

RESTifAI: LLM-Based Workflow for Reusable REST API Testing

Leon Kogler
Maximilian Ehrhart
leon.kogler@casablanca.at
maximilian.ehrhart@casablanca.at
CASABLANCA hotelsoftware GmbH
Schönwies, Austria

Benedikt Dornauer
benedikt.dornauer@uibk.ac.at
University of Innsbruck
Innsbruck, Austria

Eduard Paul Enoiu
eduard.paul.enoiu@mdh.se
Mälardalen University
Mälardalen, Sweden

Abstract

With this paper, we introduce RESTifAI, an LLM-driven approach for generating reusable, CI/CD-ready REST API tests, following the happy-path approach. Unlike existing tools that often focus primarily on internal server errors, RESTifAI systematically constructs valid test scenarios (happy paths) and derives negative cases to verify both intended functionality (2xx responses) and robustness against invalid inputs or business-rule violations (4xx responses). The results indicate that RESTifAI performs on par with the latest LLM tools, i.e., AutoRestTest and LogiAgent, while addressing limitations related to reusability, oracle complexity, and integration. To support this, we provide common comparative results and demonstrate the tool's applicability in industrial services. For tool demonstration, please refer to <https://www.youtube.com/watch?v=2vtQo0T0Lo4>. RESTifAI is publicly available at <https://github.com/casablancahotelsoftware/RESTifAI>.

CCS Concepts

• **Software and its engineering** → *API languages*; **Acceptance testing**; *Software testing and debugging*.

Keywords

REST API Testing, OpenAPI, Automated Test Generation, Large Language Model

ACM Reference Format:

Leon Kogler, Maximilian Ehrhart, Benedikt Dornauer, and Eduard Paul Enoiu. 2026. RESTifAI: LLM-Based Workflow for Reusable REST API Testing. In *2026 IEEE/ACM 48th International Conference on Software Engineering (ICSE-Companion '26)*, April 12–18, 2026, Rio de Janeiro, Brazil. ACM, New York, NY, USA, 4 pages. <https://doi.org/10.1145/3774748.3787598>

1 Introduction

Migrating from monolithic architectures to microservices, alongside the transition from native software to web-based solutions, has intensified reliance on REST APIs [7]. This trend is further motivated by the growing reliance on external services that expose APIs. Many services now follow the *OpenAPI Specification (OAS)* standard for consistency. Although this standard effectively documents service capabilities and simplifies integration, it does not guarantee correctness. Hence, problems may emerge during integration due to

incorrect API behaviors, or, more critically, during operation when unannounced background modifications, such as updates to code or infrastructure, trigger unforeseen failures and disruptions. Thus, these risks emphasize the importance of systematically and continuously testing REST APIs [2, 6]. In this way, several tools were developed, which were reviewed (till 2023) by Golmohammadi et al. [2], showing that fuzzing is the predominant technique, besides rule-based, property-based, and search-based approaches. More recent tools incorporate Large Language Models (LLMs), demonstrating new capabilities and improved performance. In particular, *AutoRestTest* [9] (or a derivative: *LlamaRestTest* [4] leveraging Llama models), *LogiAgent* [11], and *APITestGenie* [8].

Before initiating the development of our tool, we first examined the limitations of existing approaches. One of the practical limitations is that existing tools are often not designed to generate test suites that can be systematically re-executed. For example, *AutoRestTest* incorporates mutators that create a large number of request variations, making the approach less suitable for repeated testing. To the best of our knowledge, *APITestGenie* is the only other LLM-based tool explicitly targeting reusable, CI/CD-ready test suites. However, since it requires additional business requirements in natural language, we excluded it from our evaluation. Apart from that, *APITestGenie*'s authors [8] note that only 57.3% of the *APITestGenie*-generated cases execute successfully, due to hallucinations, setup issues, and misinterpretation of OAS. Another limitation of tests generated by tools is the lack of support for incorporating API-specific information, such as authentication credentials or API keys. Moreover, tests generated by most current API testing tools are often limited by the capabilities of their test oracles. Tools, such as the often-used *EvoMaster* [1] or *AutoRestTest*, primarily adopt random and mutation-based parameter generation strategies and rely on automated 500 response code oracles [2, 5]. While this strategy appears to be effective in identifying server crashes [5, 9], it might be insufficient for validating realistic scenarios and intended functional behavior. Such tools do not check that valid inputs yield successful (2xx) responses or that invalid inputs are correctly rejected with client-error (4xx) responses.

