

SeedAIchemy: LLM-Driven Seed Corpus Generation for Fuzzing

Aidan Wen*, Norah A. Alzahrani^{†§}, Jingzhi Jiang^{*§}, Andrew Joe^{*§}, Karen Shieh^{*§}, Andy Zhang^{*§},
Basel Alomair[†], David Wagner*

*University of California, Berkeley

[†]King Abdulaziz City for Science and Technology

[‡]Humain, Saudi Arabia

[§]These authors contributed equally to this research

Abstract—We introduce SeedAIchemy, an automated LLM-driven corpus generation tool that makes it easier for developers to implement fuzzing effectively. SeedAIchemy consists of five modules which implement different approaches at collecting publicly available files from the internet. Four of the five modules use large language model (LLM) workflows to construct search terms designed to maximize corpus quality. Corpora generated by SeedAIchemy perform significantly better than a naive corpus and similarly to a manually-curated corpus on a diverse range of target programs and libraries.

Index Terms—fuzzing, large language models, computer security, software testing

I. INTRODUCTION

Fuzz testing is a widely used method for improving software security. One of the attractions of fuzz testing is that it is relatively easy to adopt. However, one road bump with adopting fuzz testing is that, for best effectiveness, developers must provide a corpus of seed files. Ideally, these seed files would include many tricky cases and difficult inputs, and would ensure good branch coverage of the targets. Constructing such a corpus can be difficult for developers who are newly adopting fuzz testing or do not have a strong security background.

One approach that some developers take when initially adopting fuzz testing is to use a naive corpus: e.g., a corpus containing only a single trivial file. Such a naive corpus is easy for developers to collect and makes adoption easy, but it also makes fuzzing less effective. In contrast, collecting a high-quality corpus can require considerable effort and some sophistication and knowledge of computer security. This leaves developers with a tradeoff between ease of adoption and effectiveness.

In this paper, we seek to support developers and make it easier to adopt fuzz testing by automating the process of collecting a corpus. We introduce SeedAIchemy, a tool that automates the entire process of corpus collection, so adopting fuzz testing can be easy and effective, even for developers who are not deeply knowledgeable about security. SeedAIchemy builds a corpus by collecting seed files from publicly available files on the internet. It guides its searches using a large language model (LLM) workflow that constructs search terms customized for the target being tested and aims to find files useful for fuzzing.

We show in our evaluation that SeedAIchemy is effective on a wide range of file types. In our experiments on the Magma benchmark, SeedAIchemy generates a corpus that makes fuzzing significantly more effective than a naive corpus: fuzzing with AFL++ for 24 hours triggers $1.7\times$ as many unique bugs with a SeedAIchemy-generated corpus compared to a naive corpus on average (naive corpus: 20 bugs; SeedAIchemy corpus: 34 bugs). Moreover, the SeedAIchemy corpus approaches the high-quality manually-curated corpus provided with the Magma benchmark: fuzzing with the SeedAIchemy finds 97% of the bugs revealed by the Magma corpus on average (SeedAIchemy: 34 bugs, Magma corpus: 35 bugs). Of course, SeedAIchemy requires no human effort, whereas collecting the Magma corpus required considerable human curation.

Overall, we believe this demonstrates that LLM agents can be helpful for reducing barriers to adoption of fuzz testing, and specifically, for building a useful fuzzing corpus. By automatically generating seed corpora with similar quality to manually-curated corpora, SeedAIchemy provides an excellent starting point for developers who are not security experts or do not have resources for extended manual curation. In the rest of this paper, we explain the techniques used to build SeedAIchemy and we evaluate its effectiveness with detailed experiments.

II. PROBLEM STATEMENT

Fuzz testing (fuzzing) is a well-established and widely used technique for discovering bugs and security vulnerabilities in software [1]. Fuzzing software works by taking in a target intended to be tested, also known as the *fuzz target*, and running it repeatedly with a series of random inputs, with the goal of triggering unexpected behaviors [1], [2]. Over the past decade, fuzzing has uncovered tens of thousands of vulnerabilities in critical real-world software. For example, Google’s OSS-Fuzz, a fuzzing platform which contains over a thousand open-source projects, has uncovered more than 10,000 vulnerabilities and 36,000 bugs as of August 2023 [3]. As another example, the Syzbot fuzzing project has discovered over 5570 bugs in the Linux Kernel from 2018-2024, leading to over 4604 bug fixes [4].

