

1

OverHAuL

2

Harnessing Automation for C Libraries with Large Language Models

3

Konstantinos Chousos

4

July, 2025

5 Lorem ipsum odor amet, consectetur adipiscing elit. Habitasse congue tempus erat rhoncus
6 sapien interdum dolor nec. Posuere habitant metus tellus erat eu. Risus ultricies eu rhoncus,
7 conubia euismod convallis commodo per. Nam tellus quisque maximus dui eleifend; arcu aptent.
8 Nisi rutrum primis luctus tortor tempor maecenas. Donec curae cras dolor; malesuada ultricies
9 scelerisque. Molestie class tincidunt quis gravida ut proin. Consequat lacinia arcu justo leo maecenas
10 nunc neque ex. Platea eros ullamcorper nullam rutrum facilisis.

11 Preface

12 This thesis was written in Athens, Greece during the academic year 2024–2025. It is part of
13 the requirements for the Bachelor of Science program in the [Department of Informatics and](#)
14 [Telecommunications](#) at the [National and Kapodistrian University of Athens](#). The research was
15 conducted under the supervision of Prof. [Thanassis Avgerinos](#). The completed thesis is also available
16 through the university’s digital repository [Pergamos](#).

¹⁷ *To my beloved parents who, through their example, taught me patience, resilience and perseverance.*

18 Acknowledgments

19 I would like to express my gratitude to my supervisor, Prof. Thanassis Avgerinos, for his insightful
20 guidance, patience, and unwavering encouragement throughout this journey. His openness and
21 our shared passion for the subject greatly enhanced my enjoyment of the thesis process.

22 I am also thankful to my fellow group members in Prof. Avgerinos' weekly meetings, whose
23 willingness to exchange ideas and offer support was invaluable. My appreciation extends to Jorgen
24 and Phaedon, friends who provided thoughtful input and advice along the way.

25 A special *thank you* goes to my parents, Christina, and my friends for their constant support and
26 understanding. Their patience and encouragement helped me persevere through this challenging
27 period.

Table of contents

29	1 Introduction	1
30	2 Background	3
31	2.1 Fuzz Testing	3
32	2.1.1 Motivation	4
33	2.1.2 Methodology	5
34	2.1.3 Challenges in Adoption	7
35	2.2 Large Language Models	7
36	2.2.1 Biggest GPTs	8
37	2.2.2 Prompting	8
38	2.2.3 LLMs for Coding	9
39	2.2.4 LLMs for Fuzzing	10
40	2.3 Neurosymbolic AI	10
41	3 Related work	12
42	3.1 Previous Projects	12
43	3.1.1 KLEE	12
44	3.1.2 IRIS	12
45	3.1.3 FUDGE	13
46	3.1.4 UTopia	13
47	3.1.5 FuzzGen	13
48	3.1.6 IntelliGen	14
49	3.1.7 CKGFuzzer	14
50	3.1.8 PromptFuzz	15
51	3.1.9 OSS-Fuzz	16
52	3.1.10 OSS-Fuzz-Gen	16
53	3.1.11 AutoGen	17
54	3.2 Differences	17
55	4 OverHAuL's Design	19
56	4.1 Architecture	20
57	4.1.1 Project Analysis	20
58	4.1.2 Harness Creation	21
59	4.1.3 Harness Evaluation	21
60	4.2 Main Techniques	21
61	4.2.1 Feedback Loop	22
62	4.2.2 React Agents Swarm	22
63	4.2.3 Codebase Oracle	22

64	4.3	High-Level Algorithm	23
65	4.4	Installation and Usage	24
66	4.5	Scope	26
67	5	Evaluation	27
68	5.1	Experimental Benchmark	27
69	5.1.1	Local Benchmarking	28
70	6	Results	29
71	6.1	Answers to RQs	29
72	7	Implementation	32
73	7.1	Development Tools	32
74	7.2	Reproducibility	33
75	8	Discussion	34
76	8.1	Threats to Validity	34
77	9	Future Work	35
78	9.1	Enhancements to Core Features	35
79	9.2	Experimentation with Large Language Models and Data Representation	36
80	9.3	Comprehensive Evaluation and Benchmarking	36
81	9.4	Practical Deployment and Community Engagement	36
82	10	Conclusion	38
83		Bibliography	39
84		Appendices	46
85	A	Abandoned Techniques	46
86	B	DSPy Custom Signatures	48

87 List of Figures

88	4.1 OverHAuL Workflow	19
89	4.2 OverHAuL execution example	26
90	6.1 Benchmark Results	29

List of Listings

92	2.1	Fuzzing harness format	6
93	2.2	Example fuzzing harness	6
94	2.3	Compilation of harness	7
95	4.1	OverHAuL installation	25
96	6.1	Sample dateparse harness	31
97	7.1	DSPy example	32

List of Tables

98		
99	5.1	The benchmark project corpus. Each project name links to its corresponding GitHub
100		repository. Each is followed by a short description and its GitHub stars count, as of
101		July 18th, 2025. 27

1 Introduction

Memory-safe languages, like Ada or Rust [1], [2]—not absolutely safe themselves

NASA safety-critical code guidelines [3]

C by Dennis Ritchie in 1972 [4], [5]

- Motivation

- Memory unsafety is and will be prevalent
- Software is safe until it's not
- Humans make mistakes

- LLMs have become very powerful

- They are used more and more
- Humans now use Large Language Models (LLMs) to write software
- LLMs make mistakes [6]

- Result: Bugs exist

- Fuzzing is a popular software testing strategy

- LLM applications and capabilities are still explored
 - * One of which is automatic harnessing

- Goal

- A system that:
 1. Takes a bare C project as input
 2. Generates a new fuzzing harness from scratch using LLMs
 3. Compiles it
 4. Executes it and evaluates it

- Thesis Structure

- We examine the state of automatic fuzzing projects
 - * Most automatic fuzzers rely on existings harnesses [7] or preexisting client code [8]–[10]
 - * No evaluation step, it is left to the developer/user
- To address these challenges, we introduce OverHAuL
 - * We explain its design architecture, the techniques that emerged in its development and how they are implemented
 - * We construct a dataset of 10 open-source projects and evaluate OverHAuL with them

- 134 * We examine the results and try to answer our research questions
- 135 • Summary of Contributions
- 136 – We propose OverHAuL, a new automatic fuzzing framework
- 137 – Experimental benchmarks show a 92.5% chance of successful harness generation
- 138 – Code and artifacts are available in the project's repo [https://github.com/kchousos/](https://github.com/kchousos/OverHAuL)
- 139 OverHAuL

2 Background

2.1 Fuzz Testing

Fuzzing is an automated software-testing technique in which a *Program Under Test* (PUT) is executed with (pseudo-)random inputs in the hope of exposing undefined behavior. When such behavior manifests as a crash, hang, or memory-safety violation, the corresponding input constitutes a *test-case* that reveals a bug and often a vulnerability [11]. In essence, fuzzing is a form of adversarial, penetration-style testing carried out by the defender before the adversary has an opportunity to do so. Interest in the technique surged after the publication of three practitioner-oriented books in 2007–2008 [12]–[14].

Historically, the term was coined by Miller et al. in 1990, who used “fuzz” to describe a program that “generates a stream of random characters to be consumed by a target program” [15]. This informal usage captured the essence of what fuzzing aims to do: stress test software by bombarding it with unexpected inputs to reveal bugs. To formalize this concept, we adopt Manes et al.’s rigorous definitions [11]:

Definition 2.1 (Fuzzing). Fuzzing is the execution of a Program Under Test (PUT) using input(s) sampled from an input space (the *fuzz input space*) that protrudes the expected input space of the PUT.

This means fuzzing involves running the target program on inputs that go beyond those it is typically designed to handle, aiming to uncover hidden issues. An individual instance of such execution—or a bounded sequence thereof—is called a *fuzzing run*. When these runs are conducted systematically and at scale with the specific goal of detecting violations of a security policy, the activity is known as *fuzz testing* (or simply *fuzzing*):

Definition 2.2 (Fuzz Testing). Fuzz testing is the use of fuzzing to test whether a PUT violates a security policy.

This distinction highlights that fuzz testing is fuzzing with an explicit focus on security properties and policy enforcement. Central to managing this process is the *fuzzer engine*, which orchestrates the execution of one or more fuzzing runs as part of a *fuzz campaign*. A fuzz campaign represents a concrete instance of fuzz testing tailored to a particular program and security policy:

Definition 2.3 (Fuzzer, Fuzzer Engine). A fuzzer is a program that performs fuzz testing on a PUT.

Definition 2.4 (Fuzz Campaign). A fuzz campaign is a specific execution of a fuzzer on a PUT with a specific security policy.

Throughout each execution within a campaign, a *bug oracle* plays a critical role in evaluating the program’s behavior to determine whether it violates the defined security policy:

Definition 2.5 (Bug Oracle). A bug oracle is a component (often inside the fuzzer) that determines whether a given execution of the PUT violates a specific security policy.

In practice, bug oracles often rely on runtime instrumentation techniques, such as monitoring for fatal POSIX signals (e.g., SIGSEGV) or using sanitizers like AddressSanitizer (ASan) [16]. Tools like LibFuzzer [17] commonly incorporate such instrumentation to reliably identify crashes or memory errors during fuzzing.

Most fuzz campaigns begin with a set of *seeds*—inputs that are well-formed and belong to the PUT’s expected input space—called a *seed corpus*. These seeds serve as starting points from which the fuzzer generates new test cases by applying transformations or mutations, thereby exploring a broader input space:

Definition 2.6 (Seed). An input given to the PUT that is mutated by the fuzzer to produce new test cases. During a fuzz campaign (Definition 2.4) all seeds are stored in a seed *pool* or *corpus*.

The process of selecting an effective initial corpus is crucial because it directly impacts how quickly and thoroughly the fuzzer can cover the target program’s code. This challenge—studied as the *seed-selection problem*—involves identifying seeds that enable rapid discovery of diverse execution paths and is non-trivial [18]. A well-chosen seed set often accelerates bug discovery and improves overall fuzzing efficiency.

2.1.1 Motivation

The purpose of fuzzing relies on the assumption that there are bugs within every program, which are waiting to be discovered. Therefore, a systematic approach should find them sooner or later.

— OWASP Foundation [19]

Fuzz testing offers several tangible benefits:

1. **Early vulnerability discovery:** Detecting defects during development is cheaper and safer than addressing exploits in production.
2. **Adversary-parity:** Performing the same randomised stress that attackers employ allows defenders to pre-empt zero-days.
3. **Robustness and correctness:** Beyond security, fuzzing exposes logic errors and stability issues in complex, high-throughput APIs (e.g., decompressors) even when inputs are *expected* to be well-formed.
4. **Regression prevention:** Re-running a corpus of crashing inputs as part of continuous integration ensures that fixed bugs remain fixed.

2.1.1.1 Success Stories

Heartbleed (CVE-2014-0160) [20], [21] arose from a buffer over-read¹ in OpenSSL [22] introduced on 1 February 2012 and unnoticed until 1 April 2014. Post-mortem analyses showed that a simple fuzz campaign exercising the TLS heartbeat extension would have revealed the defect almost immediately [23].

Likewise, the *Shellshock* (or *Bashdoor*) family of bugs in GNU Bash [24] enabled arbitrary command execution on many UNIX systems. While the initial flaw was fixed promptly, subsequent bug variants were discovered by Google’s Michał Zalewski using his own fuzzer [25] in late 2014 [26].

On the defensive tooling side, the security tool named *Mayhem*—developed by the company of the same name—has since been adopted by the US Air Force, the Pentagon, Cloudflare, and numerous open-source communities. It has found and facilitated the remediation of thousands of previously unknown vulnerabilities [27].

These cases underscore the central thesis of fuzz testing: exhaustive manual review is infeasible, but scalable stochastic exploration reliably surfaces the critical few defects that matter most.

2.1.2 Methodology

As previously discussed, fuzz testing of a program under test (PUT) is typically conducted using a dedicated fuzzing engine (see Definition 2.3). Among the most widely adopted fuzzers for C and C++ projects and libraries are AFL [25]—which has since evolved into AFL++ [28]—and LibFuzzer [17]. Within the OverHAuL framework, LibFuzzer is preferred owing to its superior suitability for library fuzzing, whereas AFL++ predominantly targets executables and binary fuzzing.

2.1.2.1 LibFuzzer

LibFuzzer [17] is an in-process, coverage-guided evolutionary fuzzing engine primarily designed for testing libraries. It forms part of the LLVM ecosystem [29] and operates by linking directly with the library under evaluation. The fuzzer delivers mutated input data to the library through a designated fuzzing entry point, commonly referred to as the *fuzz target*.

Definition 2.7 (Fuzz target). A function that accepts a byte array as input and exercises the application programming interface (API) under test using these inputs [17]. This construct is also known as a *fuzz driver*, *fuzzer entry point*, or *fuzzing harness*.

For the remainder of this thesis, the terms presented in Definition 2.7 will be used interchangeably.

To effectively validate an implementation or library, developers are required to author a fuzzing harness that invokes the target library’s API functions utilizing the fuzz-generated inputs. This harness serves as the principal interface for the fuzzer and is executed iteratively, each time with mutated input designed to maximize code coverage and uncover defects. To comply with LibFuzzer’s interface requirements, a harness must conform to the following function signature:

¹<https://xkcd.com/1354/>

Listing 2.1 This function receives the fuzzing input via a pointer to an array of bytes (*Data*) and its associated size (*Size*). Efficiency in fuzzing is achieved by invoking the API of interest within the body of this function, thereby allowing the fuzzer to explore a broad spectrum of behavior through systematic input mutation.

```
1  int LLVMFuzzerTestOneInput(const uint8_t *Data, size_t Size) {  
2      DoSomethingInterestingWithData(Data, Size);  
3      return 0;  
4  }
```

239 A more illustrative example of such a harness is provided in Listing 2.2.

Listing 2.2 This example demonstrates a minimal harness that triggers a controlled crash upon receiving `HI!` as input.

```
1  // test_fuzzer.cpp  
2  #include <stdint.h>  
3  #include <stddef.h>  
4  
5  extern "C" int LLVMFuzzerTestOneInput(const uint8_t *data, size_t size) {  
6      if (size > 0 && data[0] == 'H')  
7          if (size > 1 && data[1] == 'I')  
8              if (size > 2 && data[2] == '!')  
9                  __builtin_trap();  
10     return 0;  
11 }
```

240 To compile and link such a harness with LibFuzzer, the Clang compiler—also part of the LLVM
241 project [29]—must be used alongside appropriate compiler flags. For instance, compiling the harness
242 in Listing 2.2 can be achieved as shown in Listing 2.3.

243 2.1.2.2 AFL and AFL++

244 *American Fuzzy Lop* (AFL) [25], developed by Michał Zalewski, is a seminal fuzzer targeting C and
245 C++ applications. Its core methodology relies on instrumented binaries to provide edge coverage
246 feedback, thereby guiding input mutation towards unexplored program paths. AFL supports several
247 emulation backends including QEMU [30]—an open-source CPU emulator facilitating fuzzing on
248 diverse architectures—and Unicorn [31], a lightweight multi-platform CPU emulator. While AFL
249 established itself as a foundational tool within the fuzzing community, its successor AFL++ [28]
250 incorporates numerous enhancements and additional features to improve fuzzing efficacy.

