

# GitChameleon 2.0: Evaluating AI Code Generation Against Python Library Version Incompatibilities

Diganta Misra<sup>1,2\*,\*</sup>, Nizar Islah<sup>3,10\*</sup>, Victor May<sup>4</sup>, Brice Rauby<sup>3,5</sup>, Zihan Wang<sup>6</sup>,  
Justine Gehring<sup>3,7,8</sup>, Antonio Orvieto<sup>1,2,9</sup>, Muawiz Chaudhary<sup>3</sup>,  
Eilif B. Muller<sup>3,10</sup>, Irina Rish<sup>3,10</sup>, Samira Ebrahimi Kahou<sup>3</sup>, Massimo Caccia<sup>11</sup>

<sup>1</sup>ELLIS Institute Tübingen, <sup>2</sup>MPI-IS Tübingen, <sup>3</sup>Mila Quebec AI Institute, <sup>4</sup>Google,  
<sup>5</sup>Polytechnique Montréal, <sup>6</sup>McGill University, Montréal, <sup>7</sup>Moderne, <sup>8</sup>Gologic,  
<sup>9</sup>Tübingen AI Center, <sup>10</sup>Université de Montréal, <sup>11</sup>ServiceNow Research

Correspondence: diganta.misra@tue.ellis.eu, nizar.islah@mila.quebec

## Abstract

The rapid evolution of software libraries poses a considerable hurdle for code generation, necessitating continuous adaptation to frequent version updates while preserving backward compatibility. While existing code evolution benchmarks provide valuable insights, they typically lack execution-based evaluation for generating code compliant with specific library versions. To address this, we introduce **GitChameleon 2.0**, a novel, meticulously curated dataset comprising 328 Python code completion problems, each conditioned on specific library versions and accompanied by executable unit tests. **GitChameleon 2.0** rigorously evaluates the capacity of contemporary large language models (LLMs), LLM-powered agents, code assistants, and RAG systems to perform version-conditioned code generation that demonstrates functional accuracy through execution. Our extensive evaluations indicate that state-of-the-art systems encounter significant challenges with this task; enterprise models achieving baseline success rates in the 48-51% range, underscoring the intricacy of the problem. By offering an execution-based benchmark emphasizing the dynamic nature of code libraries, **GitChameleon 2.0** enables a clearer understanding of this challenge and helps guide the development of more adaptable and dependable AI code generation methods.<sup>1</sup>

## 1 Introduction

Large language models (LLMs) are increasingly integral to software development, being adopted for tasks like code generation and review (Council, 2024; Lambiase et al., 2025).

Despite LLM advancements like larger context windows (Su et al., 2023), faster inference (Dao

\* \*Equal contribution. Roles: Team Leads, Data and Core Contributors, Senior Advisors

<sup>1</sup>GitChameleon 2.0 is the expanded successor to the original GitChameleon benchmark, increasing coverage from 116 to 328 Python code completion problems while preserving the same version-conditioned generation setting.

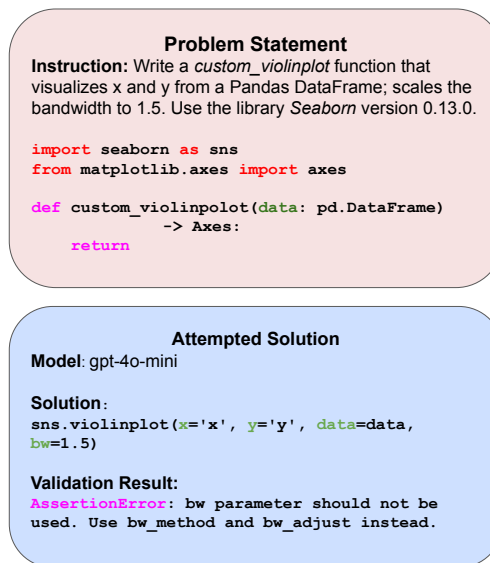


Figure 1: In this **GitChameleon 2.0** problem, the gpt-4o-mini model produced an incorrect solution due for `seaborn.violinplot` by using the deprecated `bw` parameter, instead of the appropriate `bw_method` and `bw_adjust` required by the specified library version.

et al., 2022), and high performance on general coding benchmarks (Hendrycks et al., 2021; Chen et al., 2021), a critical capability remains under-evaluated: generating code that is compliant with a specific library version. This task of version-switching, which is essential for robust development in environments with fixed or legacy dependencies, is not well-verified in contemporary LLMs.

Existing benchmarks, while valuable, often focus on migrating codebases to newer versions (i.e., code evolution) or use non-executable evaluation methods. They do not fully address the challenge of generating new, functionally correct code for a static version constraint. For instance, PyMigBench (Islam et al., 2023) provides comprehensive datasets of real-world, inter-library migrations, rather than focusing on executable,

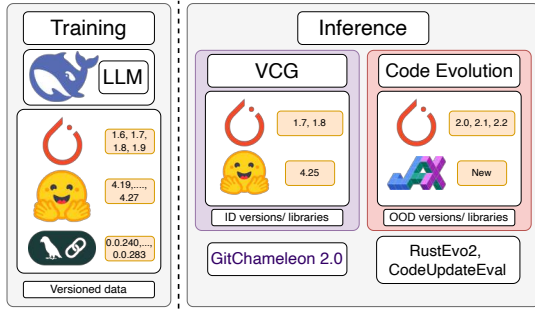


Figure 2: An illustration of two evaluation paradigms for code generation models. **Code Evolution** (right) assesses model capabilities on *out-of-distribution* (OOD) data, using library versions or new libraries not encountered during training. In contrast, **Version-Conditioned Generation (VCG)** (left) focuses on the practical ability to generate code for specific, *in-distribution* (ID) library versions that the model has seen before.

intra-library tasks conditioned on specific versions. CodeUpdateArena (Liu et al., 2025) valuably assesses LLM knowledge editing using synthetically generated API updates for functions in popular libraries, a different approach from using documented historical breaking changes. Other relevant studies, such as Wang et al. (2024b), investigate the propensity of LLMs to generate code with deprecated APIs, which does not entirely cover the broader capability of generating software that adheres to precise, user-specified library versions involving various types of API changes.

**Code Evolution vs. Version Conditioned Generation (VCG).** Existing code evaluation benchmarks often focus on assessing the code evolution or migration capabilities of LLMs, where changes occur only in the forward direction and typically involve unseen library versions or entirely new libraries. This framing inherently makes the task out-of-distribution (OOD), as illustrated in Figure 2. In contrast, version-conditioned generation (VCG)—the ability of LLMs to produce code aligned with specific, previously seen library versions—is critical for practical deployment. It enables models to function reliably in real-world production environments or constrained settings where the libraries in use may not be the latest stable versions. To better evaluate this capability, a benchmark must pose problems that are strictly *in-distribution* (ID) with respect to the relevant library version(s) required to solve them.

To bridge this gap, our work introduces

**GitChameleon 2.0**, an executable benchmark designed to assess the capability of LLMs and AI agents in generating version-aware Python code. **GitChameleon 2.0** features problems centered on documented breaking changes from popular libraries, requiring models to produce solutions for explicitly specified versions (an illustrative example is shown in Figure 1). The development of such a benchmark faces challenges in meticulously curating version-specific breaking changes from library changelogs and crafting corresponding testable scenarios. Our comprehensive evaluation of diverse LLM-based tools on **GitChameleon 2.0** reveals critical limitations in existing systems’ ability to handle library versioning.

In summary, our contributions are highlighted as follows:

- We introduce a novel code completion benchmark **GitChameleon 2.0** consisting of **328** Python-based version-conditioned problems, including visible tests for self-debugging and documentation references for Retrieval-Augmented Generation (RAG).
- We present a comprehensive empirical study on **GitChameleon 2.0**, evaluating the capabilities of a diverse range of contemporary AI code generation systems, including AI agents, IDE-integrated and CLI-based coding assistants, and RAG-based LLM pipelines.
- We reveal critical limitations in the ability of current AI systems to adhere to specific versioning constraints and highlight factors impacting their performance, thereby providing insights to steer the development of more adaptable and dependable AI code generation methods.

## 2 **GitChameleon 2.0** Benchmark

We introduce **GitChameleon 2.0**, a manually authored benchmark that comprises **328** Python-based version-conditioned problems focused on popular code libraries. To evaluate performance on **GitChameleon 2.0**, each problem is accompanied by a suite of assertion-based unit tests, enabling a thorough execution-based assessment of potential solutions. The dataset was constructed through careful manual effort, with over 350 hours invested in identifying historical breaking changes, crafting problem statements, and validating unit tests. In the following sections, we detail the dataset structure,

dataset statistics, evaluation metrics, and sample verification process.

## 2.1 Dataset Structure

Each dataset sample includes a problem related to a breaking change in a Python library.

To validate a candidate solution, we provide a suite of tests, consisting of a comprehensive suite of **Hidden Tests** to be used for model performance evaluation and ranking and a concise **Visible Test** to provide execution feedback for Self-Debugging (Chen et al., 2023) experiments.

The detailed structure of dataset samples is presented in Table 5. For a schematic of the workflow for evaluating a method against a sample from **GitChameleon 2.0**, see Figure 5.

## 2.2 Evaluation Metrics

The benchmark metric is the success rate on hidden tests, which directly penalizes version mismatches that cause runtime errors during our execution-based validation. As a secondary metric, we use the API Hit Rate (Wang et al., 2024a): the percentage of generated solutions that correctly call all APIs specified in the ground-truth solution. Note that this hit rate can be lower than the success rate, as functionally correct alternative solutions may use different APIs.

## 2.3 Statistics

**GitChameleon 2.0** consists of 328 Python-based version conditioned problems based on 26 libraries spanning scientific computing, data science and web development. The samples were collected from version releases over a period from the year 2014 to 2023 and exclude legacy and yanked version releases.

As demonstrated in Fig. 4(a), most of the samples in **GitChameleon 2.0** are from versions of libraries released in the years 2021-2023. We intentionally use versions that fall within the training window of most evaluated models. The challenge is therefore not one of data contamination, but of **control and disambiguation**: when a model has been exposed to multiple library versions, can it correctly generate code for the specific version required by the prompt.

Further details about the benchmark and its construction process are presented in Appendix A.

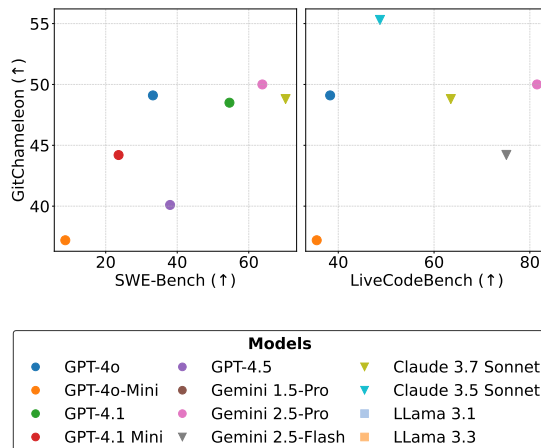


Figure 3: Can you predict **GitChameleon 2.0** performance from other code generation benchmarks? Here we present the Spearman ( $\rho$ ) and Pearson ( $r$ ) correlations between **GitChameleon 2.0**, SWE-Bench (Jimenez et al., 2024), and LiveCodeBench (Jain et al., 2024). **GitChameleon** exhibits a moderate correlation with SWE-Bench, with  $\rho$  of 0.550 and  $r$  of 0.675; and a weak correlation with LiveCodeBench, with  $\rho$  of 0.214 and  $r$  of 0.130.

## 3 Empirical Study

We evaluate **GitChameleon 2.0** in a comprehensive selection of settings, including Greedy Decoding, Chain-of-Thought (Wei et al., 2023), Self-Debugging (Chen et al., 2023), RAG (Lewis et al., 2020), Multi-Step Agents (Yao et al., 2023) and enterprise Coding Assistant software products, to assess their ability to generate version-specific executable code.

This section first presents the experimental setup, then reports the experiment results in each setting, and finally shows a breakdown of the observed results along a few key dimensions.

### 3.1 Experimental Setup

In this section, we present the experimental setup used for each of our settings. To ensure version compliance, we use a dual control mechanism: the target version is explicitly included in the model’s prompt, and the validation environment is configured with that exact library version. All prompts are shown in Appendix I. For prompt optimization, we used the Anthropic Prompt Improver<sup>2</sup>. Further automated prompt optimization efforts did not make a significant change, as described in Table 11. We evaluate models across four settings:

<sup>2</sup><https://docs.anthropic.com/en/docs/build-with-h-claude/prompt-engineering/prompt-improver>

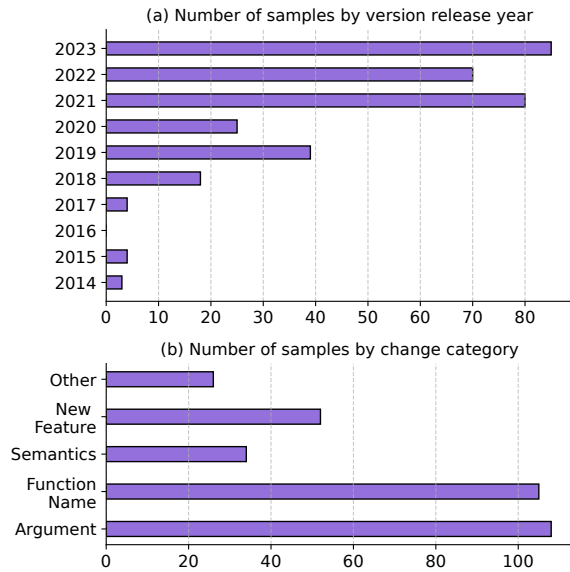


Figure 4: (a) Most versions in **GitChameleon 2.0** were released between 2021–2023, with a few in earlier years. (b) The most common type of change between versions was an argument or attribute change, while semantic or functional changes were least common.

Zero-shot Greedy, Zero-shot CoT, RAG, and Multi-step Agents. Full implementation details, prompts, and hyperparameter configurations are provided in Appendix A.

### 3.2 Qualitative Case Study: API Deprecation Failures

To illustrate the difficulty of **GitChameleon 2.0**, we highlight an example involving a breaking change in Flask 3.0.0. The task requires customizing JSON encoding for matrices.

**Gemini 2.5 Pro (Success):** Correctly subclasses `flask.json.provider.DefaultJSONProvider` and assigns it to `app.json`, fully adhering to the new Flask 3.0 API requirements.

**Claude 3.7 (Failure):** Attempts to use the outdated `app.json.encoder` API, which was deprecated in Flask 2.2 and removed in Flask 3.0.0, causing an immediate `AttributeError`.

### 3.3 Experiment Results

This section presents the benchmark results in each setting, as described in the **Experimental Setup** section (3.1). Table 1 contains the results for Greedy Decoding, Self-Debug and Zero-Shot CoT.

#### 3.3.1 Greedy Decoding

We observe that the largest Enterprise-grade models, including Claude 3.7 Sonnet, Gemini 2.5

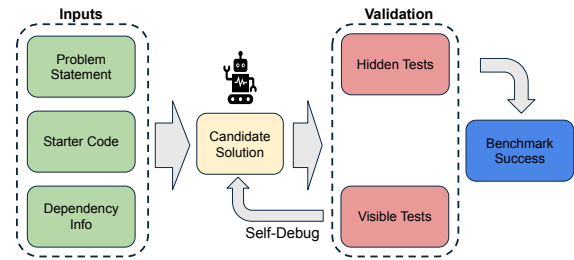


Figure 5: An illustration of the workflow for a single example within **GitChameleon 2.0**. The inputs, comprising the Problem Statement, Starter Code, and Dependency Info, are processed by an LLM or an AI agent to generate a Candidate Solution. This candidate solution then undergoes validation using the Hidden Tests to determine success on the benchmark. Results from the Visible Tests can be fed back into the solution method for self-debugging.

Pro, GPT-4.1, GPT-4o, and o1, exhibit comparable hidden success rates, generally falling within the 48–51% range. We newly benchmarked Gemini-3-Flash and found the highest performance to date (55.2%).

The open-weight Llama models are notably behind, even the recently released Llama 4 Maverick FP8 (40.8% hidden success rate). Deepseek-V3.2 achieved (43.6%).

Model size clearly impacts performance: for instance, Gemini 2.5 Flash trails its Pro counterpart by nearly 12% on hidden tests (38.1% vs. 50.0%). Similarly, the mini and nano series within the GPT family (e.g., GPT-4.1-mini, GPT-4.1-nano, GPT-4o-mini) consistently show lower performance than their larger full-size siblings, with differences on hidden tests ranging from approximately 4 to 15 points.

#### 3.3.2 Zero-Shot Chain-Of-Thought

This approach does not uniformly improve LLM performance across all models. While some models demonstrate significant gains in hidden success rates, a substantial number of enterprise-grade models and their smaller variants experience performance degradation.

For instance, notable improvements in hidden success rates are observed in models such as Llama 3.1 Instruct Turbo (from 30.2% to 36.6%, a +6.4 point increase) and o3-mini (from 45.1% to 50.9%, a +5.8 point increase).

Conversely, several models exhibit a decrease in performance with CoT. Prominent examples include Gemini 2.0 Flash (from 44.2% to 36.0%)

### Claude 3.7 (Failure)

```
def app_set_up(app: flask.Flask) -> None:
    def set_encoder(obj):
        if isinstance(obj, set):
            return sorted(list(obj))
        raise TypeError(f"Object of type {type(obj)} is
        ↪ not JSON serializable")

    app.json.encoder = lambda:
    ↪ json.JSONEncoder(default=set_encoder)
```

### Gemini 2.5 Pro (Success)

```
class
↪ MyCustomJSONHandler(flask.json.provider.DefaultJSONProvider):
    def default(self, obj):
        if isinstance(obj, np.ndarray) and
        ↪ len(obj.shape)==3 and
        ↪ obj.shape[-1]==obj.shape[-2] :
            # Calculate determinants for each matrix in
            ↪ the batch
            determinants = linalg.det(obj)
            # Convert the resulting 1D array of
            ↪ determinants to a list
            return determinants.tolist()
        # Fallback to the default serialization for other
        ↪ types
        return super().default(obj)

# Apply the custom JSON provider to the app
app.json = MyCustomJSONHandler(app)
```

Figure 6: **Case Study (Flask 3.0.0):** Claude hallucinates the removed `app.json.encoder` property, while Gemini successfully adapts to the new API by subclassing `DefaultJSONProvider`, demonstrating the challenge of version-conditioned code generation.

and even the top-performing o1 (from 51.2% to 41.2%).

### 3.3.3 LLM Self-Debugging

**Hidden Success Rate:** Across models, Self-Debugging significantly improves the hidden success rates. Observed gains range from approximately 10% to 20%. For instance, Llama 3.1’s hidden success rate increases from 30% to 52.1%, and GPT-4.1-mini shows an improvement from 44% to 68%. This demonstrates the strong capability of modern LLMs to diagnose failures and generate corrected code.

**Visible Success Rate:** As expected, the improvement is even more pronounced on visible tests, ranging from 13 to 37 points. For instance, GPT-4.1’s success rate improves from 49% to 69%, Claude 3.7 Sonnet’s success rate improves from 56% to 83% and Gemini 2.0 Flash improves from 50% to 75%.

**Visible-Hidden Gap Analysis:** In Figure 7, we present the effect of self-debugging on the size of the gap between the success rate on visible tests and the success rate on hidden tests.

