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022*, pages 24824–24837, New Orleans, LA, USA.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. 2023. AutoGen: Enabling next-gen llm applications via multi-agent conversation framework. *Computing Research Repository,*, arXiv:2308.08155.

XAgent. 2023. Xagent: An autonomous agent for complex task solving.

Xin Xia, Lingfeng Bao, David Lo, Zhenchang Xing, Ahmed E. Hassan, and Shanping Li. 2018. Measuring program comprehension: A large-scale field study with professionals. *IEEE Transactions on Software Engineering*, 44(10):951–976.

Yiheng Xu, Hongjin Su, Chen Xing, Boyu Mi, Qian Liu, Weijia Shi, Binyuan Hui, Fan Zhou, Yitao Liu, Tianbao Xie, et al. 2024. Lemur: Harmonizing natural language and code for language agents. In *Proceedings of the 12th International Conference on Learning Representations*, Vienna, Austria.

Yining Ye, Xin Cong, Shizuo Tian, Jiannan Cao, Hao Wang, Yujia Qin, Yaxi Lu, Heyang Yu, Huadong

Wang, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2023. Proagent: From robotic process automation to agentic process automation. *Computing Research Repository*, arXiv:2311.10751.

Chunyan Zhang, Junchao Wang, Qinglei Zhou, Ting Xu, Ke Tang, Hairen Gui, and Fudong Liu. 2022. A survey of automatic source code summarization. *Symmetry*, 14(3):471.

Jian Zhang, Xu Wang, Hongyu Zhang, Hailong Sun, Kaixuan Wang, and Xudong Liu. 2019. A novel neural source code representation based on abstract syntax tree. In *Proceedings of the 2019 IEEE/ACM 41st International Conference on Software Engineering*, pages 783–794, Montréal, Québec, Canada.

Junji Zhi, Vahid Garousi-Yusifoğlu, Bo Sun, Golara Garousi, Shawn Shahnewaz, and Guenther Ruhe. 2015. Cost, benefits and quality of software development documentation: A systematic mapping. *Journal of Systems and Software*, 99:175–198.

A Appendix: Experimental Details

A.1 Implementation Details

Table 3 presents the detailed statistics of the selected repositories and the token costs associated with the production of initial documentation. The inclusion of global information, such as the project's directory structure and bidirectional references, leads to significantly longer prompts, as detailed in Appendix C). Despite this, the resulting documentation is thorough yet concise, typically ranging between 0.4k and 1k tokens in length.

During the actual generation process, we addressed the issue of varying text lengths across different models. When using models with shorter context lengths (e.g., `gpt-3.5-turbo` and the LLaMA series), REPOAGENT adaptively switches to models with larger context lengths (e.g., `gpt-3.5-16k` or `gpt-4-32k`) based on the current prompt's length, to cope with the token overhead of incorporating global perspectives. In cases where even these models' limits are exceeded, REPOAGENT truncates the content by simplifying the project's directory structure and removing bidirectional reference code before reinitiating the documentation generation task. Such measures are infrequent when employing models with the longest contexts (128k), such as `gpt-4-1106` or `gpt-4-0125`. This dynamic scheduling strategy, combined with variable network conditions, may influence token consumption. Nevertheless, REPOAGENT ensures the integrity of the documentation while striving for cost-effectiveness to the greatest extent.

A.2 Settings

A.2.1 Technical Environment

All experiments were conducted within a Python 3.11.4 environment. The system had CUDA 11.7 installed and was equipped with 8 NVIDIA A100 40GB GPUs.

A.2.2 Human Evaluation Protocol

We recruited three human evaluators to assess the code documentation generated by REPOAGENT, and instructed all human evaluators to give an overall evaluation considering a set of evaluation criteria shown in Table 4. We randomly sampled 150 pieces of documentation from the repository. Subsequently, each human evaluator was assigned 50 pairs of documentation, each containing one human-authored and one model-generated documentation. The human evaluators were required to select the better documentation for each pair.

A.2.3 Reference Recall

The experiment aims to evaluate the model's ability to perceive global context, which is reflected by the recall for identifying reference relationships. The comparison methods are:

1. **ML-based method.** Iyer et al. (2016) utilized traditional machine learning and deep learning methods for generating comments describing the functionality of code objects.
2. **Long context concatenation.** The method directly concatenates the code snippets until the context length reaches 128k to let the model discover reference relationships.
3. **Single-object generation.** Sun et al. (2023) used the GPT-3.5 series to generate documentation by directly feeding code snippets of the target object. We modified the prompt on this basis, adding requirements for outputting the callers and callees.

Notably, among these methods, only the ML-based approach failed to explicitly or implicitly manifest call relationships in the final document. While it is inherently challenging for a code snippet to discern its invocation throughout the entire repository, the code typically elucidates the current object's calls explicitly. To measure the recall of callers and callees, we enhanced the original documentation by adding information about the calling functions (callers) and the called functions (callees). Then we compared the enriched documentation with our bidirectional reference data from MetaInfo.

For long context concatenation, we randomly selected 20 objects from each of the 9 repositories, culminating in a total of 180 objects. Given the intricate nature of defining context construction criteria for repository-level documentation generation tasks, we circumvented direct concatenation of adjacent and file-adjacent context content. Instead, we formulated negative samples by extracting all objects with reference relationships to fulfill the context length. Leveraging the content of objects and negative sample content, we devised context lengths for the 180 objects, spanning from 29 to 6.0k Code Lines. This approach aimed to optimize the distribution of context lengths while maximizing the utilization of the model's context length. In the case of single-object generation method, we utilized the same pool of 180 objects, providing the model with object source code snippets to generate documentation and elucidate reference relationships.

During the evaluation of both the Long Context Concatenation and Single-object Generation methods, we provided the model with tree-structured hierarchical position information for target objects and their related counterparts.

This additional information was intended to help the model in better identifying callers and delineating them in a path form. Despite this assistance, the model's misinterpretations exacerbated as the context length increased, and the Single-object Generation method yielded a substantial amount of speculative information, resulting in unstable and inaccurate caller relationship recognition.

A.2.4 Format Alignment

The experiment evaluates whether the model-generated documentation follows the defined format. LLMs generally excel in instruction following, but the complexity of our task requires models to grasp core intents within lengthy prompts, posing a challenge. We use a one-shot approach with strict output examples, enabling evaluation of model answers through format matching algorithms. Specifically, we mandate that section titles be enclosed in bold symbols, ensure clear divisions between sections, and require contents within sections to be extractable and meaningful.

We observed the shortcomings of open-source models (LLaMA series) in their ability to adhere to formatting. In contrast, the GPT-4 series models excellently achieve format integrity and stability. We also observed behavioral differences between `gpt-4-0125` and `gpt-4-1106` models, the former appeared to produce more redundant information.

Format alignment can also be achieved with perfect accuracy using hierarchical or modular generation methods. However, this approach introduces a significant token overhead since each independent module must encompass complete global information and invocation relationships. Current method has demonstrated satisfactory performance on format alignment, meeting human readability standards effectively.

A.2.5 Parameter Identification

Accurately identifying and describing parameters or attributes (depending on whether the current object is a function or a class) in code is crucial as it helps readers quickly understand the design logic and usage. We extracted recognized parameters from the Parameters section using a matching pattern: parameters follow a uniform and fixed format, with the parameter name enclosed in code identifiers followed by the parameter's descriptive text.

We organized the extracted parameters into arrays and calculated accuracy by comparing them with the values in the params field (also an array) of the Repository's MetaInfo. It is important to note that we were calculating accuracy here, not recall. This is because some models may hallucinate many nonexistent parameters based on the code snippets. These errors must be taken into consideration, otherwise they will result in biased evaluations.

Repository	Model	Prompt Tokens	Completion Tokens	Class Numbers	Function Numbers	Code Lines
unoconv	`gpt-4-0125`	4020	2550	0	1	≤1k
	`gpt-3.5-turbo`		2743			
	`Llama-2-7b`	1180	2916			
	`Llama-2-70b`		437			
simdjson	`gpt-4-0125`	45344	35068	6	55	≤ 1k
	`gpt-3.5-turbo`		29736			
	`Llama-2-7b`	49615	27562			
	`Llama-2-70b`		32961			
greenlet	`gpt-4-0125`	86587	79113	59	319	1k ≤ 10k
	`gpt-3.5-turbo`		260464			
	`Llama-2-7b`	33177	31561			
	`Llama-2-70b`		225595			
code2flow	`gpt-4-0125`	185511	134462	51	257	1k ≤ 10k
	`gpt-3.5-turbo`		234101			
	`Llama-2-7b`	354574	431761			
	`Llama-2-70b`		187835			
AutoGen	`gpt-4-0125`	4939388	516975	64	590	1k ≤ 10k
	`gpt-3.5-turbo`		288609			
	`Llama-2-7b`	889050	630139			
	`Llama-2-70b`		410256			
AutoGPT	`gpt-4-0125`	4116296	888223	318	1170	≥ 10k
	`gpt-3.5-turbo`		799380			
	`Llama-2-7b`	1838425	1893041			
	`Llama-2-70b`		927946			
ChatDev	`gpt-4-0125`	2021168	602474	183	729	≥ 10k
	`gpt-3.5-turbo`		519226			
	`Llama-2-7b`	1122400	946131			
	`Llama-2-70b`		531838			
MemGPT	`gpt-4-0125`	628482	345109	74	478	≥ 10k
	`gpt-3.5-turbo`		234101			
	`Llama-2-7b`	742591	740783			
	`Llama-2-70b`		352940			
MetaGPT	`gpt-4-0125`	154364	111159	291	885	≥ 10k
	`gpt-3.5-turbo`		134101			
	`Llama-2-7b`	1904244	2265991			
	`Llama-2-70b`		1009996			

Table 3: Statistics for the selected repositories and the token consumption for documentation generation. Note that token count calculation varies with each model's tokenizer, rendering direct comparisons between different models impractical.