Fuzzers generate random inputs by modifying a set of initial valid inputs known as the *seed corpus* [1]. A diverse and representative corpus can help fuzzers reach deeper executions and find subtle bugs [5], [6]. However, due to the difficulty and time involved, many projects in practice pay little attention to the construction of the seed corpus and use either a single seed file or none at all [2], [6].

Recent progress in large language models (LLMs) presents a new opportunity to make fuzzing easier to adopt. LLMs can be leveraged to produce creative, context-guided search queries to systematically mine large-scale information sources such as web search engines. Even if LLMs hallucinate or generate poor quality search queries, the worst case is that the generated seed files provide no additional coverage and are discarded. We show that LLMs can be similarly applied to seed corpus generation in place of human computer security experts.

We have developed SeedAchemy, an easy-to-use corpus building tool that automatically generates seed files using LLM-generated search queries. SeedAchemy generates corpora by combining the results of five modules which implement different strategies for finding seed files. By reducing the time and effort needed to construct a seed corpus, we hope more developers will use fuzzing in their development process and do so with higher effectiveness.

III. DESIGN

As discussed in Section II, a well constructed corpus can greatly improve the effectiveness of fuzzing [2]. Our goal in developing SeedAchemy was to provide a way to automatically generate seed corpora with sufficient size, diversity, and relevance for effective fuzzing on a wide range of fuzz targets.

A. General approach

To produce a large and varied seed corpus, SeedAchemy delegates seed file generation to five modules, each using a different seed gathering technique, as outlined in Fig. 1 and 2. To focus on seed files that are relevant to the fuzz target, the user provides a file extension: for instance, the user might specify “pdf” if they are fuzzing a PDF viewer. The techniques detailed in Sections B–F are then executed in parallel. Each module generates a subcorpus of seed files with the target file extension. The resulting subcorpora are combined and deduplicated, and files larger than the specified maximum file size (default: 1 MB) are filtered out to improve fuzzing efficiency.

Finally, SeedAchemy selects a subset of seed files for the final corpus. If there are more than 40,000 candidates, it selects the smallest files first and ensures a balance across modules. This corpus could be used directly for fuzzing, but for better effectiveness, we recommend using a corpus minimization tool such as afl-cmin. We provide further details in Section G.

Each module generates search terms using an LLM, then queries a particular search engine. The GitHub and feature-driven web search modules use GPT-4.1; web search and bug tracker search use GPT-4o. Common Crawl search does not use an LLM; it deterministically searches archived web content

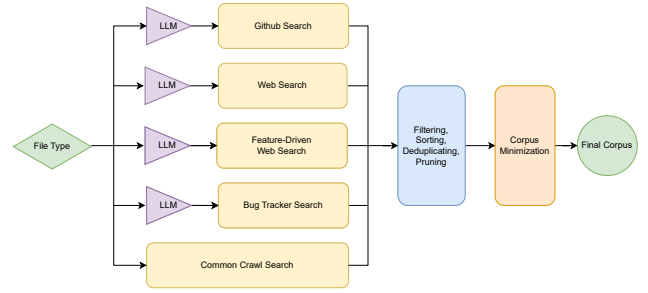


Fig. 1. Architecture of SeedAchemy. SeedAchemy combines corpora from each submodule, then applies minimization techniques to reduce the size of the final corpus.