251 AFL operates by ingesting seed inputs from a specified directory (`seeds_dir`), applying mutations,
252 and then executing the target binary to discover novel execution paths. Execution can be initiated
253 using the following command-line syntax:

Listing 2.3 This example illustrates the compilation and execution workflow necessary for deploying a LibFuzzer-based fuzzing harness.

```
1 # Compile test_fuzzer.cc with AddressSanitizer and link against LibFuzzer.
2 clang++ -fsanitize=address,fuzzer test_fuzzer.cc
3 # Execute the fuzzer without any pre-existing seed corpus.
4 ./a.out
```

```
1 ./afl-fuzz -i seeds_dir -o output_dir -- /path/to/tested/program
```

254 AFL is capable of fuzzing both black-box and instrumented binaries, employing a fork-server
255 mechanism to optimize performance. It additionally supports persistent mode execution as well
256 as modes leveraging QEMU and Unicorn emulators, thereby providing extensive flexibility for
257 different testing environments.

258 Although AFL is traditionally utilized for fuzzing standalone programs or binaries, it is also capable
259 of fuzzing libraries and other software components. In such scenarios, rather than implementing
260 the LLVMFuzzerTestOneInput style harness, AFL can use the standard main() function as the fuzzing
261 entry point. Nonetheless, AFL also accommodates integration with LLVMFuzzerTestOneInput-based
262 harnesses, underscoring its adaptability across varied fuzzing use cases.

263 2.1.3 Challenges in Adoption

264 Despite its potential for uncovering software vulnerabilities, fuzzing remains a relatively underuti-
265 lized testing technique compared to more established methodologies such as Test-Driven Develop-
266 ment (TDD). This limited adoption can be attributed, in part, to the substantial initial investment
267 required to design and implement appropriate test harnesses that enable effective fuzzing processes.
268 Furthermore, the interpretation of fuzzing outcomes—particularly the identification, diagnostic
269 analysis, and prioritization of program crashes—demands considerable resources and specialized
270 expertise. These factors collectively pose significant barriers to the widespread integration of
271 fuzzing within standard software development and testing practices.

272 2.2 Large Language Models

273 Natural Language Processing (NLP), a subfield of Artificial Intelligence (AI), has a rich and ongoing
274 history that has evolved significantly since its beginning in the 1990s [32], [33]. Among the most
275 notable—and recent—advancements in this domain are Large Language Models (LLMs), which have
276 transformed the landscape of NLP and AI in general.

277 At the core of many LLMs is the attention mechanism, which was introduced by Bahdanau et
278 al. in 2014 [34]. This pivotal innovation enabled models to focus on relevant parts of the input
279 sequence when making predictions, significantly improving language understanding and generation

tasks. Building on this foundation, the Transformer architecture was proposed by Vaswani et al. in 2017 [35]. This architecture has become the backbone of most contemporary LLMs, as it efficiently processes sequences of data, capturing long-range dependencies without being hindered by sequential processing limitations.

One of the first major breakthroughs utilizing the Transformer architecture was BERT (Bidirectional Encoder Representations from Transformers), developed by Devlin et al. in 2019 [36]. BERT’s bi-directional understanding allowed it to capture the context of words from both directions, which improved the accuracy of various NLP tasks. Following this, the Generative Pre-trained Transformer (GPT) series, initiated by OpenAI with the original GPT model in 2018 [37], further pushed the boundaries. Subsequent iterations, including GPT-2 [38], GPT-3 [39], and the most current GPT-4 [40], have continued to enhance performance by scaling model size, data, and training techniques.

In addition to OpenAI’s contributions, other significant models have emerged, such as Claude, DeepSeek-R1 and the Llama series (1 through 3) [41]–[43]. The proliferation of LLMs has sparked an active discourse about their capabilities, applications, and implications in various fields.

2.2.1 Biggest GPTs

User-facing LLMs are generally categorized between closed-source and open-source models. Closed-source LLMs like ChatGPT, Claude, and Gemini [41], [44], [45] represent commercially developed systems often optimized for specific tasks without public access to their underlying weights. In contrast, open-source models², including the Llama series [43] and Deepseek [42], provide researchers and practitioners with access to model weights, allowing for greater transparency and adaptability.

2.2.2 Prompting

Interaction with LLMs typically occurs through chat-like interfaces, a process commonly referred to as *prompting*. A critical aspect of effective engagement with LLMs is the usage of different prompting strategies, which can significantly influence the quality and relevance of the generated outputs. Various approaches to prompting have been developed and studied, including zero-shot and few-shot prompting. In zero-shot prompting, the model is expected to perform a specific task without any examples, while in few-shot prompting, the user provides a limited number of examples to guide the model’s responses [39].

To enhance performance on more complex tasks, several advanced prompting techniques have emerged. One notable strategy is the *Chain of Thought* approach [46], which entails presenting the model with sample thought processes for solving a given task. This method encourages the model to generate more coherent and logical reasoning by mimicking human-like cognitive pathways. A refined variant of this approach is the *Tree of Thoughts* technique [47], which enables the LLM

²The term “open-source” models is somewhat misleading, since these are better termed as *open-weights* models. While their weights are publicly available, their training data and underlying code are often proprietary. This terminology reflects community usage but fails to capture the limitations of transparency and accessibility inherent in these models.

to explore multiple lines of reasoning concurrently, thereby facilitating the selection of the most promising train of thought for further exploration.

In addition to these cognitive strategies, Retrieval-Augmented Generation (RAG) [48] is another innovative technique that enhances the model’s capacity to provide accurate information by incorporating external knowledge not present in its training dataset. RAG operates by integrating the LLM with an external storage system—often a vector store containing relevant documents—that the model can query in real-time. This allows the LLM to pull up pertinent and/or proprietary information in response to user queries, resulting in more comprehensive and accurate answers.

Moreover, the ReAct framework [49], which stands for Reasoning and Acting, empowers LLMs by granting access to external tools. This capability allows LLM instances to function as intelligent agents that can interact meaningfully with their environment through user-defined tools. For instance, a ReAct tool could be a function that returns a weather forecast based on the user’s current location. In this scenario, the LLM can provide accurate and truthful predictions, thereby mitigating risks associated with hallucinated responses.

2.2.3 LLMs for Coding

The impact of LLMs in software development in recent years is apparent, with hundreds of LLM-assistance extensions and Integrated Development Environments (IDEs) being published. Notable instances include tools like GitHub Copilot and IDEs such as Cursor, which leverage LLM capabilities to provide developers with coding suggestions, auto-completions, and even real-time debugging assistance [50], [51]. Such innovations have introduced a layer of interaction that enhances productivity and fosters a more intuitive coding experience. Simultaneously, certain LLMs are trained themselves with the code-generation task in mind [52]–[54].

One exemplary product of this innovation is *vibecoding* and the no-code movement, which describe the development of software by only prompting and tasking an LLM, i.e. without any actual programming required by the user. This constitutes a showcase of how LLMs can be harnessed to elevate the coding experience by supporting developers as they navigate complex programming tasks [55]. By analyzing the context of the code being written, these sophisticated models can provide contextualized insights and relevant snippets, effectively streamlining the development process. Developers can benefit from reduced cognitive load, as they receive suggestions that not only cater to immediate coding needs but also promote adherence to best practices and coding standards.

Despite these advancements, it is crucial to recognize the inherent limitations of LLMs when applied to software development. While they can help in many aspects of coding, they are not immune to generating erroneous outputs—a phenomenon often referred to as “hallucination”. Hallucinations occur when LLMs produce information that is unfounded or inaccurate, which can stem from several factors, including the limitations of their training data and the constrained context window within which they operate. As LLMs generate code suggestions based on the patterns learned from vast datasets, they may inadvertently propose solutions that do not align with the specific requirements of a task or that utilize outdated programming paradigms.

Moreover, the challenge of limited context windows can lead to suboptimal suggestions. LLMs generally process a fixed amount of text when generating responses, which can impact their ability to fully grasp the nuances of complex coding scenarios. This may result in outputs that lack the necessary depth and specificity required for successful implementation. As a consequence, developers must exercise caution and critically evaluate the suggestions offered by these models, as reliance on them without due diligence could lead to the introduction of bugs or other issues in the code.

2.2.4 LLMs for Fuzzing

While large language models (LLMs) demonstrate significant potential in enhancing the software development process, the challenges highlighted in Section 2.2.3 become even more pronounced and troublesome when these models are employed to generate fuzzing harnesses. The task of writing a fuzzing harness inherently demands an in-depth comprehension of both the library being tested and the intricate interactions expected among its various components. This level of understanding is often beyond the capabilities of LLMs, primarily due to their context window limitations, which restrict the amount of information they can effectively process and retain during code generation.

In addition to this issue, the risk of error-prone code produced by LLMs further complicates the fuzzing workflow. When a crash occurs during the fuzzing process, it becomes imperative for developers to ascertain that the root cause of the failure is not attributable to deficiencies or bugs within the harness itself. This additional layer of verification adds to the cognitive load placed upon developers, potentially detracting from their ability to focus on testing and improving the underlying software.

To enhance the reliability of LLM-generated harnesses in fuzzing contexts, it is essential that these generated artifacts undergo thorough evaluation and validation through programmatic means. This involves the implementation of systematic techniques that assess the accuracy and robustness of the generated code, ensuring that it aligns with the expected behavior of the components it is intended to interact with. This strategy can be conceptualized within the framework of Neurosymbolic AI (Section 2.3), which seeks to integrate the strengths of neural networks with symbolic reasoning capabilities. By marrying these two paradigms, it may be possible to improve the reliability and efficacy of LLMs in the creation of fuzzing harnesses, ultimately leading to a more seamless integration of automated testing methodologies into the software development lifecycle.

2.3 Neurosymbolic AI

Neurosymbolic AI (NeSy AI) represents a groundbreaking fusion of neural network methodologies with symbolic execution techniques and tools, providing a multi-faceted approach to overcoming the inherent limitations of traditional AI paradigms [56], [57]. This innovative synthesis seeks to combine the strengths of both neural networks, which excel in pattern recognition and data-driven learning, and symbolic systems, which offer structured reasoning and interpretability. By integrating these two approaches, NeSy AI aims to create cognitive models that are not only more accurate but also more robust in problem-solving contexts.

391 At its core, NeSy AI facilitates the development of AI systems that are capable of understanding
392 and interpreting feedback in real-world scenarios [58]. This characteristic is particularly significant
393 in the current landscape of artificial intelligence, where LLMs are predominant. In this context,
394 NeSy AI is increasingly viewed as a critical solution to pressing issues related to explainability,
395 attribution, and reliability in AI systems [59], [60]. These challenges are essential for ensuring
396 that AI systems can be trusted and effectively utilized in various applications, from business to
397 healthcare.

398 The burgeoning field of neurosymbolic AI is still in its nascent stages, with ongoing research and
399 development actively exploring its potential to enhance attribution methodologies within large
400 language models. By addressing these critical challenges, NeSy AI can significantly contribute to
401 the broader landscape of trustworthy AI systems, allowing for more transparent and accountable
402 decision-making processes [56], [59], [60].

403 Moreover, the application of neurosymbolic AI within the domain of fuzzing is gaining traction,
404 paving the way for innovative explorations. This integration of LLMs with symbolic systems opens
405 up new avenues for research. Currently, there are only a limited number of tools that support such
406 hybrid approaches (Chapter 3). Among these, OverHAuL constitutes a Neuro[Symbolic] tool, as
407 classified by Henry Kautz’s taxonomy [61], [62]. This means that the neural model—specifically the
408 LLM—can leverage symbolic reasoning tools—in this case a source code explorer (Chapter 7)—to
409 augment its reasoning capabilities. This symbiotic relationship enhances the overall efficacy and
410 versatility of LLMs for fuzzing harnesses generation, demonstrating the profound potential held by
411 the fusion of neural and symbolic methodologies.

3 Related work

Automated testing, automated fuzzing and automated harness creation have a long research history. Still, a lot of ground remains to be covered until true automation of these tasks is achieved. Until the introduction of transformers [35] and the 2020’s boom of commercial GPTs [44], automation regarding testing and fuzzing was mainly attempted through static and dynamic program analysis methods. These approaches are still utilized, but the fuzzing community has shifted almost entirely to researching the incorporation and employment of LLMs in the last half decade, in the name of automation [7]–[10], [63]–[68].

3.1 Previous Projects

3.1.1 KLEE

KLEE [69] is a seminal and widely cited symbolic execution engine introduced in 2008 by Cadar et al. It was designed to automatically generate high-coverage test cases for programs written in C, using symbolic execution to systematically explore the control flow of a program. KLEE operates on the LLVM [29] bytecode representation of programs, allowing it to be applied to a wide range of C programs compiled to the LLVM intermediate representation.

Instead of executing a program on concrete inputs, KLEE performs symbolic execution—that is, it runs the program on symbolic inputs, which represent all possible values simultaneously. At each conditional branch, KLEE explores both paths by forking the execution and accumulating path constraints (i.e., logical conditions on input variables) along each path. This enables it to traverse many feasible execution paths in the program, including corner cases that may be difficult to reach through random testing or manual test creation.

When an execution path reaches a terminal state (e.g., a program exit, an assertion failure, or a segmentation fault), KLEE uses a constraint solver to compute concrete input values that satisfy the accumulated constraints for that path. These values form a test case that will deterministically drive the program down that specific path when executed concretely.

3.1.2 IRIS

IRIS [63] is a 2025 open-source neurosymbolic system for static vulnerability analysis. Given a codebase and a list of user-specified Common Weakness Enumerations (CWEs), it analyzes source code to identify paths that may correspond to known vulnerability classes. IRIS combines symbolic analysis—such as control- and data-flow reasoning—with neural models trained to generalize over

code patterns. It outputs candidate vulnerable paths along with explanations and CWE references. The system operates on full repositories and supports extensible CWE definitions.

3.1.3 FUDGE

FUDGE [10] is a closed-source tool, made by Google, for automatic harness generation of C and C++ projects based on existing client code. It was used in conjunction with and in the improvement of Google’s OSS-Fuzz [70]. Being deployed inside Google’s infrastructure, FUDGE continuously examines Google’s internal code repository, searching for code that uses external libraries in a meaningful and “fuzzable” way (i.e. predominantly for parsing). If found, such code is *sliced* [71], per FUDGE, based on its Abstract Syntax Tree (AST) using LLVM’s Clang tool [29]. The above process results in a set of abstracted mostly-self-contained code snippets that make use of a library’s calls and/or API. These snippets are later *synthesized* into the body of a fuzz driver, with variables being replaced and the fuzz input being utilized. Each is then injected in an LLVMFuzzerTestOneInput function and finalized as a fuzzing harness. A building and evaluation phase follows for each harness, where they are executed and examined. Every passing harness along with its evaluation results is stored in FUDGE’s database, reachable to the user through a custom web-based UI.

3.1.4 UTopia

UTopia [8] (stylized UTOPIA) is another open-source automatic harness generation framework. Aside from the library code, It operates solely on user-provided unit tests since, according to Jeong, Jang, Yi, *et al.* [8], they are a resource of complete and correct API usage examples containing working library set-ups and tear-downs. Additionally, each of them are already close to a fuzz target, in the sense that they already examine a single and self-contained API usage pattern. Each generated harness follows the same data flow of the originating unit test. Static analysis is employed to figure out what fuzz input placement would yield the most results. It is also utilized in abstracting the tests away from the syntactical differences between testing frameworks, along with slicing and AST traversing using Clang.