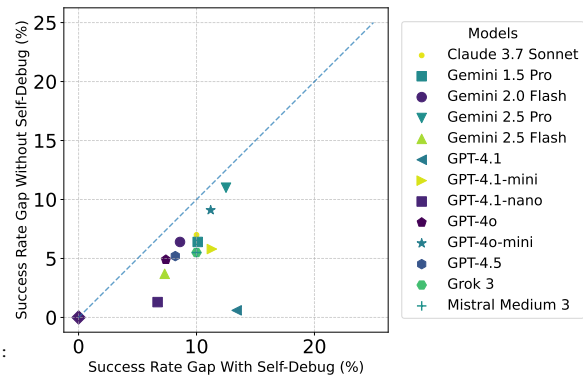


Figure 7: **Analysis of the Visible-Hidden Gap Before and After Self-Debugging.** We analyze how self-debugging affects the gap between the success rate on visible and hidden tests. We can see that for all models, the gap increases after self-debugging. This shows that self-debugging on visible tests has a limited ability to improve on the hidden tests.

### 3.3.4 Multi-Step Agent

We report the performance of Multi-Step Agents on **GitChameleon 2.0** in Table 2. A clear and significant trend is the substantial increase in success rates for all models and grounding methods when giving the agent a sandbox tool. Overall, Claude Sonnet 3.5 demonstrated the highest success rates with a sandbox, across all grounding methods, while Gemini 1.5 Pro demonstrated the best results without a sandbox.

To further evaluate open-source agents, we tested the Mini-SWE Agent framework. Evaluated headless, Gemini 2.5 Pro achieved a 54.0% hidden success rate (64.3% visible), while GPT-5 Mini achieved 36.3% (45.1% visible). These results align with our smolagents evaluation, reinforcing that state-of-the-art models in advanced open-source frameworks struggle with Version-Conditioned Generation. Full details are in Appendix C.1.

### 3.3.5 AI Coding Assistants

Table 3 presents the success rates of various CLI and IDE assistants on the visible and hidden tests in **GitChameleon 2.0**. When the problem statement is not given, Cline with GPT-4.1 achieves the best result, with a success rate of 38.4%. All assistants besides for Goose on GPT-4o demonstrate significant gains, ranging from 12 to 35 points, from including the problem statement.

| Model                        | Greedy Decoding       |                       |                       | Greedy with Self-Debug |                       |                       | Zero-shot CoT         |                       |
|------------------------------|-----------------------|-----------------------|-----------------------|------------------------|-----------------------|-----------------------|-----------------------|-----------------------|
|                              | Success Rate (%)      |                       | API Hit Rate (%)      | Success Rate (%)       |                       | API Hit Rate (%)      | Success Rate (%)      | API Hit Rate (%)      |
|                              | Hidden                | Visible               |                       | Hidden                 | Visible               |                       |                       |                       |
| <b>Open-Weights Models</b>   |                       |                       |                       |                        |                       |                       |                       |                       |
| Llama 3.3 Instruct Turbo 70B | 36.3 $\pm$ 2.7        | 43.3 $\pm$ 2.7        | 36.4 $\pm$ 2.7        | 53.0 $\pm$ 2.8         | 70.1 $\pm$ 2.5        | 37.4 $\pm$ 2.7        | 37.5 $\pm$ 2.7        | 37.2 $\pm$ 2.7        |
| DeepSeek-V3.2                | 43.6 $\pm$ 2.7        | —                     | —                     | —                      | —                     | —                     | —                     | —                     |
| Llama 4 Maverick 400B        | 40.8 $\pm$ 2.7        | 46.6 $\pm$ 2.8        | <b>49.5</b> $\pm$ 2.8 | 58.5 $\pm$ 2.7         | 72.3 $\pm$ 2.5        | <b>46.8</b> $\pm$ 2.8 | <b>46.6</b> $\pm$ 2.8 | 41.3 $\pm$ 2.7        |
| Qwen 2.5-VL Instruct 72B     | <b>48.2</b> $\pm$ 2.8 | <b>55.5</b> $\pm$ 2.7 | 43.8 $\pm$ 2.7        | <b>64.6</b> $\pm$ 2.6  | <b>77.4</b> $\pm$ 2.3 | 45.3 $\pm$ 2.7        | 45.1 $\pm$ 2.7        | <b>43.0</b> $\pm$ 2.7 |
| <b>Enterprise Models</b>     |                       |                       |                       |                        |                       |                       |                       |                       |
| Claude 3.7 Sonnet            | 48.8 $\pm$ 2.8        | 55.8 $\pm$ 2.7        | 46.0 $\pm$ 2.8        | 65.9 $\pm$ 2.6         | 75.9 $\pm$ 2.4        | 47.6 $\pm$ 2.8        | 45.1 $\pm$ 2.7        | 43.4 $\pm$ 2.7        |
| Gemini 1.5 Pro               | 45.1 $\pm$ 2.7        | 51.5 $\pm$ 2.8        | 46.8 $\pm$ 2.7        | 62.5 $\pm$ 2.8         | 72.6 $\pm$ 2.4        | 48.6 $\pm$ 2.7        | 43.3 $\pm$ 2.7        | 44.6 $\pm$ 2.8        |
| Gemini 2.0 Flash             | 44.2 $\pm$ 2.7        | 50.6 $\pm$ 2.8        | 43.8 $\pm$ 2.7        | <b>70.4</b> $\pm$ 2.7  | 79.0 $\pm$ 2.4        | 49.4 $\pm$ 2.7        | 36.0 $\pm$ 2.6        | 41.8 $\pm$ 2.7        |
| Gemini 2.5 Pro               | 50.0 $\pm$ 2.8        | <b>61.0</b> $\pm$ 2.8 | 47.7 $\pm$ 2.7        | 61.3 $\pm$ 2.8         | 73.8 $\pm$ 2.2        | 49.2 $\pm$ 2.7        | 49.4 $\pm$ 2.8        | 49.1 $\pm$ 2.8        |
| Gemini 2.5 Flash             | 38.1 $\pm$ 2.6        | 41.8 $\pm$ 2.7        | 45.4 $\pm$ 2.7        | 65.9 $\pm$ 2.8         | 73.2 $\pm$ 2.4        | 45.8 $\pm$ 2.7        | 30.8 $\pm$ 2.5        | <b>49.8</b> $\pm$ 2.8 |
| Gemini 3 Flash               | <b>55.2</b> $\pm$ 2.8 | —                     | —                     | —                      | —                     | —                     | —                     | —                     |
| GPT-4.1                      | 48.5 $\pm$ 2.8        | 49.1 $\pm$ 2.8        | 46.8 $\pm$ 2.7        | 63.4 $\pm$ 2.8         | 76.8 $\pm$ 2.1        | 48.3 $\pm$ 2.7        | 47.9 $\pm$ 2.8        | 44.5 $\pm$ 2.7        |
| GPT-4.1-mini                 | 44.2 $\pm$ 2.7        | 50.0 $\pm$ 2.8        | 44.5 $\pm$ 2.7        | 68.0 $\pm$ 2.8         | <b>79.3</b> $\pm$ 2.3 | 46.3 $\pm$ 2.7        | 24.1 $\pm$ 1.8        | 41.3 $\pm$ 2.7        |
| GPT-4.1-nano                 | 33.8 $\pm$ 2.6        | 35.1 $\pm$ 2.6        | 43.1 $\pm$ 2.7        | 67.7 $\pm$ 2.7         | 74.4 $\pm$ 2.6        | 45.8 $\pm$ 2.7        | 11.9 $\pm$ 1.8        | 32.1 $\pm$ 2.5        |
| GPT-4o                       | 49.1 $\pm$ 2.8        | 54.0 $\pm$ 2.8        | 46.5 $\pm$ 2.7        | 64.9 $\pm$ 2.8         | 72.3 $\pm$ 2.5        | 48.0 $\pm$ 2.7        | <b>50.3</b> $\pm$ 2.8 | 42.5 $\pm$ 2.7        |
| GPT-4o-mini                  | 37.2 $\pm$ 2.6        | 46.3 $\pm$ 2.7        | 38.4 $\pm$ 2.6        | 60.4 $\pm$ 2.7         | 71.6 $\pm$ 2.6        | 40.6 $\pm$ 2.7        | 36.0 $\pm$ 2.6        | 37.3 $\pm$ 2.6        |
| GPT-4.5                      | 40.8 $\pm$ 2.7        | 46.0 $\pm$ 2.7        | <b>52.8</b> $\pm$ 2.8 | 66.2 $\pm$ 2.8         | 74.4 $\pm$ 2.4        | <b>54.4</b> $\pm$ 2.7 | 39.9 $\pm$ 2.6        | 48.8 $\pm$ 2.8        |
| Grok 3                       | 48.2 $\pm$ 2.8        | 53.7 $\pm$ 2.8        | 44.8 $\pm$ 2.7        | 67.1 $\pm$ 2.8         | 77.1 $\pm$ 2.3        | 46.3 $\pm$ 2.8        | 49.4 $\pm$ 2.8        | 44.2 $\pm$ 2.7        |

Table 1: Success rate on visible and hidden tests and API hit rate under the Greedy, Self-Debug, and Zero-shot CoT settings, grouped by OSS vs. Enterprise models. Model ranking on the benchmark is determined by **Hidden Success Rate**. Visible Success Rate figures are for context on Self-Debugging. The best result in each column is in bold. For full model details and citations, please refer to Appendix J.

| Model             | Grounding Method | Success Rate (%)      |                       | API Hit Rate (%)      |                       |
|-------------------|------------------|-----------------------|-----------------------|-----------------------|-----------------------|
|                   |                  | No Sandbox            | Sandbox               | No Sandbox            | Sandbox               |
| Claude Sonnet 3.5 | DuckDuckGo       | 41.7 $\pm$ 2.7        | <b>55.3</b> $\pm$ 2.7 | 42.2 $\pm$ 2.7        | 48.9 $\pm$ 2.8        |
|                   | Perplexity       | 44.1 $\pm$ 2.7        | 51.4 $\pm$ 2.8        | 41.8 $\pm$ 2.7        | 46.0 $\pm$ 2.8        |
|                   | Grounded Gemini  | 40.0 $\pm$ 2.7        | 53.7 $\pm$ 2.8        | 41.0 $\pm$ 2.7        | 45.2 $\pm$ 2.7        |
| Gemini 1.5 Pro    | DuckDuckGo       | 46.0 $\pm$ 2.8        | 49.8 $\pm$ 2.8        | 47.4 $\pm$ 2.8        | 50.3 $\pm$ 2.8        |
|                   | Perplexity       | <b>46.5</b> $\pm$ 2.8 | 44.4 $\pm$ 2.7        | 47.2 $\pm$ 2.8        | 46.6 $\pm$ 2.8        |
|                   | Grounded Gemini  | 44.1 $\pm$ 2.7        | 49.2 $\pm$ 2.8        | <b>49.7</b> $\pm$ 2.8 | <b>51.2</b> $\pm$ 2.8 |
| GPT-4o            | DuckDuckGo       | 23.9 $\pm$ 2.4        | 33.2 $\pm$ 2.6        | 44.2 $\pm$ 2.7        | 48.1 $\pm$ 2.8        |
|                   | Perplexity       | 33.5 $\pm$ 2.6        | 41.5 $\pm$ 2.7        | 43.2 $\pm$ 2.7        | 44.7 $\pm$ 2.7        |
|                   | Grounded Gemini  | 25.4 $\pm$ 2.4        | 50.0 $\pm$ 2.8        | 46.5 $\pm$ 2.8        | 44.2 $\pm$ 2.7        |

Table 2: Multi-Step Agent performance with different models, grounding methods, and sandbox states. The best result in each column is in bold.

### 3.3.6 Retrieval-Augmented Generation

Table 4 presents the performance of various models with RAG. Many models exhibit a significant (up to 10%) boost in success rate with RAG compared to greedy decoding alone. Notably, GPT-4.1, the best performing model achieves a success rate of 58.5%, up from 48.5% with greedy decoding. These results demonstrate that the benchmark is still challenging even with access to the library documentation, with over 40% of the problems remaining unsolved in the best case.

<sup>3</sup>This version of the model is not FP8-quantized, unlike the one presented in Table 1

| Name                  | Model             | Success Rate (%)      |                       | API Hit Rate (%)      |                       |
|-----------------------|-------------------|-----------------------|-----------------------|-----------------------|-----------------------|
|                       |                   | No-prob               | Prob                  | No-prob               | Prob                  |
| <b>CLI Assistants</b> |                   |                       |                       |                       |                       |
| Claude Code           | Claude 3.7 Sonnet | 32.0 $\pm$ 2.6        | 48.8 $\pm$ 2.8        | <b>44.2</b> $\pm$ 2.7 | 45.5 $\pm$ 2.7        |
| Goose                 | GPT-4o            | <b>36.3</b> $\pm$ 2.7 | 36.9 $\pm$ 2.7        | 43.9 $\pm$ 2.7        | <b>54.5</b> $\pm$ 2.7 |
|                       | GPT-4.1           | 19.2 $\pm$ 2.2        | <b>55.5</b> $\pm$ 2.7 | 41.7 $\pm$ 2.7        | 53.0 $\pm$ 2.8        |
| <b>IDE Assistants</b> |                   |                       |                       |                       |                       |
| Cline                 | Claude 3.7 Sonnet | 32.9 $\pm$ 2.6        | 44.8 $\pm$ 2.7        | 40.5 $\pm$ 2.7        | 50.2 $\pm$ 2.8        |
|                       | GPT-4.1           | 38.4 $\pm$ 2.7        | <b>54.6</b> $\pm$ 2.7 | 42.4 $\pm$ 2.7        | 48.8 $\pm$ 2.8        |
|                       | GPT-4.1-mini      | 27.1 $\pm$ 2.5        | 42.1 $\pm$ 2.7        | 32.9 $\pm$ 2.6        | <b>52.4</b> $\pm$ 2.8 |
| Kilocode              | GPT-4.1-nano      | 38.1 $\pm$ 2.7        | 54.6 $\pm$ 2.7        | 42.4 $\pm$ 2.7        | 48.8 $\pm$ 2.8        |
|                       | GPT-4o            | <b>41.5</b> $\pm$ 2.7 | —                     | 42.7 $\pm$ 2.7        | —                     |
| Rooocode              | Claude 3.7 Sonnet | 30.2 $\pm$ 2.5        | —                     | <b>43.3</b> $\pm$ 2.7 | —                     |
|                       | Claude 3.5 Sonnet | 12.5 $\pm$ 1.8        | —                     | 41.2 $\pm$ 2.7        | —                     |

Table 3: Success and API-hit rates for CLI and IDE coding assistants, under the setting where the problem statement is given (**Prob**) and where it is not (**No-prob**), in which case we evaluate a scenario akin to tab code-completion. The results show that including the problem statement improves success rate by double-digit margins for 4 out of 5 cases evaluated.

| Model                         | Success Rate (%)      | API Hit Rate (%)      | Precision (%)         | Recall (%)            | MRR                    |
|-------------------------------|-----------------------|-----------------------|-----------------------|-----------------------|------------------------|
| <b>Open-Weights Models</b>    |                       |                       |                       |                       |                        |
| Deepseek V3                   | <b>48.9</b> $\pm 2.8$ | 48.5 $\pm 2.8$        | 41.6 $\pm 2.2$        | 50.4 $\pm 2.8$        | 0.62 $\pm 0.03$        |
| Llama 4 Maverick <sup>3</sup> | 45.1 $\pm 2.7$        | <b>50.5</b> $\pm 2.8$ | 41.2 $\pm 2.2$        | 49.8 $\pm 2.8$        | 0.61 $\pm 0.03$        |
| Qwen3                         | 41.8 $\pm 2.7$        | 39.6 $\pm 2.7$        | 36.3 $\pm 2.0$        | 46.9 $\pm 2.8$        | 0.56 $\pm 0.03$        |
| Jamba 1.6 Large               | 41.8 $\pm 2.7$        | 47.1 $\pm 2.8$        | <b>41.9</b> $\pm 2.2$ | <b>50.7</b> $\pm 2.8$ | <b>0.62</b> $\pm 0.03$ |
| <b>Enterprise Models</b>      |                       |                       |                       |                       |                        |
| Claude 3.7 Sonnet             | 56.1 $\pm 2.7$        | 53.0 $\pm 2.8$        | 41.9 $\pm 2.2$        | 50.7 $\pm 2.8$        | 0.62 $\pm 0.03$        |
| Claude 4 Sonnet               | <b>59.4</b> $\pm 2.8$ | <b>55.8</b> $\pm 2.8$ | <b>41.9</b> $\pm 2.2$ | <b>50.7</b> $\pm 2.8$ | <b>0.62</b> $\pm 0.03$ |
| Gemini 2.5 Pro                | 56.7 $\pm 2.7$        | 51.1 $\pm 2.8$        | 41.9 $\pm 2.2$        | 50.7 $\pm 2.8$        | 0.62 $\pm 0.03$        |
| GPT-4.1                       | 58.5 $\pm 2.7$        | 51.8 $\pm 2.8$        | 41.2 $\pm 2.2$        | 50.1 $\pm 2.8$        | 0.61 $\pm 0.03$        |
| Grok3                         | 54.3 $\pm 2.7$        | 55.2 $\pm 2.8$        | 41.6 $\pm 2.2$        | 50.4 $\pm 2.8$        | 0.62 $\pm 0.03$        |
| Mistral Medium 3              | 52.4 $\pm 2.7$        | 51.2 $\pm 2.8$        | 41.6 $\pm 2.2$        | 50.4 $\pm 2.8$        | 0.62 $\pm 0.03$        |
| Devstral Small                | 43.3 $\pm 2.7$        | 45.1 $\pm 2.8$        | 41.6 $\pm 2.2$        | 50.4 $\pm 2.8$        | 0.62 $\pm 0.03$        |
| Nova Pro                      | 44.2 $\pm 2.7$        | 42.4 $\pm 2.7$        | 40.7 $\pm 2.2$        | 49.6 $\pm 2.8$        | 0.60 $\pm 0.03$        |

Table 4: RAG performance for a subset of models when retrieving  $k = 3$  most relevant documents. The best success rate and API hit rate results for each model group are in bold. An extended version of the RAG experiment results is presented in Section C.

### 3.4 In-Depth Analysis of Findings

This section provides a detailed analysis of the experimental results, focusing on model performance across several key dimensions. These dimensions include the impact of different API change types, a comparison between success rate and API hit rate, and the effectiveness of self-debugging across various error types.

#### Comparison of Success Rate and API Hit Rate

API hit rate shows a moderate positive Pearson correlation with hidden-test success under Greedy Decoding with the Pearson correlation coefficient ( $r = 0.392$ ,  $p = 0.097$ ,  $N = 19$ ), indicating that models which invoke the ground truth APIs more often tend to perform better on hidden tests in the Greedy setting, but falls just short of statistical significance at 5% level. Under Zero-Shot CoT, the correlation remains similar in magnitude ( $r = 0.483$ ) and is statistically significant ( $p = 0.036$ ,  $N = 19$ ). In the Self-Debug regime, however, the association becomes both stronger and highly significant ( $r = 0.615$ ,  $p = 0.011$ ,  $N = 16$ ), demonstrating that when models can iteratively refine their outputs, invoking ground truth APIs becomes an especially reliable predictor of hidden-test performance.

#### Analysis of Performance by Type of API Change

Figure 8 illustrates the performance of models across various API change types within the **GitChameleon 2.0** benchmark, revealing notable variations in success rates. Semantic changes were the most tractable, with success rates ranging from 60–80% with Self-Debug and 55–65% without.

New-feature additions proved to be the most challenging, with success rates between 25–50% for Greedy Decoding and 50–65% for Self-Debug. Notably, the Code Assistant Goose exhibited a substantial discrepancy in its performance on semantic and function-name changes compared to argument changes and new features. This suggests a heightened sensitivity to change category for Goose, a characteristic not observed in the enterprise models or the Claude-powered tool-calling agent.

**Self-Debug Error Categorization** Figure 9 shows that self-debugging consistently lowers the rate of every class of traceback error, both in absolute numbers and relative terms:

(a) **Raw Counts:** We observe that for all error categories—from the most common (AssertionError and TypeError) down to the rarest (RuntimeError)—applying Self-Debugging significantly lowers the total number of failures.