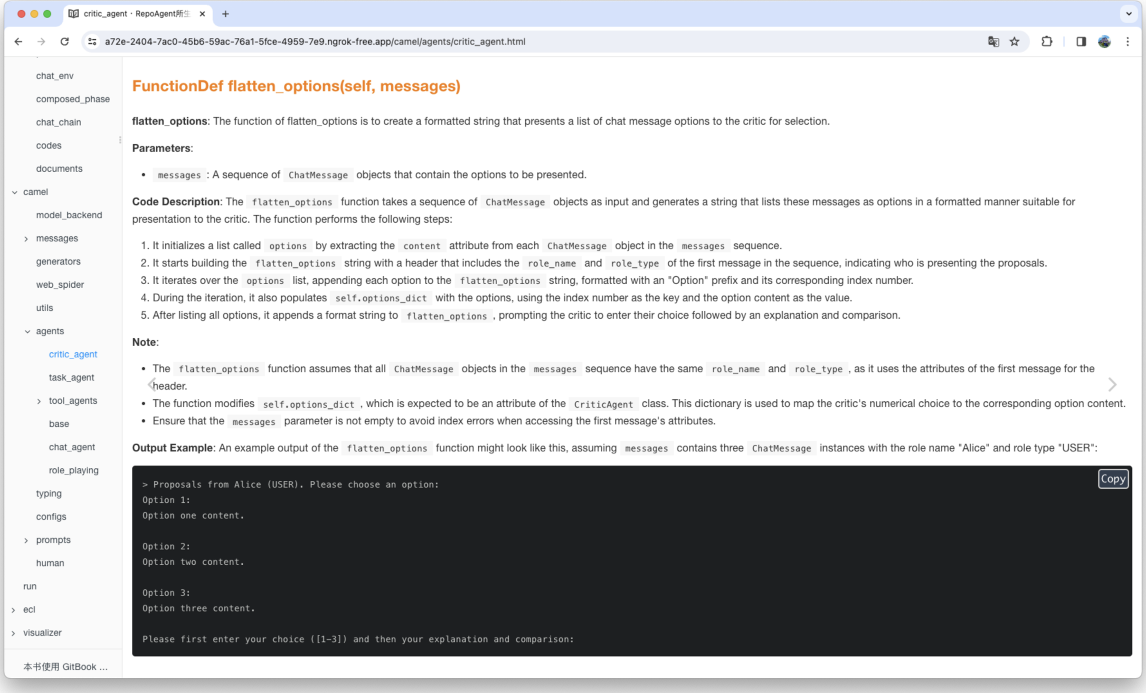
Criteria	Details
Accuracy	**Correctness**: Verify if the documentation accurately describes the code's functionality, algorithms, and expected behavior under various conditions.
	Precision: Assess whether the documentation provides precise and unambiguous information regarding the code's operations, parameters, and expected outcomes.
	Alignment with Codebase: Ensure that the documentation aligns closely with the actual implementation of the code, including any updates or changes made to the codebase.
Completeness	**Coverage**: Evaluate if the documentation comprehensively covers all significant aspects of the code, including inputs, outputs, error handling, edge cases, and any potential exceptions.
	In-depth Explanation: Determine if the documentation delves into detailed explanations of complex functionalities or algorithms, providing insights into the underlying logic.
	Documentation of External Dependencies: Check if the documentation adequately addresses any external libraries, modules, or APIs used within the codebase.
Understandability	**Clarity**: Assess the clarity and readability of the documentation, ensuring that it is easily understandable by developers of varying expertise levels.
	Conciseness: Determine if the documentation conveys information concisely without unnecessary verbosity or technical jargon that might hinder comprehension.
	Structured Organization: Evaluate if the documentation is logically organized, with clear headings, sections, and navigation aids for easy reference and comprehension.
Consistency	**Formatting Consistency**: Ensure consistency in the formatting, styling, and layout of the documentation across all sections and pages.
	Terminology Consistency: Verify that consistent terminology and naming conventions are used throughout the documentation to maintain coherence and clarity.
	Style Guide Adherence: Assess if the documentation adheres to any predefined style guides or conventions established by the project or organization.
Relevance	**Content Relevance**: Determine if the information provided in the documentation is directly relevant to the code's functionality, purpose, and usage scenarios.
	Avoidance of Redundancy: Check for redundancy or repetition within the documentation, eliminating any extraneous or irrelevant details that do not contribute to understanding the code.
Examples and Usage	**Code Samples**: Evaluate if the documentation includes sufficient code samples, snippets, or examples to illustrate the usage and implementation of key functionalities.
	Use Cases: Assess if the documentation provides real-world use cases or scenarios where the code can be applied, demonstrating its practical utility and versatility.
	Step-by-Step Instructions: Determine if the documentation offers clear, step-by-step instructions or tutorials for integrating, configuring, and utilizing the code in different environments or applications.

Table 4: Detailed criteria for human evaluation.

B Appendix: More Cases of Generated Documentation

B.1 Documentation Showcases

In this section, we showcase additional generated documentation to validate the practical application of REPOAGENT . The included images are direct screenshots from the documentation of two open-source projects, ChatDev and AutoGen. Our intent is to provide readers with a detailed and panoramic view of how our method is utilized in real-world scenarios, thereby offering a deeper understanding of its effectiveness and versatility.



FunctionDef flatten_options(self, messages)

flatten_options: The function of flatten_options is to create a formatted string that presents a list of chat message options to the critic for selection.

Parameters:

- `messages` : A sequence of `ChatMessage` objects that contain the options to be presented.

Code Description: The `flatten_options` function takes a sequence of `ChatMessage` objects as input and generates a string that lists these messages as options in a formatted manner suitable for presentation to the critic. The function performs the following steps:

1. It initializes a list called `options` by extracting the `content` attribute from each `ChatMessage` object in the `messages` sequence.
2. It starts building the `flatten_options` string with a header that includes the `role_name` and `role_type` of the first message in the sequence, indicating who is presenting the proposals.
3. It iterates over the `options` list, appending each option to the `flatten_options` string, formatted with an "Option" prefix and its corresponding index number.
4. During the iteration, it also populates `self.options_dict` with the options, using the index number as the key and the option content as the value.
5. After listing all options, it appends a format string to `flatten_options` , prompting the critic to enter their choice followed by an explanation and comparison.

Note:

- The `flatten_options` function assumes that all `ChatMessage` objects in the `messages` sequence have the same `role_name` and `role_type` , as it uses the attributes of the first message for the header.
- The function modifies `self.options_dict` , which is expected to be an attribute of the `CriticAgent` class. This dictionary is used to map the critic's numerical choice to the corresponding option content.
- Ensure that the `messages` parameter is not empty to avoid index errors when accessing the first message's attributes.

Output Example: An example output of the `flatten_options` function might look like this, assuming `messages` contains three `ChatMessage` instances with the role name "Alice" and role type "USER":

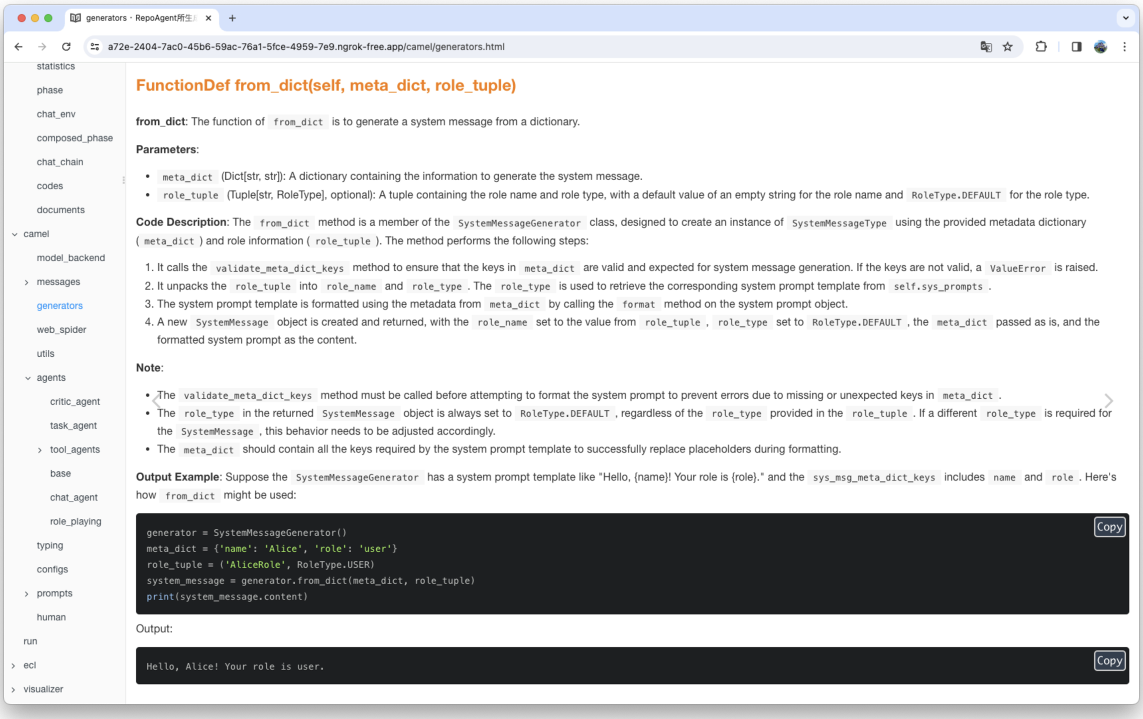
```
> Proposals from Alice (USER). Please choose an option:
Option 1:
Option one content.

Option 2:
Option two content.

Option 3:
Option three content.

Please first enter your choice ([1-3]) and then your explanation and comparison:
```

(a)



FunctionDef from_dict(self, meta_dict, role_tuple)

from_dict: The function of `from_dict` is to generate a system message from a dictionary.

Parameters:

- `meta_dict` (Dict[str, str]): A dictionary containing the information to generate the system message.
- `role_tuple` (Tuple[str, RoleType], optional): A tuple containing the role name and role type, with a default value of an empty string for the role name and `RoleType.DEFAULT` for the role type.

Code Description: The `from_dict` method is a member of the `SystemMessageGenerator` class, designed to create an instance of `SystemMessageType` using the provided metadata dictionary (`meta_dict`) and role information (`role_tuple`). The method performs the following steps:

1. It calls the `validate_meta_dict_keys` method to ensure that the keys in `meta_dict` are valid and expected for system message generation. If the keys are not valid, a `ValueError` is raised.
2. It unpacks the `role_tuple` into `role_name` and `role_type` . The `role_type` is used to retrieve the corresponding system prompt template from `self.sys_prompts` .
3. The system prompt template is formatted using the metadata from `meta_dict` by calling the `format` method on the system prompt object.
4. A new `SystemMessage` object is created and returned, with the `role_name` set to the value from `role_tuple` , `role_type` set to `RoleType.DEFAULT` , the `meta_dict` passed as is, and the formatted system prompt as the content.

Note:

- The `validate_meta_dict_keys` method must be called before attempting to format the system prompt to prevent errors due to missing or unexpected keys in `meta_dict` .
- The `role_type` in the returned `SystemMessage` object is always set to `RoleType.DEFAULT` , regardless of the `role_type` provided in the `role_tuple` . If a different `role_type` is required for the `SystemMessage` , this behavior needs to be adjusted accordingly.
- The `meta_dict` should contain all the keys required by the system prompt template to successfully replace placeholders during formatting.