3.1.5 FuzzGen

Another project of Google is FuzzGen [9], this time open-source. Like FUDGE, it leverages existing client code of the target library to create fuzz targets for it. FuzzGen uses whole-system analysis, through which it creates an *Abstract API Dependence Graph* (A^2DG). It uses the latter to automatically generate LibFuzzer-compatible harnesses. For FuzzGen to work, the user needs to provide both client code and/or tests for the API and the API library’s source code as well. FuzzGen uses the client code to infer the *correct usage* of the API and not its general structure, in contrast to FUDGE. FuzzGen’s workflow can be divided into three phases: 1. *API usage inference*. By consuming and analyzing client code and tests that concern the library under test, FuzzGen recognizes which functions belong to the library and learns its correct API usage patterns. This process is done with the help of Clang. To test if a function is actually a part of the library, a sample program is created that uses it. If the program compiles successfully, then the function is indeed a valid API call. 2. *A^2DG construction mechanism*. For all the existing API calls, FuzzGen builds an A^2DG to record

the API usages and infers its intended structure. After completion, this directed graph contains all the valid API call sequences found in the client code corpus. It is built in a two-step process: First, many smaller A²DGs are created, one for each root function per client code snippet. Once such graphs have been created for all the available client code instances, they are combined to formulate the master A²DG. This graph can be seen as a template for correct usage of the library. 3. *Fuzzer generator*. Through the A²DG, a fuzzing harness is created. Contrary to FUDGE, FuzzGen does not create multiple “simple” harnesses but a single complex one with the goal of covering the whole of the A²DG. In other words, while FUDGE fuzzes a single API call at a time, FuzzGen’s result is a single harness that tries to fuzz the given library all at once through complex API usage.

3.1.6 IntelliGen

IntelliGen [72] is a system for automatically synthesizing fuzz drivers by statically identifying potentially vulnerable entry-point functions within C projects. Implemented using LLVM [29], IntelliGen focuses on improving fuzzing efficiency by targeting code more likely to contain memory safety issues, rather than exhaustively fuzzing all available functions.

The system comprises two main components: the *Entry Function Locator* and the *Fuzz Driver Synthesizer*. The Entry Function Locator analyzes the project’s abstract syntax tree (AST) and classifies functions based on heuristics that indicate vulnerability. These include pointer dereferencing, calls to memory-related functions (e.g., `memcpy`, `memset`), and invocation of other internal functions. Functions that score highly on these metrics are prioritized for fuzz driver generation. The guiding insight is that entry points with fewer argument checks and more direct memory operations expose more useful program logic for fuzz testing.

The Fuzz Driver Synthesizer then generates harnesses for these entry points. For each target function, it synthesizes an `LLVMFuzzerTestOneInput` function that invokes the target with arguments derived from the fuzz input. This process involves inferring argument types from the source code and ensuring that runtime behavior does not violate memory safety—thus avoiding invalid inputs that would cause crashes unrelated to genuine bugs.

IntelliGen stands out by integrating static vulnerability estimation into the driver generation pipeline. Compared to prior tools like FuzzGen and FUDGE, it uses a more targeted, heuristic-based selection of functions, increasing the likelihood that fuzzing will exercise meaningful and vulnerable code paths.

3.1.7 CKGFuzzer

CKGFuzzer [73] is a fuzzing framework designed to automate the generation of effective fuzz drivers for C/C++ libraries by leveraging static analysis and large language models. Its workflow begins by parsing the target project along with any associated library APIs to construct a code knowledge graph. This involves two primary steps: first, parsing the abstract syntax tree (AST), and second, performing inter-procedural program analysis. Through this process, CKGFuzzer extracts essential program elements such as data structures, function signatures, function implementations, and call relationships.

518 Using the knowledge graph, CKGFuzzer then identifies and queries meaningful API combinations,
519 focusing on those that are either frequently invoked together or exhibit functional similarity.
520 It generates candidate fuzz drivers for these combinations and attempts to compile them. Any
521 compilation errors encountered during this phase are automatically repaired using heuristics and
522 domain knowledge. A dynamically updated knowledge base, constructed from prior library usage
523 patterns, guides both the generation and repair processes.

524 Once the drivers are successfully compiled, CKGFuzzer executes them while monitoring code
525 coverage at the file level. It uses coverage feedback to iteratively mutate underperforming API
526 combinations, refining them until new execution paths are discovered or a preset mutation budget
527 is exhausted.

528 Finally, any crashes triggered during fuzzing are subjected to a reasoning process based on chain-
529 of-thought prompting (Section 2.2.2). To help determine their severity and root cause, CKGFuzzer
530 consults an LLM-generated knowledge base containing real-world examples of vulnerabilities
531 mapped to known Common Weakness Enumeration (CWE) entries.

532 3.1.8 PromptFuzz

533 PromptFuzz [74] constitutes a framework for automatically generating fuzz drivers using LLMs,
534 with a novel focus on *prompt mutation* to improve coverage. The system is implemented in Rust
535 and targets C libraries, aiming to explore more of the API surface with each iteration.

536 The workflow begins with the random selection of API functions, extracted from header file
537 declarations. These functions are used to construct initial prompts that instruct the LLM to generate
538 a simple program utilizing the API. Each generated program is compiled, executed, and monitored
539 for code coverage. Programs that fail to compile or violate runtime checks (e.g. sanitizers) are
540 discarded.

541 A key innovation in PromptFuzz is *coverage-guided prompt mutation*. Instead of mutating generated
542 code directly, PromptFuzz mutates the LLM prompts—selecting new combinations of API functions
543 to target unexplored code paths. This process is guided by a *power scheduling* strategy that prioritizes
544 underused or promising API functions based on feedback from previous runs.

545 Once an effective program is produced, it is transformed into a fuzz driver by replacing constants
546 and arguments with variables derived from the fuzzer input. Multiple such drivers are embedded
547 into a single harness, where the input determines which program variant to execute, typically via a
548 case-switch construct.

549 Overall, PromptFuzz demonstrates that prompt-level mutation enables more effective exploration
550 of complex APIs and achieves better coverage than direct code mutations, offering a compelling
551 direction for LLM-based automated fuzzing systems.

3.1.9 OSS-Fuzz

OSS-Fuzz [70], [75] is a continuous, scalable and distributed cloud fuzzing solution for critical and prominent open-source projects. Developers of such software can submit their projects to OSS-Fuzz’s platform, where its harnesses are built and constantly executed. This results in multiple bug findings that are later disclosed to the primary developers and are later patched.

OSS-Fuzz started operating in 2016, an initiative in response to the Heartbleed vulnerability [20], [21], [23]. Its hope is that through more extensive fuzzing such errors could be caught and corrected before having the chance to be exploited and thus disrupt the public digital infrastructure. So far, it has helped uncover over 10,000 security vulnerabilities and 36,000 bugs across more than 1,000 projects, significantly enhancing the quality and security of major software like Chrome, OpenSSL, and systemd.

A project that’s part of OSS-Fuzz must have been configured as a ClusterFuzz [76] project. ClusterFuzz is the fuzzing infrastructure that OSS-Fuzz uses under the hood and depends on Google Cloud Platform services, although it can be hosted locally. Such an integration requires setting up a build pipeline, fuzzing jobs and expects a Google Developer account. Results are accessible through a web interface. ClusterFuzz, and by extension OSS-Fuzz, supports fuzzing through LibFuzzer, AFL++, Honggfuzz and FuzzTest—successor to Centipede—with the last two being Google projects [17], [28], [77], [78]. C, C++, Rust, Go, Python and Java/JVM projects are supported.

3.1.10 OSS-Fuzz-Gen

OSS-Fuzz-Gen (OFG) [7], [79] is Google’s current State-Of-The-Art (SOTA) project regarding automatic harness generation through LLMs. It’s purpose is to improve the fuzzing infrastructure of open-source projects that are already integrated into OSS-Fuzz. Given such a project, OSS-Fuzz-Gen uses its preexisting fuzzing harnesses and modifies them to produce new ones. Its architecture can be described as follows: 1. With an OSS-Fuzz project’s GitHub repository link, OSS-Fuzz-Gen iterates through a set of predefined build templates and generates potential build scripts for the project’s harnesses. 2. If any of them succeed they are once again compiled, this time through fuzz-introspector [80]. The latter constitutes a static analysis tool, with fuzzer developers specifically in mind. 3. Build results, old harness and fuzz-introspector report are included in a template-generated prompt, through which an LLM is called to generate a new harness. 4. The newly generated fuzz target is compiled and if it is done so successfully it begins execution inside OSS-Fuzz’s infrastructure.

This method proved meaningful, with code coverage in fuzz campaigns increasing thanks to the new generated fuzz drivers. In the case of [81], line coverage went from 38% to 69% without any manual interventions [79].

In 2024, OSS-Fuzz-Gen introduced an experimental feature for generating harnesses in previously unfuzzed projects [82]. The code for this feature resides in the `experimental/from_scratch` directory of the project’s GitHub repository [7], with the latest known working commit being 171aac2 and the latest overall commit being four months ago.

3.1.11 AutoGen

AutoGen [64] is a closed-source tool that generates new fuzzing harnesses, given only the library code and documentation. It works as following: The user specifies the function for which a harness is to be generated. AutoGen gathers information for this function—such as the function body, used header files, function calling examples—from the source code and documentation. Through specific prompt templates containing the above information, an LLM is tasked with generating a new fuzz driver, while another is tasked with generating a compilation command for said driver. If the compilation fails, both LLMs are called again to fix the problem, whether it was on the driver’s or command’s side. This loop iterates until a predefined maximum value or until a fuzz driver is successfully generated and compiled. If the latter is the case, it is then executed. If execution errors exist, the LLM responsible for the driver generation is used to correct them. If not, the pipeline has terminated and a new fuzz driver has been successfully generated.

3.2 Differences

OverHAuL differs, in some way, with each of the aforementioned works. Firstly, although KLEE and IRIS [63], [69] tackle the problem of automated testing and both IRIS and OverHAuL can be considered neurosymbolic AI tools, the similarities end there. None of them utilize LLMs the same way we do—with KLEE not utilizing them by default, as it precedes them chronologically—and neither are automating any part of the fuzzing process.

When it comes to FUDGE, FuzzGen and UTopia [8]–[10], all three depend on and demand existing client code and/or unit tests. On the other hand, OverHAuL requires only the bare minimum: the library code itself. Another point of difference is that in contrast with OverHAuL, these tools operate in a linear fashion. No feedback is produced or used in any step and any point failure results in the termination of the entire run.

OverHAuL challenges a common principle of these tools, stated explicitly in FUDGE’s paper [10]: “Choosing a suitable fuzz target (still) requires a human”. OverHAuL chooses to let the LLM, instead of the user, explore the available functions and choose one to target in its fuzz driver.

OSS-Fuzz-Gen [7] can be considered a close counterpart of OverHAuL, and in some ways it is. A lot of inspiration was gathered from it, like for example the inclusion of static analysis and its usage in informing the LLM. Yet, OSS-Fuzz-Gen has a number of disadvantages that make it in some cases an inferior option. For one, OFG is tightly coupled with the OSS-Fuzz platform [70], which even on its own creates a plethora of issues for the common developer. To integrate their project into OSS-Fuzz, they would need to: Transform it into a ClusterFuzz project [76] and take time to write harnesses for it. Even if these prerequisites are carried out, it probably would not be enough. Per OSS-Fuzz’s documentation [75]: “To be accepted to OSS-Fuzz, an open-source project must have a significant user base and/or be critical to the global IT infrastructure”. This means that OSS-Fuzz is a viable option only for a small minority of open-source developers and maintainers. One countermeasure of the above shortcoming would be for a developer to run OSS-Fuzz-Gen locally. This unfortunately proves to be an arduous task. As it is not meant to be used standalone, OFG is not packaged in the form of a self-contained application. This makes it hard to setup and difficult to use interactively. Like in the case of FUDGE, OFG’s actions are performed linearly.

630 No feedback is utilized nor is there graceful error handling in the case of a step’s failure. Even
631 in the case of the experimental feature for bootstrapping unfuzzed projects, OFG’s performance
632 varies heavily. During experimentation, a lot of generated harnesses were still wrapped either in
633 Markdown backticks or `<code>` tags, or were accompanied with explanations inside the generated
634 .c source file. Even if code was formatted correctly, in many cases it missed necessary headers for
635 compilation or used undeclared functions.

636 Lastly, the closest counterpart to OverHAuL is AutoGen [64]. Their similarity stands in the
637 implementation of a feedback loop between LLM and generated harness. However, most other
638 implementation decisions remain distinct. One difference regards the fuzzed function. While
639 AutoGen requires a target function to be specified by the user in which it narrows during its
640 whole run, OverHAuL delegates this to the LLM, letting it explore the codebase and decide by
641 itself the best candidate. Another difference lies in the need—and the lack of—of documentation.
642 While AutoGen requires it to gather information for the given function, OverHAuL leans into the
643 role of a developer by reading the related code and comments and thus avoiding any mismatches
644 between documentation and code. Finally, the LLMs’ input is built based on predefined prompt
645 templates, a technique also present in OSS-Fuzz-Gen. OverHAuL operates one abstraction level
646 higher, leveraging DSPy [83] for programming instead of prompting the LLMs used.

647 In conclusion, OverHAuL constitutes an *open-source* tool that offers new functionality by offering a
648 straightforward installation process, packaged as a self-contained Python package with minimal
649 external dependencies. It also introduces novel approaches compared to previous work by

- 650 1. Implementing a feedback mechanism between harness generation, compilation, and evalua-
651 tion phases,
- 652 2. Using autonomous ReAct agents capable of codebase exploration,
- 653 3. Leveraging a vector store for code consumption and retrieval.

4 OverHAuL's Design

In this thesis we present *OverHAuL* (**H**arness **A**utomation with **L**LMs), a neurosymbolic AI tool that automatically generates fuzzing harnesses for C libraries through LLM agents. In its core, OverHAuL is comprised by three LLM ReAct agents [49]—each with its own responsibility and scope—and a vector store index reserving the given project's analyzed codebase. An overview of OverHAuL's process is presented in Figure 4.1. The objective of OverHAuL is to streamline the process of fuzz testing for C libraries. Given a link to a git repository [84] of a C library, OverHAuL automatically generates a new fuzzing harness specifically designed for the project. In addition to the harness, it produces a compilation script to facilitate building the harness, generates a representative input that can trigger crashes, and logs the output from the executed harness.