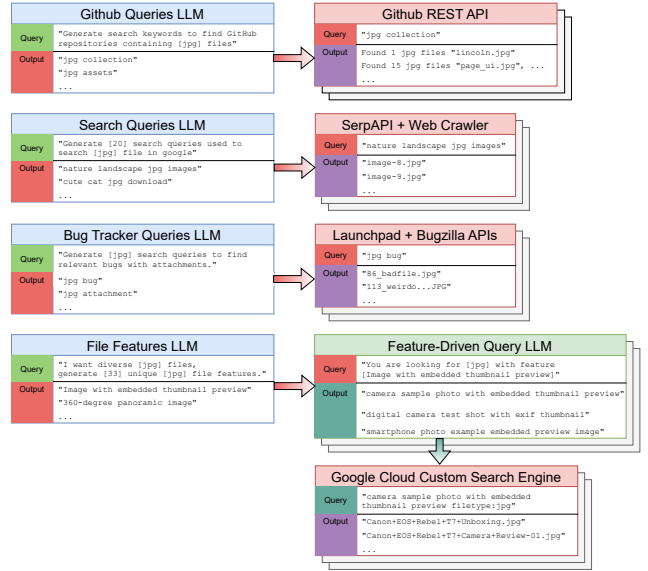


Fig. 2. Example queries and outputs of a single run of SeedAchemy for the JPG datatype. To preserve space, LLM prompts are simplified versions of the real ones used in SeedAchemy. Only modules that used an LLM to generate search queries are shown.

for files with the specified file extension. Each module runs for at most 1 hour.

B. GitHub Search

The GitHub Search module uses LLM-generated search queries to find seed files on GitHub. We prompt GPT-4.1 to generate a set of 50 diverse search queries that look for repositories that are broadly used or directly relevant to software testing. For each search query, we find the top 10 matches and shallow-clone the associated repository to minimize bandwidth and storage usage. Cloning repositories rather than downloading individual files reduces the number of API calls and allows users to stay within GitHub’s rate limits. For each cloned file, we check the file extension and magic number to determine whether the file is the correct format. Files satisfying either condition are retained.

C. Web search

The Web Search module looks for seed files using a general search on the internet. We ask GPT-4o to generate a mix of 20 distinct queries that capture the characteristics of the desired file type (e.g. “small PNG clipart” or “PDF ebooks”) as well as queries that explicitly look for existing seed files (e.g. “fuzzing seed JPG files”). We then search Google with those queries, download each search result, and follow links recursively to collect all files matching the specified extension. We use a Scrapy spider [7] to process multiple webpages in parallel, which speeds up the scraping process and allows us to find more files within the 1 hour timeout period.

D. Feature-driven Web Search

Another way to increase the diversity of a seed corpus is to use seed files that contain different features or structures. For example, a PDF file can contain images, Javascript code, or forms, and a high quality corpus should include as many of these features as possible. This module aims to find seed files that cover a wide range of available features for a given file type.

The Feature-driven Web Search module uses GPT-4.1 to identify features relevant to a file type and generate corresponding search queries. We use one-shot prompting to instruct GPT-4.1 to produce 33 distinct feature descriptors (e.g., “PNG with a visible watermark”, “PNG containing embedded GPS/location metadata”). We then ask it to expand each feature keyword into three complete search queries (e.g., “sample stock photo png with watermark download”, “technical diagram png export with dpi metadata”). We then search Google with each query, fetch the top 10 results, and crawl them recursively with a Scrapy spider without depth restriction [7]. Downloaded files are included in the final corpus if they match the correct file extension or magic number [8].

E. Bug Tracker Search

Bug trackers provide a good source of diverse, high-quality input files [9]. Many bug reports include a reproducing test case as an attachment alongside the bug description and commit. When manually crafting seed corpora, security experts often use files that have triggered bugs in the past [10]. This is the motivation behind the Bug Tracker Search module, which searches popular bug tracker and reporting sites for files that have triggered bugs in the past.

The Bug Tracker Search module works by prompting an LLM to generate search queries to find bug reports that are relevant to the target file type and have associated attachments. It uses the generated queries to search Ubuntu Launchpad and Red Hat Bugzilla using the launchpadlib API and the Bugzilla REST API. The search results are then crawled to collect attachments that match the target file type by checking for matches with the file extension or the magic number. We focus on these two bug trackers because of their popularity, extensive use in open-source ecosystems, and comprehensive API support.