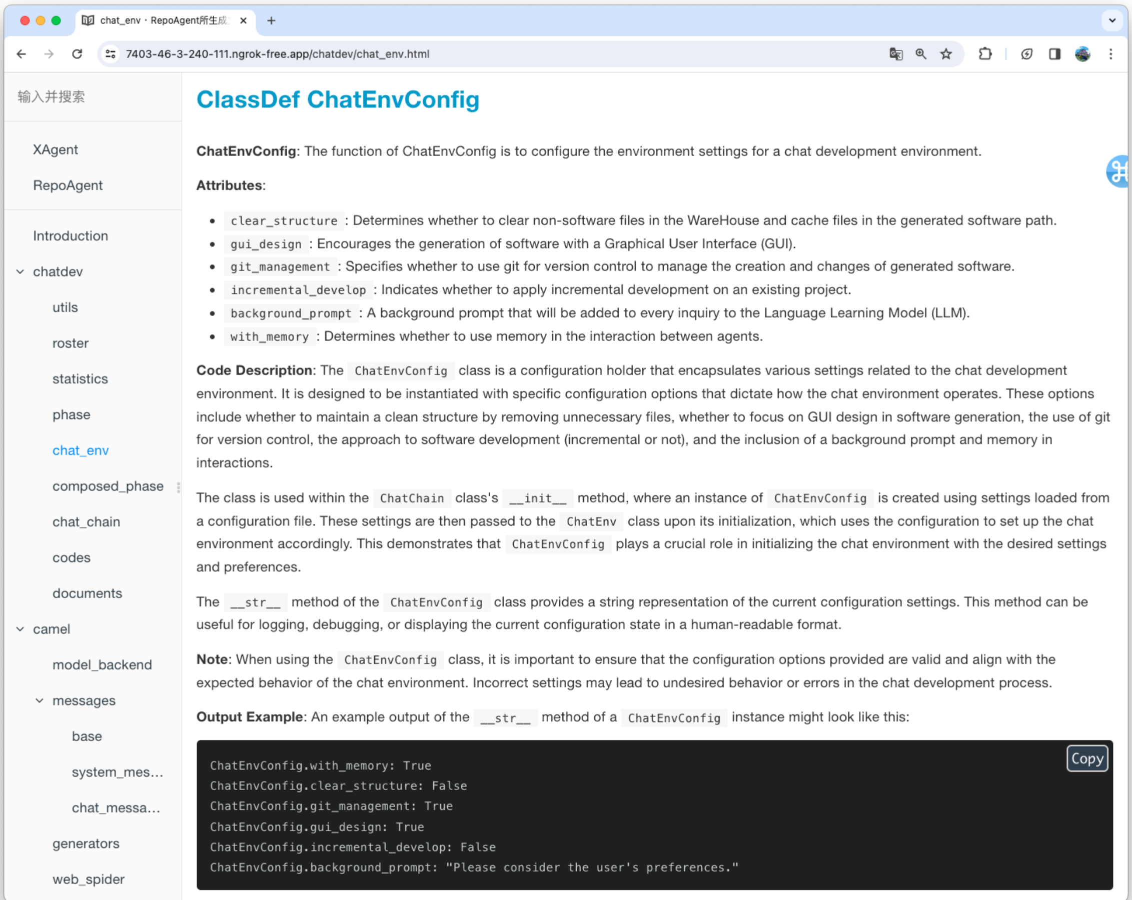
Output Example: Suppose the `SystemMessageGenerator` has a system prompt template like "Hello, {name}! Your role is {role}." and the `sys_msg_meta_dict_keys` includes `name` and `role` . Here's how `from_dict` might be used:

```
generator = SystemMessageGenerator()
meta_dict = {'name': 'Alice', 'role': 'user'}
role_tuple = ('AliceRole', RoleType.USER)
system_message = generator.from_dict(meta_dict, role_tuple)
print(system_message.content)
```

Output:

```
Hello, Alice! Your role is user.
```

(b)



ClassDef ChatEnvConfig

ChatEnvConfig: The function of ChatEnvConfig is to configure the environment settings for a chat development environment.

Attributes:

- `clear_structure` : Determines whether to clear non-software files in the WareHouse and cache files in the generated software path.
- `gui_design` : Encourages the generation of software with a Graphical User Interface (GUI).
- `git_management` : Specifies whether to use git for version control to manage the creation and changes of generated software.
- `incremental_develop` : Indicates whether to apply incremental development on an existing project.
- `background_prompt` : A background prompt that will be added to every inquiry to the Language Learning Model (LLM).
- `with_memory` : Determines whether to use memory in the interaction between agents.

Code Description: The `ChatEnvConfig` class is a configuration holder that encapsulates various settings related to the chat development environment. It is designed to be instantiated with specific configuration options that dictate how the chat environment operates. These options include whether to maintain a clean structure by removing unnecessary files, whether to focus on GUI design in software generation, the use of git for version control, the approach to software development (incremental or not), and the inclusion of a background prompt and memory in interactions.

The class is used within the `ChatChain` class's `__init__` method, where an instance of `ChatEnvConfig` is created using settings loaded from a configuration file. These settings are then passed to the `ChatEnv` class upon its initialization, which uses the configuration to set up the chat environment accordingly. This demonstrates that `ChatEnvConfig` plays a crucial role in initializing the chat environment with the desired settings and preferences.

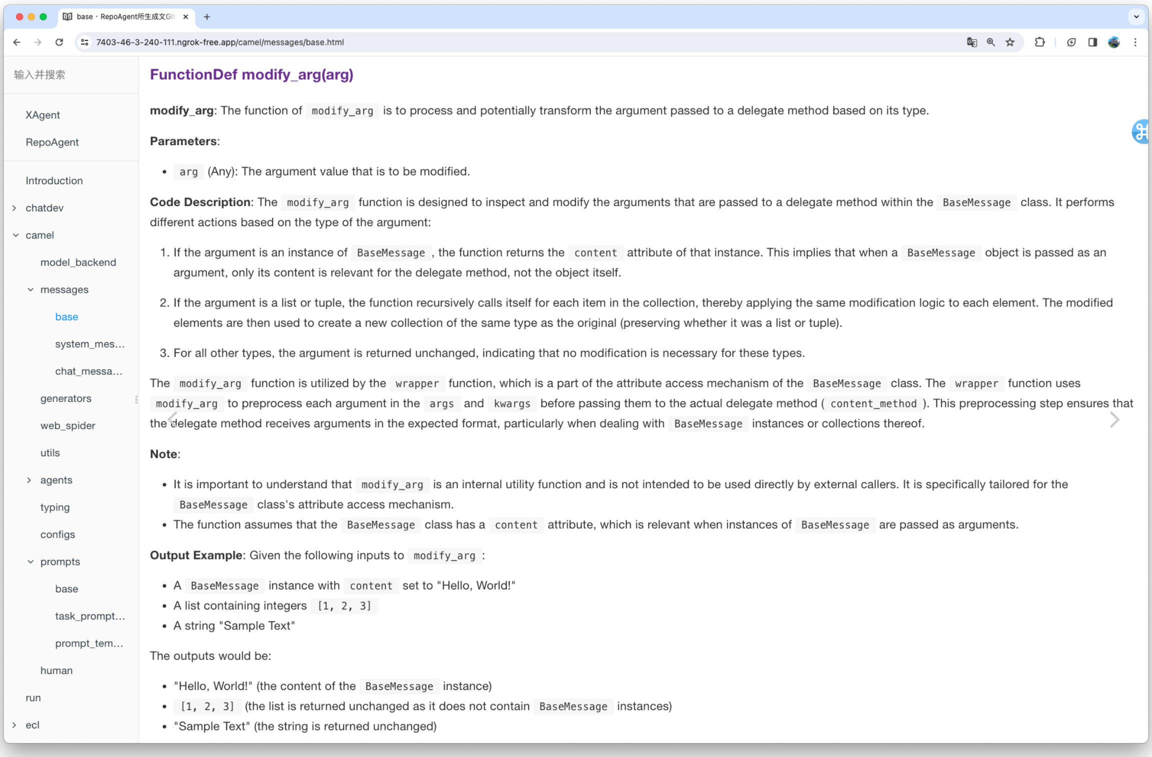
The `__str__` method of the `ChatEnvConfig` class provides a string representation of the current configuration settings. This method can be useful for logging, debugging, or displaying the current configuration state in a human-readable format.

Note: When using the `ChatEnvConfig` class, it is important to ensure that the configuration options provided are valid and align with the expected behavior of the chat environment. Incorrect settings may lead to undesired behavior or errors in the chat development process.

Output Example: An example output of the `__str__` method of a `ChatEnvConfig` instance might look like this:

```
ChatEnvConfig.with_memory: True
ChatEnvConfig.clear_structure: False
ChatEnvConfig.git_management: True
ChatEnvConfig.gui_design: True
ChatEnvConfig.incremental_develop: False
ChatEnvConfig.background_prompt: "Please consider the user's preferences."
```

(c)



FunctionDef modify_arg(arg)

modify_arg: The function of `modify_arg` is to process and potentially transform the argument passed to a delegate method based on its type.

Parameters:

- `arg` (Any): The argument value that is to be modified.

Code Description: The `modify_arg` function is designed to inspect and modify the arguments that are passed to a delegate method within the `BaseMessage` class. It performs different actions based on the type of the argument:

1. If the argument is an instance of `BaseMessage` , the function returns the `content` attribute of that instance. This implies that when a `BaseMessage` object is passed as an argument, only its content is relevant for the delegate method, not the object itself.
2. If the argument is a list or tuple, the function recursively calls itself for each item in the collection, thereby applying the same modification logic to each element. The modified elements are then used to create a new collection of the same type as the original (preserving whether it was a list or tuple).
3. For all other types, the argument is returned unchanged, indicating that no modification is necessary for these types.

The `modify_arg` function is utilized by the `wrapper` function, which is a part of the attribute access mechanism of the `BaseMessage` class. The `wrapper` function uses `modify_arg` to preprocess each argument in the `args` and `kwargs` before passing them to the actual delegate method (`content_method`). This preprocessing step ensures that the delegate method receives arguments in the expected format, particularly when dealing with `BaseMessage` instances or collections thereof.

Note:

- It is important to understand that `modify_arg` is an internal utility function and is not intended to be used directly by external callers. It is specifically tailored for the `BaseMessage` class's attribute access mechanism.
- The function assumes that the `BaseMessage` class has a `content` attribute, which is relevant when instances of `BaseMessage` are passed as arguments.