(b) **Percentage Reduction:** When normalized by the Greedy Decoding baseline, reductions span roughly 50% up to about 90%. The biggest relative improvements appear in the infrequent categories—such as RuntimeError and SyntaxError—while the common AssertionError and TypeError still see decrease in the range of 60-70%.

## 4 Related Work

The continuous evolution of software libraries presents significant challenges for AI-driven code generation. This section reviews existing benchmarks designed to evaluate model performance in this context. Specialized frameworks developed to address the challenge are presented in section D.2

The challenge of evaluating large language models (LLMs) in the context of evolving software libraries and their versions has been approached by several benchmarks. These benchmarks, while valuable, often differ in scope, methodology, or evaluation techniques compared to **GitChameleon 2.0**.

**PyMigBench** Focusing on Python library migration, this benchmark uses 321 real-world instances, evaluating both individual code transformations and the functional correctness of entire migrated segments via unit tests (Islam et al., 2023). PyMigBench revealed that LLMs often handle individual changes well but struggle with achieving full functional correctness, especially for complex argument transformations.

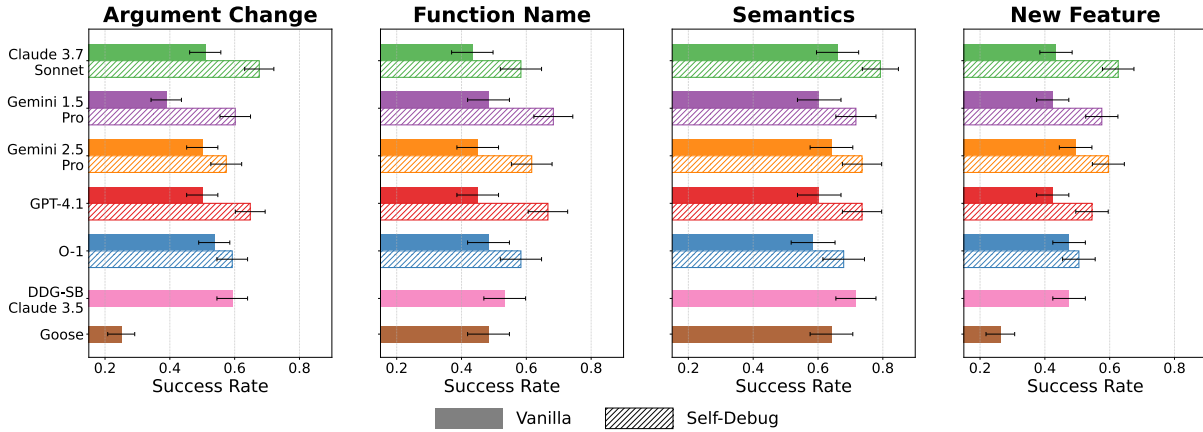


Figure 8: **Success Rate Breakdown by Type of Change:** We analyze success rates with and without self-debugging, grouped by the type of change. Light shaded bars represent values obtained from self-debugging. Standard error is drawn as a black line. We include DDG-SB, a Multi-Step Agent variant where DuckDuckGo is used for grounding and access to a sandbox is enabled, and the Coding Assistant Goose. Self-Debug results for these are omitted.

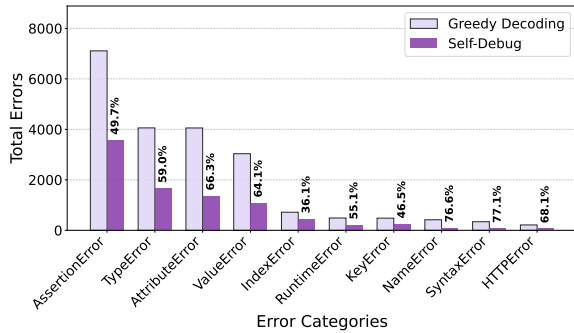


Figure 9: **Total error count for each category under Greedy decoding versus Self-Debug.** Self-Debug yields substantial decreases all types of errors.

**VersiCode** (Wu et al., 2024) and the dataset by Wang et al. (Wang et al., 2024b) address library evolution but primarily depend on string matching for evaluation.

**CodeUpdateArena** (Liu et al., 2025) investigates model adaptation to synthetically generated API updates for functions in popular libraries.

**GitChameleon 2.0** distinguishes itself by focusing on the real-world scenario where developers are often constrained to specific library versions due to technical debt. Unlike CodeUpdateArena’s synthetic changes, **GitChameleon 2.0** evaluates LLMs on their ability to generate code for actual, documented historical breaking changes within library versions they were likely exposed to during training. Furthermore, diverging from the string-matching evaluations of VersiCode and Wang et al. (Wang et al., 2024b), **GitChameleon 2.0** is based on executable tests. This provides a more practical and rigorous assessment of functional accuracy in version-specific code generation. For an

extended discussion of how **GitChameleon 2.0** is differentiated from existing work, please see Appendix D.2.

## 5 Conclusion

The rapid evolution of software libraries presents a critical challenge for LLM-powered AI systems in generating functionally correct, version-conditioned code. To address this, we introduce **GitChameleon 2.0**, a novel Python-based benchmark meticulously curated with version-conditioned problems and executable tests. Our extensive evaluation reveals that state-of-the-art LLMs, agents and code assistants currently struggle significantly with this task, achieving modest success rates.

By shedding light on current limitations and facilitating execution-based evaluation, **GitChameleon 2.0** aims to foster the development of more robust and adaptable code generation models for evolving software environments.

## Limitations

While we aim to provide a comprehensive and holistic evaluation of LLMs on the task of version-conditioned generation, our benchmark is currently limited to Python and a small set of libraries. Automated sample construction could help facilitate an update to the benchmark in the future. Moreover, we focus solely on code generation from natural language instructions, and do not evaluate version-to-version translation—i.e., converting code from one library version to another—even when both versions are in-distribution relative to the model’s training. For instance, if a model has been trained

on PyTorch versions 1.7, 1.8, and 1.9, it would be valuable to assess whether it performs better when given a solution in 1.8 and asked to upgrade to 1.9 or downgrade to 1.7. Finally, we do not include human evaluations, which could provide a baseline for estimating average human performance on this task.

## References

- Abubakar Abid, Ali Abdalla, Ali Abid, Dawood Khan, Abdulrahman Alfozan, and James Zou. 2019. [Gradio: Hassle-free sharing and testing of ML models in the wild](#). *Preprint*, arXiv:1906.02569.
- Meta AI. 2025. Everything we announced at our first-ever LlamaCon. <https://ai.meta.com/blog/llamacon-llama-news/>. Discusses Llama 3.3 Instruct Turbo and Llama 4 Maverick.
- Mohannad Alhanahnah, Yazan Boshmaf, and Benoit Baudry. 2024. [DepsRAG: Towards managing software dependencies using large language models](#). *arXiv preprint arXiv:2405.20455v2*.
- Anthropic. 2025. Claude 3.7 Sonnet and Claude Code. <https://www.anthropic.com/news/claude-3-7-sonnet>.
- Arcee. Model Selection | Arcee AI Documentation — docs.arcee.ai. <https://docs.arcee.ai/arcee-conductor/arcee-small-language-models/model-selection#caller-large-tool-use-and-function-call>. [Accessed 15-07-2025].
- Farnaz Behrang, Zhizhou Zhang, Georgian-Vlad Saioc, Peng Liu, and Milind Chabbi. 2025. [Dr.fix: Automatically fixing data races at industry scale](#). *Preprint*, arXiv:2504.15637.
- Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Andreas Joly, Bertrand Druillette, Gael Varoquaux, and Marion Gramfort. 2013. API design for machine learning software: experiences from the scikit-learn project. *arXiv preprint arXiv:1309.0238*.
- Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri 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, and 39 others. 2021. [Evaluating large language models trained on code](#). *ArXiv*.
- Xinyun Chen, Maxwell Lin, Nathanael Schärli, and Denny Zhou. 2023. [Teaching large language models to self-debug](#). *Preprint*, arXiv:2304.05128.
- Keyuan Cheng, Xudong Shen, Yihao Yang, Tengyue Wang, Yang Cao, Muhammad Asif Ali, Hanbin Wang, Lijie Hu, and Di Wang. 2025. [Codemenv: Benchmarking large language models on code migration](#). *Preprint*, arXiv:2506.00894.
- Matteo Ciniselli, Alberto Martin-Lopez, and Gabriele Bavota. 2024. [On the generalizability of deep learning-based code completion across programming language versions](#). *Preprint*, arXiv:2403.15149.
- Google Cloud. 2025. Gemini 2.5 on Vertex AI: Pro, Flash & Model Optimizer Live. <https://cloud.google.com/blog/products/ai-machine-learning/gemini-2-5-pro-flash-on-vertex-ai>. Discusses Gemini 2.5 Pro and Gemini 2.5 Flash.
- Team Cohere, :, Aakanksha, Arash Ahmadian, Marwan Ahmed, Jay Alammari, Milad Alizadeh, Yazeed Alnumay, Sophia Althammer, Arkady Arkhangorodsky, Viraat Aryabumi, Dennis Aumiller, Raphaël Avalos, Zahara Aviv, Sammie Bae, Saurabh Baji, Alexandre Barbet, Max Bartolo, Björn Bebensee, and 211 others. 2025. [Command a: An enterprise-ready large language model](#). *Preprint*, arXiv:2504.00698.
- Forbes Technology Council. 2024. Revolutionizing software development with large language models. <https://www.forbes.com/councils/forbestechcouncil/2024/03/20/revolutionizing-software-development-with-large-language-models/>.
- Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- DeepSeek-AI. 2025. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#). *Preprint*, arXiv:2501.12948.
- DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025. [Deepseek-v3 technical report](#). *Preprint*, arXiv:2412.19437.
- DuckDuckGo. 2025. DuckDuckGo: Privacy, simplified. <https://duckduckgo.com/>.
- Lishui Fan, Mouxiang Chen, and Zhongxin Liu. 2024. [Self-explained keywords empower large language models for code generation](#). *Preprint*, arXiv:2410.15966.
- Google. 2025. Grounding with Google Search | Gemini API. <https://ai.google.dev/gemini-api/docs/grounding>.
- Aric A Hagberg, Daniel A Schult, and Pieter J Swart. 2008. Exploring network structure, dynamics, and function using NetworkX. In *Proceedings of the 7th Python in Science Conference*, pages 11–15.

- Charles R Harris, K Jarrod Millman, Stéfan J van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J Smith, Robert Kern, Matti Picus, Changqing Hoyer, Marten H van Kerkwijk, Alex Brett, Andrew Wen, Pete Zhang, Joe Igoe, Keith Featherstone, and Travis E Oliphant. 2020. Array programming with NumPy. *Nature*, 585(7825):357–362.
- Dan Hendrycks, Steven Basart, Saurav Kadavath, Mantas Mazeika, Akul Arora, Ethan Guo, Collin Burns, Samir Puranik, Horace He, Dawn Song, and Jacob Steinhardt. 2021. Measuring coding challenge competence with apps. *NeurIPS*.
- J. D. Hunter. 2007. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3):90–95.
- Amazon Artificial General Intelligence. 2024. **The amazon nova family of models: Technical report and model card**. *Amazon Technical Reports*.
- Mohayeminul Islam, Ajay Kumar Jha, Sarah Nadi, and Ildar Akhmetov. 2023. **Pymigbench: A benchmark for python library migration**. In *2023 IEEE/ACM 20th International Conference on Mining Software Repositories (MSR)*, pages 511–515.
- Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. 2024. Livecodebench: Holistic and contamination free evaluation of large language models for code. *arXiv preprint arXiv:2403.07974*.
- Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R Narasimhan. 2024. **SWE-bench: Can language models resolve real-world github issues?** In *The Twelfth International Conference on Learning Representations*.
- Haolin Jin, Zechao Sun, and Huaming Chen. 2024. **Rgd: Multi-llm based agent debugger via refinement and generation guidance**. *Preprint*, arXiv:2410.01242.
- Kelsey Jordahl, Joris Van den Bossche, Martin Fleischmann, Jacob Wasserman, James McBride, Jeffrey Gerard, Jeff Tratner, Matthew Perry, Adrian Garcia Badaracco, Carson Farmer, Geir Arne Hjelle, Alan D. Snow, Micah Cochran, Sean Gillies, Lucas Culbertson, Matt Bartos, Nick Eubank, maxalbert, Aleksey Bilogur, and 11 others. 2020. **geopandas/geopandas: v0.8.1**.
- Kat Kampf. 2025. Create and edit images with Gemini 2.0 in preview. <https://developers.googleblog.com/en/generate-images-gemini-2-0-flash-preview/>. Discusses Gemini 2.0 Flash.
- Paul Kassianik, Baturay Saglam, Alexander Chen, Blaine Nelson, Anu Vellore, Massimo Aufiero, Fraser Burch, Dhruv Kedia, Avi Zohary, Sajana Weerawardhena, Aman Priyanshu, Adam Swanda, Amy Chang, Hyrum Anderson, Kojin Oshiba, Omar Santos, Yaron Singer, and Amin Karbasi. 2025. **Llama-3.1-FoundationAI-SecurityLLM-Base-8B Technical Report**. *arXiv preprint arXiv:2504.21039*. Cited for Llama 3.1 Instruct Turbo.
- Sachit Kuhar, Wasi Uddin Ahmad, Zijian Wang, Nihal Jain, Haifeng Qian, Baishakhi Ray, Murali Krishna Ramanathan, Xiaofei Ma, and Anoop Deoras. 2024. **Libevolutioneval: A benchmark and study for version-specific code generation**. *Preprint*, arXiv:2412.04478.
- Stefano Lambiase, Gemma Catolino, Fabio Palomba, Filomena Ferrucci, and Daniel Russo. 2025. **Exploring individual factors in the adoption of llms for specific software engineering tasks**. *Preprint*, arXiv:2504.02553.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, volume 33.
- James Li. 2024. ReAct vs Plan-and-Execute: A Practical Comparison of LLM Agent Patterns. <https://dev.to/jamesli>.
- Linxi Liang, Jing Gong, Mingwei Liu, Chong Wang, Guangsheng Ou, Yanlin Wang, Xin Peng, and Zibin Zheng. 2025. **Rustevo: An evolving benchmark for api evolution in llm-based rust code generation**. *Preprint*, arXiv:2503.16922.
- Opher Lieber, Barak Lenz, Hofit Bata, Gal Cohen, Jhonathan Osin, Itay Dalmedigos, Erez Safahi, Shaked Meir, Yonatan Belinkov, Shai Shalev-Shwartz, Omri Abend, Raz Alon, Tomer Asida, Amir Bergman, Roman Glozman, Michael Gokhman, Avshalom Manevich, Nir Ratner, Noam Rozen, and 3 others. 2024. **Jamba: A hybrid transformer-mamba language model**. *Preprint*, arXiv:2403.19887.
- Yue Liu, Chakkrit Tantithamthavorn, Yonghui Liu, Patanamon Thongtanunam, and Li Li. 2024. **Automatically recommend code updates: Are we there yet?** *Preprint*, arXiv:2209.07048.
- Zeyu Leo Liu, Shrey Pandit, Xi Ye, Eunsol Choi, and Greg Durrett. 2025. **Codeupdatearena: Benchmarking knowledge editing on API updates**.
- Edward Loper and Steven Bird. 2002. **NLTK: The natural language toolkit**. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, pages 63–70, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Stephan Lukasczyk and Gordon Fraser. 2022. Pynquin: Automated unit test generation for python. In *Proceedings of the ACM/IEEE 44th International*

- Conference on Software Engineering: Companion Proceedings*, pages 168–172.
- Ziyang Luo, Can Xu, Pu Zhao, Qingfeng Sun, Xiubo Geng, Wenxiang Hu, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. 2023. **WizardCoder: Empowering code large language models with Evol-Instruct**.
- Wes McKinney. 2010. Data Structures for Statistical Computing in Python. In *Proceedings of the 9th Python in Science Conference*, pages 51–56.
- Mistral AI. 2025. Medium is the new large: Introducing mistral medium 3. <https://mistral.ai/news/mistral-medium-3>. Accessed: 2025-05-17.
- OpenAI. 2024. **GPT-4o System Card**. *arXiv preprint arXiv:2410.21276*. Cited for GPT-4o.
- OpenAI. 2024. New embedding models and api updates. <https://openai.com/index/new-embedding-models-and-api-updates/>. Accessed: 2025-07-28.
- OpenAI. 2024. OpenAI o1 System Card. <https://openai.com/index/openai-o1-system-card/>. Discusses the o1 model series, including o1 and mentioning o3-mini.
- OpenAI. 2025a. Introducing GPT-4.1 in the API. <https://openai.com/index/gpt-4-1/>. Discusses GPT-4.1, GPT-4.1 mini, and GPT-4.1 nano.
- OpenAI. 2025b. Introducing GPT-4.5. <https://openai.com/index/introducing-gpt-4-5/>.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, and 2 others. 2019. **PyTorch: An imperative style, high-performance deep learning library**. *Preprint*, arXiv:1912.01703.
- Perplexity AI. 2024. Getting started with Perplexity. <https://www.perplexity.ai/hub/blog/getting-started-with-perplexity>.
- Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. **Qwen2.5 technical report**. *Preprint*, arXiv:2412.15115.
- Rafiqul Rabin, Jesse Hostetler, Sean McGregor, Brett Weir, and Nick Judd. 2025. **Sandboxeval: Towards securing test environment for untrusted code**. *Preprint*, arXiv:2504.00018.
- Reka. RekaAI/reka-flash-3 · Hugging Face — huggingface.co. <https://huggingface.co/RekaAI/reka-flash-3>. [Accessed 15-07-2025].
- Aymeric Roucher, Albert Villanova del Moral, Thomas Wolf, Leandro von Werra, and Erik Kaunismäki. 2025. ‘smolagents’: a smol library to build great agentic systems. <https://github.com/huggingface/smolagents>.
- Aditi Singh, Abul Ehtesham, Saket Kumar, and Tala Talei Khoei. 2025. **Agentic retrieval-augmented generation: A survey on agentic rag**. *Preprint*, arXiv:2501.09136.
- Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. 2023. **Roformer: Enhanced transformer with rotary position embedding**. *Preprint*, arXiv:2104.09864.
- Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, and 1118 others. 2024. **Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context**. *Preprint*, arXiv:2403.05530.
- The pandas development team. 2020. **pandas-dev/pandas: Pandas**.
- Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, and 1 others. 2020. **Scipy 1.0: fundamental algorithms for scientific computing in python**. *Nature methods*, 17(3):261–272.
- Chaozheng Wang, Shuzheng Gao, Cuiyun Gao, Wenxuan Wang, Chun Yong Chong, Shan Gao, and Michael R. Lyu. 2024a. **A systematic evaluation of large code models in api suggestion: When, which, and how**. *Preprint*, arXiv:2409.13178.
- Chong Wang, Kaifeng Huang, Jian Zhang, Yebo Feng, Lyuye Zhang, Yang Liu, and Xin Peng. 2024b. **How and Why LLMs Use Deprecated APIs in Code Completion? an Empirical Study**. *arXiv preprint arXiv:2312.14617*.
- Chong Wang, Kaifeng Huang, Jian Zhang, Yebo Feng, Lyuye Zhang, Yang Liu, and Xin Peng. 2025a. **LLMs Meet Library Evolution: Evaluating Deprecated API Usage in LLM-based Code Completion**. In *2025 IEEE/ACM 47th International Conference on Software Engineering (ICSE)*, pages 781–781, Los Alamitos, CA, USA. IEEE Computer Society.
- Chong Wang, Kaifeng Huang, Jian Zhang, Yebo Feng, Lyuye Zhang, Yang Liu, and Xin Peng. 2025b. **Llms meet library evolution: Evaluating deprecated api usage in llm-based code completion**. *Preprint*, arXiv:2406.09834.
- Xingyao Wang. 2025. **Introducing openhands lm 32b – a strong, open coding agent model**. *All Hands AI Blog*.

- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. [Chain-of-thought prompting elicits reasoning in large language models](#). *Preprint*, arXiv:2201.11903.
- Tongtong Wu, Weigang Wu, Xingyu Wang, Kang Xu, Suyu Ma, Bo Jiang, Ping Yang, Zhenchang Xing, Yuan-Fang Li, and Gholamreza Haffari. 2024. [Versi-Code: Towards version-controllable code generation](#).
- xAI. 2025. [Grok-3](#). Official xAI announcement. Accessed May 17, 2025.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2023. [React: Synergizing reasoning and acting in language models](#). In *The Eleventh International Conference on Learning Representations*.
- Sixiang Ye, Zeyu Sun, Guoqing Wang, Liwei Guo, Qingyuan Liang, Zheng Li, and Yong Liu. 2025. [Prompt alchemy: Automatic prompt refinement for enhancing code generation](#). *Preprint*, arXiv:2503.11085.
- Shuyan Zhou, Uri Alon, Frank F Xu, Zhiruo Wang, Zhengbao Jiang, and Graham Neubig. 2022. [DocPrompting: Generating code by retrieving the docs](#).

## A Benchmark Details

This appendix provides additional details on the **GitChameleon 2.0** benchmark. We provide details on the dataset construction process, the structure of the dataset samples, on the processes for validating the examples and constructing the hidden tests, and finally present additional statistics regarding the dataset.

### A.1 Dataset Construction Process

The examples were created by the authors, which took roughly 350 human hours. To construct that dataset, we compiled a list of popular Python libraries, focusing on those that had more than 1000 stars on Github as well as detailed documentation of changes between versions. For each library, we reviewed the change logs to identify breaking changes: deprecated functions, argument changes, alterations in behavior, and newly introduced functions.

For each identified change, we wrote a concise problem statement, starter code, expected solution and a suite of tests, consisting of a comprehensive suite of hidden tests to be used for model performance evaluation and ranking and a manually written concise visible test to be used for self-debugging experiments. We also added a ground-truth set of relevant documents for RAG experiments.

**NOTE:** Low-level changes—such as backend optimizations that do not alter the surface-level API—are not considered valid changes for our benchmark. For example, if between Torch 1.7 and Torch 1.8 the `torch.nn.Softmax()` function received a CUDA-based numerical stability improvement, this does not modify the API usage of `Softmax()` and is therefore not labeled as a change in our benchmark. Since most changes in mature libraries primarily impact backend functionality, collecting 328 valid samples required significant effort.

### A.2 Structure of Dataset Samples

The main fields of each sample are given in Table 5. Additionally, each problem in **GitChameleon 2.0** is associated with metadata to assist in the analysis of the results, as described in Table 6. Each problem is classified with a type of API evolution change among the categories defined in Table 7.

|                            |   |
|----------------------------|---|
| <b>Library</b>             | The software library under test.  |
| <b>Library Version</b>     | The exact version of that library.  |
| <b>Task Description</b>    | A problem centered on a particular library change.  |
| <b>Initial Code</b>        | The Python snippet provided as a starting point.  |
| <b>Extra Dependencies</b>  | Any additional packages required to solve the task.   |
| <b>Hidden Tests</b>        | Comprehensive unit tests designed to maximize coverage. The success rate on these is the benchmark metric.      |
| <b>Visible Test</b>        | A concise test that validates the specific target behavior, intended to be used for Self-Debugging experiments. |
| <b>Reference Solution</b>  | A correct, ground-truth implementation.   |
| <b>Reference Documents</b> | A set of version-specific reference documents, to be used for RAG experiments.                                  |

Table 5: Problem column definitions for the **GitChameleon 2.0** dataset.

|                           |  |
|---------------------------|--|
| <b>Change Category</b>    | The type of library-evolution changes, as defined in table 7.                              |
| <b>Target Entity</b>      | The specific function or class under test.   |
| <b>Solution Style</b>     | “Functional” if only a function body is expected, or “Full” for a general code completion. |
| <b>Web Framework Task</b> | “Yes” if the problem exercises a web-development framework, otherwise “No.”                |

Table 6: Metadata column definitions.

### A.3 Dataset Validation

To ensure the validity of the dataset examples, we followed the following process: First, we created a clean Docker container for each problem and installed the required dependencies into it. Then, we executed the visible and hidden validation tests to ensure that all are successful.

### A.4 Hidden Test Construction

This section presents how we generated the hidden tests for each dataset example. These tests were generated by instructing the Zencoder AI Coding Agent <sup>4</sup> to create test files for each example, incorporating the appropriate dependency versions. The Zencoder agent, built on the GPT-4.1 base model, operated with internet search enabled and was granted execution access, allowing it to self-correct outputs that initially failed during runtime. Further errors encountered during verification were resolved by supplying error traces back to Zencoder or through an isolated instance of GPT-4o, supplemented with manual inspection and intervention where necessary. This process enabled us to construct a robust and comprehensive test suite, achieving a coverage of **96.5%**. The decision to use ZENCODER was motivated by limitations observed

<sup>4</sup><https://zencoder.ai>

| Change Category  | Description   |
|--|---|
| <b>Argument or Attribute change</b>                      | The API call to a function, method, or class has a change in arguments (e.g. name, order, new, deprecated argument) between versions.   |
| <b>Function Name change</b>                              | The name of the API call has changed between versions (e.g. <code>pandas.append</code> to <code>pandas.concat</code> ).   |
| <b>Semantics or Function Behavior change</b>             | The semantic / runtime behavior of the API call changed between versions (e.g. returning a different type).   |
| <b>New feature or additional dependency-based change</b> | A feature was introduced in a specific version; therefore, to execute the same functionality, a model using an older version should make use of an additional dependency (e.g. <code>torch.special</code> was introduced in TORCH 1.10, previously one could use NUMPY for the same). |

Table 7: Categories of API Evolution Changes

in alternative unit test generation approaches. Rule-based generators such as Pynguin (Lukasczyk and Fraser, 2022) fail to account for version differences among samples that share the same or similar problem statements. Meanwhile, AI-based unit test generators like Claude Code and EarlyAI<sup>5</sup> were not suitable: the former typically generated test classes where each sub-function was populated only with `pass()` statements, while the latter was restricted to functional-style problems and could not handle the more complex, class-based structures prevalent in **GitChameleon 2.0**.

### A.5 Additional Dataset Statistics

Figure 10 presents the number of unique versions per library and the number of samples per library.

### A.6 Detailed Experimental Setup

#### A.6.1 Greedy Decoding

We configured the generation parameters with a sampling temperature of 0 and a `top_p` value of 0.95. We had specified a structured output schema that specifies the fields `Answer` and `Explanation`, where both are of type `string`.

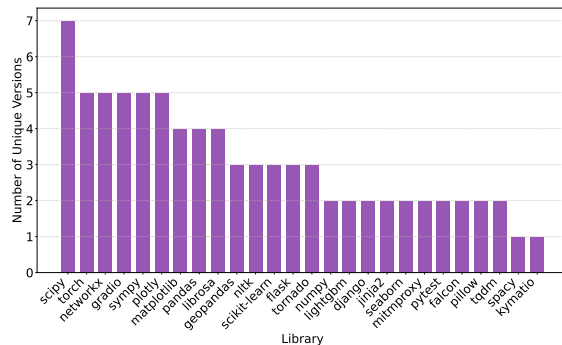
#### A.6.2 Zero-Shot Chain-Of-Thought (CoT)

We had used the same generation parameters as for Greedy Decoding and an output schema that specifies the fields `Answer` and `Steps`, where the former is a of type `string` and the latter is a list of `string`.

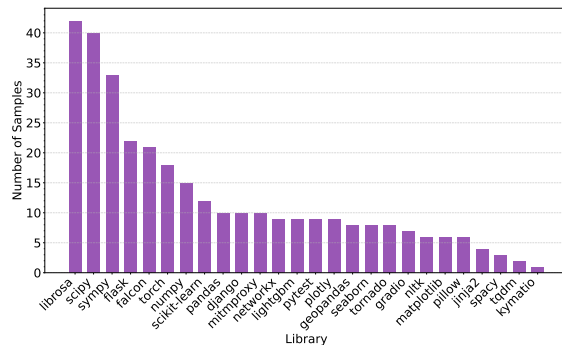
#### A.6.3 Self-Debugging

On examples that failed with Greedy Decoding, we employed the method described in (Chen et al., 2023) to feed the visible test error trace along with

<sup>5</sup><https://www.startearly.ai/>



(a) Number of unique versions per library.



(b) Number of samples per library.

Figure 10: Dataset library statistics. (a) The count of distinct versions identified for each library, presented in decreasing order of uniqueness. (b) The total frequency of samples containing each library, ordered by their occurrence count.

the model’s explanation of its output back to the model.

#### A.6.4 Retrieval-Augmented Generation

We designed a RAG (Lewis et al., 2020) pipeline where we first constructed a vectorized database (VectorDB) by embedding each sample’s relevant API documentation with the OpenAI text-embedding-3 large model (OpenAI, 2024). The corpus used for constructing the VectorDB included 536 documents, with 140 samples having 1 associated document, 168 having 2 associated documents and 20 having 3 documents.

Subsequently, we used DocPrompting (Zhou et al., 2022) to query the VectorDB to generate solutions.

#### A.6.5 Multi-Step Agent

We conducted experiments with a tool-calling agent, as implemented by the smolagents (Roucher et al., 2025) framework. This agent implementation mostly follows the ReAct (Yao et al., 2023) method,

but, it alternates between acting and planning (Li, 2024) steps.

Following the Agentic RAG approach (Singh et al., 2025), we had equipped the agent with a grounding tool in order to assess its capability to independently fetch relevant info for solving the benchmark problems. To this end, we had experimented with the following grounding tools: DuckDuckGo Search (DuckDuckGo, 2025), Perplexity (Perplexity AI, 2024), and Gemini with Grounding (Google, 2025).

Additionally, we examined agentic multi-step self-debugging (Jin et al., 2024) by including or omitting a code execution sandbox tool (Rabin et al., 2025), which provides the needed dependencies for each example. The sandbox takes a Python program as input and outputs the standard output from the program.

#### A.6.6 AI Coding Assistants

In addition to evaluating a generic agentic framework endowed with basic tools, we also analyze the performance of specialized AI coding assistant software.

For this setting, we examine both Command-Line Interface (CLI), such as Claude Code<sup>6</sup> coding assistants and Integrated Development Environment (IDE) coding assistants, such as Cline<sup>7</sup>.

The input to the assistants is given as a Python file which consists of the required library, version and extra dependencies as in-line comments and subsequently the starter code. NOTE: All assistants had internet and terminal commands execution access.

We had furthermore ablated this setting versus giving the full problem statement as input.

### B Extra Methodologies: Reasoning, Sampling and Prompting

This section presents results from additional experimental methodologies:

- **Temperature Sampling:** Results are shown in Table 9. We evaluate sampling at temperature  $T = 0.8$  across 10 seeds using both the OpenAI and Gemini model suites. The performance difference compared to greedy decoding is minimal.

<sup>6</sup><https://docs.anthropic.com/en/docs/claude-code/overview>

<sup>7</sup><https://cline.bot/>

- **Reasoning Models:** Performance results for the OpenAI o-series reasoning models are provided in Table 8.
- **Self-Explained Keywords (SEK) Prompting:** We evaluate the SEK prompting method proposed by Fan et al. (2024), applied to both OpenAI and Gemini models. SEK involves a two-stage process: (1) *Keyword Extraction*, where the model generates relevant keywords for the coding task, and (2) *Keyword Categorization*, where keywords are ranked and classified into (a) Function, (b) General, and (c) Abstract categories. TF-IDF ranking is performed using a 50,000-document subset of the EVOL-CODEALPACA-V1 corpus (Luo et al., 2023). As shown in our empirical analysis, SEK does not yield significant improvements over greedy sampling, and in several cases underperforms relative to it. NOTE: Temperature  $T = 0$  is used in both stages of SEK prompting.

### C Extended Experiment Results and Analysis

This section contains the following additional experimental results:

- An experiment on Automatic Prompt Optimization of the system prompt for Greedy Decoding is described in Table 11.
- An experiment on static analysis based generated solutions fixing to ensure model failures are not attributed to confounding factors like indentation problems and unused imports or variable declarations. Refer to Table 13 for further details.
- Table 12 contains an extended set of RAG results, including both additional models and the setting where only a single document is retrieved.

We also present the following additional analyses:

- A comparison of success rates between Self-Debug and Greedy Decoding, when broken down by version release year (Figure 11) and by library (Figure 12).
- A comparison of success rates between RAG and Greedy Decoding by library is shown in Figure 13.

| Model                              | Vanilla Decoding      |                       |                       | Vanilla with Self-Debug |                       |                       | Zero-shot CoT         |                       |
|------------------------------------|-----------------------|-----------------------|-----------------------|-------------------------|-----------------------|-----------------------|-----------------------|-----------------------|
|                                    | Success Rate (%)      |                       | API Hit Rate (%)      | Success Rate (%)        |                       | API Hit Rate (%)      | Success Rate (%)      | API Hit Rate (%)      |
|                                    | Hidden                | Visible               |                       | Hidden                  | Visible               |                       | Hidden                |                       |
| <b>Additional Evaluated Models</b> |                       |                       |                       |                         |                       |                       |                       |                       |
| o1                                 | <b>51.2</b> $\pm 2.8$ | <b>60.1</b> $\pm 2.7$ | 42.1 $\pm 2.7$        | 57.6 $\pm 2.7$          | 68.6 $\pm 2.6$        | <b>49.2</b> $\pm 2.8$ | 41.2 $\pm 2.7$        | <b>41.3</b> $\pm 2.7$ |
| o3-mini                            | 44.5 $\pm 2.7$        | 52.7 $\pm 2.8$        | 40.6 $\pm 2.7$        | <b>66.8</b> $\pm 2.6$   | <b>76.5</b> $\pm 2.3$ | 45.7 $\pm 2.8$        | <b>50.9</b> $\pm 2.8$ | 40.7 $\pm 2.7$        |
| o4-mini                            | 48.2 $\pm 2.8$        | 57.0 $\pm 2.7$        | <b>48.3</b> $\pm 2.8$ | 63.1 $\pm 2.7$          | 75.0 $\pm 2.4$        | 45.4 $\pm 2.7$        | –                     | –                     |
| codex-mini                         | 48.5 $\pm 2.8$        | 58.2 $\pm 2.7$        | 47.5 $\pm 2.8$        | –                       | –                     | –                     | 32.0 $\pm 2.6$        | 37.9 $\pm 2.7$        |
| Llama 3.1 Instruct Turbo           | 30.2 $\pm 2.5$        | 38.1 $\pm 2.7$        | 39.7 $\pm 2.7$        | 52.1 $\pm 2.8$          | 69.2 $\pm 2.5$        | 41.5 $\pm 2.7$        | 36.6 $\pm 2.7$        | 35.3 $\pm 2.6$        |
| Mistral Medium 3                   | 43.6 $\pm 2.7$        | 49.1 $\pm 2.8$        | 44.2 $\pm 2.7$        | 61.3 $\pm 2.8$          | 71.3 $\pm 2.5$        | 45.4 $\pm 2.7$        | 44.2 $\pm 2.7$        | 44.1 $\pm 2.7$        |

Table 8: Success rate on visible and hidden tests and API hit rate under the Vanilla, Self-Debug, and Zero-shot CoT settings, for additional models. Model ranking on the benchmark is determined by **Hidden Success Rate**. Visible Success Rate figures are for context on Self-Debugging. The best result in each column is in bold. For full model details and citations, please refer to Appendix J.

- Figure 14 analyzes the intra-model sample agreement rates in the Greedy Decoding, Zero-Shot CoT and RAG settings.

### C.1 Evaluating Open-Source Agentic Systems: Mini-SWE Agent

A key methodological challenge in evaluating agentic frameworks on large-scale benchmarks is that many promising systems are designed primarily for interactive demonstrations and lack a non-interactive, headless mode. To systematically evaluate performance within an open-source agentic environment, we conducted experiments using the Mini-SWE Agent framework.

**Methodology** We utilized a one-shot prompting strategy. The agent was first verified on a single example, which was then provided as an in-context exemplar (`example_0.py`) for the remaining tasks. The prompt explicitly instructed the agent to process problems sequentially to prevent context window overload:

*"Follow these instructions carefully. Solve each example\_{i}.py in the problems directory, by following the instructions in the commented lines very carefully. Do not assume that the problem has already been solved. Start from example\_i.py, read example\_0.py to see an example of a completed solution. Carefully, meticulously craft a solution based on the instructions provided and place the solution in the target directory. Once you have checked your answer, move to*

*the next example\_{i+1}.py. Do not attempt to solve all the problems in the folder at the same time; it will overload your context window. Keep a running note of these instructions so that you do not forget."*

**Results** Consistent with our primary findings, Gemini 2.5 Pro achieved a 53.96% hidden success rate (64.33% visible), representing the upper bound of performance. Conversely, GPT-5 Mini—a new state-of-the-art small model—scored only 36.3% on hidden tests (45.1% visible). This underscores the difficulty and relevance of the Version-Conditioned Generation (VCG) problem even for highly capable agentic frameworks.

**Qualitative Failure Case** To illustrate the challenge of version-sensitivity within the Mini-SWE Agent, we highlight a failure case involving the `lightgbm` library (version 3.0.0). The agent was tasked with writing a function to decode a byte string.

```
# Metadata: "library": "lightgbm", "version": "3.0.0",
↪ "release_date": "2020-08"
# Problem: "Write a function to decode a byte string."

import lightgbm.compat as compat
def decode_string(string: bytes) -> str:
    # Agent's Incorrect Generation:
    return compat.decode(string)
```

The agent produced code calling `lightgbm.compat.decode()`, resulting in an `AttributeError`. The root cause is a direct consequence of the agent not adhering to the specified library version: the `decode()` function does not exist in version 3.0.0 of `lightgbm`. The correct, version-aware API call is `lightgbm.compat.decode_string()`. This

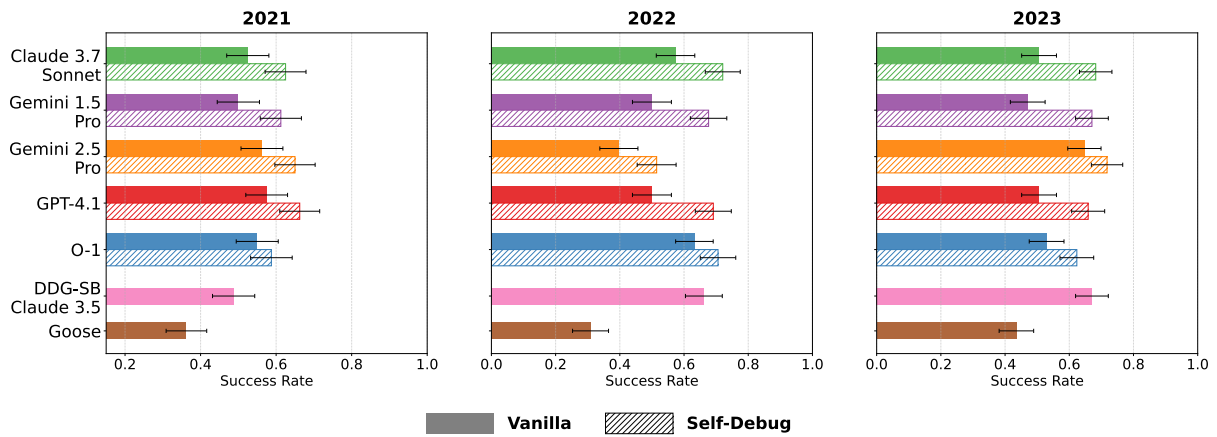


Figure 11: **Success Rate Breakdown by Version Release Year.** Lighter and darker shaded bars represent values obtained with and without Self-Debugging, respectively. Standard error is drawn as a black line. This plot shows that the release year does not significantly impact the results for most evaluated settings.

perfectly illustrates how an agent’s failure to condition its generation on the explicit version metadata leads to critical execution failures.

| Model            | Hidden Success Rate (%) | API Hit Rate (%)      |
|------------------|-------------------------|-----------------------|
| o1               | 50.5 $\pm$ 0.8          | 44.0 $\pm$ 0.8        |
| o3-mini          | 46.4 $\pm$ 1.6          | 42.5 $\pm$ 0.6        |
| GPT-4.1          | 48.9 $\pm$ 1.4          | 48.1 $\pm$ 1.0        |
| GPT-4.1-mini     | 45.9 $\pm$ 1.3          | 46.9 $\pm$ 0.6        |
| GPT-4.1-nano     | 33.8 $\pm$ 1.1          | 43.8 $\pm$ 0.8        |
| GPT-4o           | 47.2 $\pm$ 1.2          | 45.1 $\pm$ 0.9        |
| GPT-4o-mini      | 40.2 $\pm$ 1.2          | 41.0 $\pm$ 1.1        |
| Gemini 1.5 Pro   | 45.4 $\pm$ 1.2          | 45.5 $\pm$ 0.7        |
| Gemini 2.5 Pro   | 41.0 $\pm$ 3.4          | <b>48.3</b> $\pm$ 1.7 |
| Gemini 2.0 Flash | 43.4 $\pm$ 3.1          | 42.5 $\pm$ 0.9        |
| Gemini 2.5 Flash | 46.4 $\pm$ 0.8          | 46.8 $\pm$ 1.2        |

Table 9: **Hidden Success Rate using temperature sampling ( $T = 0.8$ ), averaged over 10 seeds.** A comparison to the greedy decoding baseline in Table 1 reveals that the changes in performance between greedy decoding and temperature sampling are mixed. For most models, the differences are small, but for a few specific models, the changes are big and noteworthy. For the majority of models evaluated (8 out of 11), the performance change is minor, typically within  $\pm 2$  percentage points. For example, Gemini-2.5-pro, shows a notable decrease in success rate (-9.0 points).