To address these limitations, we present RESTifAI, a LLM-based workflow whose novelty derives from automatically generating positive tests (happy paths), which confirm correct system behavior under valid inputs, and systematically deriving negative tests from these happy paths, that validate robustness under invalid or unexpected conditions.

2 Approach and Architecture

Unlike fully autonomous agentic systems such as *LogiAgent*, RESTifAI adopts a workflow-based approach, using LLMs selectively



This work is licensed under a Creative Commons Attribution 4.0 International License. *ICSE-Companion '26, Rio de Janeiro, Brazil*
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2296-7/2026/04
<https://doi.org/10.1145/3774748.3787598>

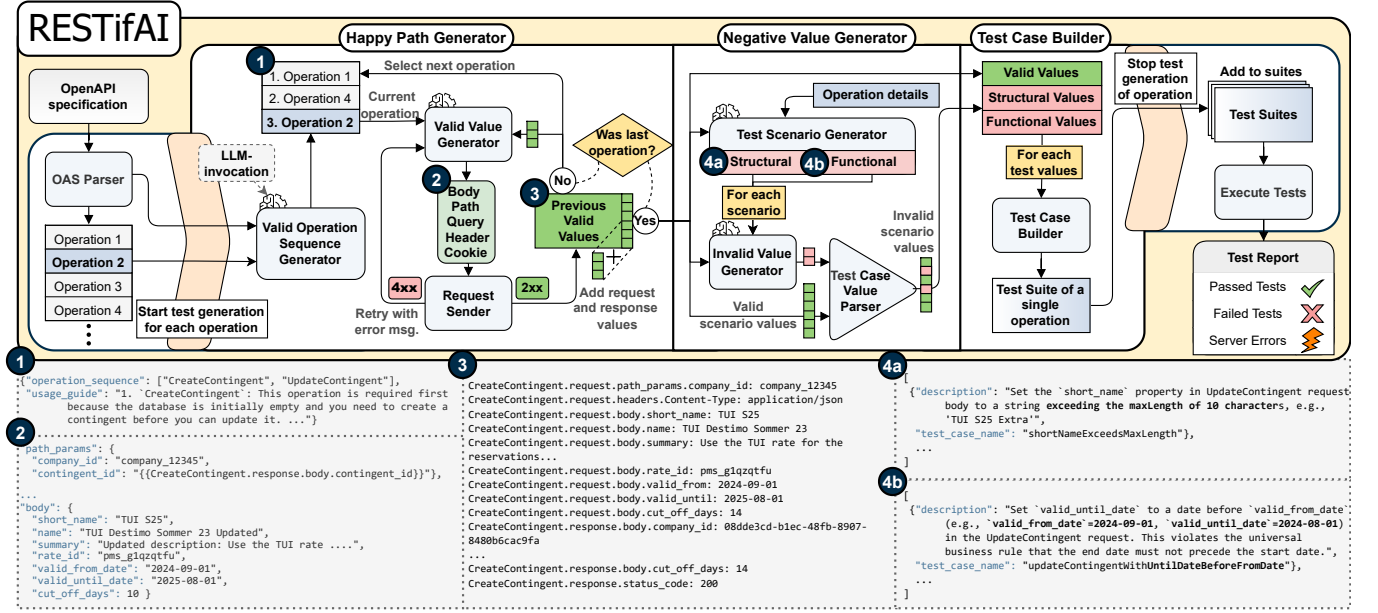


Figure 1: RESTifAI workflow with illustrative outputs from four intermediate steps. ①, ②, ④ are structured LLM outputs.

to ensure more reliable test generation [3, 10]. Overall, it can be seen as exploratory testing, yet also as requirements-based, since it relies on the OAS specification. The main components of RESTifAI include a *Happy Path Generator*, a *Negative Value Generator*, and a *Test Case Builder*, as illustrated in Figure 1. As input, our tool takes the OAS of the *Service Under Test* (SUT) and, as a result, generates a test suite (logically structured test cases) for each operation, executes the tests, and returns a test report including passed tests, failed tests, and highlighted server errors. The steps are detailed further:

Happy Path Generation. The **Happy Path Generator** aims to create a valid scenario that results in a 2xx response for the operation under test, while tracking parameters and values of request and response properties, as well as their dependencies. Therefore, the Happy Path Generator utilizes an LLM to generate a sequence of dependent operations required to produce a valid request for the final operation (the operation under test). The LLM also provides a natural-language usage guide outlining these dependencies and guiding the **Value Generator** on how parameter values should be derived ①. Next, in an iterative process, RESTifAI attempts to generate valid values ② for each operation in the sequence, based on the operation’s specification, the previously generated usage guide, and the execution trace of previously successfully executed dependent operations ③. This execution trace is represented by key-value pairs, where the keys identify the exact location of values within the test scenario. Values may be either **GENERATED**, representing newly created inputs, or **DEPENDENT**, referencing keys of previously stored values. Starting with an initial empty execution trace, the **Value Generator** generates Body, Path, Query, Header, and Cookie parameters for the first operation in the sequence. The **Request Sender** then uses the generated values to

send a request to the service endpoint. If the request yields a 4xx response, the Value Generator retries, incorporating the error message into the regeneration process. Upon receiving a successful 2xx response, the request and response values are parsed into key-value pairs and stored in the execution trace. This procedure repeats for each operation in the sequence until the final operation is executed successfully, at which point RESTifAI proceeds to generate negative values.

Negative Value Generator. Using the final “Happy Path” data and the specification of the operation under test, the **Test Scenario Generator** employs an LLM to produce oracles that define invalid parameter values expected to result in 4xx client error responses. These are split into two categories of oracles: **Structural oracles** ④a test the SUT against invalid inputs derived from constraints in the OAS, such as incorrect data types or formats, while **Functional oracles** ④b, in contrast, verify business logic constraints not captured in the OAS by leveraging the LLM’s domain-specific knowledge to define logical limitations of the SUT. For each generated test scenario, the **Invalid Value Generator** invokes the LLM to generate invalid parameter values using the parameter keys recorded in the valid execution trace from the **Happy Path Generator**. The **Test Case Value Parser** then replaces the valid values in the original scenario with the corresponding invalid values. This ensures that the negative test case triggers a 4xx response due to the substituted values while all other parameters remain valid.

Test Case Builder. Based on the positive scenario values generated by the **Happy Path Generator** and the multiple negative scenario values generated by the **Negative Value Generator**, the **Test Case Builder** generates an executable Postman collection for each of them. By maintaining the scenario values in a standardized

key-value format the testing framework can be chosen independently, requiring modification only to the **Test Case Builder** to adapt to different test frameworks.

To ensure reproducibility and isolation RESTifAI allows the integration of environment initialization scripts in various formats. Such scripts are executed before generating a new Happy Path and before executing each individual test case, thereby maintaining the *SUT* in a controlled and consistent test environment. For instance, an initialization script may reset the database to a predefined state, ensuring test repeatability. Additionally, RESTifAI can be guided with human-provided natural language during Happy Path generation by adding information not specified in the *OAS*, such as authentication credentials or pre-configured database identifiers, to the LLM’s prompt.

RESTifAI includes a command-line interface (CLI) that allows users to customize the test generation process by selecting the type of test cases, specifying environment initialization scripts, or providing user input. For enhanced usability and a more interactive experience, a Python-based graphical frontend is provided, allowing users to efficiently configure test scenarios, execute them, and visually inspect the resulting test reports.

3 Results and Comparison to Current Tools

To assess the quality of RESTifAI in automatic REST-API testing, we conducted a comparison with AutoRestTest and LogiAgent, all leveraging the same Azure OpenAI GPT-4.1-mini LLM¹.

Table 1: Number of Operations Covered (#OC) and the Server Errors (#SE) detected by the tools on the services.

Service	AutoRestTest		LogiAgent		RESTifAI	
	#OC	#SE	#OC	#SE	#OC	#SE
FDIC	6	41	6	0	6	0
Genome Nexus	23	0	19	0	23	0
Languagetool	2	0	1	0	2	1
OhSome	33	0	4	3	128	50
Restcountries	22	83	13	0	20	0

The evaluation was performed on a set of standard benchmarks such as open-source services (Languagetool, Genome Nexus, and Restcountries), which we run locally to enable code-coverage analysis using the JaCoCo library, and online services (OhSome and FDIC), for which source code is inaccessible and code-coverage metrics cannot be obtained. References to the services can be found in our Replication Package. OhSome was explicitly included because it is considered one of the most complex services in the AutoRestTest study, and AutoRestTest was the only tool to generate 12 successful requests for it.

Due to conceptual differences between the tools, we standardized test generation using execution time to ensure a fairer comparison. Our tool stops after a set number of test suites, matching the number of API operations, while tools like LogiAgent and AutoRestTest keep generating test cases indefinitely, potentially leading to longer

runtimes under the same conditions. The execution time of RESTifAI is then applied to AutoRestTest and LogiAgent for each service (see replication package for execution time).