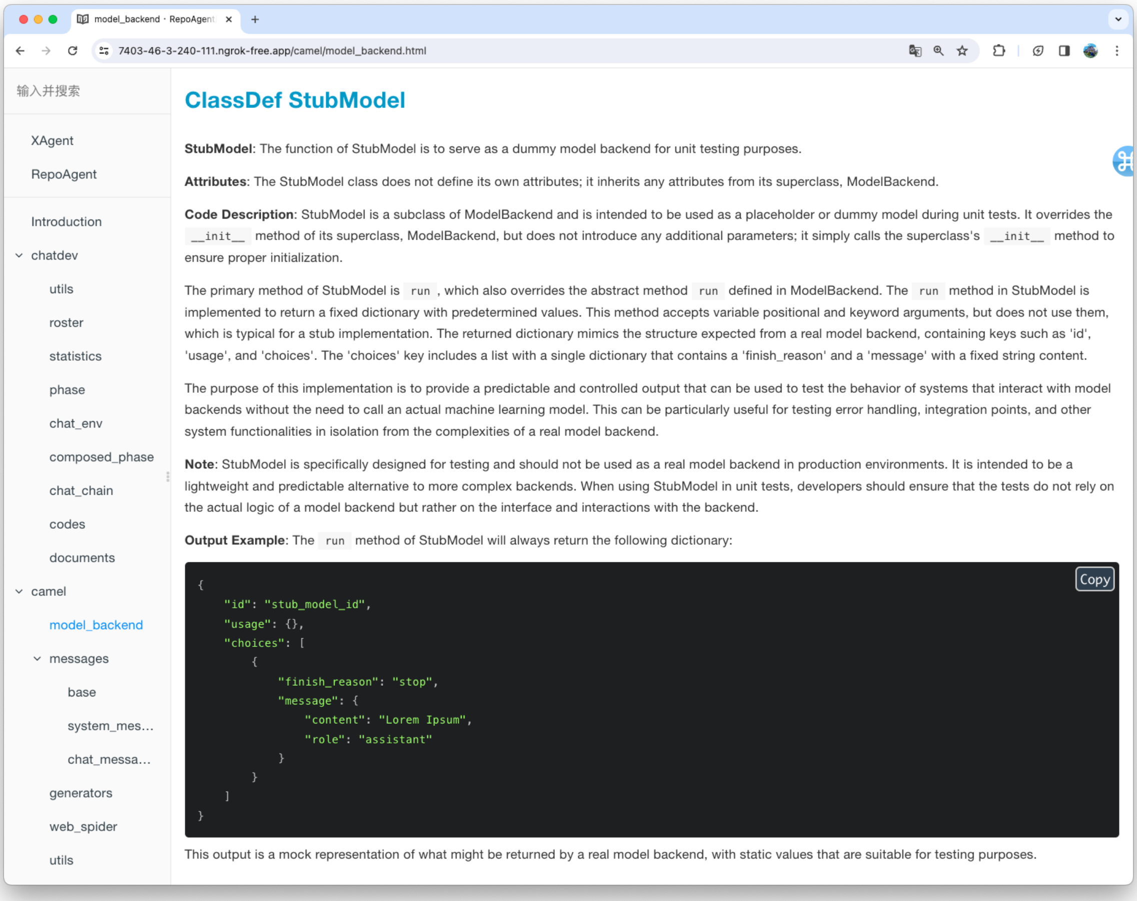
Output Example: Given the following inputs to `modify_arg` :

- A `BaseMessage` instance with `content` set to "Hello, World!"
- A list containing integers `[1, 2, 3]`
- A string "Sample Text"

The outputs would be:

- "Hello, World!" (the content of the `BaseMessage` instance)
- `[1, 2, 3]` (the list is returned unchanged as it does not contain `BaseMessage` instances)
- "Sample Text" (the string is returned unchanged)

(d)



ClassDef StubModel

StubModel: The function of StubModel is to serve as a dummy model backend for unit testing purposes.

Attributes: The StubModel class does not define its own attributes; it inherits any attributes from its superclass, ModelBackend.

Code Description: StubModel is a subclass of ModelBackend and is intended to be used as a placeholder or dummy model during unit tests. It overrides the `__init__` method of its superclass, ModelBackend, but does not introduce any additional parameters; it simply calls the superclass's `__init__` method to ensure proper initialization.

The primary method of StubModel is `run`, which also overrides the abstract method `run` defined in ModelBackend. The `run` method in StubModel is implemented to return a fixed dictionary with predetermined values. This method accepts variable positional and keyword arguments, but does not use them, which is typical for a stub implementation. The returned dictionary mimics the structure expected from a real model backend, containing keys such as 'id', 'usage', and 'choices'. The 'choices' key includes a list with a single dictionary that contains a 'finish_reason' and a 'message' with a fixed string content.

The purpose of this implementation is to provide a predictable and controlled output that can be used to test the behavior of systems that interact with model backends without the need to call an actual machine learning model. This can be particularly useful for testing error handling, integration points, and other system functionalities in isolation from the complexities of a real model backend.

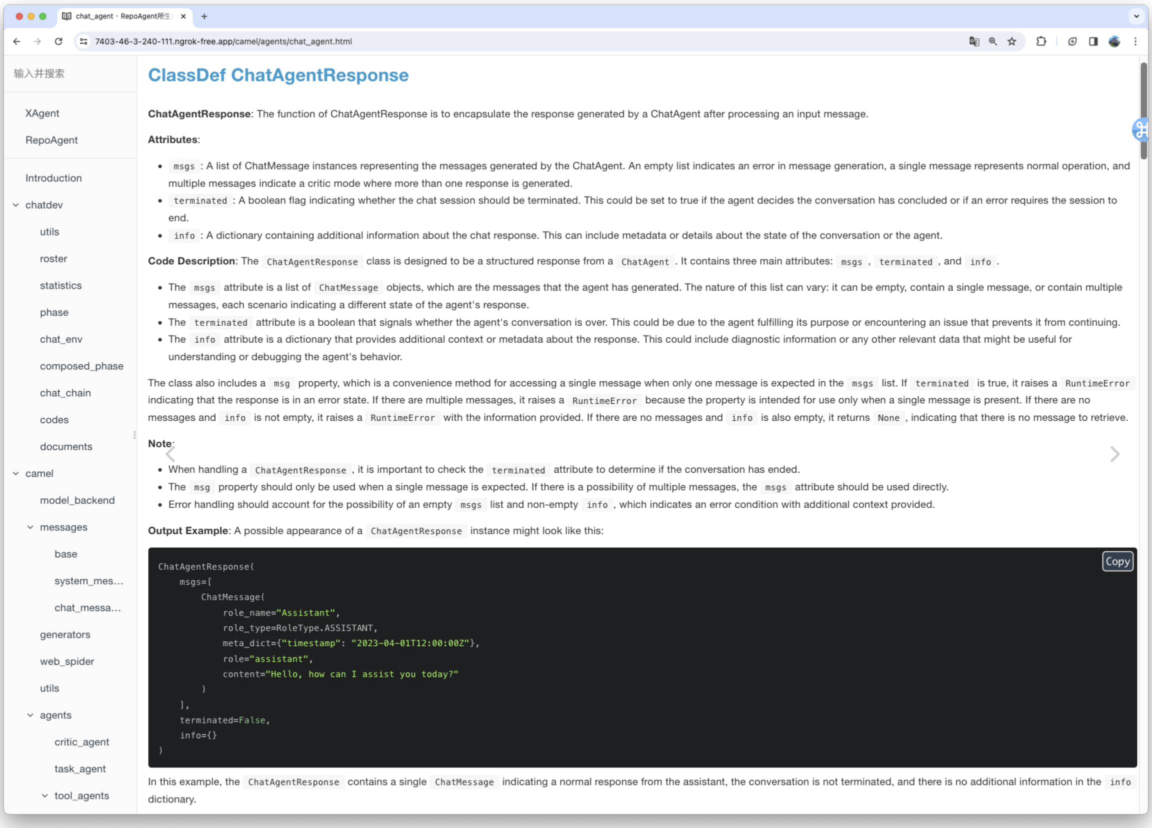
Note: StubModel is specifically designed for testing and should not be used as a real model backend in production environments. It is intended to be a lightweight and predictable alternative to more complex backends. When using StubModel in unit tests, developers should ensure that the tests do not rely on the actual logic of a model backend but rather on the interface and interactions with the backend.

Output Example: The `run` method of StubModel will always return the following dictionary:

```
{
    "id": "stub_model_id",
    "usage": {},
    "choices": [
        {
            "finish_reason": "stop",
            "message": {
                "content": "Lorem Ipsum",
                "role": "assistant"
            }
        }
    ]
}
```

This output is a mock representation of what might be returned by a real model backend, with static values that are suitable for testing purposes.

(e)



ClassDef ChatAgentResponse

ChatAgentResponse: The function of ChatAgentResponse is to encapsulate the response generated by a ChatAgent after processing an input message.

Attributes:

- `msgs` : A list of ChatMessage instances representing the messages generated by the ChatAgent. An empty list indicates an error in message generation, a single message represents normal operation, and multiple messages indicate a critic mode where more than one response is generated.
- `terminated` : A boolean flag indicating whether the chat session should be terminated. This could be set to true if the agent decides the conversation has concluded or if an error requires the session to end.
- `info` : A dictionary containing additional information about the chat response. This can include metadata or details about the state of the conversation or the agent.

Code Description: The `ChatAgentResponse` class is designed to be a structured response from a `ChatAgent`. It contains three main attributes: `msgs`, `terminated`, and `info`.

- The `msgs` attribute is a list of `ChatMessage` objects, which are the messages that the agent has generated. The nature of this list can vary: it can be empty, contain a single message, or contain multiple messages, each scenario indicating a different state of the agent's response.
- The `terminated` attribute is a boolean that signals whether the agent's conversation is over. This could be due to the agent fulfilling its purpose or encountering an issue that prevents it from continuing.
- The `info` attribute is a dictionary that provides additional context or metadata about the response. This could include diagnostic information or any other relevant data that might be useful for understanding or debugging the agent's behavior.

The class also includes a `msg` property, which is a convenience method for accessing a single message when only one message is expected in the `msgs` list. If `terminated` is true, it raises a `RuntimeError` indicating that the response is in an error state. If there are multiple messages, it raises a `RuntimeError` because the property is intended for use only when a single message is present. If there are no messages and `info` is not empty, it raises a `RuntimeError` with the information provided. If there are no messages and `info` is also empty, it returns `None`, indicating that there is no message to retrieve.