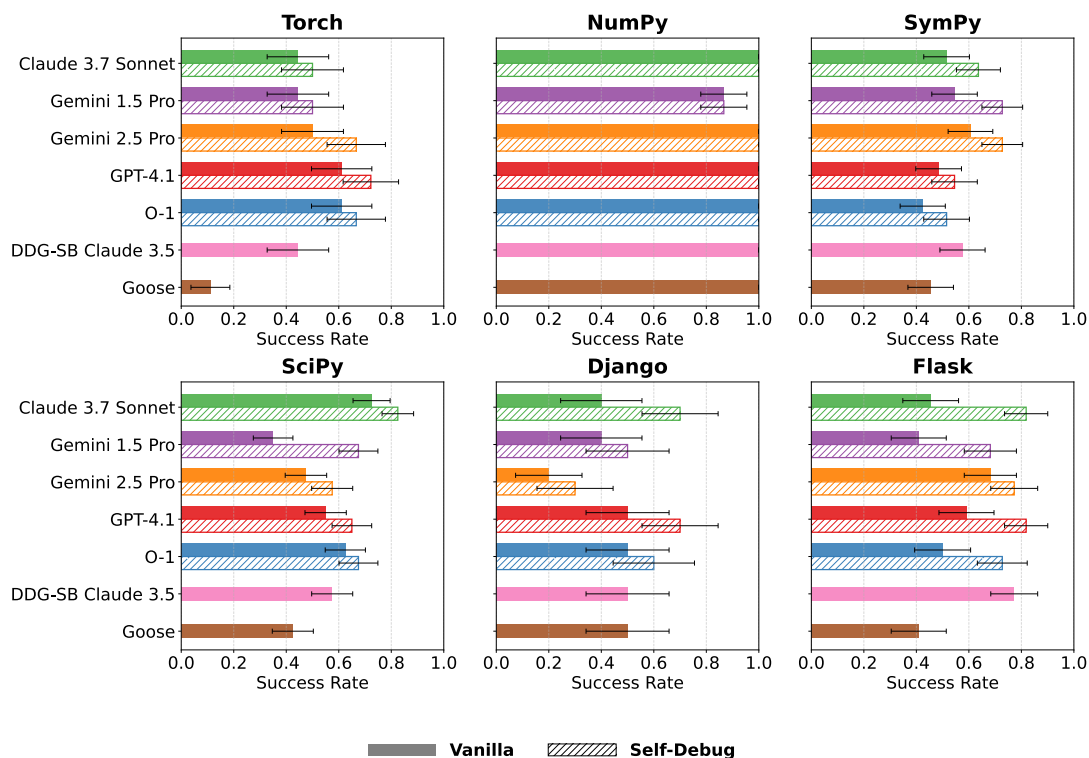


Figure 12: **Success Rate Breakdown by Library**. This figure shows the differences in success rate between the libraries included in **GitChameleon 2.0**. All evaluated settings do very well on NumPy, which is to be expected given the popularity of the library and the subsequent abundance of code that uses it. The success rates on the web development frameworks are notably lower than on the scientific computing libraries, perhaps due to having more complex abstractions.

| Model            | Hidden Success Rate (%) | API Hit Rate (%)      |
|------------------|-------------------------|-----------------------|
| GPT-4o           | 29.6 $\pm$ 2.5          | 43.6 $\pm$ 2.7        |
| GPT-4o-mini      | 27.7 $\pm$ 2.5          | 40.3 $\pm$ 2.7        |
| GPT-4.1          | 43.6 $\pm$ 2.7          | 49.4 $\pm$ 2.8        |
| GPT-4.1-mini     | 41.2 $\pm$ 2.7          | 44.0 $\pm$ 2.7        |
| GPT-4.1-nano     | 32.9 $\pm$ 2.6          | 43.8 $\pm$ 2.7        |
| GPT-4.5          | 33.8 $\pm$ 2.6          | <b>58.0</b> $\pm$ 2.7 |
| Gemini 1.5 Pro   | 44.5 $\pm$ 2.7          | 45.7 $\pm$ 2.8        |
| Gemini 2.0 Flash | 41.2 $\pm$ 2.7          | 43.4 $\pm$ 2.7        |
| Gemini 2.5 Pro   | 47.3 $\pm$ 2.8          | 50.0 $\pm$ 2.8        |
| Gemini 2.5 Flash | <b>48.2</b> $\pm$ 2.8   | 43.4 $\pm$ 2.7        |

Table 10: **Success and API hit rates under the SEK setting**. While SEK, being a two-round prompting scheme, is expected to outperform greedy decoding, we observe that it does not yield significant improvements. For example, with GPT-4.1, the success rate actually drops by 4.9% when using SEK compared to greedy decoding.

| Model        | Best Round | Success Rate (%)      | $\Delta$ (%) |
|--------------|------------|-----------------------|--------------|
| GPT-4.1-mini | 1          | 42.1 $\pm$ 2.7        | -2.1         |
| GPT-4.1-nano | 3          | 37.5 $\pm$ 2.7        | <b>+3.7</b>  |
| GPT-4.1      | 1          | <b>50.0</b> $\pm$ 2.8 | +1.5         |
| GPT-4o       | 0          | 49.1 $\pm$ 2.8        | 0.0          |

Table 11: **Automatic System Prompt Optimization results**. The prompt was optimized for at most 5 rounds using the method described in (Ye et al., 2025), with early stopping if the improvement over previous round is less than 1.5%. We used GPT-4.1 as the mutation model and a random fixed 20% subset of the dataset for the optimization process. For the initial prompt, we use the same system prompt that we had used for our Greedy Decoding experiments, as given in Figure 18. We report the delta of the hidden test success rate, in comparison to the Greedy Decoding baseline. The results demonstrate the limited utility of further optimizing the prompts we had used in our experiments.

| Model                      | $k = 1$               |                       | $k = 3$               |                       |                       |                       |                        |
|----------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|------------------------|
|                            | Success Rate (%)      | API Hit Rate (%)      | Success Rate (%)      | API Hit Rate (%)      | Precision (%)         | Recall (%)            | MRR                    |
| <b>Open-Weights Models</b> |                       |                       |                       |                       |                       |                       |                        |
| CommandA                   | 43.6 $\pm$ 2.7        | 43.9 $\pm$ 2.7        | <b>48.2</b> $\pm$ 2.8 | 45.4 $\pm$ 2.7        | <b>41.9</b> $\pm$ 2.7 | <b>50.7</b> $\pm$ 2.8 | <b>0.63</b> $\pm$ 0.03 |
| CommandR 7B                | 23.2 $\pm$ 2.3        | 36.3 $\pm$ 2.7        | 23.2 $\pm$ 2.3        | 35.6 $\pm$ 2.6        | 41.6 $\pm$ 2.7        | 50.4 $\pm$ 2.8        | 0.62 $\pm$ 0.03        |
| Deepseek R1                | <b>50.9</b> $\pm$ 2.8 | 44.8 $\pm$ 2.7        | 51.2 $\pm$ 2.8        | <b>47.9</b> $\pm$ 2.8 | 41.5 $\pm$ 2.7        | 50.1 $\pm$ 2.8        | 0.62 $\pm$ 0.03        |
| Reka Flash-3               | 8.5 $\pm$ 1.5         | 34.5 $\pm$ 2.6        | 11.6 $\pm$ 1.8        | 31.9 $\pm$ 2.6        | 29.9 $\pm$ 2.5        | 39.6 $\pm$ 2.8        | 0.47 $\pm$ 0.03        |
| Jamba 1.6 Mini             | 18.0 $\pm$ 2.1        | 35.4 $\pm$ 2.6        | 29.3 $\pm$ 2.5        | 40.4 $\pm$ 2.7        | 41.6 $\pm$ 2.7        | 50.1 $\pm$ 2.8        | 0.62 $\pm$ 0.03        |
| OpenHands LM 32B v0.1      | 34.8 $\pm$ 2.6        | 41.0 $\pm$ 2.7        | 28.9 $\pm$ 2.5        | 36.5 $\pm$ 2.7        | 25.9 $\pm$ 2.4        | 33.7 $\pm$ 2.7        | 0.42 $\pm$ 0.03        |
| Llama 4 Scout              | 38.7 $\pm$ 2.7        | <b>45.1</b> $\pm$ 2.7 | 39.3 $\pm$ 2.7        | 43.6 $\pm$ 2.7        | 41.3 $\pm$ 2.7        | 50.4 $\pm$ 2.8        | 0.62 $\pm$ 0.03        |
| <b>Enterprise Models</b>   |                       |                       |                       |                       |                       |                       |                        |
| Arcee CoderL               | 46.3 $\pm$ 2.8        | 47.3 $\pm$ 2.8        | 36.6 $\pm$ 2.7        | 40.4 $\pm$ 2.7        | 31.1 $\pm$ 2.6        | 41.0 $\pm$ 2.8        | 0.49 $\pm$ 0.03        |
| Claude 3.5 Haiku           | 43.6 $\pm$ 2.7        | 47.9 $\pm$ 2.8        | 43.0 $\pm$ 2.7        | 47.5 $\pm$ 2.8        | 41.9 $\pm$ 2.7        | 50.7 $\pm$ 2.8        | 0.62 $\pm$ 0.03        |
| Claude 3.5 Sonnet          | 8.5 $\pm$ 1.5         | 18.6 $\pm$ 2.1        | 49.4 $\pm$ 2.8        | 51.5 $\pm$ 2.8        | 41.9 $\pm$ 2.7        | 50.7 $\pm$ 2.8        | 0.62 $\pm$ 0.03        |
| Codestral                  | 44.2 $\pm$ 2.7        | 47.3 $\pm$ 2.8        | 46.0 $\pm$ 2.8        | 48.5 $\pm$ 2.8        | 41.9 $\pm$ 2.7        | 50.7 $\pm$ 2.8        | 0.62 $\pm$ 0.03        |
| CommandR+                  | 32.0 $\pm$ 2.6        | 43.0 $\pm$ 2.7        | 36.6 $\pm$ 2.7        | 41.9 $\pm$ 2.7        | 41.6 $\pm$ 2.7        | 50.4 $\pm$ 2.8        | 0.62 $\pm$ 0.03        |
| Gemini 2.5 Flash           | <b>54.3</b> $\pm$ 2.8 | <b>50.5</b> $\pm$ 2.8 | <b>55.2</b> $\pm$ 2.8 | <b>51.2</b> $\pm$ 2.8 | <b>41.9</b> $\pm$ 2.7 | <b>50.7</b> $\pm$ 2.8 | <b>0.62</b> $\pm$ 0.03 |
| GPT-4.1-mini               | 46.9 $\pm$ 2.8        | 50.0 $\pm$ 2.8        | 48.8 $\pm$ 2.8        | 50.0 $\pm$ 2.8        | 41.3 $\pm$ 2.7        | 50.4 $\pm$ 2.8        | 0.62 $\pm$ 0.03        |
| GPT-4.1-nano               | 38.1 $\pm$ 2.7        | 45.1 $\pm$ 2.7        | 37.8 $\pm$ 2.7        | 45.0 $\pm$ 2.7        | 41.3 $\pm$ 2.7        | 50.4 $\pm$ 2.8        | 0.62 $\pm$ 0.03        |
| GPT-4o-mini                | 41.5 $\pm$ 2.8        | 45.4 $\pm$ 2.7        | 43.3 $\pm$ 2.8        | 46.8 $\pm$ 2.8        | 41.0 $\pm$ 2.7        | 50.1 $\pm$ 2.8        | 0.62 $\pm$ 0.03        |
| GPT-4o                     | 48.2 $\pm$ 2.8        | 47.0 $\pm$ 2.7        | 52.1 $\pm$ 2.8        | 49.4 $\pm$ 2.8        | 40.6 $\pm$ 2.7        | 49.5 $\pm$ 2.8        | 0.61 $\pm$ 0.03        |
| Inflection 3 Productivity  | 24.7 $\pm$ 2.8        | 42.0 $\pm$ 2.6        | 21.9 $\pm$ 2.7        | 44.2 $\pm$ 2.7        | 41.9 $\pm$ 2.7        | 50.7 $\pm$ 2.8        | 0.62 $\pm$ 0.03        |
| LFM 40B MoE                | 30.8 $\pm$ 2.7        | 38.3 $\pm$ 2.7        | 20.7 $\pm$ 2.7        | 34.0 $\pm$ 2.7        | 33.8 $\pm$ 2.7        | 44.8 $\pm$ 2.8        | 0.53 $\pm$ 0.03        |

Table 12: RAG performance of additional models when retrieving  $k = 1$  and  $k = 3$  most relevant documents. Precision is shown only for  $k = 3$  as it is equivalent to Recall in the  $k = 1$  case. This table shows that retrieving three documents is better in almost all cases than retrieving a single document, despite the incurred false positives that arise due to most of the examples having less than three relevant documents.

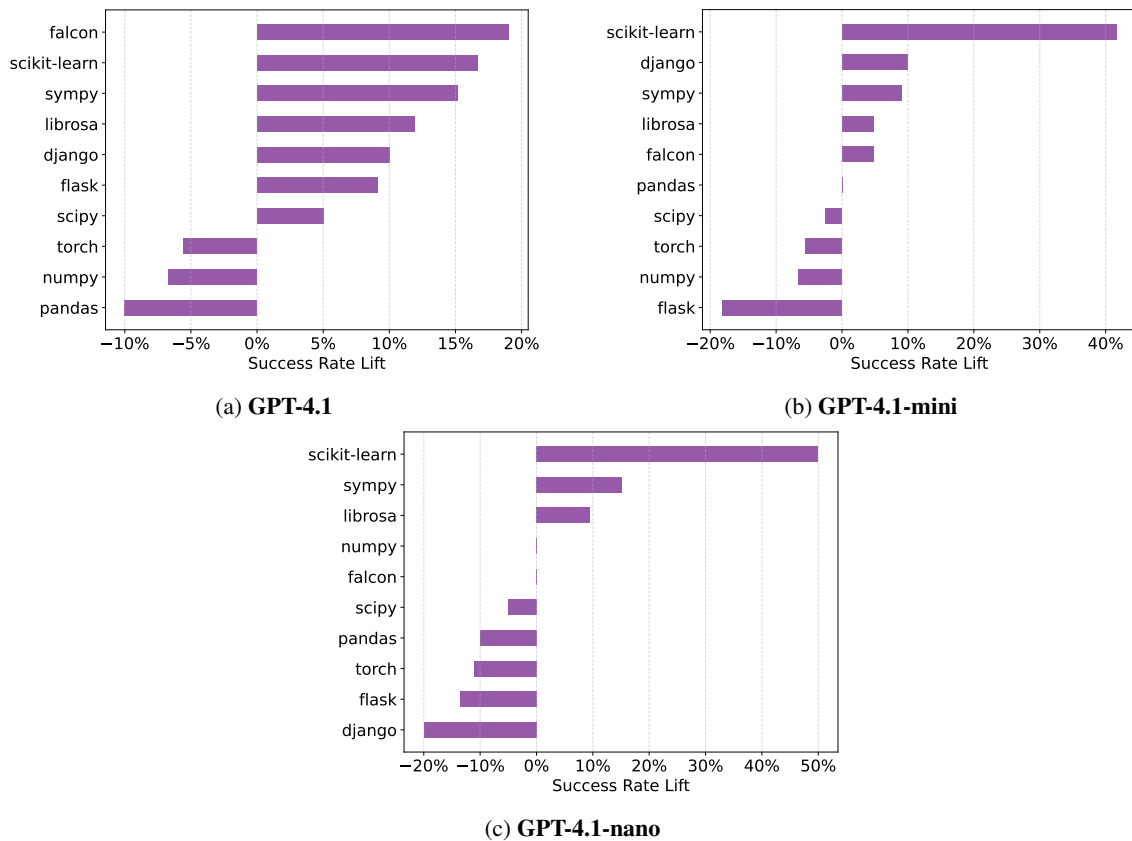
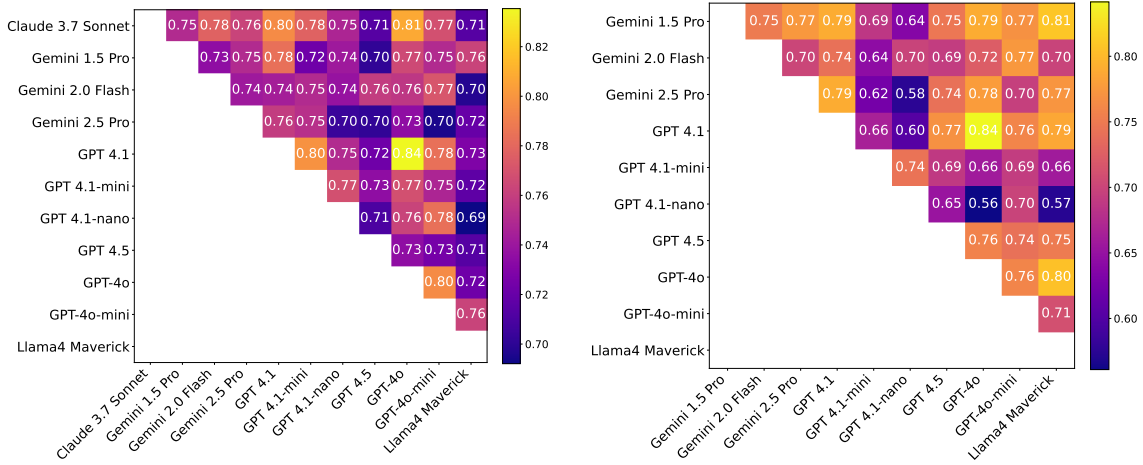
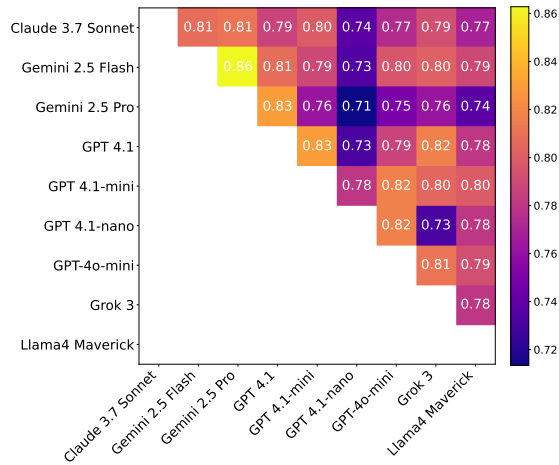


Figure 13:  $\Delta$  **Success Rate of RAG over Greedy Decoding, per library**. The 10 most frequent libraries in **GitChameleon 2.0** are shown here. The plots demonstrate a trend where smaller models are less effective at using RAG, with the full-size GPT-4.1 improving on 7 libraries, the mini version improving on 5 and the nano version improving only on 3.