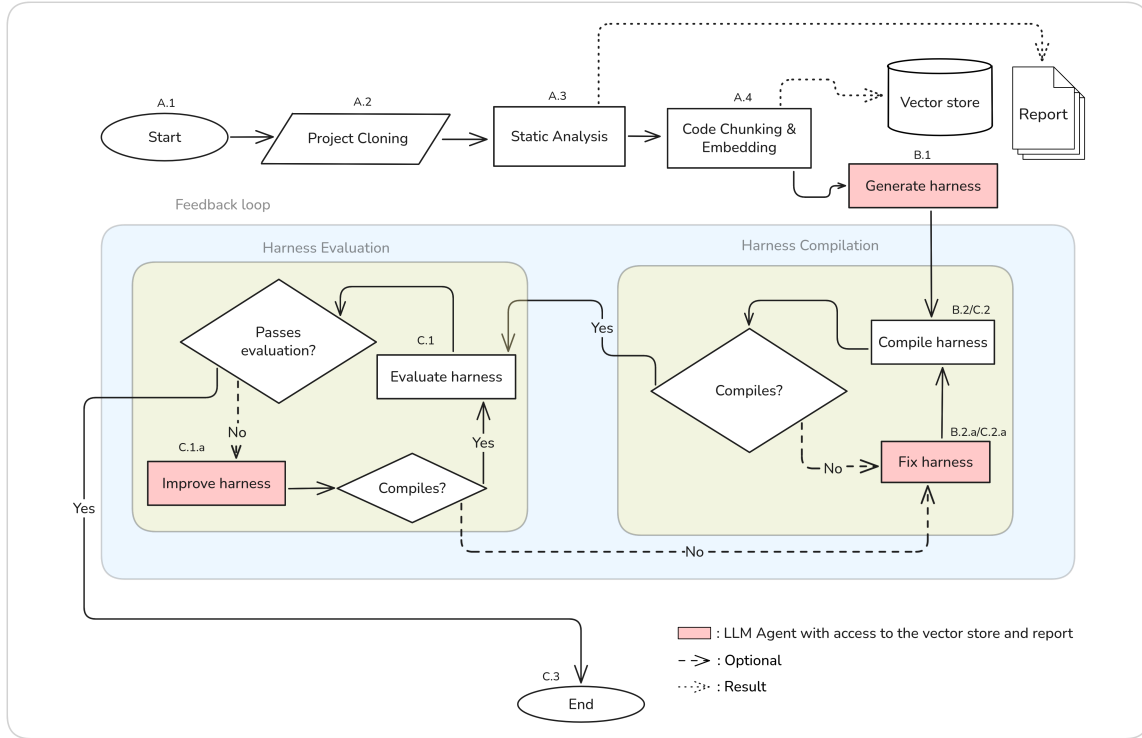


Figure 4.1: Overview of OverHAuL's automatic harnessing process.

As commented in Section 3.2, OverHAuL does not expect and depend on the existence of client code or unit tests [8]–[10] *nor* does it require any preexisting fuzzing harnesses [7] or any documentation present [64]. Also importantly, OverHAuL is decoupled from other fuzzing projects, thus lowering the barrier to entry for new projects [7], [70]. Lastly, the user isn't mandated to specify manually

the function which the harness-to-be-generated must fuzz. Instead, OverHAuL’s agents examine and assess the provided codebase, choosing after evaluation the most optimal targeted function.

OverHAuL utilizes autonomous ReAct agents [49] which inspect and analyze the project’s source code. The latter is stored and interacted with as a set of text embeddings [85], kept in a vector store. Both approaches are, to the best of our knowledge, novel in the field of automatic fuzzing harnesses generation. OverHAuL also implements an evaluation component that assesses in real-time all generated harnesses, making the results tenable, reproducible and well-founded. Ideally, this methodology provides a comprehensive and systematic framework for identifying previously unknown software vulnerabilities in projects that have not yet been fuzz tested.

Finally, OverHAuL excels in its user-friendliness, as it constitutes a simple and easily-installable Python package with minimal external dependencies—only real dependency being Clang, a prevalent compiler available across all primary operating systems. This contrasts most other comparable systems, which are typically characterized by their limited documentation, lack of extensive testing, and a focus primarily on experimental functionality.¹

4.1 Architecture

OverHAuL can be compartmentalized in three stages: First, the project analysis stage (Section 4.1.1), the harness creation stage (Section 4.1.2) and the harness evaluation stage (Section 4.1.3).

4.1.1 Project Analysis

In the project analysis stage (steps A.1–A.4), the project to be fuzzed is ran through a static analysis tool and is sliced into function-level chunks, which are stored in a vector store. The results of this stage are a static analysis report and a vector store containing embeddings of function-level code chunks, both of which are later available to the LLM agents.

The static analysis tool Flawfinder [86] is executed with the project directory as input and is responsible for the static analysis report. This report is considered a meaningful resource, since it provides the LLM agent with some starting points to explore, regarding the occurrences of potentially vulnerable functions and/or unsafe code practices.

The vector store is created in the following manner: The codebase is first chunked in function-level pieces by traversing the code’s Abstract Syntax Tree (AST) through Clang. Each chunk is represented by an object with the function’s signature, the corresponding filepath and the function’s body. Afterwards, each function body is turned into a vector embedding through an embedding model. Each embedding is stored in the vector store. This structure is created and used for easier and more semantically meaningful code retrieval, and to also combat context window limitations present in the LLMs.

¹I.e. “research code”.

4.1.2 Harness Creation

Second is the harness creation stage (steps B.1–B.2). In this part, a “generator” ReAct LLM agent is tasked with creating a fuzzing harness for the project. The agent has access to a querying tool that acts as an interface between it and the vector store. When the agent makes queries like “functions containing strcpy()”, the querying tool turns the question into an embedding and through similarity search returns the top $k = 5$ most similar results—in this case, functions of the project. With this approach, the agent is able to explore the codebase semantically and pinpoint potentially vulnerable usage patterns easily.

The harness generated by the agent is then compiled using Clang and linked with the AddressSanitizer, LeakSanitizer, and UndefinedBehaviorSanitizer. The compilation command used is generated programmatically, according to the rules described in Section 4.5. If the compilation fails for any reason, e.g. a missing header include, then the generated faulty harness and its compilation output are passed to a new “fixer” agent tasked with repairing any errors in the harness (step B.2.a). This results in a newly generated harness, presumably free from the previously shown flaws. This process is iterated until a compilable harness has been obtained. After success, a script is also exported in the project directory, containing the generated compilation command.

4.1.3 Harness Evaluation

Third comes the evaluation stage (steps C.1–C.3). During this step, the compiled harness is executed and its results evaluated. Namely, a generated harness passes the evaluation phase if and only if:

1. The harness has no memory leaks during its execution This is inferred by the existence of leak-<hash> files.
2. A new testcase was created *or* the harness executed for at least MIN_EXECUTION_TIME (i.e. did not crash on its own) When a crash happens, and thus a testcase is created, it results in a crash-<hash> file.
3. The created testcase is not empty This is examined through xxd’s output given the crash-file.

Similarly to the second stage’s compilation phase (steps B.2–B.2.a), if a harness does not pass the evaluation for whatever reason it is sent to an “improver” agent. This agent is instructed to refine it based on its code and cause of failing the evaluation. This process is also iterative. If any of the improved harness versions fail to compile, the aforementioned “fixer” agent is utilized again (steps C.2–C.2.a). All produced crash files and the harness execution output are saved in the project’s directory.

4.2 Main Techniques

The fundamental techniques that distinguish OverHAuL in its approach and enhance its effectiveness in achieving its objectives are: The implementation of an iterative feedback loop between the LLM agents, the distribution of responsibility across a swarm of distinct agents and the employment of a “codebase oracle” for interacting with the given project’s source code.

4.2.1 Feedback Loop

The initial generated harness produced by OverHAuL is unlikely to be successful from the get-go. The iterative feedback loop implemented facilitates its enhancement, enabling the harness to be tested under real-world conditions and subsequently refined based on the results of these tests. This approach mirrors the typical workflow employed by developers in the process of creating and optimizing fuzz targets.

In this iterative framework, the development process continues until either an acceptable and functional harness is realized or the defined *iteration budget* is exhausted. The iteration budget $N = 10$ is initialized at the onset of OverHAuL’s execution and is shared between both the compilation and evaluation phases of the harness development process. This means that the iteration budget is decremented each time a dashed arrow in the flowchart illustrated in Figure 4.1 is followed. Such an approach allows for targeted improvements while maintaining oversight of resource allocation throughout the harness development cycle.

4.2.2 React Agents Swarm

An integral design decision in our framework is the implementation of each agent as a distinct LLM instance, although all utilizing the same underlying model. This approach yields several advantages, particularly in the context of maintaining separate and independent contexts for each agent throughout each OverHAuL run.

By assigning individual contexts to the agents, we enable a broader exploration of possibilities during each run. For instance, the “improver” agent can investigate alternative pathways or strategies that the “generator” agent may have potentially overlooked or internally deemed inadequate inaccurately. This separation not only fosters a more diverse range of solutions but also enhances the overall robustness of the system by allowing for iterative refinement based on each agent’s unique insights.

Furthermore, this design choice effectively addresses the limitations imposed by context window sizes. By distributing the “cognitive” load across multiple agents, we can manage and mitigate the risks associated with exceeding these constraints. As a result, this architecture promotes efficient utilization of available resources while maximizing the potential for innovative outcomes in multi-agent interactions. This layered approach ultimately contributes to a more dynamic and exploratory research environment, facilitating a comprehensive examination of the problem space.

4.2.3 Codebase Oracle

The third central technique employed is the creation and utilization of a codebase oracle, which is effectively realized through a vector store. This oracle is designed to contain the various functions within the project, enabling it to return the most semantically similar functions upon querying it. Such an approach serves to address the inherent challenges associated with code exploration difficulties faced by LLM agents, particularly in relation to their limited context window.

773 By structuring the codebase into chunks at the level of individual functions, LLM agents can engage
774 with the code more effectively by focusing on its functional components. This methodology not
775 only allows for a more nuanced understanding of the codebase but also ensures that the responses
776 generated do not consume an excessive portion of the limited context window available to the agents.
777 In contrast, if the codebase were organized and queried at the file level, the chunks of information
778 would inevitably become larger, leading to an increase in noise and a dilution of meaningful content
779 in each chunk [87]. Given the constant size of the embeddings used in processing, each progressively
780 larger chunk would be less semantically significant, ultimately compromising the quality of the
781 retrieval process.

782 Defining the function as the primary unit of analysis represents the most proportionate balance
783 between the size of the code segments and their semantic significance. It serves as the ideal “zoom-
784 in” level for the exploration of code, allowing for greater clarity and precision in understanding the
785 functionality of individual code segments. This same principle is widely recognized in the training
786 of code-specific LLMs, where a function-level approach has been shown to enhance performance
787 and comprehension [88]. By adopting this methodology, we aim to foster a more robust interaction
788 between LLM agents and the underlying codebase, ultimately facilitating a more effective and
789 efficient exploration process.

790 4.3 High-Level Algorithm

791 A pseudocode version of OverHAuL’s main function can be seen in Algorithm 4.1. It represents
792 the workflow presented in Figure 4.1 and uses the techniques described in sections 4.1 and 4.2. It
793 is important to emphasize that, within the context of this algorithm, the `HarnessAgents()` function
794 serves as an interface that bridges the “generator”, “fixer” and “improver” LLM agents. The agent
795 that is used upon each function call depends on the values of the function’s arguments. This results
796 in the *harness* variable representing all generated, fixed or improved harnesses. This approach is
797 adopted for making the abstract algorithm simpler and easier to understand.

Algorithm 4.1 OverHAuL

Require: *repository***Ensure:** *harness, compilation_script, crash_input, execution_log*

```
1: path ← REPOCLONE(repository)
2: report ← STATICANALYSIS(path)
3: vector_store ← CREATEORACLE(path)
4: acceptable ← False
5: compiled ← False
6: error ← None
7: violation ← None
8: output ← None
9: for i = 1 to MAX_ITERATIONS do
10:   harness ← HARNESSAGENTS(path, report, vector_store, error, violation, output)
11:   error, compiled ← BUILDHARNESS(path, harness)
12:   if  $\neg$ compiled then
13:     continue ▷ Fix harness
14:   end if
15:   output, accepted ← EVALUATEHARNESS(path, harness)
16:   if  $\neg$ accepted then
17:     continue ▷ Improve harness
18:   else
19:     acceptable ← True
20:     break
21:   end if
22: end for
23: return compiled  $\wedge$  acceptable
```

798 4.4 Installation and Usage

799 The source code of OverHAuL is available in <https://github.com/kchousos/OverHAuL>. OverHAuL
800 can be installed by cloning the git repository locally, creating and enabling a Python3.10 virtual
801 environment [89] and installing it inside the environment using Python’s PIP package installer [90],
802 like in Listing 4.1.

803 To use OverHAuL, you need to provide a secret key for using OpenAI’s API service. This key can
804 be either stored in a .env file in the root directory, like so:

```
1 # cat .env
2 OPENAI_API_KEY=<API-key-here>
```

805 Or it can be exported in the shell environment:

Listing 4.1 OverHAuL’s installation process.

```
1 $ git clone https://github.com/kchousos/overhaul; cd overhaul
2 ...
3 $ python3.10 -m venv .venv
4 $ source ./venv/bin/activate
5 $ pip install .
6 ...
7 $ overhaul --help
8 usage: overhaul [-h] [-c COMMIT] [-m MODEL] [-f FILES [FILES ...]] [-o OUTPUT_DIR] repo
9
10 Generate fuzzing harnesses for C/C++ projects
11
12 positional arguments:
13   repo                  Link of a project's git repo, for which to generate a harness.
14
15 options:
16   -h, --help            show this help message and exit
17   -c COMMIT, --commit COMMIT
18                       A specific commit of the project to check out
19   -m MODEL, --model MODEL
20                       LLM model to be used. Available: o3-mini, o3, gpt-4o,
21                       gpt-4o-mini, gpt-4.1, gpt-4.1-mini, gpt-3.5-turbo, gpt-4
22   -f FILES [FILES ...], --files FILES [FILES ...]
23                       File patterns to include in analysis (e.g. *.c *.h)
24   -o OUTPUT_DIR, --output-dir OUTPUT_DIR
25                       Directory to clone the project into. Defaults to "output"
26 $
```

```
1 $ export OPENAI_API_KEY=<API-key-here>
2 $ overhaul <repo-link>
```

806 Once these preliminary steps are completed, OverHAuL can be executed. The primary argument
807 required by OverHAuL is the repository link of the library that is to be fuzzed. Additionally, users
808 have the option to specify certain command-line flags, which allow them to control the checked-out
809 commit of the cloned project, select the OpenAI LLM model from a predefined list, define specific
810 file patterns for OverHAuL to search for, and determine the directory in which the project will be
811 cloned. A sample successful execution can be presented in Figure 4.2.