Note:

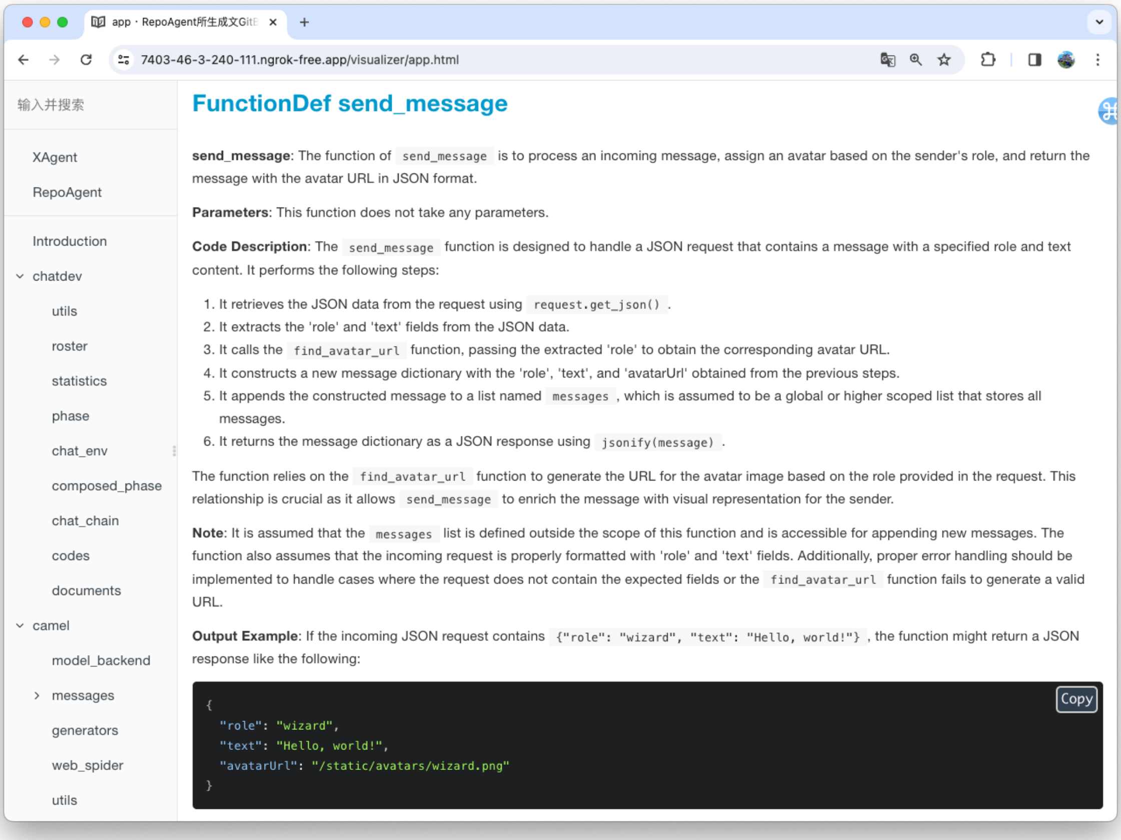
- When handling a `ChatAgentResponse`, it is important to check the `terminated` attribute to determine if the conversation has ended.
- The `msg` property should only be used when a single message is expected. If there is a possibility of multiple messages, the `msgs` attribute should be used directly.
- Error handling should account for the possibility of an empty `msgs` list and non-empty `info`, which indicates an error condition with additional context provided.

Output Example: A possible appearance of a `ChatAgentResponse` instance might look like this:

```
ChatAgentResponse(
    msgs=[
        ChatMessage(
            role_name="Assistant",
            role_type=RoleType.ASSISTANT,
            meta_dict={"timestamp": "2023-04-01T12:00:00Z"},
            role="assistant",
            content="Hello, how can I assist you today?"
        )
    ],
    terminated=False,
    info={}
)
```

In this example, the `ChatAgentResponse` contains a single `ChatMessage` indicating a normal response from the assistant, the conversation is not terminated, and there is no additional information in the `info` dictionary.

(f)



FunctionDef send_message

send_message: The function of `send_message` is to process an incoming message, assign an avatar based on the sender's role, and return the message with the avatar URL in JSON format.

Parameters: This function does not take any parameters.

Code Description: The `send_message` function is designed to handle a JSON request that contains a message with a specified role and text content. It performs the following steps:

1. It retrieves the JSON data from the request using `request.get_json()`.
2. It extracts the 'role' and 'text' fields from the JSON data.
3. It calls the `find_avatar_url` function, passing the extracted 'role' to obtain the corresponding avatar URL.
4. It constructs a new message dictionary with the 'role', 'text', and 'avatarUrl' obtained from the previous steps.
5. It appends the constructed message to a list named `messages`, which is assumed to be a global or higher scoped list that stores all messages.
6. It returns the message dictionary as a JSON response using `jsonify(message)`.

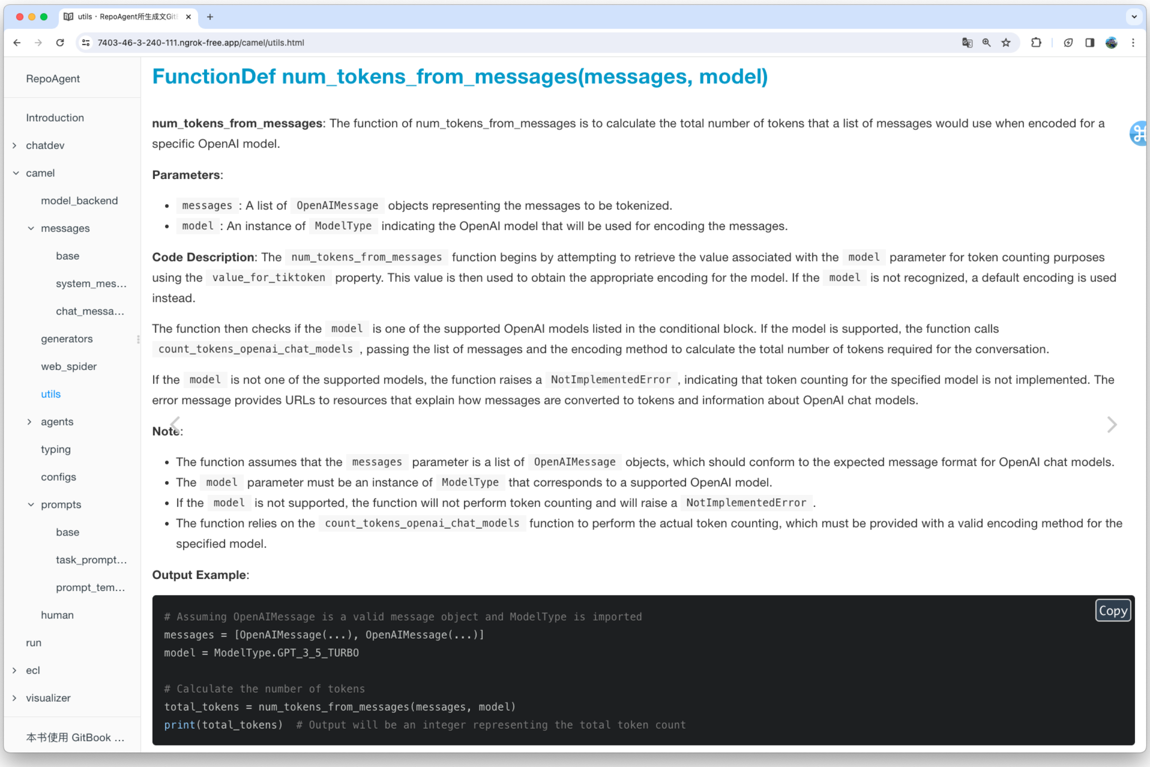
The function relies on the `find_avatar_url` function to generate the URL for the avatar image based on the role provided in the request. This relationship is crucial as it allows `send_message` to enrich the message with visual representation for the sender.

Note: It is assumed that the `messages` list is defined outside the scope of this function and is accessible for appending new messages. The function also assumes that the incoming request is properly formatted with 'role' and 'text' fields. Additionally, proper error handling should be implemented to handle cases where the request does not contain the expected fields or the `find_avatar_url` function fails to generate a valid URL.

Output Example: If the incoming JSON request contains `{"role": "wizard", "text": "Hello, world!"}`, the function might return a JSON response like the following:

```
{
  "role": "wizard",
  "text": "Hello, world!",
  "avatarUrl": "/static/avatars/wizard.png"
}
```

(g)



FunctionDef num_tokens_from_messages(messages, model)

num_tokens_from_messages: The function of num_tokens_from_messages is to calculate the total number of tokens that a list of messages would use when encoded for a specific OpenAI model.

Parameters:

- `messages` : A list of `OpenAIMessage` objects representing the messages to be tokenized.
- `model` : An instance of `ModelType` indicating the OpenAI model that will be used for encoding the messages.

Code Description: The `num_tokens_from_messages` function begins by attempting to retrieve the value associated with the `model` parameter for token counting purposes using the `value_for_tiktoken` property. This value is then used to obtain the appropriate encoding for the model. If the `model` is not recognized, a default encoding is used instead.

The function then checks if the `model` is one of the supported OpenAI models listed in the conditional block. If the model is supported, the function calls `count_tokens_openai_chat_models`, passing the list of messages and the encoding method to calculate the total number of tokens required for the conversation.

If the `model` is not one of the supported models, the function raises a `NotImplementedError`, indicating that token counting for the specified model is not implemented. The error message provides URLs to resources that explain how messages are converted to tokens and information about OpenAI chat models.

Note:

- The function assumes that the `messages` parameter is a list of `OpenAIMessage` objects, which should conform to the expected message format for OpenAI chat models.
- The `model` parameter must be an instance of `ModelType` that corresponds to a supported OpenAI model.
- If the `model` is not supported, the function will not perform token counting and will raise a `NotImplementedError`.
- The function relies on the `count_tokens_openai_chat_models` function to perform the actual token counting, which must be provided with a valid encoding method for the specified model.

Output Example:

```
# Assuming OpenAIMessage is a valid message object and ModelType is imported
messages = [OpenAIMessage(...), OpenAIMessage(...)]
model = ModelType.GPT_3_5_TURBO

# Calculate the number of tokens
total_tokens = num_tokens_from_messages(messages, model)
print(total_tokens)  # Output will be an integer representing the total token count
```

(h)

Figure 8: Samples of code documentation generated by REPOAGENT for the ChatDev repository.