F. Common Crawl Search

Common Crawl [11] is an open source project started in 2007 that crawls the web and public archive data. Each crawl retrieves tens to hundreds of terabytes of web data, making it one of the largest openly available projects of web crawl data. The Common Crawl Search module searches the Common Crawl archive hosted on Amazon Web Services [12], specifically, the *CC-MAIN-2025-08* crawl, by performing an exact match search on the *content_mime_type* metadata. The *CC-MAIN-2025-08* archive contains files with over 2,000 different types, providing a rich resource for fuzzing seeds of various formats.

G. Corpus minimization

SeedAIchemy can find up to several hundred thousand files depending on the prevalence of the file type. However, fuzzing performance is best when the seed corpus is small because smaller corpora increase execution speed and allow fuzzers to quickly discard uninteresting inputs [6]. One way to reduce the size of a corpus is to use a minimization tool, which finds the smallest subset of seed files that maintains the same code coverage as the original corpus [13]. Thus, the final step of SeedAIchemy involves running the corpus through afl-cmin, which is a well regarded minimization tool that has been shown to improve fuzzing performance [6], [13].

afl-cmin minimizes corpora with edge coverage as its metric for fuzzing performance [14]. Measuring coverage can be slow, so afl-cmin can be slow. To ensure that afl-cmin completes in a reasonable amount of time, we discard all files larger than 1 MB and limit the output corpus to at most 40,000 files by preferentially selecting the 40,000 smallest files before minimization.

H. Files without extensions

SeedAIchemy allows users to search for files using a file description or phrase (e.g., “php_serialize”). This feature works similarly to the file extension based corpus generation, with two modifications: LLMs generate search queries using a file description rather than a file extension, and file extension and magic number filtering are disabled. This feature allows SeedAIchemy to generate seed corpora for fuzz targets whose input files do not have any standard file extensions (e.g., binary blobs), which makes it useful for a wider range of targets.

IV. EVALUATION

We evaluate SeedAIchemy on the Magma fuzzing benchmark. Magma provides a high-quality manually-curated corpus for each target, mostly gathered from the original library repositories and other sources. We compare fuzzing with a SeedAIchemy-generated corpus to fuzzing with a naive corpus, the Magma corpus, and a corpus generated by G²FUZZ, an existing LLM-based corpus generation tool [15].

We follow robust evaluation and benchmarking principles proposed in prior work [2], [16], [17]. For example, Klees et al. suggest adhering to the following guidelines:

- Run multiple fuzzing trials and evaluate results using statistical tests.
- Use fuzzing timeouts of at least 24 hours.
- Have a diverse set of benchmark programs.
- Test multiple seed corpora.
- Measure performance in terms of known bugs.

The first two conditions ensure that the evaluation procedure accounts for the randomness of fuzzing, the next two ensure that the proposed fuzzing technique is effective in a diverse range of fuzzing conditions, and the last condition ensures that the evaluation uses a concrete metric for fuzzing performance. Our evaluation follows these guidelines by carrying out the following:

- Run 10 trials for each corpus and fuzz target pair. We compute a 95% confidence interval for each aggregated statistic and compare results using Wilcoxon signed-rank tests.
- Use a fuzzing timeout of 24 hours.
- Test fuzz targets from the Magma benchmark.
- Compare SeedAIChecky to three corpora of differing quality.
- Measure performance in terms of bugs reached, bugs triggered, and code coverage.

A. Experimental method

Our evaluation was conducted using AFL++, a widely used coverage-guided mutation-based greybox fuzzer [18]. We measured the performance of AFL++ running for 24 hours on fuzz targets seeded with four different corpora as shown in Table I:

- **Naive**: a corpus consisting of a single file randomly sampled from the Magma benchmark corpus. This represents a corpus that a developer with minimal fuzzing experience could possibly use.
- **G²FUZZ**: a corpus generated using the G²FUZZ Input Generator Synthesis subpart, an LLM-based corpus generation tool proposed by Zhang et al. [15]. G²FUZZ generates seed files by asking an LLM to produce a Python script that generates seed files with the correct format. SeedAIChecky differs from this approach by asking the LLM to generate search queries rather than the file itself.
- **SeedAIChecky**: our corpus generated with a maximum runtime of 1 hour, then minimized with afl-cmin using a timeout of 1 second per seed file.
- **Magma**: the corpus from the Magma benchmark [19], representing a corpus manually crafted by a domain expert with knowledge of the fuzz target.