812 In this example, the dateparse repository is cloned into the ./output/dateparse directory, which is
813 relative to the root directory of OverHAuL. Following a successful execution, this directory will
814 contain a new folder named harnesses, which will house all the generated harnesses formatted as
815 harness_n.c—where n ranges from 1 to $N - 1$, with N representing the total number of harnesses
816 produced. The most recent and verifiably correct harness will be designated simply as harness.c.

```

$ overhaul https://github.com/dvhar/dateparse -m gpt-4.1
2025-07-17 00:55:18.007 | INFO | overhaul.cli:parse_arguments:184 - Cloning project's repo in the output/dateparse directory...
2025-07-17 00:55:18.976 | INFO | overhaul.core.harnesser:_init_:167 - Initializing LLM...
2025-07-17 00:55:18.991 | INFO | overhaul.cli:main:243 - Iteration 1 of harnessing...
2025-07-17 00:55:18.991 | INFO | overhaul.core.harnesser:harness:207 - Calling LLM to generate a harness...
2025-07-17 00:55:22.837 | DEBUG | overhaul.rag.ragger:rag_tool:88 - Agent using RAG: "Show the functions in dateparse.c that contain a strcpy call, especially near lines 405, 800, and 1242."
2025-07-17 00:55:25.807 | DEBUG | overhaul.rag.ragger:rag_tool:88 - Agent using RAG: "Provide the full implementation (signature and body) of the dateparse function in dateparse.c."
2025-07-17 00:55:30.207 | DEBUG | overhaul.rag.ragger:rag_tool:88 - Agent using RAG: "Show the full signature and body of the function int dateparse(const char* datestr, date_t* t, int *offset, int stringlen) in dateparse.c."
2025-07-17 00:55:34.176 | DEBUG | overhaul.rag.ragger:rag_tool:88 - Agent using RAG: "Show the definition of struct parser and the typedef of date_t as used in dateparse.c and dateparse.h."
2025-07-17 00:55:38.581 | DEBUG | overhaul.rag.ragger:rag_tool:88 - Agent using RAG: "Show the definition of struct parser and the typedef of date_t from dateparse.h."
2025-07-17 00:56:01.897 | INFO | overhaul.io.file_manager:write_harness:60 - Writing harness to project...
2025-07-17 00:56:01.898 | INFO | overhaul.io.file_manager:write_harness:92 - Harness written to output/dateparse/harnesses/harness.c
2025-07-17 00:56:01.898 | INFO | overhaul.core.builder:build_harness:63 - Building harness...
2025-07-17 00:56:01.899 | INFO | overhaul.core.builder:build_harness:140 - Starting compilation of harness: harnesses/harness.c
2025-07-17 00:56:02.345 | INFO | overhaul.core.builder:build_harness:149 - Harness compiled successfully
2025-07-17 00:56:02.345 | INFO | overhaul.core.evaluator:evaluate_harness:81 - Evaluating harness...
2025-07-17 00:56:02.345 | INFO | overhaul.core.evaluator:evaluate_harness:90 - Starting execution of harness...
2025-07-17 00:56:02.417 | INFO | overhaul.core.evaluator:evaluate_harness:119 - Harness execution completed in 0.07 seconds.
2025-07-17 00:56:02.419 | INFO | overhaul.core.evaluator:evaluate_harness:181 - New testcases created (1): [{'crash-dfaa340e98889cd82d2cd680cf96fd04552a2b4', 1752702962.4113252}]
2025-07-17 00:56:02.419 | SUCCESS | overhaul.cli:main:282 - All done!

```

Figure 4.2: A successful execution of OverHAuL, harnessing [dvhar's dateparsing C library](#), using OpenAI's gpt-4.1 model. Debug statements are printed to showcase the interaction between the LLM agents and the codebase oracle (Section 4.2.3).

Additionally, the dateparse directory will include an executable script named `overhaul.sh`, which contains the compilation command necessary for the harness. A log file titled `harness.out` will also be present, documenting the output from the latest harness execution. Lastly and most importantly, there will be at least one non-empty crash file included, serving as a witness to the harness's correctness.

4.5 Scope

Currently, OverHAuL is designed to generate new harnesses specifically for medium-sized C libraries. Given the inherent complexity of dealing with C++ projects, this is not a feature yet supported within the system.

The compilation command utilized by OverHAuL is created programmatically. It incorporates the root directory along with all subdirectories that conform to a predefined set of common naming conventions. Additionally, the compilation process uses all C source files identified within these directories. Crucially, it is important that no `main()` function is present in any of the files to ensure successful compilation. For this reason any files or directories that include “test”, “main”, “example”, “demo”, or “benchmark” in their paths are systematically excluded from the compilation process. This exclusion also decreases the “noise” in the oracle, as these files do not constitute part of the core library and would therefore not contain any functions meaningful to the LLM agents.

Lastly, No support for build systems such as Make or CMake [91], [92] is yet implemented. Such functionality would exponentially increase the complexity of the build step and is beyond the scope of this thesis.

5 Evaluation

To thoroughly assess the performance and effectiveness of OverHAuL, we established four *research questions* to direct our investigative efforts. These questions are designed to provide a structured framework for our inquiry and to ensure that our research remains focused on the key aspects of OverHAuL’s functionality and impact within its intended domain. By addressing these questions, we aim to uncover valuable insights that will contribute to a deeper understanding of OverHAuL’s capabilities and its position in contemporary automatic fuzzing applications:

- **RQ1:** Can OverHAuL generate working harnesses for unfuzzed C projects?
- **RQ2:** What characteristics do these harnesses have? Are they similar to man-made harnesses?
- **RQ3:** How do LLM usage patterns influence the generated harnesses?
- **RQ4:** How do different symbolic techniques affect the generated harnesses?

5.1 Experimental Benchmark

To evaluate OverHAuL, a benchmarking script was implemented¹ and a corpus of ten open-source C libraries was assembled. This collection comprises of: Firstly, GitHub user dhvar’s dateparse library, which is also used as a running example in OSS-Fuzz-Gen’s [7] experimental from-scratch harnessing feature (Section 3.1.10). Secondly, nine other libraries chosen randomly² from the package catalog of Clib, a “package manager for the C programming language” [93], [94]. All libraries can be seen Table 5.1, along with their descriptions.

Table 5.1: The benchmark project corpus. Each project name links to its corresponding GitHub repository. Each is followed by a short description and its GitHub stars count, as of July 18th, 2025.

Project	Description	Stars
dhvar/dateparse	A library that allows parsing dates without knowing the format in advance.	2
clibs/buffer	A string manipulation library.	204
jwerle/libbeaufort	A library implementation of the Beaufort cipher [95].	13
jwerle/libbacon	A library implementation of the Baconian cipher [96].	8

¹<https://github.com/kchousos/OverHAuL/blob/master/benchmarks/benchmark.sh>

²From the subset of libraries that do not have exotic external dependencies, like the X11 development toolchain.

Project	Description	Stars
jwerle/chfreq.c	A library for computing the character frequency in a string.	5
jwerle/progress.c	A library for displaying progress bars in the terminal.	76
willemt/cbuffer	A circular buffer implementation.	261
willemt/torrent-reader	A torrent-file reader library.	6
orangeduck/mpc	A type-generic parser combinator library.	2,753
h2non/semver.c	A semantic version v2.0 parsing and rendering library [97].	190

855 5.1.1 Local Benchmarking

856 To run the benchmark locally, one would need to follow the installation instructions in Section 4.4
857 and then execute the benchmarking script, like so:

```
1 $ ./benchmarks/benchmark.sh
```

858 The cloned repositories with their corresponding harnesses will then be located in a subdirectory of
859 `benchmark_results`, which will have the name format of `mini__<timestamp>__ReAct__<llm-model>__<max-
860 exec-time>__<iter-budget>`. “Mini” corresponds to the benchmark project corpus described above,
861 since a 30-project corpus was initially created and is now coined as “full” benchmark. Both the mini
862 and full benchmarks are located in `benchmarks/repos.txt` and `benchmarks/repos-mini.txt` respectively.
863 To execute the benchmark for the “full” corpus, users can add the `-b full` flag in the script’s
864 invocation. Also, the LLM model used can be defined with the `-m` command-line flag.

6 Results

OverHAuL was evaluated through the experimental benchmark (Section 5.1) from 6th of June, 2025 to 18th of July, 2025, using OpenAI’s gpt-4.1-mini model [98]. For these runs, each OverHAuL execution was configured with a 5 minute harness execution timeout and an iteration budget of 10. Each benchmark run was executed as a GitHub Actions workflow, and the result directory (described in Section 5.1.1) for each is available as a downloadable artifact in the corresponding GitHub Actions entry. In Figure 6.1, the results of these benchmark runs are showcased.

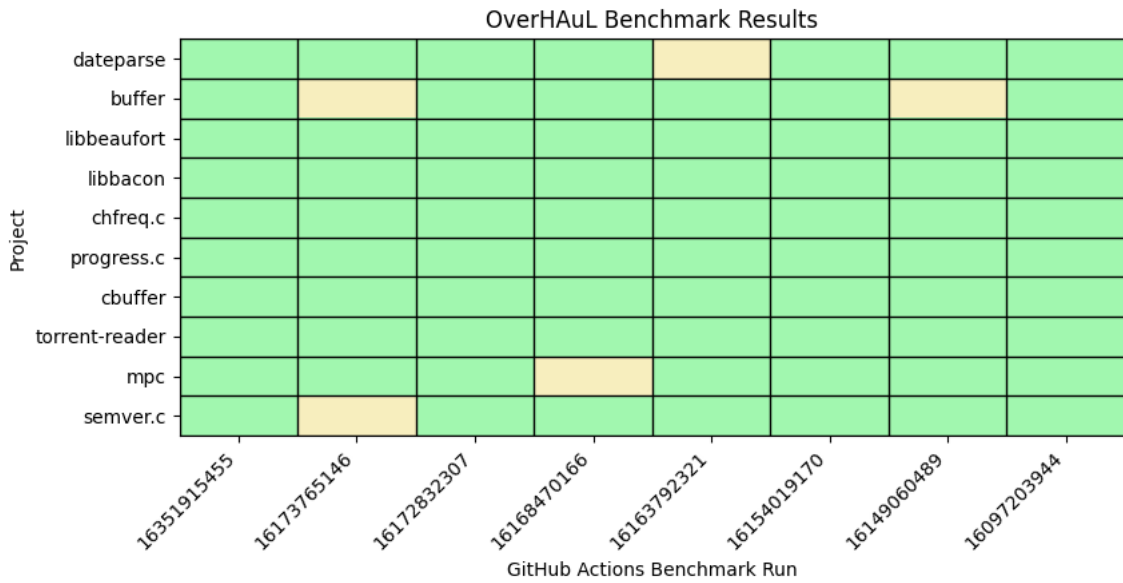


Figure 6.1: The benchmark results for OverHAuL are illustrated with the y-axis depicting the ten-project corpus outlined in Section 5.1. The x-axis represents the various benchmark runs. Each label constitutes a unique hash identifier corresponding to a specific GitHub Actions workflow run, which can be accessed at <https://github.com/kchousos/OverHAuL/actions/runs/HASH>. An overview of all benchmark runs is available at <https://github.com/kchousos/OverHAuL/actions/workflows/benchmarks.yml>. In this matrix, a green block indicates that OverHAuL successfully generated a new harness for the project and was able to find a crash input. On the other hand, a yellow block indicates that while a compilable harness was produced, no crash input was found within the five-minute execution period. Importantly, there are no red blocks, which would indicate cases where a compilable harness could not be generated.

6.1 Answers to RQs

We can deduce the following facts from Figure 6.1: Firstly, OverHAuL has a very high success rate in finding crash inputs for the given programs, specifically one of 92.5%. Secondly, although

875 according strictly to these results OverHAuL never generates an un-compilable harness, we can err
876 on the side of caution and infer that OverHAuL not being able to generate a compilable harness is a
877 rare occurrence. These conclusions can safely answer RQ1 with a resounding yes.

878 From sampling OverHAuL’s generated harnesses, the answer to RQ2 remains unclear. Most of the
879 time, the fuzz targets that are produced are understandable and similar to something a software
880 engineer might program. Take for example Listing 6.1. Nonetheless, sometimes generated harnesses
881 contain usage of inexplicable or arbitrary constants and peculiar control flow checks. This makes
882 them harder to understand and quite possibly incorrect in many cases, thus diverging from seeming
883 human-written. RQ2’s answer remains an unclear “it depends”, given the variance in OverHAuL’s
884 results.

885 In exploring the utilization of LLMs, two critical dimensions warrant examination: the selection
886 of the LLM model itself and the prompting techniques employed. Within the context of model
887 selection, all benchmark tests conducted on GitHub’s infrastructure utilized OpenAI’s gpt-4.1-mini.
888 Additionally, local testing involved gpt-4.1, gpt-4o, gpt-4, and gpt-3.5-turbo. The results indicate
889 that both gpt-4.1 and gpt-4.1-mini demonstrated comparable positive outcomes, while gpt-4o
890 produced semi-average results. In contrast, gpt-4 and gpt-3.5-turbo exhibited significantly poorer
891 performance, averaging approximately 2 out of 10 successfully harnessed projects per benchmark
892 run. This underscores the considerable impact that the size and capabilities of the underlying LLM
893 model have on OverHAuL’s effectiveness. Consequently, gpt-4o emerges as a contemporary cut-off
894 point in LLM development concerning OverHAuL’s performance, suggesting that advancements in
895 LLM technologies will likely enhance OverHAuL’s capabilities rapidly.

896 Regarding the prompting techniques adopted, ReAct prompting has yielded the most favorable
897 outcomes in the current version of OverHAuL [49]. As detailed in Appendix A, both zero-shot
898 prompting and Chain-of-Thought (COT) prompting [46] produced similarly unsatisfactory results.
899 A primary challenge in automatic harness generation is ensuring that the generated harness aligns
900 with real-world conditions, particularly in terms of compilation success and effective runtime
901 behavior. This alignment can only be validated through LLM–environment interaction, i.e. within
902 agentic workflows [99]. Furthermore, the superior results associated with ReAct prompting can be
903 attributed to its inherent capacity for more sophisticated code exploration.

904 In summary, the response to RQ3 comes to be that ongoing advancements in LLM models will enable
905 systems like OverHAuL to generate increasingly effective outcomes. Additionally, architectures
906 that incorporate agentic modules capable of environment assessment and feedback gathering will
907 surpass more traditional applications of LLMs, particularly in the domain of automatic fuzzing.

908 Throughout the development of OverHAuL and its various iterations, numerous programming tech-
909 niques were assessed in pursuit of answering RQ4 (Appendix A). Simple source code concatenation
910 and its subsequent injection into LLM prompts revealed significant limitations, primarily due to the
911 constraints of context windows. Conversely, the usage of tools capable of retrieving file contents
912 marked a meaningful advancement. Nonetheless, this approach still encountered challenges, such
913 as inaccessible code blocks and exploration that lacked semantic relevance. In response to these
914 difficulties, the implementation of a function-level vector store functioning as a codebase oracle is
915 proposed as a highly scalable solution. This strategy not only enhances the organization of larger
916 files but also accommodates expanding project sizes, facilitating more semantically meaningful
917 code examination.

Listing 6.1 Sample harness for dateparse, generated by OverHAuL.

```
1  #include <stddef.h>
2  #include <stdint.h>
3  #include <string.h>
4  #include <stdlib.h>
5  #include "dateparse.h"
6
7  // No artificial size limit to allow fuzzer to explore full input size for boundaries
8
9  int LLVMFuzzerTestOneInput(const uint8_t *data, size_t size) {
10     // Allocate buffer with extra byte for null termination
11     char *input_str = (char *)malloc(size + 1);
12     if (!input_str) {
13         return 0;
14     }
15
16     memcpy(input_str, data, size);
17     input_str[size] = '\0';
18
19     date_t parsed_date = 0;
20     int offset = 0;
21
22     // Array of string lengths targeting boundary conditions (including 0 = internal strlen)
23     size_t test_lens[] = {0, size, size > 0 ? size - 1 : 0, 12, 13, 14};
24
25     for (size_t i = 0; i < sizeof(test_lens) / sizeof(test_lens[0]); i++) {
26         size_t len = test_lens[i];
27         if (len ≤ size) {
28             dateparse(input_str, &parsed_date, &offset, (int)len);
29         }
30     }
31
32     free(input_str);
33     return 0;
34 }
```

7 Implementation

In creating the codebase oracle, we employ the “libclang” Python package [100] to slice functions based on the AST capability provided by Clang. As detailed in Section 4.2.3, the intermediate output consists of a list of Python dictionaries, with each dictionary storing a function’s body, signature, and corresponding file path. Each chunk of function code is then converted into an embedding using OpenAI’s “text-embedding-3-small” model [101] and stored in a FAISS vector store index [102]. This index is mapped to a metadata structure that contains the aforementioned function data—specifically the actual function body, signature, and file path. When a search is conducted on the index, the results returned are the embeddings. The responses that the LLM agent receives are derived from the corresponding metadata entries of each embedding.