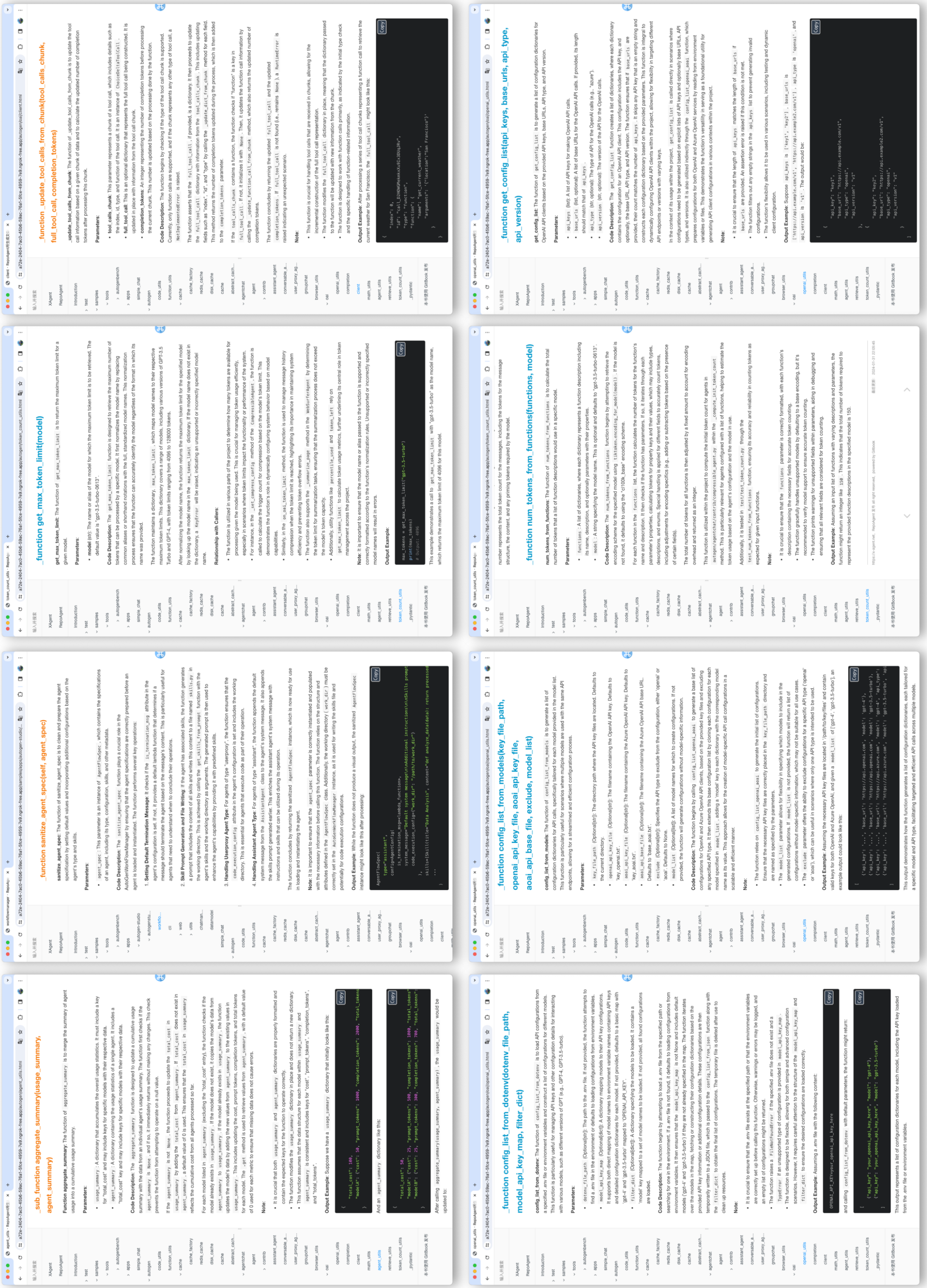


Figure 9: Samples of code documentation generated by RepoAgent for the AutoGen repository.

C Appendix: Full Prompts

C.1 A full prompt of ask_user method in AutoGPT

```
You are an AI documentation assistant, and your task is to generate documentation based on the given
code of an object. The purpose of the documentation is to help developers and beginners understand
the function and specific usage of the code.

Currently, you are in a project, and the related hierarchical structure of this project is as follows
(The current object is marked with an *):
autogpts
    autogpt
        autogpt
            commands
                user_interaction.py
                    ask_user
                        *ask_user


The path of the document you need to generate in this project is:
autogpts/autogpt/autogpt/commands/user_interaction.py/ask_user.

Now you need to generate a document for a Function, whose name is "ask_user".

The content of the code is as follows:

async def ask_user(question: str, agent: Agent) -> str:
    print(f"\nQ: {question}")
    resp = await clean_input(agent.legacy_config, "A:")
    return f"The user's answer: '{resp}'"


As you can see, the code calls the following objects, their code and docs are as following:

OBJ_NAME: clean_input
OBJ_PATH: autogpts/autogpt/autogpt/app/utils.py/clean_input

Document:
**Function Name**: clean_input

**Parameters**:
- `config`: An instance of the `Config` class, which holds the configuration settings for the
application.
- `prompt`: A string that represents the prompt to be displayed to the user. It defaults to an empty
string if not provided.

**Code Description**:
The `clean_input` function is an asynchronous function designed to process user input in a
chat-enabled application environment. It performs several key operations based on the application's
configuration and the plugins that are enabled.

1. **Plugin Handling**: If chat messages are enabled in the application configuration
(`config.chat_messages_enabled`), the function iterates through each plugin configured in the
application. For each plugin, it checks if the plugin can handle user input
(`can_handle_user_input`). If so, it invokes the plugin's `user_input` method with the user's
prompt. Depending on the plugin's response, the function may return specific keys
(`config.authorise_key` or `config.exit_key`) or the plugin's response directly.

2. **User Input Collection**: If the application is not handling the input through plugins or if
chat messages are not enabled, the function logs a debug message indicating that it is waiting for
user input via the keyboard. It then asynchronously waits for the user's input, ensuring that the
application's signal handler for interrupt signals (`SIGINT`) is not overridden. This is crucial for
allowing the application to handle interrupt signals as configured elsewhere in the application,
particularly in `autogpt/main.py`.

3. **Interrupt Handling**: The function is equipped to handle a `KeyboardInterrupt` exception, which
occurs when the user attempts to interrupt the application (typically by pressing Ctrl+C). In such
cases, it logs a message indicating that the user has interrupted AutoGPT and that the application
is quitting, and then exits the application with a status code of 0.
```

Note:
- The function's ability to handle plugins dynamically allows for a flexible and extensible architecture where different plugins can be developed and integrated to process user input in various ways.
- The use of asynchronous I/O (`await session.prompt_async`) for collecting user input makes the function suitable for asynchronous applications, ensuring that the application remains responsive.
- Proper handling of `KeyboardInterrupt` ensures a graceful shutdown of the application when the user decides to interrupt the execution.

Output Example:
- If a plugin processes the input and responds with "yes", the function might return the value of `config.authorise_key`.
- If the user inputs "Hello" and no plugins are handling the input, the function returns "Hello".
- If the user interrupts the application, no output is returned as the application exits.

[Code begin of clean_input]

~~~
```
async def clean_input(config: Config, prompt: str = ""):
    try:
        if config.chat_messages_enabled:
            for plugin in config.plugins:
                if not hasattr(plugin, "can_handle_user_input"):
                    continue
                if not plugin.can_handle_user_input(user_input=prompt):
                    continue
                plugin_response = plugin.user_input(user_input=prompt)
                if not plugin_response:
                    continue
                if plugin_response.lower() in [
                    "yes",
                    "yeah",
                    "y",
                    "ok",
                    "okay",
                    "sure",
                    "alright",
                ]:
                    return config.authorise_key
                elif plugin_response.lower() in [
                    "no",
                    "nope",
                    "n",
                    "negative",
                ]:
                    return config.exit_key
                return plugin_response

        # ask for input, default when just pressing Enter is y
        logger.debug("Asking user via keyboard...")

        # handle_sigint must be set to False, so the signal handler in the
        # autogpt/main.py could be employed properly. This referes to
        # https://github.com/Significant-Gravitas/AutoGPT/pull/4799/files/
        3966cdfd694c2a80c0333823c3bc3da090f85ed3#r1264278776
        answer = await session.prompt_async(ANSI(prompt + " "), handle_sigint=False)
        return answer
    except KeyboardInterrupt:
        logger.info("You interrupted AutoGPT")
        logger.info("Quitting...")
        exit(0)

```==========
~~~

[Code end of clean_input]

Also, the code has been called by the following objects, their code and docs are as following:

OBJ_NAME: execute_step
OBJ_PATH: autogpts/autogpt/autogpt/app/agent_protocol_server.py/ AgentProtocolServer/execute_step
Document:
None

[Code begin of execute_step]

~~~~
```
    async def execute_step(self, task_id: str, step_request: StepRequestBody) -> Step:
        """Create a step for the task."""
        logger.debug(f"Creating a step for task with ID: {task_id}...")

        # Restore Agent instance
        task = await self.get_task(task_id)
        agent = configure_agent_with_state(
            state=self.agent_manager.retrieve_state(task_agent_id(task_id)),
            app_config=self.app_config,
            llm_provider=self._get_task_llm_provider(task),
        )

        # According to the Agent Protocol spec, the first execute_step request contains
        #  the same task input as the parent create_task request.
        # To prevent this from interfering with the agent's process, we ignore the input
        #  of this first step request, and just generate the first step proposal.
        is_init_step = not bool(agent.event_history)
        execute_command, execute_command_args, execute_result = None, None, None
        execute_approved = False

        # HACK: only for compatibility with AGBenchmark
        if step_request.input == "y":
            step_request.input = ""

        user_input = step_request.input if not is_init_step else ""

        if (
            not is_init_step
            and agent.event_history.current_episode
            and not agent.event_history.current_episode.result
        ):
            execute_command = agent.event_history.current_episode.action.name
            execute_command_args = agent.event_history.current_episode.action.args
            execute_approved = not user_input

            logger.debug(
                f"Agent proposed command"
                f" {execute_command}({fmt_kwargs(execute_command_args)})."
                f" User input/feedback: {repr(user_input)}"
            )

        # Save step request
        step = await self.db.create_step(
            task_id=task_id,
            input=step_request,
            is_last=execute_command == finish.__name__ and execute_approved,
        )
        agent.llm_provider = self._get_task_llm_provider(task, step.step_id)

        # Execute previously proposed action
        if execute_command:
            assert execute_command_args is not None
            agent.workspace.on_write_file = lambda path: self._on_agent_write_file(
                task=task, step=step, relative_path=path
            )

            if step.is_last and execute_command == finish.__name__:
                assert execute_command_args
                step = await self.db.update_step(
                    task_id=task_id,
                    step_id=step.step_id,
                    output=execute_command_args["reason"],
                )
                logger.info(
                    f"Total LLM cost for task {task_id}: "
                    f"${round(agent.llm_provider.get_incurred_cost(), 2)}"
                )
                return step
~~~~

```
    if execute_command == ask_user.__name__: # HACK
        execute_result = ActionSuccessResult(outputs=user_input)
        agent.event_history.register_result(execute_result)
    elif not execute_command:
        execute_result = None
    elif execute_approved:
        step = await self.db.update_step(
            task_id=task_id,
            step_id=step.step_id,
            status="running",
        )
        # Execute previously proposed action
        execute_result = await agent.execute(
            command_name=execute_command,
            command_args=execute_command_args,
        )
    else:
        assert user_input
        execute_result = await agent.execute(
            command_name="human_feedback", # HACK
            command_args={},
            user_input=user_input,
        )