Each corpus was evaluated on fuzz targets from Magma v1.2 [19], a ground-truth fuzzing benchmark that forward-ports real bugs into modern versions of widely used software programs. Because Magma contains known bugs, it measures bug coverage without stack hashes or other bug deduplication techniques that are known to be unreliable [2]. Magma includes nine real-world targets spanning diverse domains and input formats shown in Table II. It measures performance in terms of bugs

TABLE I
EVALUATION CORPORA

Corpus	Description
Naive	Randomly sampled file from the Magma corpus
G ² FUZZ	LLM generated corpus [15]
SeedAIChecky	SeedAIChecky corpus minimized with afl-cmin
Magma	Magma benchmark corpus

TABLE II
MAGMA TARGETS AND INPUT FILE TYPES

Libraries	Programs	File type(s)
libpng	libpng_read_fuzzer	.png
libsndfile	sndfile_fuzzer	audio files
libtiff	tiff_read_rgba_fuzzer	.tiff
	tiffcp	.tiff
libxml2	libxml2_xml_read_memory_fuzzer	.xml
	xmllint	.xml
lua	lua	.lua
openssl	asn1	binary blobs
	asn1parse	binary blobs
	bignum	binary blobs
	client	binary blobs
	server	binary blobs
	x509	binary blobs
php	exif	.jpeg, .tiff
	json	.json
	parser	.php
	unserialize	serialized php
poppler	pdf_fuzzer	.pdf
	pdfimages	.pdf
	pdftoppm	.pdf
sqlite3	sqlite3_fuzzer	.sql

reached and bugs triggered; a bug is “reached” if the associated lines of code are executed and “triggered” if the conditions for revealing the bug are satisfied. To support fuzzing with custom corpora, we ran our experiments on a fork of Magma commit 75d1ae7 that contains our own modifications.

For each fuzz target and corpus pair, we performed 10 independent fuzzing trials and evaluated performance using three metrics: (a) average number of bugs triggered, (b) average number of bugs reached, and (c) average code coverage as measured by AFL++ bitmaps. For each bug, we also measured the average time to reach and trigger that bug. Because LLMs produce nondeterministic output, we generated a new corpus for each SeedAIChecky fuzzing trial. To ensure consistency across trials, all experiments were run on a machine with two Intel(R) Xeon(R) Gold 5320 CPUs (26 physical cores each) and 504 GB of RAM, running Ubuntu 20.04.6 LTS.

B. Cost analysis

SeedAIChecky is inexpensive to deploy, costing only a few cents for each generated corpus. A 1-hour trial consumes approximately 0.5k input tokens and 0.7k output tokens on GPT-4.1, and consumes 12.5k input tokens and 2.3k output tokens on GPT-4o. This costs roughly \$0.06 based on pricing at the time of evaluation (GPT-4.1: \$2.00 per million uncached input tokens and \$8.00 per million output tokens; GPT-4o: \$2.50 per million uncached input tokens and \$10.00 per

TABLE III
AVERAGE BUGS REACHED, BUGS TRIGGERED, AND NORMALIZED CODE
COVERAGE FOR EVALUATED CORPORA

Corpus	Reached	Triggered	Coverage
Naive	44.5	19.9	0.68
G ² FUZZ	55.8	25.1	0.76
SeedAIchemy	75.9	33.9	0.91
Magma	76.7	35.3	1.00

million output tokens) [20]. The cost of AWS Athena/S3 is harder to estimate as it depends on the input file extension. However, based on pricing at the time of evaluation [21] [22], generating a corpus costs only a few cents. SeedAIchemy is compatible with the free tiers of Google Search API and SerpAPI, and the cost of CPU and GPU usage, internet, and storage is negligible at this scale.