All LLM agents and components are developed using the DSPy library, a declarative Python framework for LLM programming created by Stanford’s NLP research team [83]. DSPy offers built-in modules and abstractions that facilitate the composition of LLMs and prompting techniques, such as Chain of Thought and ReAct (Listing 7.1). Each agent within OverHAuL is an instance of DSPy’s ReAct module [103], accompanied by a custom Signature [104]—displayed in Appendix B. DSPy was selected over other contemporary LLM libraries, such as LangChain and Llamaindex [105], [106], because of its user-friendliness, logical abstractions, and efficient development process—qualities that are often lacking in these alternative libraries [107]–[109].

Listing 7.1 Sample DSPy program.

```
1 import dspy
2 lm = dspy.LM('openai/gpt-4o-mini', api_key='YOUR_OPENAI_API_KEY')
3 dspy.configure(lm=lm)
4
5 math = dspy.ChainOfThought("question → answer: float")
6 math(question="Two dice are tossed. What is the probability that the sum equals two?")
```

Repository cloning is executed using the `--depth 1` flag to minimize disk storage usage and reduce the size of artifacts.

7.1 Development Tools

The development of OverHAuL incorporates a variety of tools aimed at enhancing functionality and efficiency. Notably, “uv” is a Python package and project manager written in Rust that serves as a replacement for Poetry. Additionally, “Ruff,” a code linter and formatter also developed in Rust,

942 contributes to code quality by enforcing consistent formatting standards. The project also employs
943 “MyPy,” the widely-used static type checker for Python, to ensure type correctness. Testing is
944 facilitated through “PyTest,” a robust Python testing framework. Lastly, “pdoc” is utilized as a Static
945 Site Generator (SSG) to automate the creation of API documentation¹ [110]–[114].

946 7.2 Reproducibility

947 OverHAuL’s source code is available at <https://github.com/kchousos/OverHAuL>. Each benchmark
948 run was conducted within the framework of a GitHub Actions workflow, resulting in a detailed
949 summary accompanied by an artifact containing all cloned repositories. These artifacts are the
950 compressed result directories described in Section 5.1.1 and provide the essential components
951 necessary for the reproducibility each project’s results, as described in Section 4.4. All benchmark
952 runs can be conveniently accessed at [https://github.com/kchousos/OverHAuL/actions/workflows/](https://github.com/kchousos/OverHAuL/actions/workflows/benchmarks.yml)
953 [benchmarks.yml](https://github.com/kchousos/OverHAuL/actions/workflows/benchmarks.yml).

¹<https://kchousos.github.io/OverHAuL/>

8 Discussion

As discussed in Section Chapter 6, the capabilities and effectiveness of OverHAuL are closely tied to the choice of the underlying large language model. OverHAuL’s modular architecture ensures that advances in LLM research will directly enhance its performance. Each release of a new, more capable model can be readily integrated, thereby amplifying OverHAuL’s effectiveness without the need for substantial redesign.

A noteworthy consideration in our benchmarking setup is the possibility that some of the open-source libraries evaluated may have been included in the LLM’s training data. This introduces a risk of overestimating OverHAuL’s performance on code that is unseen or proprietary. Results for closed-source or less widely available libraries could therefore be weaker. Nonetheless, this potential limitation can theoretically be addressed through targeted fine-tuning of the LLM [115].

8.1 Threats to Validity

Our evaluation of OverHAuL was conducted on ten relatively obscure open-source C libraries representing a range of application domains and functionalities. While this selection reduces the likelihood that these projects were used in LLM training and thus minimizes potential bias, it remains uncertain how transferable our results are to larger, more complex, or structurally different codebases. Factors such as varying design paradigms, architectural patterns, or real-world deployment contexts may pose new challenges for OverHAuL’s scalability and effectiveness.

Additionally, the risk that LLMs could hallucinate constitutes an internal threat to validity. Such hallucinations may require multiple attempts or occasional manual adjustments to produce valid and useful fuzz drivers. However, because LLMs—and thus OverHAuL—operate in a non-deterministic manner, it is possible to rerun the process and obtain alternative results. The inherent stochasticity of the underlying LLMs thus allows users to recover from initial failures, ensuring that the impact of hallucinations remains limited to efficiency rather than undermining the core applicability of the approach.

In summary, while our findings demonstrate the potential of OverHAuL, they also highlight important limitations and directions for future work, especially in improving robustness and evaluating performance across a broader spectrum of software projects.

9 Future Work

The prototype implementation of OverHAuL offers a compelling demonstration of its potential to automate the fuzzing process for open-source libraries, providing tangible benefits to developers and maintainers alike. This initial version successfully validates the core design principles underpinning OverHAuL, showcasing its ability to streamline and enhance the software testing workflow through automated generation of fuzz drivers using large language models. Nevertheless, while these foundational capabilities lay a solid groundwork, numerous avenues exist for further expansion, refinement, and rigorous evaluation to fully realize the tool’s potential and adapt to evolving challenges in software quality assurance.

9.1 Enhancements to Core Features

Enhancing OverHAuL’s core functionality represents a primary direction for future development. First, expanding support to encompass a wider array of build systems commonly employed in C and C++ projects—such as GNU Make, CMake, Meson, and Ninja [91], [92], [116], [117]—would significantly broaden the scope of libraries amenable to automated fuzzing using OverHAuL. This advancement would enable OverHAuL to scale effectively and be applied to larger, more complex codebases, thereby increasing its practical utility and impact.

Second, integrating additional fuzzing engines beyond LibFuzzer stands out as a strategic enhancement. Incorporation of widely adopted fuzzers like AFL++ [28] could diversify the fuzzing strategies available to OverHAuL, while exploring more experimental tools such as GraphFuzz [66] may pioneer specialized approaches for certain code patterns or architectures. Multi-engine support would also facilitate extending language coverage, for instance by incorporating fuzzers tailored to other programming ecosystems—for example, Google’s Atheris for Python projects [118]. Such versatility would position OverHAuL as a more universal fuzzing automation platform.

Third, the evaluation component of OverHAuL presents an opportunity for refinement through more sophisticated analysis techniques. Beyond the current criteria, incorporating dynamic metrics such as differential code coverage tracking between generated fuzz harnesses would yield deeper insights into test quality and coverage completeness. This quantitative evaluation could guide iterative improvements in fuzz driver generation and overall testing effectiveness.

Finally, OverHAuL’s methodology could be extended to leverage existing client codebases and unit tests in addition to the library source code itself, resources that for now OverHAuL leaves untapped. Inspired by approaches like those found in FUDGE and FuzzGen [9], [10], this enhancement would enable the tool to exploit programmer-written usage scenarios as seeds or contexts, potentially generating more meaningful and targeted fuzz inputs. Incorporating these richer information sources would likely improve the efficacy of fuzzing campaigns and uncover subtler bugs.

9.2 Experimentation with Large Language Models and Data Representation

OverHAuL’s reliance on large language models (LLMs) invites comprehensive experimentation with different providers and architectures to assess their comparative strengths and limitations. Conducting empirical evaluations across leading models—such as OpenAI’s o1 and o3 families and Anthropic’s Claude Opus 4—will provide valuable insights into their capabilities, cost-efficiency, and suitability for fuzz driver synthesis. Additionally, specialized code-focused LLMs, including generative and fill-in models like Codex-1 and CodeGen [52]–[54], merit exploration due to their targeted optimization for source code generation and understanding.

Another dimension worthy of investigation concerns the granularity of code chunking employed during the given project’s code processing stage. Whereas the current approach partitions code at the function level, experimenting with more nuanced segmentation strategies—such as splitting per step inside a function, as a finer-grained technique—could improve the semantic coherence of stored representations and enhance retrieval relevance during fuzz driver generation. This line of inquiry has the potential to optimize model input preparation and ultimately improve output quality.

9.3 Comprehensive Evaluation and Benchmarking

To thoroughly establish OverHAuL’s effectiveness, extensive large-scale evaluation beyond the initial 10-project corpus is imperative. Applying the tool to repositories indexed in the clib package manager [93], which encompasses hundreds of C libraries, would test scalability and robustness across diverse real-world settings. Such a broad benchmark would also enable systematic comparisons against state-of-the-art automated fuzzing frameworks like OSS-Fuzz-Gen and AutoGen, elucidating OverHAuL’s relative strengths and identifying areas for improvement [7], [64].

Complementing broad benchmarking, detailed ablation and matrix studies dissecting the contributions of individual pipeline components and algorithmic choices will yield critical insights into what drives OverHAuL’s performance. Understanding the impact of each module will guide targeted optimizations and support evidence-based design decisions.

Furthermore, an economic analysis exploring resource consumption—such as API token usage and associated monetary costs—relative to fuzzing effectiveness would be valuable for assessing the practical viability of integrating LLM-based fuzz driver generation into continuous integration processes.

9.4 Practical Deployment and Community Engagement

From a usability perspective, embedding OverHAuL within a GitHub Actions workflow represents a practical and impactful enhancement, enabling seamless integration with developers’ existing toolchains and continuous integration pipelines. This would promote wider adoption by reducing barriers to entry and fostering real-time feedback during code development cycles.

1052 Additionally, establishing a mechanism to generate and submit automated pull requests (PRs) to the
1053 maintainers of fuzzed libraries—highlighting detected bugs and proposing patches—would not only
1054 validate OverHAuL’s findings but also contribute tangible improvements to open-source software
1055 quality. This collaborative feedback loop epitomizes the symbiosis between automated testing tools
1056 and the open-source community. As an initial step, developing targeted PRs for the projects where
1057 bugs were discovered during OverHAuL’s development would help facilitate practical follow-up
1058 and improvements.

1059 **10 Conclusion**

1060 Recap

1061 Performed a literature review of similar projects.

1062 Presented the algorithm *and* the implementation.

1063 generative AI disclaimer à la ACM?

Bibliography

- [1] Ada Developers. “Ada Reference Manual, 2022 Edition,” Ada Information Clearinghouse. (2022), [Online]. Available: https://www.adaic.org/resources/add_content/standards/22rm/html/RM-TTL.html.
- [2] Rust Project Developers. “Rust Programming Language.” (2025), [Online]. Available: <https://www.rust-lang.org/>.
- [3] G. J. Holzmann, “The Power of 10: Rules for Developing Safety-Critical Code,” Jun. 2006. [Online]. Available: <https://web.eecs.umich.edu/~imarkov/10rules.pdf>.
- [4] B. W. Kernighan and D. M. Ritchie, *The C Programming Language* (Prentice-Hall Software Series). Englewood Cliffs, NJ: Prentice-Hall, 1978, 228 pp., ISBN: 978-0-13-110163-0.
- [5] D. M. Ritchie, S. C. Johnson, M. E. Lesk, and B. W. Kernighan, “The C programming language,” *Bell Sys. Tech. J.*, vol. 57, no. 6, pp. 1991–2019, 1978. [Online]. Available: https://www.academia.edu/download/67840358/1978.07_Bell_System_Technical_Journal.pdf#page=85.
- [6] N. Perry, M. Srivastava, D. Kumar, and D. Boneh. “Do Users Write More Insecure Code with AI Assistants?” arXiv: [2211.03622](https://arxiv.org/abs/2211.03622). (Dec. 18, 2023), [Online]. Available: <http://arxiv.org/abs/2211.03622>, pre-published.
- [7] D. Liu, O. Chang, J. metzman, M. Sablotny, and M. Maruseac, *OSS-fuzz-gen: Automated fuzz target generation*, version <https://github.com/google/oss-fuzz-gen/tree/v1.0>, May 2024. [Online]. Available: <https://github.com/google/oss-fuzz-gen>.
- [8] B. Jeong, J. Jang, H. Yi, *et al.*, “UTopia: Automatic Generation of Fuzz Driver using Unit Tests,” in *2023 IEEE Symposium on Security and Privacy (SP)*, May 2023, pp. 2676–2692. DOI: [10.1109/SP46215.2023.10179394](https://doi.org/10.1109/SP46215.2023.10179394). [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/10179394>.
- [9] K. Ispoglou, D. Austin, V. Mohan, and M. Payer, “FuzzGen: Automatic fuzzer generation,” in *29th USENIX Security Symposium (USENIX Security 20)*, 2020, pp. 2271–2287. [Online]. Available: <https://www.usenix.org/conference/usenixsecurity20/presentation/ispoglou>.
- [10] D. Babić, S. Bucur, Y. Chen, *et al.*, “FUDGE: Fuzz driver generation at scale,” in *Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, Tallinn Estonia: ACM, Aug. 12, 2019, pp. 975–985, ISBN: 978-1-4503-5572-8. DOI: [10.1145/3338906.3340456](https://doi.org/10.1145/3338906.3340456). [Online]. Available: <https://dl.acm.org/doi/10.1145/3338906.3340456>.
- [11] V. J. M. Manes, H. Han, C. Han, *et al.* “The Art, Science, and Engineering of Fuzzing: A Survey.” arXiv: [1812.00140 \[cs\]](https://arxiv.org/abs/1812.00140). (Apr. 7, 2019), [Online]. Available: <http://arxiv.org/abs/1812.00140>, pre-published.
- [12] A. Takanen, J. DeMott, C. Miller, and A. Kettunen, *Fuzzing for Software Security Testing and Quality Assurance* (Information Security and Privacy Library), Second edition. Boston London Norwood, MA: Artech House, 2018, 1 p., ISBN: 978-1-63081-519-6.
- [13] M. Sutton, A. Greene, and P. Amini, *Fuzzing: Brute Force Vulnerability Discovery*. Upper Saddle River, NJ: Addison-Wesley, 2007, 543 pp., ISBN: 978-0-321-44611-4.
- [14] N. Rathaus and G. Evron, *Open Source Fuzzing Tools*, G. Evron, Ed. Burlington, MA: Syngress Pub, 2007, 199 pp., ISBN: 978-1-59749-195-2.