# Propose next action
try:
    next_command, next_command_args, raw_output = await agent.propose_action()
    logger.debug(f"AI output: {raw_output}")
except Exception as e:
    step = await self.db.update_step(
        task_id=task_id,
        step_id=step.step_id,
        status="completed",
        output=f"An error occurred while proposing the next action: {e}",
    )
    return step

# Format step output
output = (
    (
        f"`{execute_command}({fmt_kwargs(execute_command_args)})` returned:"
        + ("\n\n" if "\n" in str(execute_result) else " ")
        + f"{execute_result}\n\n"
    )
    if execute_command_args and execute_command != ask_user.__name__
    else ""
)
output += f"{raw_output['thoughts']['speak']}\n\n"
output += (
    f"Next Command: {next_command}({fmt_kwargs(next_command_args)})"
    if next_command != ask_user.__name__
    else next_command_args["question"]
)

additional_output = {
    **(
        {
            "last_action": {
                "name": execute_command,
                "args": execute_command_args,
                "result": (
                    orjson.loads(execute_result.json())
                    if not isinstance(execute_result, ActionErrorResult)
                    else {
                        "error": str(execute_result.error),
                        "reason": execute_result.reason,
                    }
                ),
            },
```

~~~
                }
                if not is_init_step
                else {}
            ),
            **raw_output,
        }

        step = await self.db.update_step(
            task_id=task_id,
            step_id=step.step_id,
            status="completed",
            output=output,
            additional_output=additional_output,
        )

        logger.debug(
            f"Running total LLM cost for task {task_id}: "
            f"${round(agent.llm_provider.get_incurred_cost(), 3)}"
        )
        agent.state.save_to_json_file(agent.file_manager.state_file_path)
        return step

```==========
[Code end of execute_step]

Please generate a detailed explanation document for this object based on the code of the target
object itself and combine it with its calling situation in the project.

Please write out the function of this Function in bold plain text, followed by a detailed analysis
in plain text (including all details), in language English to serve as the documentation for this
part of the code.

The standard format is as follows:

**ask_user**: The function of ask_user is XXX
**parameters**: The parameters of this Function.
- parameter1: XXX
- parameter2: XXX
- ...
**Code Description**: The description of this Function.
(Detailed and CERTAIN code analysis and description...None)
**Note**: Points to note about the use of the code
**Output Example**: Mock up a possible appearance of the code's return value.

Please note:
- Any part of the content you generate SHOULD NOT CONTAIN Markdown hierarchical heading and divider
syntax.
- Write mainly in the desired language. If necessary, you can write with some English words in the
analysis and description to enhance the document's readability because you do not need to translate
the function name or variable name into the target language.

Keep in mind that your audience is document readers, so use a deterministic tone to generate precise
content and don't let them know you're provided with code snippet and documents. AVOID ANY
SPECULATION and inaccurate descriptions! Now, provide the documentation for the target object in
English in a professional way.
~~~

D Appendix: Chat With Repo

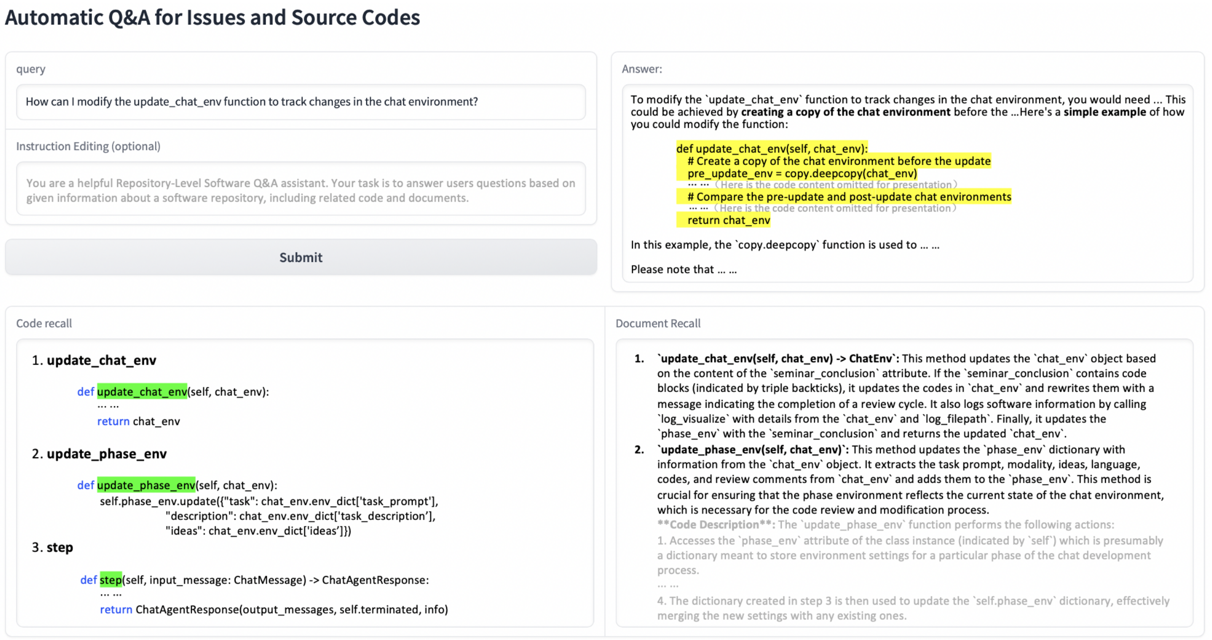
Moving beyond documentation generation, we are actively exploring how best to use REPOAGENT and examining its potential for a broader range of downstream applications in the future. We categorize these applications as:

- README.md Generation
- Automatic Q&A for Issues and Source Codes
- Unit Test Generation
- Automated Development of New Features

- Repo Level Debugging
- Generation of Public Tutorial Documentation

We conceptualize "**Chat With Repo**" as a unified gateway for these downstream applications, acting as a connector that links REPOAGENT to human users and other AI agents. Our future research will focus on adapting the interface to various downstream applications and customizing it to meet their unique characteristics and implementation requirements.

Here we demonstrate a preliminary prototype of **Automatic Q&A for Issues and Code Explanation**. A running example is shown in Figure 10. The program begins by pre-vectorizing code documentation and storing it in a vector database. When a query request is received, it is transformed into an embedding vector for fetching relevant documentation information from the database. This is followed by using the documentation's MetaInfo to locate the pertinent source code, effectively retrieving relevant sections of both documentation text and source code. Moreover, beyond embedding search, a multi-way recall mechanism has been developed, incorporating entity recognition with keyword search. This involves extracting code entities from the user's question using a LLM, and conducting searches across documentation and code repositories to match the top K returned documentation and code blocks. A weighting module has been developed for recalling the most relevant information. Additionally, we input directory tree information to help the model better understand the entire repository. The final step is to concatenate documentation and code blocks retrieved through both mechanisms, along with the target object's parent code, referencing code, and directory tree information, into a prompt for the LLM to generate answers. This sophisticated RAG-based retrieval system bridges human natural language with code language, enabling precise recall at the repository level and paving the way for downstream applications.


Automatic Q&A for Issues and Source Codes
query
Instruction Editing (optional)
Submit
Answer:
Code recall
Document Recall

How can I modify the update_chat_env function to track changes in the chat environment?

You are a helpful Repository-Level Software Q&A assistant. Your task is to answer users questions based on given information about a software repository, including related code and documents.

To modify the `update_chat_env` function to track changes in the chat environment, you would need ... This could be achieved by **creating a copy of the chat environment** before the ...Here's a **simple example** of how you could modify the function:

```
def update_chat_env(self, chat_env):
   # Create a copy of the chat environment before the update
   pre_update_env = copy.deepcopy(chat_env)
   ... ... （Here is the code content omitted for presentation）
   # Compare the pre-update and post-update chat environments
   ... ... （Here is the code content omitted for presentation）
   return chat_env
```

In this example, the `copy.deepcopy` function is used to ... ...

Please note that ... ...

1. **update_chat_env**

```
def update_chat_env(self, chat_env):
     ... ...
     return chat_env
```

2. **update_phase_env**

```
def update_phase_env(self, chat_env):
     self.phase_env.update({"task": chat_env.env_dict['task_prompt'],
                     "description": chat_env.env_dict['task_description'],
                     "ideas": chat_env.env_dict['ideas']})
```

3. **step**

```
def step(self, input_message: ChatMessage) -> ChatAgentResponse:
     ... ...
     return ChatAgentResponse(output_messages, self.terminated, info)
```

1. **`update_chat_env(self, chat_env) -> ChatEnv`:** This method updates the `chat_env` object based on the content of the `seminar_conclusion` attribute. If the `seminar_conclusion` contains code blocks (indicated by triple backticks), it updates the codes in `chat_env` and rewrites them with a message indicating the completion of a review cycle. It also logs software information by calling `log_visualize` with details from the `chat_env` and `log_filepath`. Finally, it updates the `phase_env` with the `seminar_conclusion` and returns the updated `chat_env`.
2. **`update_phase_env(self, chat_env)`:** This method updates the `phase_env` dictionary with information from the `chat_env` object. It extracts the task prompt, modality, ideas, language, codes, and review comments from `chat_env` and adds them to the `phase_env`. This method is crucial for ensuring that the phase environment reflects the current state of the chat environment, which is necessary for the code review and modification process.
   ****Code Description**:** The `update_phase_env` function performs the following actions:
   1. Accesses the `phase_env` attribute of the class instance (indicated by `self`) which is presumably a dictionary meant to store environment settings for a particular phase of the chat development process.
   ... ...
   4. The dictionary created in step 3 is then used to update the `self.phase_env` dictionary, effectively merging the new settings with any existing ones.

Figure 10: Demonstration of Automatic Q&A for Issues and Code Explanation by the "Chat With Repo" interface.

A real world “Chat With Repo” example with input and output is shown as follows.