(a) Greedy Decoding

(b) Zero-Shot Chain-Of-Thought



(c) RAG (k=3)

Figure 14: **Intra model sample agreement rates.** These plots show the rate of samples that have the same pass/fail result among all pairs of models, under the Greedy Decoding, Zero-Shot CoT and RAG settings. Each cell in these plots represents the agreement rate of a pair of models, with the rate also being color-coded. The high agreement rates in all three subfigures show that ensembling different models would have a limited effect on the success rates.

## D Related Work

This section discusses additional lines of related work and provides a detailed comparison of the differentiation of **GitChameleon 2.0** versus previous efforts.

### D.1 Code Evolution Datasets

While the main text provides a high-level overview of the most similar benchmarks, this section offers a more detailed differentiation between **GitChameleon 2.0** and other relevant works. We categorize these benchmarks based on several key dimensions, including their evaluation method (execution-based vs. non-executable) and, most importantly, their core **task format (instruction-based generation vs. completion- or repair-based tasks)**. This distinction is critical as it tests different capabilities of language models.

#### D.1.1 Task Format: Instruction-Based Generation

**GitChameleon 2.0** is fundamentally an **instruction-based** benchmark. For each problem, the model is given a natural language "Problem Statement" and starter code. The core challenge is to comprehend the user's intent and generate a new, functionally correct solution that adheres to specific version constraints. This tests a

| Assistant         | Model             | Linter        | Pylint Score $\uparrow$ | Success Rate (%) |
|-------------------|-------------------|---------------|-------------------------|------------------|
| Cline (IDE)       | GPT-4.1           | N/A           | 1.06                    | 54.6 $\pm$ 2.8   |
|                   |                   | Black + Isort | 1.69                    | 54.6 $\pm$ 2.8   |
|                   |                   | Ruff          | <b>2.64</b>             | 54.6 $\pm$ 2.8   |
| Goose (CLI)       | GPT-4o            | N/A           | 0.53                    | 36.3 $\pm$ 2.7   |
|                   |                   | Black + Isort | 1.82                    | 36.3 $\pm$ 2.7   |
|                   |                   | Ruff          | <b>2.92</b>             | 36.3 $\pm$ 2.7   |
| Claude Code (CLI) | Claude 3.7 Sonnet | N/A           | 0.00                    | 48.8 $\pm$ 2.8   |
|                   |                   | Black + Isort | 1.92                    | 48.8 $\pm$ 2.8   |
|                   |                   | Ruff          | <b>2.60</b>             | 48.8 $\pm$ 2.8   |

Table 13: **Static Analysis and Auto-linting/Formatting.** Pylint<sup>8</sup> scores are averaged across code samples and are scored out of 10. The success rate numbers presented are the same as in Table 3 wherein Goose has no access to problem statement while Cline and Claude are provided with the same. We observe that the original generated solutions via coding assistants do not meet minimum quality standard requirements, however when improved via auto-linters like Black<sup>9</sup>, ISort<sup>10</sup> and Ruff<sup>11</sup>, their code quality improves but with no impact to the success rate. This demonstrates that there are no confounding errors like indentation issues, unused imports and other formatting issues influencing our evaluation results observed. NOTE: For Ruff formatting, we used the already formatted/ linted solutions via Black and ISort.

model's ability to translate human requirements into code.

#### D.1.2 Task Format: Code Update, Repair, and Completion

In contrast, many other benchmarks focus on tasks where the primary input is existing code, not a natural language instruction. The model's goal is to modify, repair, or complete a given code snippet.

**Code Update and Repair Benchmarks** A significant body of work evaluates a model's ability to modify or repair existing code.

- **CodeUpdateEval** (Liu et al., 2024) and **JavaVersionGenBench** (Ciniselli et al., 2024) are code modification benchmarks for Python and Java, respectively. They provide a model with a working piece of code and require it to be updated to a newer library version.
- **RustEvo2** (Liang et al., 2025) is a code repair benchmark for Rust. It provides a model with code that is broken due to a dependency update and asks it to generate a fix based on compiler errors.

<sup>8</sup><https://pylint.pycqa.org/en/latest/index.html>

<sup>9</sup><https://black.readthedocs.io/en/stable/>

<sup>10</sup><https://pycqa.github.io/isort/>

<sup>11</sup><https://docs.astral.sh/ruff/>

| Benchmark               | Language | Evaluation Method | Core Task   | Source of Changes                                     | Key Differentiator from <b>GitChameleon 2.0</b>  |
|-------------------------|----------|-------------------|---|---|--|
| <b>GitChameleon 2.0</b> | Python   | Execution-Based   | <b>Generation for a static version:</b> Writes new code for a specific, often older, library version.         | Real, documented historical breaking changes.         | <i>(Baseline for comparison)</i>   |
| CodeUpdateEval          | Python   | Execution-Based   | <b>Code Updating:</b> Modifies existing code to work with a newer library version.                            | Real-world software update commits.                   | Focuses on migrating code <b>forward</b> to a newer version, not generating for a static one.              |
| JavaVersionGenBench     | Java     | Execution-Based   | <b>Code Updating:</b> Modifies existing Java code to handle version updates.                                  | Real-world Java projects.                             | Focuses on the <b>Java</b> ecosystem and its specific language/tooling challenges.                         |
| LLM-Deprecated-API      | Python   | Non-Executable    | <b>Deprecation Fixing:</b> Identifies and replaces specific deprecated API calls.                             | A curated list of deprecated APIs.                    | Uses a <b>non-executable</b> evaluation method and has a narrow scope focused only on API deprecation.     |
| LibEvolutionEval        | Python   | Non-Executable    | <b>Code Completion:</b> Fills in a missing part of a code snippet based on context.                           | API documentation and release notes.                  | Is a <b>completion-based</b> task that does not test functional correctness through execution.             |
| RustEvo2                | Rust     | Execution-Based   | <b>Code Repair:</b> Fixes existing code that fails to compile after a dependency update.                      | Real breaking changes from Rust libraries ("crates"). | Focuses on the <b>Rust</b> ecosystem and a reactive, compiler-error-driven repair task.                    |
| CODEMENV                | Python   | Execution-Based   | <b>Environment Compatibility:</b> Generates code that is compatible with a complex environment specification. | A broad set of environment configurations.            | Has a broader focus on overall environment compatibility, not specifically on historical breaking changes. |

Table 14: Detailed comparison of **GitChameleon 2.0** with related benchmarks across several key dimensions, highlighting differences in evaluation methodology, core task, and primary programming language.

These tasks are distinct from **GitChameleon 2.0**'s, as they test a reactive, corrective capability rather than the proactive generation of new code from a specification.

**Completion-Based and Non-Executable Benchmarks** Another category of benchmarks uses non-executable metrics or focuses on code completion.

- **LibEvolutionEval** (Kuhar et al., 2024) is a non-executable benchmark structured as a "fill-in-the-middle" **completion-based task**. Its evaluation is based on textual similarity metrics (e.g., F1 score), not the functional correctness of the code.
- **LLM-Deprecated-API** (Wang et al., 2025b), which we note in our introduction, focuses on replacing deprecated APIs. This is a specific type of repair task that is evaluated using non-executable string matching.
- **CODEMENV** (Cheng et al., 2025) evaluates a model's ability to generate code compatible with a complex environment specification. While execution-based, its task is primarily driven by satisfying technical constraints rather than implementing a distinct, high-level natural language instruction.

For a detailed breakdown, Table 14 contrasts **GitChameleon 2.0** with these related benchmarks across several key methodological dimensions.

## D.2 Specialized Frameworks and Repair Techniques

Recognizing the unique challenges of library evolution, researchers and practitioners are developing specialized frameworks and automated repair techniques that often combine LLMs with other methods.

### D.2.1 DepsRAG

This framework utilizes a multi-agent system built around RAG and Knowledge Graphs specifically for reasoning about software dependencies (Alhanahnah et al., 2024). It employs distinct agents managed by an LLM: one to construct and query the dependency KG, another for web searches, and a critic agent to review and refine the generated responses, aiming for higher accuracy in complex dependency analysis tasks.

### D.2.2 Dr.Fix

This tool represents a family of approaches using LLMs, often combined with program analysis and RAG, for automated program repair. It focuses on fixing API misuse in LLM-generated code based on

the taxonomy of misuse types. It employs a detect-reason-fix pipeline and demonstrates substantial improvements in repair accuracy metrics such as BLEU and Exact Match (Behrang et al., 2025).

### D.2.3 ReplaceAPI / InsertPrompt

These are lightweight, targeted techniques designed specifically to mitigate the use of deprecated APIs in LLM-based code completion. ReplaceAPI performs a direct, post-generation substitution of known deprecated API calls with their replacements, achieving high fix rates in evaluations (Wang et al., 2025b). InsertPrompt modifies the input prompt to discourage the generation of deprecated APIs in the first place. They serve as valuable baseline approaches for this specific problem (Wang et al., 2025a).

### D.2.4 Conclusion

These works indicate a trend towards hybrid and agentic systems, moving beyond single LLM calls to more sophisticated architectures that integrate LLMs with other methods for handling library evolution. **GitChameleon 2.0** serves as an essential resource for evaluating such systems.

In the subsequent sections we present qualitative sample analyses and model generation differences.

## E Case Study: Code Assistant Failure With Search

Here we present an example where a coding assistant (Zencoder) fails at solving a problem in our benchmark even when it can search the internet for the correct API call. The task required using the `torch.special` module or equivalent `scipy.special` to calculate the modified Bessel function of the first kind of order 0. This can be computed using `torch.special.i0` or `scipy.special.i0` if the version pre-dates the introduction of `torch.special`.

### E.1 Inputs

**Problem Statement:** Calculate the modified Bessel function of the first kind, order 0 using pytorch's special functions if available in this version, otherwise you may use another library.

#### Environment

- Python Version: 3.7
- Pytorch Version: 1.10

#### Starting Code

---

```
import torch
def bessel_i0(input_tensor: torch.Tensor) -> torch.Tensor:
```

---

Listing 1: Initial Python function signature provided.

### E.2 Model Attempt and Failure

The model's attempt involved manually implementing the modified Bessel function, handling small and large inputs as separate cases.

#### Assistant Trace

---

```
[truncated]
**Web search**
Done
Let me check if the torch.special module
    with Bessel functions is
specifically available in PyTorch
1.10.0.
```

```
**Web search**
Done
The search results dont specifically
confirm if torch.special.i0 is
available in PyTorch 1.10.0. Let me
search more specifically.
```

```
**Web search**
Done
Since Im not getting clear information
about the availability of torch.
special.i0 in PyTorch 1.10.0, Ill
implement the function manually
using the mathematical definition of
the modified Bessel function of the
first kind of order 0.
Now, lets implement the solution:
```

---

Listing 2: The assistant's trace while designing the solution.

#### Proposed Solution

---

```
def bessel_i0(input_tensor: torch.Tensor) -> torch.Tensor:
    result = torch.zeros_like(input_
        tensor)
    abs_x = torch.abs(input_tensor)

    # For small values (|x| < 15), use
    the series expansion
    small_mask = abs_x < 15.0
    small_x = abs_x[small_mask]

    if small_x.numel() > 0:
        small_result = torch.ones_like(
            small_x)
        term = torch.ones_like(small_x)
        x_squared_over_4 = (small_x /
            2.0) ** 2

    # Sum the series until
    convergence or max
    iterations
```

```

for k in range(1, 30): # 30
    terms should be enough for
    good precision
    term = term * x_squared_over
    _4 / (k ** 2)
    small_result = small_result
    + term

    # Check for convergence
    if torch.all(term < 1e-10 *
        small_result):
        break

result[small_mask] = small_
result

# For large values (|x| ≥ 15),
use the asymptotic approximation
large_mask = ~small_mask
large_x = abs_x[large_mask]

if large_x.numel() > 0:
    large_result = torch.exp(large_x
        ) / torch.sqrt(2 * math.pi *
        large_x)
    result[large_mask] = large_
    result

return result}

```

Listing 3: The solution proposed by the assistant.

#### Error Traceback:

```

TestBesselI0.test_large_values -----
def test_large_values(self):
    """Test bessel_i0 with large
    positive and negative values."""
    input_tensor = torch.tensor([10.0,
    -10.0, 20.0, -20.0], dtype=torch
    .float32)
    result = bessel_i0(input_tensor)
    expected = torch.from_numpy(scipy_i0
    (input_tensor.numpy()))

> torch.testing.assert_close(
    result, expected)
E AssertionError: Tensor-likes are
    not close!

```

Listing 4: Error trace for the implementation of `bessel_i0`.

**Explanation:** The failure in the manual implementation arises because the asymptotic approximation

$$I_0(x) \approx \frac{e^x}{\sqrt{2\pi x}}$$

has only  $O(1/x)$  accuracy and omits higher-order correction terms, so for  $x \gtrsim 20$  its relative error exceeds the test tolerance of  $10^{-5}$ . To fix this, either use the built-in `torch.special.i0(x)` or include at least the first correction term in the expansion:

$$I_0(x) \approx \frac{e^x}{\sqrt{2\pi x}} \left(1 - \frac{1}{8x}\right),$$

which reduces the large- $x$  error below the required threshold.

## F Case Study: Self-Debugging in Batched Matrix Exponential Computation

This appendix details an instance where self-debugging helped rectify an initially incorrect solution from GPT-4.5 for computing the matrix exponential of batched matrices. The task required leveraging the `scipy.linalg.expm` function while ensuring compatibility with batched inputs.

### F.1 Inputs

**Problem Statement:** Complete the function that computes the matrix exponential of batched matrices, non specified parameters should use the default value

#### Environment:

- Python Version: 3.10
- SciPy Version: 1.8.1

#### Starting Code

```

from scipy import linalg
import numpy as np

def compute_matrix_exponential(A: np.
    ndarray) -> np.ndarray:
    # Model needs to complete this
    function
    pass

```

Listing 5: Initial Python function signature provided.

### F.2 First Model Attempt and Failure

The model's first attempt involved reshaping the input array and applying `scipy.linalg.expm` directly.

#### Initial Model Output

```

from scipy import linalg
import numpy as np

def compute_matrix_exponential(A: np.
    ndarray) -> np.ndarray:
    return linalg.expm(A)

```

Listing 6: First attempt by the model to define `compute_matrix_exponential`.

**Model's Explanation (Initial):** *The function directly returns `linalg.expm(A)` because SciPy's `expm` computes the matrix exponential for a given matrix. Given that  $A$  is a batch of matrices, SciPy should naturally handle this batch input and apply the exponential operation to each matrix independently, producing the correct batched result.*

**Result:** Failure

**Error Traceback Failure:** Calling `linalg.expm(A)` on a 3D array  $A \in \mathbb{R}^{b \times n \times n}$  makes SciPy treat it as one “matrix,” and raises

`ValueError: expected a square matrix`

because it only supports 2D inputs.

---

```
Traceback (most recent call last):
  File "/app/repo/eval_venvs/gcham_venv_111/lib/
python3.10/site-packages/scipy/sparse/linalg/
_matfuncs.py", line 601, in _expm
    raise ValueError('expected a square matrix')
ValueError: expected a square matrix
```

---

Listing 7: Error traceback from the initial model output.

### F.3 Self-Debugging Process and Correction

The self-debugging process recognized that simply calling `linalg.expm` was insufficient. Instead, `linalg.expm` must be applied individually to each 2D matrix within the batch. Additionally, a crucial pre-condition is that each matrix within the batch must be square.

**Second Model Attempt (Post-Debugging)** The model regenerated the solution, incorporating input validation and iterating over the individual matrices in the batch.

#### Regenerated Model Output

---

```
from scipy import linalg
import numpy as np

def compute_matrix_exponential(A: np.
ndarray) -> np.ndarray:
    return np.stack([linalg.expm(A[i])
        for i in range(A.shape[0]) if A[
            i].shape[0] == A[i].shape[1]],
            axis=0)
```

---

Listing 8: Corrected `compute_matrix_exponential` after self-debugging.

**Model’s Explanation (Regenerated):** *The corrected solution applies `linalg.expm` separately to each 2D square matrix in the batch using a list comprehension, then re-stacks the results into a 3D array. This ensures each call to `expm` receives a valid input, resolving the error.*

**Result:** Success

### F.4 Analysis of the Correction

The regenerated output effectively addressed the shortcomings of the initial attempt:

- **Fix:** Apply `expm` to each 2D slice  $A_i \in \mathbb{R}^{n \times n}$  and re-stack, ensuring every call gets a proper square matrix.

This revised approach correctly computes the matrix exponential for each matrix in the batch and handles the data dimensions appropriately, thus resolving the `ValueError` and fulfilling the problem requirements.

## G Qualitative Analysis

This section presents additional qualitative studies of attempted solutions in the Greedy Decoding and the Chain-of-Thought settings.

### G.1 Greedy Decoding

#### G.1.1 Example 1: (PyTorch)

**Problem Statement:** Calculate the logarithm of the cumulative distribution function of the standard normal distribution using available functions. If not available in PyTorch, use another library.

**Environment:**

- Python Version: 3.7
- Torch Version: 1.9.0
- Extra Dependencies: `scipy==1.7.3,`  
`numpy==1.21.6`

#### Starting Code

---

```
import torch
def log_ndtr(input_tensor: torch.Tensor)
-> torch.Tensor:
    pass
```

---

Listing 9: Greedy Example 1 Starting Code

#### Solution

---

```
import numpy as np
from scipy.stats import norm
output = torch.from_numpy(norm.
    logcdf(input_tensor.numpy()))
return output
```

---

Listing 10: Greedy Example 1 Solution

#### Claude 3.7 Failure Output

---

```
import torch
import math
from torch.distributions import Normal

def log_ndtr(input_tensor: torch.Tensor)
-> torch.Tensor:
    normal = Normal(0, 1)
    return normal.cdf(input_tensor).log
    ()
```

---

Listing 11: Claude 3.7 Output

#### Gemini 2.5 Pro Failure Output

---

```

import torch
from torch.distributions import Normal

def log_ndtr(input_tensor: torch.Tensor)
    -> torch.Tensor:
    loc = torch.tensor(0.0, device=input_
        tensor.device, dtype=input_tensor.
        dtype)
    scale = torch.tensor(1.0, device=input_
        tensor.device, dtype=input_tensor.
        dtype)

    standard_normal = Normal(loc=loc,
        scale=scale)

    log_cdf_values = standard_normal.log_
        cdf(input_tensor)

    return log_cdf_values

```

---

Listing 12: Gemini 2.5 Pro Output

### Analysis of Model Outputs

Both Claude 3.7 and Gemini 2.5 Pro incorrectly attempt to compute the log-CDF of the standard normal distribution using PyTorch’s Normal distribution. Claude 3.7 computes `normal.cdf(input_tensor).log()`, which is numerically unstable for large negative inputs since the CDF approaches zero and its logarithm can underflow to  $-\infty$ , leading to invalid results. Gemini 2.5 Pro mistakenly calls a non-existent `log_cdf` method on the PyTorch `Normal` object, causing an immediate `AttributeError` in PyTorch 1.9. This reflects a confusion between PyTorch and other libraries (such as SciPy or TensorFlow Probability) that do provide this method. In summary, neither model provides a robust or correct implementation; the only reliable approach in this environment is to use `scipy.stats.norm.logcdf` as shown in the reference solution, or to implement a numerically stable approximation in PyTorch when SciPy is unavailable.