- [15] B. P. Miller, L. Fredriksen, and B. So, “An empirical study of the reliability of UNIX utilities,” *Commun. ACM*, vol. 33, no. 12, pp. 32–44, Dec. 1, 1990, issn: 0001-0782. DOI: [10.1145/96267.96279](https://doi.org/10.1145/96267.96279). [Online]. Available: <https://dl.acm.org/doi/10.1145/96267.96279>.
- [16] K. Serebryany, D. Bruening, A. Potapenko, and D. Vyukov, “AddressSanitizer: A fast address sanity checker,” in *2012 USENIX Annual Technical Conference (USENIX ATC 12)*, 2012, pp. 309–318. [Online]. Available: <https://www.usenix.org/conference/atc12/technical-sessions/presentation/serebryany>.
- [17] LLVM Project. “libFuzzer – a library for coverage-guided fuzz testing. – LLVM 21.0.0git documentation.” (2025), [Online]. Available: <https://llvm.org/docs/LibFuzzer.html>.
- [18] A. Rebert, S. K. Cha, T. Avgerinos, *et al.*, “Optimizing seed selection for fuzzing,” in *Proceedings of the 23rd USENIX Conference on Security Symposium*, ser. SEC’14, USA: USENIX Association, Aug. 20, 2014, pp. 861–875, ISBN: 978-1-931971-15-7.
- [19] OWASP Foundation. “Fuzzing.” (), [Online]. Available: <https://owasp.org/www-community/Fuzzing>.
- [20] Blackduck, Inc. “Heartbleed Bug.” (Mar. 7, 2025), [Online]. Available: <https://heartbleed.com/>.
- [21] CVE Program. “CVE - CVE-2014-0160.” (2014), [Online]. Available: <https://cve.mitre.org/cgi-bin/cvename.cgi?name=cve-2014-0160>.
- [22] The OpenSSL Project, *Openssl/openssl*, OpenSSL, Jul. 15, 2025. [Online]. Available: <https://github.com/openssl/openssl>.
- [23] D. Wheeler. “How to Prevent the next Heartbleed.” (2014), [Online]. Available: <https://dwheeler.com/essays/heartbleed.html>.
- [24] GNU Project. “Bash - GNU Project - Free Software Foundation.” (), [Online]. Available: <https://www.gnu.org/software/bash/>.
- [25] M. Zalewski. “American fuzzy lop.” (), [Online]. Available: <https://lcamtuf.coredump.cx/afl/>.
- [26] J. Saarinen. “Further flaws render Shellshock patch ineffective,” iTnews. (Sep. 29, 2014), [Online]. Available: <https://www.itnews.com.au/news/further-flaws-render-shellshock-patch-ineffective-396256>.
- [27] T. Simonite, “This Bot Hunts Software Bugs for the Pentagon,” *Wired*, Jun. 1, 2020, issn: 1059-1028. [Online]. Available: <https://www.wired.com/story/bot-hunts-software-bugs-pentagon/>.
- [28] M. Heuse, H. Eißfeldt, A. Fioraldi, and D. Maier, *AFL++*, version 4.00c, Jan. 2022. [Online]. Available: <https://github.com/AFLplusplus/AFLplusplus>.
- [29] LLVM Project. “The LLVM Compiler Infrastructure Project.” (2025), [Online]. Available: <https://llvm.org/>.
- [30] F. Bellard, P. Maydell, and QEMU Team, *QEMU*, version 10.0.2, May 29, 2025. [Online]. Available: <https://www.qemu.org/>.
- [31] Unicorn Engine, *Unicorn-engine/unicorn*, Unicorn Engine, Jul. 15, 2025. [Online]. Available: <https://github.com/unicorn-engine/unicorn>.
- [32] H. Li, “Language models: Past, present, and future,” *Commun. ACM*, vol. 65, no. 7, pp. 56–63, Jun. 21, 2022, issn: 0001-0782. DOI: [10.1145/3490443](https://doi.org/10.1145/3490443). [Online]. Available: <https://dl.acm.org/doi/10.1145/3490443>.
- [33] Z. Wang, Z. Chu, T. V. Doan, S. Ni, M. Yang, and W. Zhang, “History, development, and principles of large language models: An introductory survey,” *AI and Ethics*, vol. 5, no. 3, pp. 1955–1971, Jun. 1, 2025, issn: 2730-5961. DOI: [10.1007/s43681-024-00583-7](https://doi.org/10.1007/s43681-024-00583-7). [Online]. Available: <https://doi.org/10.1007/s43681-024-00583-7>.
- [34] D. Bahdanau, K. Cho, and Y. Bengio. “Neural Machine Translation by Jointly Learning to Align and Translate.” arXiv: [1409.0473](https://arxiv.org/abs/1409.0473) [cs, stat]. (May 19, 2016), [Online]. Available: <http://arxiv.org/abs/1409.0473>, pre-published.

- [35] A. Vaswani, N. Shazeer, N. Parmar, *et al.* “Attention Is All You Need.” arXiv: 1706.03762 [cs]. (Aug. 1, 2023), [Online]. Available: <http://arxiv.org/abs/1706.03762>, pre-published.
- [36] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” arXiv: 1810.04805 [cs]. (May 24, 2019), [Online]. Available: <http://arxiv.org/abs/1810.04805>, pre-published.
- [37] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, “Improving language understanding by generative pre-training,” 2018. [Online]. Available: <https://www.mikecaptain.com/resources/pdf/GPT-1.pdf>.
- [38] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, “Language models are unsupervised multitask learners,” *OpenAI blog*, vol. 1, no. 8, p. 9, 2019. [Online]. Available: <https://storage.prod.researchhub.com/uploads/papers/2020/06/01/language-models.pdf>.
- [39] T. B. Brown, B. Mann, N. Ryder, *et al.* “Language Models are Few-Shot Learners.” arXiv: 2005.14165 [cs]. (Jul. 22, 2020), [Online]. Available: <http://arxiv.org/abs/2005.14165>, pre-published.
- [40] OpenAI, J. Achiam, S. Adler, *et al.* “GPT-4 Technical Report.” arXiv: 2303.08774 [cs]. (Mar. 4, 2024), [Online]. Available: <http://arxiv.org/abs/2303.08774>, pre-published.
- [41] Anthropic. “Claude.” (2025), [Online]. Available: <https://claude.ai/new>.
- [42] DeepSeek-AI, D. Guo, D. Yang, *et al.* “DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning.” arXiv: 2501.12948 [cs]. (Jan. 22, 2025), [Online]. Available: <http://arxiv.org/abs/2501.12948>, pre-published.
- [43] A. Grattafiori, A. Dubey, A. Jauhri, *et al.* “The Llama 3 Herd of Models.” arXiv: 2407.21783 [cs]. (Nov. 23, 2024), [Online]. Available: <http://arxiv.org/abs/2407.21783>, pre-published.
- [44] OpenAI. “ChatGPT.” (2025), [Online]. Available: <https://chatgpt.com>.
- [45] Google. “Google Gemini,” Gemini. (2025), [Online]. Available: <https://gemini.google.com>.
- [46] J. Wei, X. Wang, D. Schuurmans, *et al.* “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models.” arXiv: 2201.11903 [cs]. (Jan. 10, 2023), [Online]. Available: <http://arxiv.org/abs/2201.11903>, pre-published.
- [47] S. Yao, D. Yu, J. Zhao, *et al.* “Tree of Thoughts: Deliberate Problem Solving with Large Language Models.” arXiv: 2305.10601 [cs]. (Dec. 3, 2023), [Online]. Available: <http://arxiv.org/abs/2305.10601>, pre-published.
- [48] P. Lewis, E. Perez, A. Piktus, *et al.* “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.” arXiv: 2005.11401 [cs]. (Apr. 12, 2021), [Online]. Available: <http://arxiv.org/abs/2005.11401>, pre-published.
- [49] S. Yao, J. Zhao, D. Yu, *et al.* “ReAct: Synergizing Reasoning and Acting in Language Models.” arXiv: 2210.03629. (Mar. 10, 2023), [Online]. Available: <http://arxiv.org/abs/2210.03629>, pre-published.
- [50] Anysphere. “Cursor - The AI Code Editor.” (2025), [Online]. Available: <https://cursor.com/>.
- [51] Microsoft. “GitHub Copilot · Your AI pair programmer,” GitHub. (2025), [Online]. Available: <https://github.com/features/copilot>.
- [52] E. Nijkamp, H. Hayashi, C. Xiong, S. Savarese, and Y. Zhou, “CodeGen2: Lessons for training llms on programming and natural languages,” *ICLR*, 2023.
- [53] E. Nijkamp, B. Pang, H. Hayashi, *et al.*, “CodeGen: An open large language model for code with multi-turn program synthesis,” *ICLR*, 2023.
- [54] OpenAI. “Introducing Codex.” (May 16, 2025), [Online]. Available: <https://openai.com/index/introducing-codex/>.

- [55] A. Sarkar and I. Drosos. “Vibe coding: Programming through conversation with artificial intelligence.” arXiv: 2506.23253 [cs]. (Jun. 29, 2025), [Online]. Available: <http://arxiv.org/abs/2506.23253>, pre-published.
- [56] A. Sheth, K. Roy, and M. Gaur. “Neurosymbolic AI – Why, What, and How.” arXiv: 2305.00813 [cs]. (May 1, 2023), [Online]. Available: <http://arxiv.org/abs/2305.00813>, pre-published.
- [57] A. d’Avila Garcez and L. C. Lamb. “Neurosymbolic AI: The 3rd Wave.” arXiv: 2012.05876. (Dec. 16, 2020), [Online]. Available: <http://arxiv.org/abs/2012.05876>, pre-published.
- [58] D. Ganguly, S. Iyengar, V. Chaudhary, and S. Kalyanaraman. “Proof of Thought : Neurosymbolic Program Synthesis allows Robust and Interpretable Reasoning.” arXiv: 2409.17270. (Sep. 25, 2024), [Online]. Available: <http://arxiv.org/abs/2409.17270>, pre-published.
- [59] M. Gaur and A. Sheth. “Building Trustworthy NeuroSymbolic AI Systems: Consistency, Reliability, Explainability, and Safety.” arXiv: 2312.06798. (Dec. 5, 2023), [Online]. Available: <http://arxiv.org/abs/2312.06798>, pre-published.
- [60] D. Tilwani, R. Venkataramanan, and A. P. Sheth. “Neurosymbolic AI approach to Attribution in Large Language Models.” arXiv: 2410.03726. (Sep. 30, 2024), [Online]. Available: <http://arxiv.org/abs/2410.03726>, pre-published.
- [61] M. K. Sarker, L. Zhou, A. Eberhart, and P. Hitzler, “Neuro-symbolic artificial intelligence: Current trends,” *AI Communications*, vol. 34, no. 3, pp. 197–209, Mar. 4, 2022, ISSN: 1875-8452, 0921-7126. DOI: 10.3233/aic-210084. [Online]. Available: <https://journals.sagepub.com/doi/full/10.3233/AIC-210084>.
- [62] H. Kautz, “The Third AI Summer,” Lecture, presented at the 34th Annual Meeting of the Association for the Advancement of Artificial Intelligence (New York, NY, USA), Feb. 10, 2020. [Online]. Available: https://www.youtube.com/watch?v=_cQITY0SPiw.
- [63] Z. Li, S. Dutta, and M. Naik. “IRIS: LLM-Assisted Static Analysis for Detecting Security Vulnerabilities.” arXiv: 2405.17238 [cs]. (Apr. 6, 2025), [Online]. Available: <http://arxiv.org/abs/2405.17238>, pre-published.
- [64] Y. Sun, “Automated Generation and Compilation of Fuzz Driver Based on Large Language Models,” in *Proceedings of the 2024 9th International Conference on Cyber Security and Information Engineering*, ser. ICCSIE ’24, New York, NY, USA: Association for Computing Machinery, Dec. 3, 2024, pp. 461–468, ISBN: 979-8-4007-1813-7. DOI: 10.1145/3689236.3689272. [Online]. Available: <https://doi.org/10.1145/3689236.3689272>.
- [65] D. Wang, G. Zhou, L. Chen, D. Li, and Y. Miao. “ProphetFuzz: Fully Automated Prediction and Fuzzing of High-Risk Option Combinations with Only Documentation via Large Language Model.” arXiv: 2409.00922 [cs]. (Sep. 1, 2024), [Online]. Available: <http://arxiv.org/abs/2409.00922>, pre-published.
- [66] H. Green and T. Avgerinos, “GraphFuzz: Library API fuzzing with lifetime-aware dataflow graphs,” in *Proceedings of the 44th International Conference on Software Engineering*, Pittsburgh Pennsylvania: ACM, May 21, 2022, pp. 1070–1081. DOI: 10.1145/3510003.3510228. [Online]. Available: <https://dl.acm.org/doi/10.1145/3510003.3510228>.
- [67] Y. Deng, C. S. Xia, C. Yang, S. D. Zhang, S. Yang, and L. Zhang. “Large Language Models are Edge-Case Fuzzers: Testing Deep Learning Libraries via FuzzGPT.” arXiv: 2304.02014 [cs]. (Apr. 4, 2023), [Online]. Available: <http://arxiv.org/abs/2304.02014>, pre-published.
- [68] Y. Deng, C. S. Xia, H. Peng, C. Yang, and L. Zhang, “Large Language Models Are Zero-Shot Fuzzers: Fuzzing Deep-Learning Libraries via Large Language Models,” in *Proceedings of the 32nd ACM SIGSOFT International Symposium on Software Testing and Analysis*, ser. ISSTA 2023, New York, NY, USA: Association for Computing Machinery, Jul. 13, 2023, pp. 423–435, ISBN: 979-8-4007-0221-1. DOI: 10.1145/3597926.3598067. [Online]. Available: <https://dl.acm.org/doi/10.1145/3597926.3598067>.

- [69] C. Cadar, D. Dunbar, and D. Engler, “KLEE: Unassisted and Automatic Generation of High-Coverage Tests for Complex Systems Programs,” presented at the USENIX Symposium on Operating Systems Design and Implementation, Dec. 8, 2008. [Online]. Available: <https://www.semanticscholar.org/paper/KLEE%3A-Unassisted-and-Automatic-Generation-of-Tests-Cadar-Dunbar/0b93657965e506dfbd56fbc1c1d4b9666b1d01c8>.
- [70] A. Arya, O. Chang, J. Metzman, K. Serebryany, and D. Liu, *OSS-Fuzz*, Apr. 8, 2025. [Online]. Available: <https://github.com/google/oss-fuzz>.
- [71] N. Sasirekha, A. Edwin Robert, and M. Hemalatha, “Program Slicing Techniques and its Applications,” *International Journal of Software Engineering & Applications*, vol. 2, no. 3, pp. 50–64, Jul. 31, 2011, ISSN: 09762221. DOI: 10.5121/ijsea.2011.2304. [Online]. Available: <http://www.airccse.org/journal/ijsea/papers/0711ijsea04.pdf>.
- [72] M. Zhang, J. Liu, F. Ma, H. Zhang, and Y. Jiang. “IntelliGen: Automatic Driver Synthesis for FuzzTesting.” arXiv: 2103.00862 [cs]. (Mar. 1, 2021), [Online]. Available: <http://arxiv.org/abs/2103.00862>, pre-published.
- [73] H. Xu, W. Ma, T. Zhou, *et al.* “CKGFuzzer: LLM-Based Fuzz Driver Generation Enhanced By Code Knowledge Graph.” arXiv: 2411.11532 [cs]. (Dec. 20, 2024), [Online]. Available: <http://arxiv.org/abs/2411.11532>, pre-published.
- [74] Y. Lyu, Y. Xie, P. Chen, and H. Chen. “Prompt Fuzzing for Fuzz Driver Generation.” arXiv: 2312.17677 [cs]. (May 29, 2024), [Online]. Available: <http://arxiv.org/abs/2312.17677>, pre-published.
- [75] OSS-Fuzz. “OSS-Fuzz Documentation,” OSS-Fuzz. (2025), [Online]. Available: <https://google.github.io/oss-fuzz/>.
- [76] Google, *Google/clusterfuzz*, Google, Apr. 9, 2025. [Online]. Available: <https://github.com/google/clusterfuzz>.
- [77] Google, *Google/fuzztest*, Google, Jul. 10, 2025. [Online]. Available: <https://github.com/google/fuzztest>.
- [78] Google, *Google/honggfuzz*, Google, Jul. 10, 2025. [Online]. Available: <https://github.com/google/honggfuzz>.
- [79] D. Liu, J. Metzman, O. Chang, and G. O. S. S. Team. “AI-Powered Fuzzing: Breaking the Bug Hunting Barrier,” Google Online Security Blog. (Aug. 16, 2023), [Online]. Available: <https://security.googleblog.com/2023/08/ai-powered-fuzzing-breaking-bug-hunting.html>.
- [80] Open Source Security Foundation (OpenSSF), *OSSF/fuzz-introspector*, Open Source Security Foundation (OpenSSF), Jun. 30, 2025. [Online]. Available: <https://github.com/ossf/fuzz-introspector>.
- [81] L. Thomason, *Leethomason/tinysql2*, Jul. 10, 2025. [Online]. Available: <https://github.com/leethomason/tinysql2>.
- [82] OSS-Fuzz Maintainers. “Introducing LLM-based harness synthesis for unfuzzed projects,” OSS-Fuzz blog. (May 27, 2024), [Online]. Available: <https://blog.oss-fuzz.com/posts/introducing-llm-based-harness-synthesis-for-unfuzzed-projects/>.
- [83] O. Khattab, A. Singhvi, P. Maheshwari, *et al.* “DSPy: Compiling Declarative Language Model Calls into Self-Improving Pipelines.” arXiv: 2310.03714 [cs]. (Oct. 5, 2023), [Online]. Available: <http://arxiv.org/abs/2310.03714>, pre-published.
- [84] L. Torvalds, *Git*, Apr. 7, 2005. [Online]. Available: <https://git-scm.com/>.
- [85] T. Mikolov, K. Chen, G. Corrado, and J. Dean. “Efficient Estimation of Word Representations in Vector Space.” arXiv: 1301.3781 [cs]. (Sep. 6, 2013), [Online]. Available: <http://arxiv.org/abs/1301.3781>, pre-published.
- [86] D. A. Wheeler. “Flawfinder Home Page,” Flawfinder. (), [Online]. Available: <https://dwheeler.com/flawfinder/>.