Query

````
How would you modify the update\_chat\_env function to track changes in the chat environment?
````

Answer

~~~
To modify the `update_chat_env` function to track changes in the chat environment, you would need to
add a mechanism to store the state of the chat environment before and after the update. This could
be achieved by creating a copy of the chat environment before the update and comparing it with the
state of the chat environment after the update. Here's a simple example of how you could modify the
function:

```python
def update_chat_env(self, chat_env):
    # Create a copy of the chat environment before the update
    pre_update_env = copy.deepcopy(chat_env)

    # Update the chat environment as per your requirements
    # This is where you would add your update logic

    # Compare the pre-update and post-update chat environments
    if pre_update_env != chat_env:
        print(”Chat environment has been updated.”)
    else:
        print(”Chat environment has not been changed.”)

    return chat_env
```

In this example, the `copy.deepcopy` function is used to create a deep copy of the `chat_env`
object. This ensures that changes to the `chat_env` object after the copy do not affect the
`pre_update_env` object. After the update, the `pre_update_env` and `chat_env` objects are
compared. If they are not equal, this means that the chat environment has been updated.

Please note that this is a simple example and the actual implementation may vary depending on the
specific requirements of your project. For instance, you might want to log the changes in a file
instead of printing them, or you might want to use a more sophisticated method for comparing the
chat environments.
~~~

Embedding Recall

````
1.
**Function Name**: `update_phase_env`

**Function of update_phase_env**: The function `update_phase_env` is designed to update the
environment settings of a particular phase in a chat development context by incorporating various
elements from a given chat environment object.

**Parameters**: The parameters of this function include:
- `self`: Represents the instance of the class where the `update_phase_env` function resides. It
allows the function to access the attributes and methods of the class instance.
- `chat_env`: An object that contains the environment settings for the chat. It is expected to have
an attribute `env_dict` which is a dictionary containing keys such as 'task_prompt', 'modality',
'ideas', and 'language'. Additionally, `chat_env` should have methods `get_codes` and
`get_requirements` that return specific information related to codes and requirements.

**Code Description**: The `update_phase_env` function performs the following actions:
1. Accesses the `phase_env` attribute of the class instance (indicated by `self`) which is
presumably a dictionary meant to store environment settings for a particular phase of the chat
development process.
2. Calls the `update` method on `self.phase_env` to modify its contents.
3. Constructs a new dictionary with the following keys and values:
   - `”task”`: Set to the value associated with the `'task_prompt'` key in the `chat_env.env_dict`
   dictionary.
````

   - `"modality"`: Set to the value associated with the `'modality'` key in the `chat_env.env_dict` dictionary.
   - `"ideas"`: Set to the value associated with the `'ideas'` key in the `chat_env.env_dict` dictionary.
   - `"language"`: Set to the value associated with the `'language'` key in the `chat_env.env_dict` dictionary.
   - `"codes"`: Obtained by calling the `get_codes` method on the `chat_env` object.
   - `"requirements"`: Obtained by calling the `get_requirements` method on the `chat_env` object.

4. The dictionary created in step 3 is then used to update the `self.phase_env` dictionary, effectively merging the new settings with any existing ones.

Note: When using the `update_phase_env` function, it is important to ensure that the `chat_env` object passed as a parameter is properly initialized and contains all the necessary attributes and methods. The `env_dict` attribute must include the keys 'task_prompt', 'modality', 'ideas', and 'language', and the object must have the `get_codes` and `get_requirements` methods implemented. Failure to meet these conditions may result in a `KeyError` or `AttributeError`. Additionally, this function mutates the state of the `self.phase_env` dictionary, so any changes made will persist in the instance of the class where this function is called.
2.
CodeReviewModification: The function of the `CodeReviewModification` class is to manage the phase of code review modifications within a chat development environment. It inherits from a base `Phase` class and is designed to update the environment with information relevant to the code review and modification process.

Attributes: The attributes of this Class are inherited from the `Phase` class and any that are added or modified within this class.
- `phase_env`: A dictionary attribute inherited from `Phase` that stores environment variables relevant to the current phase.
- `seminar_conclusion`: Although not explicitly defined in the provided code, it is implied that this attribute is used to store the conclusion of a seminar or review session.

Code Description: The `CodeReviewModification` class contains two main methods:

1. `update_phase_env(self, chat_env)`: This method updates the `phase_env` dictionary with information from the `chat_env` object. It extracts the task prompt, modality, ideas, language, codes, and review comments from `chat_env` and adds them to the `phase_env`. This method is crucial for ensuring that the phase environment reflects the current state of the chat environment, which is necessary for the code review and modification process.

2. `update_chat_env(self, chat_env) -> ChatEnv`: This method updates the `chat_env` object based on the content of the `seminar_conclusion` attribute. If the `seminar_conclusion` contains code blocks (indicated by triple backticks), it updates the codes in `chat_env` and rewrites them with a message indicating the completion of a review cycle. It also logs software information by calling `log_visualize` with details from the `chat_env` and `log_filepath`. Finally, it updates the `phase_env` with the `seminar_conclusion` and returns the updated `chat_env`.

Note: Points to note about the use of the code:
- The `chat_env` parameter is expected to be an object that contains an `env_dict` with keys such as 'task_prompt', 'modality', 'ideas', 'language', and 'review_comments', as well as methods like `get_codes()` and `update_codes()`.
- The `seminar_conclusion` attribute must be set before calling `update_chat_env` as it uses this attribute to update the `chat_env`.
- The `log_visualize` function and `get_info` function are not defined within the provided code snippet, so they should be implemented elsewhere in the project or imported from a module.
- The `ChatEnv` return type suggests that there is a `ChatEnv` class defined elsewhere in the project, which should be used in conjunction with this class.

3.
chatting: The function of `chatting` is to conduct a simulated chat session between two roles within a software development environment, with the goal of reaching a conclusion on a specific phase of the project.

Parameters:
- `chat_env`: The global chat environment which contains configurations and context for the chat session.
- `task_prompt`: A string representing the user's query or task that needs to be addressed during the chat.
- `assistant_role_name`: The name of the role assumed by the assistant in the chat.
- `user_role_name`: The name of the role assumed by the user initiating the chat.
- `phase_prompt`: A string containing the prompt for the current phase of the chat.

- `phase_name`: The name of the current phase of the chat.
- `assistant_role_prompt`: The prompt associated with the assistant's role.
- `user_role_prompt`: The prompt associated with the user's role.
- `task_type`: An enumeration value representing the type of task being simulated in the chat.
- `need_reflect`: A boolean indicating whether the chat session requires reflection to generate a conclusion.
- `with_task_specify`: A boolean indicating whether the task needs to be specified within the chat.
- `model_type`: An enumeration value indicating the type of language model to be used for generating responses.
- `placeholders`: A dictionary containing placeholders that can be used to fill in the phase environment for generating the phase prompt.
- `chat_turn_limit`: An integer representing the maximum number of turns the chat session can have.

Code Description:
The `chatting` function starts by ensuring that the `placeholders` argument is not `None` and that the `chat_turn_limit` is within an acceptable range (1 to 100). It then checks if the roles specified by `assistant_role_name` and `user_role_name` exist within the `chat_env`.

A `RolePlaying` session is initialized with the provided role names, prompts, task type, and model type. The function then begins the chat session by initializing the first user message using the `init_chat` method of the `RolePlaying` session.

The chat session proceeds in turns, where each turn consists of the user sending a message to the assistant and the assistant responding. The messages and responses are generated by interacting with a language model (LLM). The conversation is logged using a `log_visualize` function, which is not defined within the provided code snippet.

During the chat, the function looks for a special `<INFO>` marker in the conversation, which indicates a significant conclusion has been reached. If such a conclusion is found, or if the chat is terminated, the loop ends.

If the `need_reflect` flag is set, the function may call `self_reflection` to generate a conclusion if one has not been reached during the chat session. The reflection is based on the entire conversation history and the context of the phase.

Finally, the function logs the seminar conclusion, extracts the relevant part after the `<INFO>` marker, and returns it as the result of the chat session.

Note:
- The function assumes that the `chat_env` has methods `exist_employee` to check for the existence of roles.
- The `RolePlaying` class is used to simulate the chat session and is expected to have methods like `init_chat` and `step`.
- The `log_visualize` function is used for logging purposes but is not defined within the provided code snippet.
- The function raises a `ValueError` if the specified roles are not found within the `chat_env`.
- The `self_reflection` method is used for generating reflections and is assumed to be a member of the same class.

Output Example:
If the chat session concludes with a marked conclusion, the function might return something like:
```
"PowerPoint is the best choice for our presentation needs."
```
If the chat session does not reach a marked conclusion but requires reflection, the `self_reflection` method might return:
```
"Yes"
```
In cases where the chat is terminated without a marked conclusion and no reflection is needed, the last message from the assistant might be returned as is.
4.
Function Name: execute

Purpose: The function `execute` is designed to handle a phase of a chat development environment by updating the environment, checking for module not found errors, resolving them if present, and conducting a chat session if no such errors are found.

Parameters:
- `chat_env`: An instance of `ChatEnv`, which represents the current chat environment.

```
- `chat_turn_limit`: An integer indicating the maximum number of turns allowed in the chat session.
- `need_reflect`: A boolean indicating whether reflection is needed in the chat session.

**Code Description**:
The `execute` function begins by updating the phase environment with the current `chat_env`. It
then checks if there is a "ModuleNotFoundError" in the `test_reports` of the `phase_env`. If such
an error is present, it attempts to fix the error by calling `chat_env.fix_module_not_found_error`
and logs the error for visualization.

The function then uses a regular expression to find all instances of the error message indicating a
missing module and constructs a string `pip_install_content` that contains the commands to install
the missing modules using `pip`. This string is also logged for visualization.

If no "ModuleNotFoundError" is found, the function proceeds to conduct a chat session by calling the
`chatting` method with various parameters such as `chat_env`, `task_prompt`, `need_reflect`, role
names, prompts, `chat_turn_limit`, and placeholders. The result of this chat session is stored in
`self.seminar_conclusion`.

After handling the error or conducting the chat session, the function updates the chat environment
with the potentially modified `chat_env` and returns it.

**Note**:
- The function assumes that the `chat_env` object has the methods `fix_module_not_found_error` and
`update_chat_env` implemented.
- The `chatting` method is also assumed to be implemented and is responsible for conducting the chat
session.
- The function uses regular expressions to parse error messages, so it is important that the error
messages follow the expected format for the regular expressions to work correctly.
- The function logs actions for visualization, which implies that a logging mechanism should be in
place for the output to be meaningful.

**Output Example**:
A possible appearance of the code's return value could be an updated `ChatEnv` object with
modifications based on the error handling and chat session conducted within the `execute` function.
The object would reflect the new state of the chat environment after the execution of this function.
```

Key Words Recall

```
Key words:
modify, update_chat_env, track

Code:
['\n```python\n  def update_chat_env(self, chat_env):\n    return chat_env\n\n```']
```

Code Recall

```
def update_chat_env(self, chat_env) -> ChatEnv:
    chat_env._update_requirements(self.seminar_conclusion)
    chat_env.rewrite_requirements()
    log_visualize(
        "**[Software Info]**:\n\n {}".format(get_info(chat_env.env_dict['directory'],
        self.log_filepath)))
    return chat_env

def update_chat_env(self, chat_env) -> ChatEnv:
    chat_env.update_codes(self.seminar_conclusion)
    if len(chat_env.codes.codebooks.keys()) == 0:
        raise ValueError("No Valid Codes.")
    chat_env.rewrite_codes("Code Complete #" + str(self.phase_env["cycle_index"]) + " Finished")
    log_visualize(
        "**[Software Info]**:\n\n {}".format(get_info(chat_env.env_dict['directory'],
        self.log_filepath)))
    return chat_env
```