Operation Coverage is a key metric for evaluating automated API testing tools [2, 5, 9] and plays a fundamental role for RESTifAI, as test cases are generated only if a valid request for the operation under test can be produced. While AutoRestTest optimizes exploration of successful 2xx responses, RESTifAI achieved 128 out of 134 successful operations on the OhSome service, compared to 33 operations covered by AutoRestTest, as shown in Table 1. These results highlight the effectiveness of targeted value generation over mutation-based approaches, particularly as the number of operation parameters increases.

The number of detected **Server Errors** is commonly used to compare fuzzing-based testing approaches [2]. While AutoRestTest is primarily optimized for uncovering Server Errors, RESTifAI is designed to generate realistic input data that effectively tests both 2xx success responses and 4xx client errors. As shown for OhSome in Table 1, RESTifAI also performs well in uncovering Server Errors, especially in scenarios requiring the precise generation of real-world edge cases.

Table 2: Comparison of BC (Branch-Coverage), LC (Line-Coverage), and MC (Method-Coverage) for local services.

Service		AutoRestTest	LogiAgent	RESTifAI
BC	Genome Nexus	0%	0%	0%
	Languagetool	16%	7%	19%
	Restcountries	88%	40%	74%
LC	Genome Nexus	65%	60%	65%
	Languagetool	27%	16%	28%
	Restcountries	78%	49%	72%
MC	Genome Nexus	45%	39%	45%
	Languagetool	28%	16%	30%
	Restcountries	85%	68%	83%

Table 3: Number of Test Cases (# TC) generated and the number of Tokens per Test Case (Tokens/TC)

Service	LogiAgent		RESTifAI	
	#TC	Tokens/TC	#TC	Tokens/TC
FDIC	26	37001	133	88762
Genome Nexus	38	62071	326	22258
Languagetool	6	46689	35	3585
OhSome	234	48036	2213	37947
Restcountries	27	62143	191	9296
Average		51188		32370

We match AutoRestTest’s performance in line, branch, and method coverage across various services (see Table 2). In general, it can be seen that higher OC tends to lead to higher **Code Coverage**.

¹The replication package, along with a detailed execution description, is available on <https://github.com/casablancahotelsoftware/RESTifAI>.

RESTifAI creates individual oracles for each new test case similar to LogiAgent, which is not possible with AutoRestTest. The workflow-based approach combined with selective LLM utilization results in a significantly larger *Number of Test Cases*, while using significantly fewer *Tokens per Test Case*, as shown in Table 3. We required more tokens on FDIC because of storing the large payloads returned from the service in our key-value store.

Application to a Real-World Use Case

Beyond our benchmarks on well-established API services, RESTifAI was applied to two microservices from CASABLANCA hotelsoftware: the *Guest-Review Service* and the *Inventory Service*. Real-world services often rely on external systems or require specific parameters like a fixed 'company_id', so automated testing tools must handle these constraints. RESTifAI allows users to provide such parameters or define a custom sequence of operations via the CLI to generate a successful happy path, thereby offering a clear advantage over existing tools.

Table 4: Classification of reusable test cases

	Bug	Enhancement	Invalid	Passed
Guest-Review	2	12	9	51
Inventory	11	6	30	91

In Table 4, we classified the failed generated test cases with domain-experts into Bug, Enhancement and Invalid test cases following the metrics defined in [11]. The high number of invalid test cases in the Inventory Service can be attributed to the fact that certain business rules are not explicitly captured in the API contract, leading to invalid test cases. In the Guest-Review Service, 60.87% of failed test cases revealed real bugs or enhancements.

For our services, we identified, e.g. the following circumstances:

Case 1: Structural Error The test case 'roomtypedWrong-Type_ST' revealed a structural inconsistency where an integer was provided for the 'room_type_id' parameter instead of the expected string.

Case 2: Logical Error In the Inventory Service, our tool identified an issue with date-range constraints: the API allowed the 'until' date to be set before the 'from' date, which violates the expected business logic; see Fig. 1.

Both cases highlight a key advantage of RESTifAI over approaches that rely solely on automated 5xx response code oracles. In each case, the services erroneously returned a 200 response code instead of the appropriate 400 Bad Request response code. Tools such as AutoRestTest, which detect only 5xx response codes, are not able to identify such issues.