#### G.1.2 Greedy Example 2 (SciPy)

**Problem Statement:** complete the following function that check if all the batch of matrices are invertible, using numpy 1.25.1.

**Environment:**

- Python Version: 3.10
- Scipy Version: 1.11.1

#### Starting Code

---

```

import warnings
from scipy.linalg import det
import numpy as np
warnings.filterwarnings('error')

```

---

```

def check_invertibility(matrices: np.
    ndarray) -> np.bool_:
    pass

```

---

Listing 13: Greedy Example 3 Starting Code

#### Solution

---

```

return np.all(det(matrices))

```

---

Listing 14: Greedy Example 3 Solution

#### Gemini 2.5 Correct Output

---

```

import warnings
from scipy.linalg import det
import numpy as np
warnings.filterwarnings('error')

def check_invertibility(matrices: np.
    ndarray) -> np.bool_:
    try:
        # Check if matrices is a batch
        # of matrices
        if matrices.ndim < 3:
            # If it's a single matrix,
            # add a batch dimension
            matrices = np.expand_dims(
                matrices, axis=0)

        # Check if all matrices are
        # square
        batch_size, rows, cols =
            matrices.shape
        if rows != cols:
            return np.bool_(False)

        # Calculate determinants for all
        # matrices in the batch
        determinants = np.array([det(
            matrix) for matrix in
            matrices])

        # A matrix is invertible if its
        # determinant is not zero
        # Check if all determinants are
        # non-zero
        return np.bool_(np.all(
            determinants != 0))

    except Exception:
        # If any error occurs during
        # computation, return False
        return np.bool_(False)

```

---

Listing 15: Gemini 2.5 Pro Output

#### GPT 4.1 Failure Output

---

```

import warnings
from scipy.linalg import det
import numpy as np
warnings.filterwarnings('error')

def check_invertibility(matrices: np.
    ndarray) -> np.bool_:
    for matrix in matrices:
        try:
            d = det(matrix)

```

---

```

    if np.isclose(d, 0):
        return np.bool_(False)
    except Warning:
        return np.bool_(False)
    return np.bool_(True)

```

Listing 16: GPT 4.1 Output

### Analysis of Model Outputs

Gemini 2.5 Pro produces a robust solution that handles both single matrices and batches, verifies that matrices are square, computes the determinant for each matrix, and correctly checks if all determinants are non-zero to determine invertibility, returning False on any exception to ensure safe failure. In contrast, GPT 4.1's output contains a subtle but critical error in exception handling: the code attempts to catch Warning as an exception, but Python treats warnings and exceptions differently, so this approach is ineffective—especially since warnings.filterwarnings('error') escalates warnings to exceptions that should be caught as Exception. As a result, GPT 4.1's function may fail or produce unpredictable results in the presence of numerical issues or shape mismatches, despite using the correct logic for checking determinant values.

#### G.1.3 Greedy Example 3 (SymPy)

**Problem Statement:** Write a custom jacobi symbols function that compute the Jacobi symbol.

**Environment:**

- Python Version: 3.9
- SymPy Version: 1.13

#### Starting Code

```

import sympy

def custom_jacobi_symbols(a: int, n: int) -> int:
    pass

```

Listing 17: Greedy Example 4 Starting Code

#### Solution

```

return sympy.jacobi_symbol(a, n)

```

Listing 18: Greedy Example 4 Solution

### GPT 4.1 Correct Output

```

import sympy

def custom_jacobi_symbols(a: int, n: int) -> int:
    return sympy.ntheory.jacobi_symbol(a, n)

```

Listing 19: GPT 4.1 Output

### O1 Failure Output

```

import sympy

def custom_jacobi_symbols(a: int, n: int) -> int:
    return sympy.ntheory.residues.jacobi_symbol(a, n)

```

Listing 20: O1 Output

### Analysis of Model Outputs

GPT-4.1 provides the correct solution by calling `sympy.ntheory.jacobi_symbol`, which is the proper and documented API for computing the Jacobi symbol in SymPy 1.13. In contrast, O1 incorrectly uses `sympy.ntheory.residues.jacobi_symbol`, assuming that the function is nested in the residues submodule; however, in this version of SymPy, `jacobi_symbol` is not found there, resulting in an `AttributeError` on all tests. This error highlights a lack of up-to-date API knowledge or reliance on outdated documentation.

#### G.1.4 Greedy Example 4 (Flask)

**Problem Statement:** Complete the app set-up so that, when given a batch of matrix, the json encoding compute the determinants of each matrix, before flattening and converting the result to a list, we are using `scipy 1.11.1`. Do not run the app in your code.

**Environment:**

- Python Version: 3.10
- Flask Version: 3.0.0
- Extra Dependences: `scipy==1.11.1`

#### Starting Code

```

import flask
import numpy as np
from scipy import linalg

app = flask.Flask('test1')
@app.route('/data')
def data(num_list):
    return flask.jsonify({'numbers': num_list})

def eval_app(app, data_fn, num_arr):
    with app.test_request_context():
        response = data_fn(num_arr)
    return response.get_data(as_text=True)

class MyCustomJSONHandler(flask.json.provider.DefaultJSONProvider):
    def default(self, obj):
        if isinstance(obj, np.ndarray) and len(obj.shape)==3 and obj.shape[-1]==obj.shape[-2]:

```

```
pass
```

Listing 21: Greedy Example 5 Starting Code

### Solution

```
res = linalg.det(obj)
    return res.tolist()
    return super().default(obj)

app.json_provider_class =
    MyCustomJSONHandler
app.json = app.json_provider_class(
    app)
```

Listing 22: Greedy Example 5 Solution

### Gemini 2.5 Pro Correct Output

```
import flask
import numpy as np
from scipy import linalg

app = flask.Flask('test1')
@app.route('/data')
def data(num_list):
    return flask.jsonify({'numbers': num_list})

def eval_app(app, data_fn, num_arr):
    with app.test_request_context():
        response = data_fn(num_arr)
        return response.get_data(as_text=True)

class MyCustomJSONHandler(flask.json.provider.DefaultJSONProvider):
    def default(self, obj):
        if isinstance(obj, np.ndarray)
            and len(obj.shape)==3 and
            obj.shape[-1]==obj.shape[-2]
            :
            # Calculate determinants for
            # each matrix in the
            # batch
            determinants = linalg.det(
                obj)
            # Convert the resulting 1D
            # array of determinants to
            # a list
            return determinants.tolist()
        # Fallback to the default
        # serialization for other
        # types
        return super().default(obj)

# Apply the custom JSON provider to the
# app
app.json = MyCustomJSONHandler(app)
```

Listing 23: Gemini 2.5 Pro Output

### Claude 3.7 Failure Output

```
import flask
import json

app = flask.Flask('test')
@app.route('/data')
def data(num_set):
    return flask.jsonify({'numbers': num_set})
```

```
def eval(app, data_fn, num_set):
    with app.test_request_context():
        response = data_fn(num_set)
        return response.get_data(as_text=False)
```

```
def app_set_up(app: flask.Flask) -> None:
    :
    def set_encoder(obj):
        if isinstance(obj, set):
            return sorted(list(obj))
        raise TypeError(f"Object of type
            {type(obj)} is not JSON
            serializable")
```

```
app.json.encoder = lambda: json.
    JSONEncoder(default=set_encoder)
```

Listing 24: Claude 3.7 Output

### Analysis of Model Outputs

Gemini 2.5 Pro provides a solution that correctly subclasses `flask.json.provider.DefaultJSONProvider` and overrides the default method to check for batches of square matrices, compute their determinants using `scipy.linalg.det`, and serialize the result as a list, fully aligning with both the requirements and Flask 3.0.0's JSON provider API. In contrast, Claude 3.7's output incorrectly tries to customize JSON serialization by modifying `app.json.encoder`—a property that is not available in Flask 3.0.0—and focuses on encoding sets rather than NumPy arrays, neglecting the determinant computation altogether. Consequently, Claude's approach fails with an `AttributeError` and does not satisfy the task requirements even if the encoder logic were correct, while Gemini's implementation is robust and functional for the specified environment.

## G.2 Zero-Shot Chain-Of-Thought

### G.2.1 CoT Example 1 (Torch)

**Problem Statement:** Calculate the modified Bessel function of the first kind, order 1 using PyTorch's special functions if available in this version, otherwise you may use another library.

**Environment:**

- Python Version: 3.7
- Torch Version: 1.9.0
- Extra Dependencies: `scipy==1.7.3,`  
`numpy==1.21.6`

### Starting Code

---

```
import torch
def bessel_i1(input_tensor: torch.Tensor
) -> torch.Tensor:
    pass
```

---

Listing 25: CoT Example 1 Starting Code

### Solution

---

```
import numpy as np
from scipy.special import i1 as
    scipy_i1
output = torch.from_numpy(scipy_i1(
    input_tensor.numpy()))
return output
```

---

Listing 26: CoT Example 1 Solution

### Claude 3.7 Correct Output

---

```
import torch
from scipy import special

def bessel_i1(input_tensor: torch.Tensor
) -> torch.Tensor:
    # Convert PyTorch tensor to NumPy
    array
    input_numpy = input_tensor.cpu().
        numpy()

    # Apply the modified Bessel function
    of the first kind, order 1
    result_numpy = special.i1(input_
        numpy)

    # Convert back to PyTorch tensor
    result_tensor = torch.from_numpy(
        result_numpy).to(input_tensor.
        device)

    # Ensure the output has the same
    dtype as the input
    result_tensor = result_tensor.to(
        dtype=input_tensor.dtype)

    return result_tensor
```

---

Listing 27: Claude 3.7 Output

### Gemini 2.5 Pro Failure Output

---

```
import torch
def bessel_i1(input_tensor: torch.Tensor
) -> torch.Tensor:
    """Calculates the modified Bessel
    function of the first kind,
    order 1.

    Args:
        input_tensor: The input tensor.

    Returns:
        A tensor containing the element-
        wise modified Bessel
        function of the first kind,
        order 1.
    """
    # Check if torch.special.i1 is
    available (it is in torch 1.9.0)
    if hasattr(torch.special, 'i1'):
```

```
        return torch.special.i1(input_
            tensor)
    else:
        # Fallback or error handling if
        needed for other versions
        # For torch 1.9.0, this else
        block should not be reached.
        raise NotImplementedError("torch
        .special.i1 is not available
        in this PyTorch version.")
```

---

Listing 28: Gemini 2.5 Pro Output

### Analysis of Model Outputs

Claude 3.7 provides a robust solution by correctly identifying that `torch.special.i1` is not available in PyTorch 1.9.0 and instead leveraging `scipy.special.i1` to compute the modified Bessel function, converting the result back to a PyTorch tensor and ensuring proper device and dtype handling; this approach is fully compatible with the given environment and passes all tests. In contrast, Gemini 2.5 Pro mistakenly assumes that `torch.special.i1` exists in PyTorch 1.9.0, checks for its presence using `hasattr`, and raises a `NotImplementedError` when unavailable; however, since this function is absent in the specified version, the code fails every test due to the raised exception.

### G.2.2 CoT Example 2 (Scikit-learn)

**Problem Statement:** Adapt the use of ‘manhattan distances’ to obtain a pairwise distance matrix.

**Environment:**

- Python Version: 3.10
- Scikit-learn Version: 1.1
- Extra Dependencies: `numpy==1.23.5`

### Starting Code

---

```
from sklearn.metrics.pairwise import
    manhattan_distances
import numpy as np
def get_pairwise_dist(X: np.ndarray, Y:
    np.ndarray) -> np.ndarray:
    distances = manhattan_distances(X, Y
        , sum_over_features=False)
    pass
```

---

Listing 29: CoT Example 2 Starting Code

### Solution

---

```
return np.sum(distances, axis=1)
```

---

Listing 30: CoT Example 2 Solution

### GPT 4.1 Correct Output

---

```

from sklearn.metrics.pairwise import
    manhattan_distances
import numpy as np
def get_pairwise_dist(X: np.ndarray, Y:
    np.ndarray) -> np.ndarray:
    distances = manhattan_distances(X, Y
    )
    return distances

```

---

Listing 31: GPT 4.1 Output

### Claude 3.7 Failure Output

---

```

from sklearn.metrics.pairwise import
    manhattan_distances
import numpy as np
def get_pairwise_dist(X: np.ndarray, Y:
    np.ndarray) -> np.ndarray:
    distances = manhattan_distances(X, Y
    , sum_over_features=False)
    return np.sum(distances, axis=2)

```

---

Listing 32: Claude 3.7 Pro Output

### Analysis of Model Outputs

In this example, GPT 4.1 provides a correct implementation by calling `manhattan_distances(X, Y)` without the `sum_over_features` argument, which by default returns the desired pairwise distance matrix of shape  $(n_{\text{samples}_X}, n_{\text{samples}_Y})$ . In contrast, Claude 3.7 incorrectly sets `sum_over_features=False`, causing the output to be a 3D array of shape, and then applies `np.sum(distances, axis=2)`, which does not match the problem’s intent and produces incorrect results if further summed incorrectly. The misuse of the `sum_over_features` parameter and the erroneous axis for reduction demonstrates a misunderstanding of the scikit-learn API and leads to a failure, while GPT 4.1’s output is concise and correct for all valid input shapes.

### G.2.3 CoT Example 3 (Falcon)

**Problem Statement:** Create a class named `CustomRouter` to manage your application’s routes. The class should maintain an internal dictionary named `routes` for storing the mapping between URI templates and their associated resources. Implement an `add_route` method that accepts three arguments: a URI template, a resource, and additional keyword arguments. This method should use Falcon’s routing utility—specifically, the `map_http_methods` function from the `falcon.routing` module—to generate a mapping of HTTP methods to resource handlers. If a `'fallback'` parameter is provided in the keyword arguments, it should be passed to `map_http_methods`. The

method should then store a tuple consisting of the resource and the generated method mapping in the routes dictionary, using the URI template as the key, and finally return the generated method mapping. You must not modify the provided starting code, which includes the empty `CustomRouter` class. Instead, complete the solution by dynamically adding the required method in the `solution()` function.

#### Environment:

- Python Version: 3.10
- Falcon Version: 2.0

#### Starting Code

---

```

class CustomRouter:
    def __init__(self):
        self.routes = {}

def solution() -> None:

    def add_route(

```

---

Listing 33: CoT Example 3 Starting Code

#### Solution

---

```

self, uri_template, resource, **kwargs):
    from falcon.routing import map_
        http_methods
    method_map = map_http_methods(
        resource, kwargs.get('
        fallback', None))
    self.routes[uri_template] = (
        resource, method_map)
    return method_map

CustomRouter.add_route = add_route

```

---

Listing 34: CoT Example 3 Solution

### O1 Correct Output

---

```

import falcon
from falcon.routing import map_http_
    methods

class CustomRouter:
    def __init__(self):
        self.routes = {}

def solution() -> None:
    def add_route(self, uri_template,
        resource, **kwargs):
        fallback = kwargs.pop('fallback'
        , None)
        if fallback is not None:
            method_map = map_http_
                methods(resource,
                fallback=fallback)
        else:
            method_map = map_http_
                methods(resource)
        self.routes[uri_template] = (
            resource, method_map)

```

```

    return method_map
CustomRouter.add_route = add_route

```

Listing 35: O1 Output

### GPT 4.1 Failure Output

```

from falcon.routing import map_http_
    methods

def solution() -> None:
    def add_route(self, uri_template,
        resource, **kwargs):
        fallback = kwargs.pop('fallback',
            None)
        if fallback is not None:
            method_map = map_http_
                methods(resource,
                    fallback=fallback)
        else:
            method_map = map_http_
                methods(resource)
        self.routes[uri_template] = (
            resource, method_map)
    return method_map

CustomRouter.add_route = add_route

```

Listing 36: GPT 4.1 Output

### Analysis of Model Outputs

In this case, the O1 output provides a fully correct solution by dynamically attaching an `add_route` method to the `CustomRouter` class, properly handling the optional `fallback` argument, and updating the routes dictionary according to the problem statement. The method is inserted via the `solution()` function without modifying the provided class definition, and it uses Falcon’s `map_http_methods` utility to construct the method mapping as required. In contrast, the GPT 4.1 output omits the explicit definition of the `CustomRouter` class in its solution, violating the requirement to use the existing starting code. Although the logic within the `solution()` function is correct, the absence of a `CustomRouter` definition in the completed module would lead to a `NameError` or otherwise prevent the expected dynamic method attachment. The critical distinction is that O1 respects all constraints including not modifying the class definition directly, while GPT 4.1 provides an incomplete module, failing to meet the initialization requirements set by the problem.

## H Logic vs. Knowledge Retention

The goal of our proposed benchmark, **GitChameleon**, is to evaluate a model’s ability to retain version-specific knowledge—specifically, whether it can recall the functionalities associated

with particular library versions it has been trained on. Notably, this capability is distinct from the ability to generate logically correct code. While we do not explicitly disentangle whether model failures on our evaluation suite stem from incorrect logic generation or incorrect API version usage, our benchmark is intentionally designed so that most problems primarily test knowledge retention rather than complex logic reasoning. For each problem in our dataset, we compute the number of logic-related nodes in the Abstract Syntax Tree (AST) of the ground-truth solution and present their distribution in Figure 15. As shown, most ground-truth solutions contain fewer than five logic-related AST nodes. This supports our claim that the benchmark is primarily designed to assess version-specific knowledge retention rather than complex logic-based code generation.

Table 15: Criteria for classifying AST nodes as logic-related.

| Condition   | Classification |
|---|----------------|
| Calling a user-defined function                                 | ✓              |
| Calling built-in Python operators (e.g., +)                     | ✓              |
| Calling a math or utility function with non-obvious purpose     | ✓              |
| Calling a library method (e.g., <code>torch.from_numpy</code> ) | ✗              |
| Composing multiple calls together                               | ✓              |

The criteria for classifying AST nodes as logic-related are provided in Table 15, and we include visualizations of the ASTs for two example ground-truth solutions for further illustration in Figures 16 and 17 respectively.

1. Sample ID: 0, Logic Nodes: 3

```

import torch
def log_ndtr(input_tensor: torch.
    Tensor) -> torch.Tensor:
    import numpy as np
    from scipy.stats import norm
    output = torch.from_numpy(norm.
        logcdf(input_tensor.numpy())
    )
    return output

```

Listing 37: Sample 0 Ground Truth Solution

2. Sample ID: 329, Logic Nodes: 0

```

import matplotlib.pyplot as plt
def use_seaborn() -> None:
    plt.style.use("seaborn")

```

Listing 38: Sample 329 Ground Truth Solution

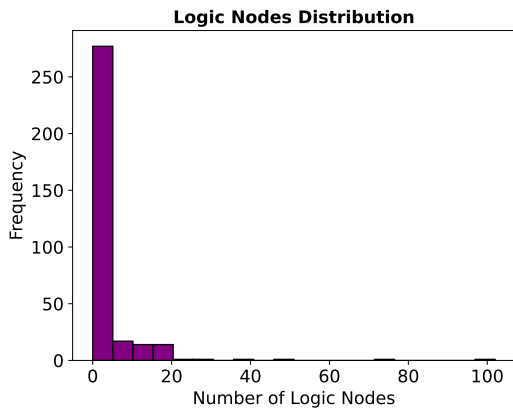


Figure 15: **Logic Nodes Distribution over samples' ground truth solutions' ASTs.** Most ground truth solutions have less than **five** logic nodes.

## I Prompt Templates

This appendix contains all the prompts we had used for our experiments:

- The prompts for greedy sampling are given in Figure 18.
- The prompts for self-debugging are given in Figure 19.
- The prompt for the multi-step agent is given in Figure 20.
- The prompt for RAG is given in Figure 21.
- The prompt and file format for Coding Assistants are given in Figure 22.
- The prompt for SEK is given in Figure 23 (for keywords generation) and Figure 24 (for code generation).