- [87] S. Zhao, Y. Yang, Z. Wang, Z. He, L. K. Qiu, and L. Qiu. “Retrieval Augmented Generation (RAG) and Beyond: A Comprehensive Survey on How to Make your LLMs use External Data More Wisely.” arXiv: 2409.14924 [cs]. (Sep. 23, 2024), [Online]. Available: <http://arxiv.org/abs/2409.14924>, pre-published.
- [88] M. Chen, J. Tworek, H. Jun, *et al.* “Evaluating Large Language Models Trained on Code.” arXiv: 2107.03374 [cs]. (Jul. 14, 2021), [Online]. Available: <http://arxiv.org/abs/2107.03374>, pre-published.
- [89] Python Software Foundation. “Venv — Creation of virtual environments,” Python documentation. (Jul. 17, 2025), [Online]. Available: <https://docs.python.org/3/library/venv.html>.
- [90] pip developers. “Pip documentation v25.1.1.” (2025), [Online]. Available: <https://pip.pypa.io/en/stable/>.
- [91] A. Cedilnik, B. Hoffman, B. King, K. Martin, and A. Neundorff, *CMake - Upgrade Your Software Build System*, 2000. [Online]. Available: <https://cmake.org/>.
- [92] S. I. Feldman, “Make — a program for maintaining computer programs,” *Software: Practice and Experience*, vol. 9, no. 4, pp. 255–265, 1979, ISSN: 1097-024X. DOI: 10.1002/spe.4380090402. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1002/spe.4380090402>.
- [93] Clibs Project. “Clib Packages,” GitHub. (2025), [Online]. Available: <https://github.com/clibs/club/wiki/Packages>.
- [94] Clibs Project, *Clibs/club*, clibs, Jul. 1, 2025. [Online]. Available: <https://github.com/clibs/club>.
- [95] O. I. Franksen, “Babbage and cryptography. Or, the mystery of Admiral Beaufort’s cipher,” *Mathematics and Computers in Simulation*, vol. 35, no. 4, pp. 327–367, 1993. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/037847549390063Z>.
- [96] F. Bacon, *Of the Proficiency and Advancement of Learning... Edited by the Rev. GW Kitchin*. Bell & Daldy, 1861.
- [97] T. Preston-Werner. “Semantic Versioning 2.0.0,” Semantic Versioning. (), [Online]. Available: <https://semver.org/>.
- [98] OpenAI Docs. “GPT-4.1 mini - Open AI API.” (2025), [Online]. Available: <https://platform.openai.com>.
- [99] D. Giannone. “Demystifying AI Agents: ReAct-Style Agents vs Agentic Workflows,” Medium. (Feb. 9, 2025), [Online]. Available: <https://medium.com/@DanGiannone/demystifying-ai-agents-react-style-agents-vs-agentic-workflows-cedca7e26471>.
- [100] T. He, *Sighingnow/libclang*, Jul. 3, 2025. [Online]. Available: <https://github.com/sighingnow/libclang>.
- [101] OpenAI Docs. “Text-embedding-3-small - OpenAI API.” (2025), [Online]. Available: <https://platform.openai.com>.
- [102] M. Douze, A. Guzhva, C. Deng, *et al.* “The Faiss library.” arXiv: 2401.08281 [cs]. (Feb. 11, 2025), [Online]. Available: <http://arxiv.org/abs/2401.08281>, pre-published.
- [103] Stanford NLP Team. “Signatures - DSPy Documentation.” (2025), [Online]. Available: <https://dspy.ai/learn/programming/signatures/>.
- [104] Stanford NLP Team. “ReAct - DSPy Documentation.” (2025), [Online]. Available: <https://dspy.ai/api/modules/ReAct/>.
- [105] H. Chase, *LangChain*, Oct. 2022. [Online]. Available: <https://github.com/langchain-ai/langchain>.
- [106] J. Liu, *LlamaIndex*, Nov. 2022. DOI: 10.5281/zenodo.1234. [Online]. Available: https://github.com/jerryliu/llama_index.
- [107] F. Both. “Why we no longer use LangChain for building our AI agents.” (2024), [Online]. Available: <https://octomind.dev/blog/why-we-no-longer-use-langchain-for-building-our-ai-agents>.
- [108] M. Woolf. “The Problem With LangChain.” (Jul. 14, 2023), [Online]. Available: <https://minimaxir.com/2023/07/langchain-problem/>.

- 1324 [109] Woyera. “6 Reasons why Langchain Sucks,” Medium. (Sep. 8, 2023), [Online]. Available: <https://medium.com/@woyera/6-reasons-why-langchain-sucks-b6c99c98efbe>.
1325
- 1326 [110] Astral, *Astral-sh/ruff*, Astral, Jul. 18, 2025. [Online]. Available: <https://github.com/astral-sh/ruff>.
1327
- 1328 [111] Astral, *Astral-sh/uv*, Astral, Jul. 18, 2025. [Online]. Available: <https://github.com/astral-sh/uv>.
1329
- 1330 [112] A. Cortesi, M. Hils, and T. Kriechbaumer, *Mitmproxy/pdoc*, mitmproxy, Jul. 18, 2025. [Online]. Available: <https://github.com/mitmproxy/pdoc>.
1331
- 1332 [113] PyTest Dev Team, *Pytest-dev/pytest*, pytest-dev, Jul. 18, 2025. [Online]. Available: <https://github.com/pytest-dev/pytest>.
1333
- 1334 [114] Python Software Foundation, *Python/mypy*, Python, Jul. 18, 2025. [Online]. Available: <https://github.com/python/mypy>.
1335
- 1336 [115] OpenAI Docs. “Model optimization - OpenAI API.” (2025), [Online]. Available: <https://platform.openai.com>.
1337
- 1338 [116] E. Martin, *Ninja-build/ninja*, ninja-build, Jul. 14, 2025. [Online]. Available: <https://github.com/ninja-build/ninja>.
1339
- 1340 [117] J. Pakkanen, *Mesonbuild/meson*, The Meson Build System, Jul. 14, 2025. [Online]. Available: <https://github.com/mesonbuild/meson>.
- [118] Google, *Google/atheris*, Google, Apr. 9, 2025. [Online]. Available: <https://github.com/google/atheris>.

A Abandoned Techniques

During its development, OverHAuL went through several iterations. A number of approaches were implemented and evaluated, with some being replaced for better alternatives. These are:

1. One-shot harness generation

Before the iterative feedback loop (Section 4.2.1) was implemented, OverHAuL attempted to operate in a straightforward pipeline, with just a “generator” agent being tasked to generate the harness. This meant that at either the compilation step or evaluation step, any failure resulted in the execution being terminated. This approach put too much responsibility in the response of a single LLM query, with results more often than not being unsatisfactory.

2. Chain-of-Thought LLM instances

The current implementation of ReAct agents has effectively supplanted the less effective Chain of Thought (COT) LLM modules [46]. This shift underscores a critical realization in the harness generation process: the primary challenge lies not in the creation of the harness itself, but rather in the necessity for real-time feedback during execution. This is the reason why first employing COT prompting offered limited observed improvements.

COT techniques are particularly advantageous when the task assigned to the LLM demands a more reflective, in-depth analysis. However, when it comes to tasks such as knowledge extraction from a codebase oracle and taking live feedback from the environment into consideration, ReAct agents demonstrate greater efficiency and effectiveness.

3. Source code concatenation

Initially, there was no implementation of a codebase oracle. Instead, the LLM agents operated with a Python string that contained a concatenation of all the collected source code. While this method proved effective for smaller and simpler projects, it encountered significant limitations when applied to more complex codebases. The primary challenge was the excessive consumption of the LLM’s context window, which hindered its ability to process and analyze larger codebases effectively. As a result, this approach became increasingly unsustainable as project complexity grew, underscoring the need for a more robust solution.

4. {index, read}_tool usage for ReAct agents

The predecessor of the oracle comprised a dual-system approach for code exploration, integrating the index_tool and the read_tool. The index_tool offered the LLM agent a structured JSON object that delineated the project’s architecture, including all relevant file paths. On the other hand, the read_tool required a file path as input and returned the file’s content, albeit truncated to a maximum of 4000 characters. While this methodology presented an improvement in scalability over earlier systems, several limitations persisted.

1375 Firstly, the LLM was constrained to searching through the codebase strictly in file-specific
1376 terms, which limited its efficacy in understanding the broader context of code relationships.
1377 Furthermore, the imposed character limit on the `read_tool` meant that certain portions of the
1378 codebase remained inaccessible, impeding the agent's analytical capabilities. Even if this
1379 character limit were to be lifted, the resultant output would still occupy a significant portion
1380 of the context window, particularly in larger and more intricate projects. As such, while this
1381 approach offered advancements in code exploration, it still fell short.

B DSPy Custom Signatures

```

1  class GenerateHarness(dspy.Signature):
2      """
3      You are an experienced C/C++ security testing engineer. You must write a
4      libFuzzer-compatible `int LLVMFuzzerTestOneInput(const uint8_t *data, size_t
5      size)` harness for a function of the given C project. Your goal is for the
6      harness to be ready for compilation and for it to find successfully a bug in
7      the function-under-test. Write verbose (within reason) and helpful comments
8      on each step/decision you take/make, especially if you use "weird" constants
9      or values that have something to do with the project.
10
11     You have access to a rag_tool, which contains a vector store of
12     function-level chunks of the project. Use it to write better harnesses. Keep
13     in mind that it can only reply with function chunks, do not ask it to
14     combine stuff.
15
16     The rag_tool does not store any information on which lines the functions
17     are. So do not ask questions based on lines.
18
19     Make sure that you only fuzz an existing function. You will know that a
20     functions exists when the rag_tool returns to you its signature and body.
21     """
22
23     static: str = dspy.InputField(
24         desc=""" Output of static analysis tools for the project. If you find it
25         helpful, write your harness so that it leverages some of the potential
26         vulnerabilities described below. """
27     )
28     new_harness: str = dspy.OutputField(
29         desc=""" C code for a libFuzzer-compatible harness. Output only the C
30         code, DO NOT format it in a markdown code cell with backticks, so
31         that it will be ready for compilation.
32
33         <important>
34
35         Add all the necessary includes, either project-specific or standard
36         libraries like <string.h>, <stdint.h> and <stdlib.h>. Also include any
37         header files that are part of the project and are probably useful. Most

```

```

38     projects have a header file with the same name as the project at the
39     root.
40
41     **The function to be fuzzed absolutely must be part of the source
42     code**, do not write a harness for your own functions or speculate about
43     existing ones. You must be sure that the function that is fuzzed exists
44     in the source code. Use your rag tool to query the source code.
45
46     Do not try to fuzz functions of the project that are static, since they
47     are only visible in the file that they were declared. Choose other
48     user-facing functions instead.
49
50     </important>
51
52     **Do not truncate the input to a smaller size than the original**,
53     e.g. for avoiding large stack usage or to avoid excessive buffers. Opt
54     to using the heap when possible to increase the chance of exposing
55     memory errors of the library, e.g. mmap instead of declaring
56     buf[1024]. Any edge cases should be handled by the library itself, not
57     the harness. On the other hand, do not write code that will most
58     probably crash irregardless of the library under test. The point is for
59     a function of the library under test to crash, not the harness
60     itself. Use and take advantage of any custom structs that the library
61     declares.
62
63     Do not copy function declarations inside the harness. The harness will
64     be compiled in the root directory of the project. """
65 )
66
67
68 class FixHarness(dspy.Signature):
69     """
70     You are an experienced C/C++ security testing engineer. Given a
71     libFuzzer-compatible harness that fails to compile and its compilation
72     errors, rewrite it so that it compiles successfully. Analyze the compilation
73     errors carefully and find the root causes. Add any missing #includes like
74     <string.h>, <stdint.h> and <stdlib.h> and #define required macros or
75     constants in the fuzz target. If needed, re-declare functions or struct
76     types. Add verbose comments to explain what you changed and why.
77     """
78
79     old_harness: str = dspy.InputField(desc="The harness to be fixed.")
80     error: str = dspy.InputField(desc="The compilation error of the harness.")
81     new_harness: str = dspy.OutputField(
82         desc="The newly created harness with the necessary modifications for

```

```

83         correct compilation."""
84     )
85
86
87     class ImproveHarness(dspy.Signature):
88         """
89         You are an experienced C/C++ security testing engineer. Given a
90         libFuzzer-compatible harness that does not find any bug/does not crash (even
91         after running for {Config.EXECUTION_TIMEOUT} seconds) or has memory leaks
92         (generates leak files), you are called to rewrite it and improve it so that
93         a bug can be found more easily and/or memory is managed correctly. Determine
94         the information you need to write an effective fuzz target and understand
95         constraints and edge cases in the source code to do it more
96         effectively. Reply only with the source code --- without backticks. Add
97         verbose comments to explain what you changed and why.
98         """
99
100         old_harness: str = dspy.InputField(
101             desc="The harness to be improved so it can find a bug more quickly."
102         )
103         output: str = dspy.InputField(desc="The output of the harness' execution.")
104         new_harness: str = dspy.OutputField(
105             desc="""The newly created harness with the necessary modifications for
106             quicker bug-finding. If the provided harness has unnecessary input
107             limitations regarding size or format etc., remove them.""
108         )

```