4 Discussion and Conclusion

RESTifAI advances LLM-based API testing by coupling happy-path generation with systematic negative derivation. Thereby, the

tool complements fuzzers by prioritizing correctness (2xx/4xx responses), emphasizing OAS-conformant behavior and business-rule validation, whereas fuzzers excel at broad input exploration and crash detection. The results indicate that we compete with the latest REST API test generation tools. Additionally, compared to existing approaches, RESTifAI produces fully automated and re-executable CI/CD-ready tests, while also offering simple integration benefits, as demonstrated in the use case or demo video. A current limitation of RESTifAI is the absence of assertions that validate the functional behavior of the underlying business logic, a recurring shortcoming in existing tools, which we aim to address in future work.

As RESTifAI is designed to generate more complex oracles, there is currently no universal method to automatically measure their correctness. At present, domain experts must verify the generated tests, which introduces observer bias and complicates validity comparisons across tools, especially when *SUT* requirements are rarely defined. One promising direction may be to employ LLMs as judges to mitigate observer bias. Currently, RESTifAI reuses the same operation sequences for both happy-path and negative test generation. While this aims to improve validity, it also restricts the exploration of diverse scenarios. This limitation could be mitigated through more dynamic strategies, such as multi-agent or human-in-the-loop approaches.

Acknowledgments

This work was supported by and done within the scope of the ITEA4 GENIUS project, which was nationally funded by FFG with grant 921454. Thanks for the support from Diffblue, Peter Schrammel, Grammarly and ChatGPT were used to enhance text quality.

References

- [1] Andrea Arcuri. 2018. EvoMaster: Evolutionary Multi-context Automated System Test Generation. In *2018 IEEE 11th International Conference on Software Testing, Verification and Validation (ICST)*. IEEE, Vasteras, 394–397. doi:10.1109/ICST.2018.00046
- [2] Amid Golmohammadi, Man Zhang, and Andrea Arcuri. 2024. Testing RESTful APIs: A Survey. *ACM Transactions on Software Engineering and Methodology* 33, 1 (Jan. 2024), 1–41. doi:10.1145/3617175
- [3] Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2022. Decomposed Prompting: A Modular Approach for Solving Complex Tasks. doi:10.48550/ARXIV.2210.02406
- [4] Myeongsoo Kim, Saurabh Sinha, and Alessandro Orso. 2025. LlamaRestTest: Effective REST API Testing with Small Language Models. *Proceedings of the ACM on Software Engineering* 2, FSE (June 2025), 465–488. doi:10.1145/3715737
- [5] Myeongsoo Kim, Qi Xin, Saurabh Sinha, and Alessandro Orso. 2022. Automated Test Generation for REST APIs: No Time to Rest Yet. In *Proceedings of the 31st ACM SIGSOFT International Symposium on Software Testing and Analysis*. ACM, Virtual South Korea, 289–301. doi:10.1145/3533767.3534401
- [6] Sebastian Kotstein and Justus Bogner. 2021. Which RESTful API Design Rules Are Important and How Do They Improve Software Quality? A Delphi Study with Industry Experts. In *Service-Oriented Computing*, Johanna Barzen (Ed.). Vol. 1429. Springer International Publishing, Cham, 154–173. doi:10.1007/978-3-030-87568-8_10
- [7] Maxime Lamothe, Yann-Gaël Guéhéneuc, and Weiyi Shang. 2022. A Systematic Review of API Evolution Literature. *Comput. Surveys* 54, 8 (Nov. 2022), 1–36. doi:10.1145/3470133
- [8] André Pereira, Bruno Lima, and João Pascoal Faria. 2024. APITestGenie: Automated API Test Generation through Generative AI. doi:10.48550/arXiv.2409.03838 arXiv:2409.03838 [cs]
- [9] Tyler Stennett, Myeongsoo Kim, Saurabh Sinha, and Alessandro Orso. 2025. AutoRestTest: A Tool for Automated REST API Testing Using LLMs and MARL. doi:10.48550/ARXIV.2501.08600
- [10] Barry Zhang and Erik Schluntz. 2024. How We Build Effective Agents.
- [11] Ke Zhang, Chenxi Zhang, Chong Wang, Chi Zhang, YaChen Wu, Zhenchang Xing, Yang Liu, Qingshan Li, and Xin Peng. 2025. LogiAgent: Automated Logical Testing for REST Systems with LLM-Based Multi-Agents. doi:10.48550/ARXIV.2503.15079