## J Artifacts and Model Details

This appendix provides citations for various artifacts and models mentioned in the paper.

### J.1 Libraries

This is the full list of libraries included in **GitChameleon 2.0**.

- PyTorch (Paszke et al., 2019)
- Geopandas (Jordahl et al., 2020)
- NLTK (Loper and Bird, 2002)
- NetworkX (Hagberg et al., 2008)

- GeoPy<sup>12</sup>
- Gradio (Abid et al., 2019)
- Scikit-Learn (Buitinck et al., 2013)
- Matplotlib (Hunter, 2007)
- PyCaret<sup>13</sup>
- Pandas (The pandas development team, 2020; McKinney, 2010)
- NumPy (Harris et al., 2020)
- LightGBM<sup>14</sup>
- spaCy<sup>15</sup>
- Django<sup>16</sup>
- SciPy (Virtanen et al., 2020)
- Flask<sup>17</sup>
- Jinja2<sup>18</sup>
- SymPy<sup>19</sup>
- Seaborn<sup>20</sup>
- mitmproxy<sup>21 22</sup>
- pytest<sup>23</sup>
- Falcon web framework<sup>24</sup>
- Tornado web server<sup>25</sup>
- Plotly<sup>26</sup>
- Librosa<sup>27</sup>
- Pillow<sup>28</sup>

<sup>12</sup><https://pypi.org/project/geopy/>

<sup>13</sup><https://pycaret.org/>

<sup>14</sup><https://lightgbm.readthedocs.io/>

<sup>15</sup><https://spacy.io/>

<sup>16</sup><https://www.djangoproject.com/>

<sup>17</sup><https://flask.palletsprojects.com/>

<sup>18</sup><https://jinja.palletsprojects.com/>

<sup>19</sup><https://www.sympy.org/en/index.html>

<sup>20</sup><https://seaborn.pydata.org/>

<sup>21</sup><https://mitmproxy.org/>

<sup>22</sup><https://mitmproxy.org/>

<sup>23</sup><https://pytest.org/>

<sup>24</sup><https://falconframework.org/>

<sup>25</sup><https://www.tornadoweb.org/>

<sup>26</sup><https://plotly.com/python/>

<sup>27</sup><https://librosa.org/doc/latest/index.html>

<sup>28</sup><https://python-pillow.org/>

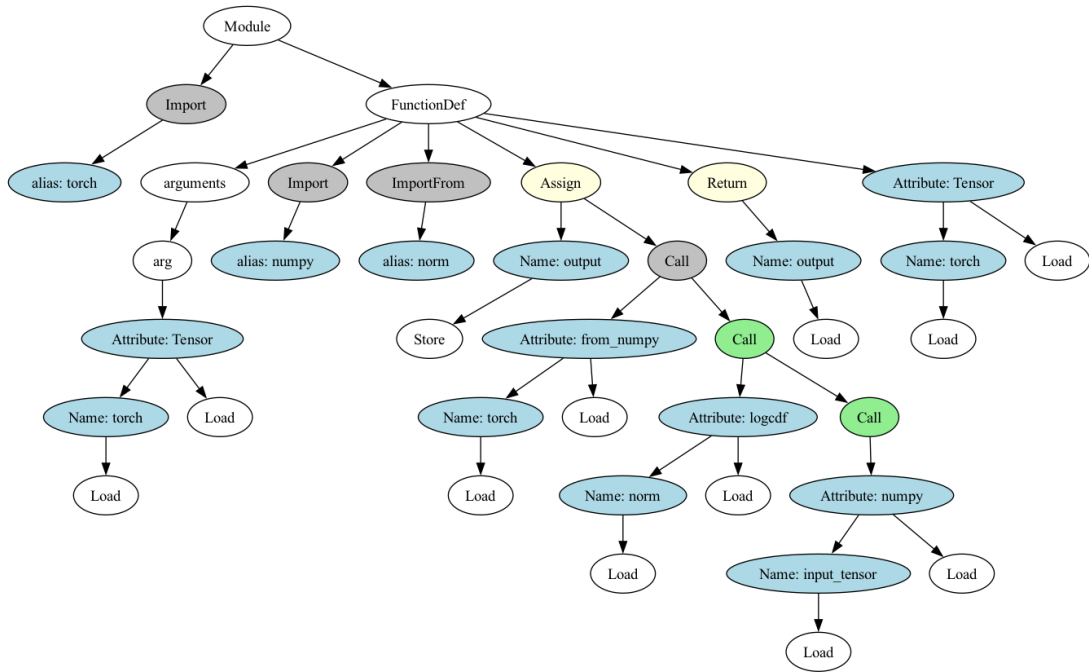


Figure 16: AST visualization for the ground-truth solution of Sample ID 0. The three color-coded call nodes (in grey and green) represent the logic-related components, classified under the “composing multiple calls together” category. The corresponding ground-truth code is shown in Code block 37 for reference.

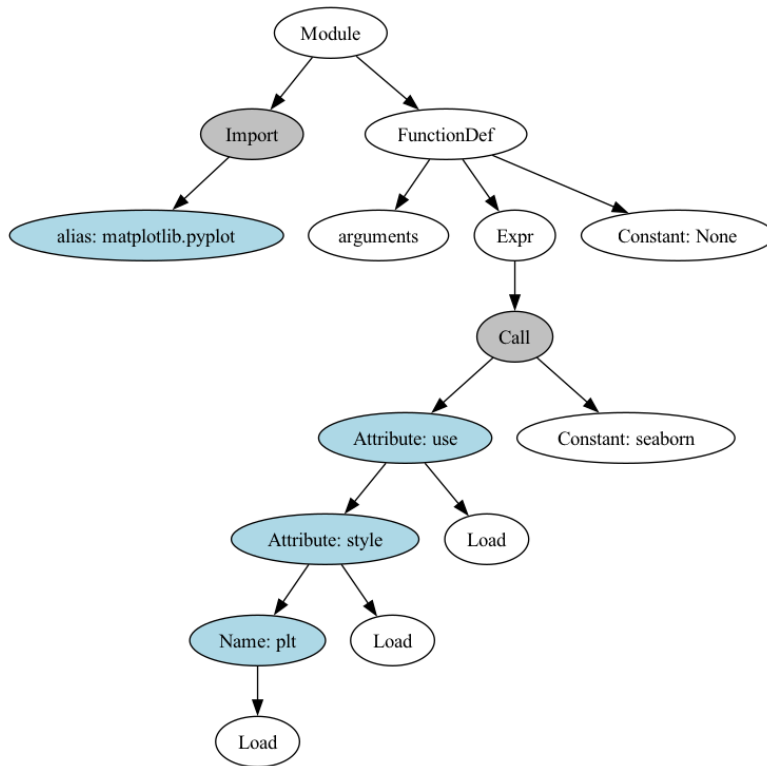


Figure 17: AST visualization for the ground-truth solution of Sample ID 329. No logic nodes are present, as the only call node corresponds to the “calling a library method” category. The ground-truth solution is provided for reference in Code block 38.

Figure 18: Prompts for Greedy Sampling

(a) System Prompt for Zero-Shot Prompting

```

You are a skilled Python programmer tasked
↳ with solving a coding problem. Your
↳ goal is to provide a clear, efficient,
↳ and correct solution that meets all the
↳ specified requirements.

Please provide your solution following
↳ these guidelines:

1. Use the required library in your
↳ solution.
2. Incorporate the provided starter code
↳ correctly.
3. Write your solution in Python.
4. Format your solution within a markdown
↳ code block.
5. Ensure your code is clean, efficient,
↳ and well-commented.
6. Output only the code block and nothing
↳ else.

Example output format:

```python
# [Your code here, incorporating the
↳ starter code]

# [Additional code and comments as needed]
```

After writing your solution, please review
↳ it to ensure all requirements are met
↳ and the code is correct and efficient.

Here are the key elements for this task:

```

(b) System Prompt for Chain-Of-Thought Prompting

```

You are a skilled Python programmer tasked
↳ with solving a coding problem. Your
↳ goal is to provide a clear, efficient,
↳ and correct solution that meets all the
↳ specified requirements.

First, let's think step-by-step. Then,
↳ please provide your solution following
↳ these guidelines:

1. Use the required library in your
↳ solution.
2. Incorporate the provided starter code
↳ correctly.
3. Write your solution in Python.
4. Format your solution within a markdown
↳ code block.
5. Ensure your code is clean, efficient,
↳ and well-commented.
6. Output nothing else after the code
↳ block.

Example output format:

[Step-by-step thinking]
```python
# [Your code here, incorporating the
↳ starter code]

# [Additional code and comments as needed]
```

After writing your solution, please review
↳ it to ensure all requirements are met
↳ and the code is correct and efficient.

Here are the key elements for this task:

```

(c) User Prompt

```

1. Required Library:
<library>
{{library}}
</library>

2. Python version:
<python>
{{python_version}}
</python>

2. Coding Problem:
<coding_problem>
{{coding_problem}}
</coding_problem>

3. Starter Code:
<starter_code>
{{starter_code}}
</starter_code>

```

- tqdm<sup>29</sup>
- Kymatio<sup>30</sup>

## J.2 Models

### Open-Weights Models

The following open-weights models were evaluated:

- Llama 3.1 Instruct Turbo: (Kassianik et al., 2025)
- Llama 3.3 Instruct Turbo 70B: (AI, 2025)
- Llama 4 Maverick 400B: (AI, 2025)
- Qwen 2.5-VL Instruct 72B: (Qwen et al., 2025)
- Qwen 3 235B: (Yang et al., 2025)
- Command A 111B: (Cohere et al., 2025)
- DeepSeek R1 685B: (DeepSeek-AI, 2025)
- DeepSeek v3: (DeepSeek-AI et al., 2025)
- Openhands LM 32B v0.1: (Wang, 2025)
- Reka Flash-3: (Reka)
- Jamba 1.6 Mini, Large: (Lieber et al., 2024)

### Enterprise Models

The following enterprise models were evaluated:

- Arcee CoderL: (Arcee)
- Claude 3.5 Haiku<sup>31</sup>
- Claude 3.5 Sonnet<sup>32</sup>
- Claude 3.7 Sonnet: (Anthropic, 2025)
- Claude 4 Sonnet<sup>33</sup>
- CommandR+<sup>34</sup>
- Gemini 1.5 Pro: (Team et al., 2024)
- Gemini 2.0 Flash: (Kampf, 2025)

<sup>29</sup><https://github.com/tqdm/tqdm>

<sup>30</sup><https://librosa.org/doc/latest/index.html>

<sup>31</sup><https://www.anthropic.com/claude/haiku>

<sup>32</sup><https://www.anthropic.com/news/claude-3-5-sonnet>

<sup>33</sup><https://www.anthropic.com/claude/sonnet>

<sup>34</sup><https://cohere.com/blog/command-r-plus-micro-soft-azure>

- Gemini 2.5 Pro: (Cloud, 2025)
- Gemini 2.5 Flash: (Cloud, 2025)
- GPT-4.1: (OpenAI, 2025a)
- GPT-4.1-mini: (OpenAI, 2025a)
- GPT-4.1-nano: (OpenAI, 2025a)
- GPT-4o: (OpenAI, 2024)
- GPT-4o-mini: (OpenAI, 2024)
- GPT-4.5: (OpenAI, 2025b)
- o1: (OpenAI, 2024)
- o3-mini: (OpenAI, 2024)
- codex-mini<sup>35</sup>
- Grok 3: (xAI, 2025)
- Mistral Medium 3: (Mistral AI, 2025)
- Devstral Small<sup>36</sup>
- Inflection 3 Productivity<sup>37</sup>
- Liquid LFM 40B MoE<sup>38</sup>
- Nova Pro: (Intelligence, 2024)

## J.3 Coding Assistants (CLI/IDE)

The following coding assistants were studied as part of the experimentation pipeline:

- Claude Code<sup>39</sup> (CLI)
- Goose<sup>40</sup> (CLI)
- Cline<sup>41</sup> (IDE-VSCode)
- RooCode<sup>42</sup> (IDE-VSCode)
- KiloCode<sup>43</sup> (IDE-VSCode)

<sup>35</sup><https://platform.openai.com/docs/models/code-x-mini-latest>

<sup>36</sup><https://mistral.ai/news/devstral>

<sup>37</sup><https://openrouter.ai/inflection/inflection-3-productivity>

<sup>38</sup><https://www.liquid.ai/blog/liquid-foundation-models-our-first-series-of-generative-ai-models>

<sup>39</sup><https://docs.anthropic.com/en/docs/claude-code/overview>

<sup>40</sup><https://block.github.io/goose/>

<sup>41</sup><https://cline.bot/>

<sup>42</sup><https://roocode.com/>

<sup>43</sup><https://kilocode.ai/>

Figure 19: Prompts for Self-Debugging

(a) System Prompt

```
You are an expert programming assistant. Your
↳ task is to fix issues in a generated
↳ Python solution for a given programming
↳ problem. You are provided with:

- A problem statement
- Starter code
- A previously generated incorrect solution
- A top-level execution trace or error message
- Dependencies information (versions,
↳ libraries).

Please generate a corrected Python solution by
↳ following these strict guidelines:

1. Use the required libraries explicitly in
↳ your code.
2. Correctly incorporate the provided starter
↳ code - do not remove or alter its
↳ structure.
3. Write in standard Python syntax.
4. Wrap your entire solution within a single
↳ Markdown code block.
5. Do not include any text outside the code
↳ block - no explanations, comments,
↳ docstrings, or usage examples.
6. Ensure the code is clean, efficient, and
↳ syntactically valid.
7. Avoid interactive, stateful, or
↳ environment-dependent constructs (e.g.,
↳ Django projects, web servers).
8. Your output must be executable in a
↳ non-interactive environment (e.g., a test
↳ harness or script runner).

Example output format:

```python
# [Your corrected code here]
```

Before submitting, carefully review your code
↳ for correctness, completeness, and
↳ adherence to all constraints.
```

(b) User Prompt

```
<Problem>
{problem}
</Problem>

<Python Version>
{python_version}
</Python Version>

<Library>
{library}
</Library>

<Version>
{version}
</Version>

<Extra Dependencies>
{additional_dependencies}
</Extra Dependencies>

<Starting Code>
{starting_code}
</Starting Code>

<Generated Solution>
{solution}
</Generated Solution>

<Trace>
{top_level_trace}
</Trace>
```

Figure 20: Tool-Calling Agent Prompt

```
You are to solve a coding problem in Python.

# Instructions:

* The coding problem requires using the
↳ library {library}=={version}. Try using
↳ the problem with only this library and the
↳ standard Python libraries.

* Do a thorough research on the web about how
↳ to solve the coding problem for the given
↳ library version. Repeat multiple times if
↳ needed.

* BEFORE FINISHING YOUR WORK, YOU MUST check
↳ your solution to the coding problem by
↳ running the `docker_problem_sandbox` tool.

* Use the `final_answer` tool to return a
↳ self-contained Python script that solves
↳ the problem. DO NOT INCLUDE ANY TEXT
↳ BESIDES FOR THE CODE IN THE FINAL ANSWER.

* The solution needs to be in a markdown code
↳ block.

* The solution needs to start with the starter
↳ code provided below.

# Coding Problem:

{problem}

# Starter Code:

```python
{starting_code}
```
```

Figure 21: RAG Prompt

```
You are an AI assistant specialized in solving
↳ Python programming problems using
↳ information derived from documentation.

Each query may specify particular libraries
↳ and version constraints. Your task is to
↳ generate a correct, efficient, and minimal
↳ Python solution that adheres strictly to
↳ these requirements.

Please follow these rules when crafting your
↳ response:

1. Use only the specified libraries and
↳ respect the given version constraints.
2. Incorporate any provided starter code as
↳ required.
3. Write only Python code- no in- line
↳ comments or usage examples. Do not provide
↳ anything in the response but the code.
4. Ensure the code is clean, minimal, and
↳ adheres to best practices.
5. The code must be executable in a
↳ non-interactive environment (e.g., avoid
↳ frameworks like Django or code requiring a
↳ web server).Context:
{context}

Based on the above, respond to the user query
↳ below.

Query: {query}
```

Here, {context} refers to the context of the top-k retrieved documents from the vectorized database for that query and {query} is the same as the User Prompt given in Figure 18(c).

Figure 22: Prompt and File Format for Coding Assistants

(a) Prompt

```
Solve each sample_{i}.py in this folder then  
↳ subsequently save your solutions as py  
↳ files with the same name in a separate  
↳ subfolder called "{assistant name}" that  
↳ just completes the starting code provided  
↳ in the sample and uses the instructions  
↳ written in the comments at the start of  
↳ each file.
```

(b) Input File Format

```
# Complete using the following libraries  
↳ and/or extra dependencies and their  
↳ versions:  
# problem statement: {problem}  
# library: {library}  
# version: {version}  
# extra_dependencies: {extra_dependencies}  
{starting_code}
```

(a) presents the prompt template we had used for our Coding Assistant experiments. (b) shows the format of the example files referenced in the prompt.

Figure 23: Prompts for SEK (Keyword Generation Stage)

(a) System Prompt

```
You are a seasoned Python developer at a  
↳ Fortune 500 company who excels at  
↳ analyzing complex code. Analyze the given  
↳ code problem from the problem statement  
↳ and starter code provided. Try to extract  
↳ the keywords from the code problem. For  
↳ each identified keyword:  
1. Provide the keyword.  
2. Give a formalized explanation of the  
↳ keyword using technical languages.  
  
Provided Format:  
Keywords:[Keywords]  
Explanations:[Formalized explanations]  
  
Guidelines:  
- Prioritize keywords that are crucial to  
↳ understanding the input parameters, return  
↳ content or supplementary information.  
- Use precise languages in explanations and  
↳ provide formalized definitions where  
↳ appropriate.  
- Ensure explanations are consistent with the  
↳ behaviors expected based on the problem  
↳ description.  
- Limit to the top 1-3 important keywords to  
↳ focus on core concepts.  
- You are supposed to output a structured JSON  
↳ output containing the extracted keywords  
↳ and their corresponding formalized  
↳ explanations in individual lists of  
↳ strings. The keys for this JSON must be  
↳ Keywords and Explanations.  
- Strictly adhere to the provided format, do  
↳ not output anything else.
```

(b) User Prompt

```
<Problem Statement>  
{problem}  
</Problem Statement>  
  
<Starting Code>  
{starting_code}  
</Starting Code>
```

Figure 24: Prompts for SEK (Code Generation Stage)

(a) System Prompt

```
You are a skilled Python programmer tasked
↪ with solving a coding problem. Your goal
↪ is to provide a clear, efficient, and
↪ correct solution that meets all the
↪ specified requirements.

Please provide your solution following these
↪ guidelines:

1. Use the required library in your solution.
2. Incorporate the provided starter code
↪ correctly.
3. Write your solution in Python.
4. Format your solution within a markdown code
↪ block.
5. Ensure your code is clean and efficient.
6. Output only the code block and nothing
↪ else. Do not add any in-line comments,
↪ documentations, references or usage
↪ examples.
7. Make sure your code is executable in a
↪ non-interactive environment. For example,
↪ do not write code which requires building
↪ a Django project or deploying a web-app.

Example output format:

```python
# [Your code here, incorporating the starter
↪ code]
```

After writing your solution, please review it
↪ to ensure all requirements are met and the
↪ code is correct and efficient.

Here are the key elements for this task:
```

(b) User Prompt

```
<Python Version>
{python_version}
</Python Version>

<Library>
{library}
</Library>

<Version>
{version}
</Version>

<Extra Dependencies>
{extra_dependencies}
</Extra Dependencies>

<Problem Statement>
{problem}
</Problem Statement>

<Keywords>
Analyze the following key terms and their
↪ relationships within the problem context:
{General_Keywords}
{Abstract_Keywords}
</Keywords>

<Starting Code>
{starting_code}
</Starting Code>
```