V. RESULTS

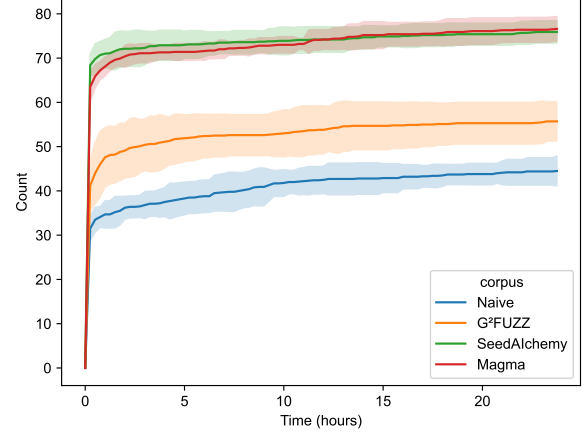
The corpora produced by SeedAIchemy outperform naive and G²FUZZ corpora on the Magma benchmark according to all three evaluation metrics. Compared to the Magma corpora, SeedAIchemy reaches 99% of the bugs that Magma does, triggers 96% as many bugs, and attains 91% of Magma’s code coverage (Table III), showing that corpora produced by SeedAIchemy are of similar quality to a manually constructed corpus.

A. Bug finding ability

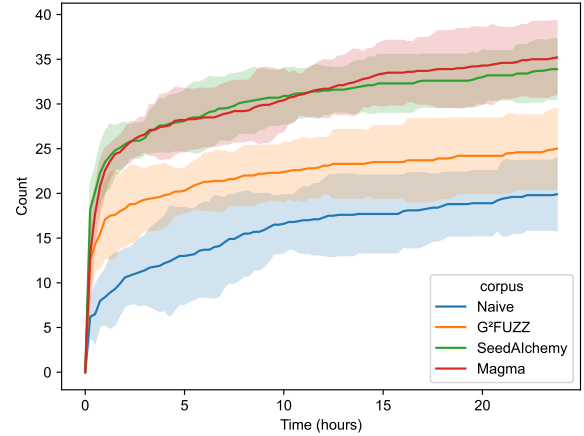
Fig. 3a and 3b show the total number of bugs reached and bugs triggered for each corpus averaged across 10 trials. SeedAIchemy reached and triggered $1.7\times$ as many bugs as a naive corpus. Also, it reached and triggered $1.4\times$ as many bugs as the G²FUZZ corpus. It reached 99% as many bugs as Magma and triggered 96% as many bugs. Overall, after 24 hours of fuzzing, SeedAIchemy and Magma achieved similar numbers of bugs reached and bugs triggered while the Naive and G²FUZZ corpora performed significantly worse.

Fig. 4a and 4b show the fuzzing performance of each corpus separated by targets. SeedAIchemy reached more bugs in 78% (7/9) of the targets and triggered more bugs in 89% (8/9) of the targets compared to the Naive and G²FUZZ corpora. Compared to the Magma corpus, SeedAIchemy had an equal or higher number of bugs reached and bugs triggered in over 50% of the targets. Overall, SeedAIchemy performed similarly to the Magma corpus on all targets except for openssl and libxml2 in the number of bugs reached and bugs triggered respectively.

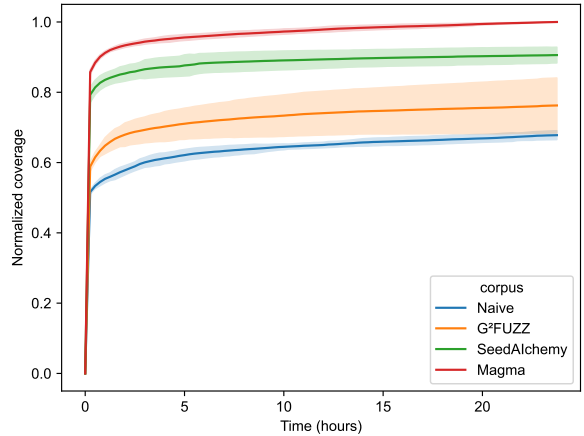
Fig. 5 and 6 show the average time to reach and time to trigger a bug for each corpus. For clarity of presentation, the reported times were rounded to the nearest second, minute, or hour, as appropriate. Across the set of bugs triggered by both the Magma and SeedAIchemy corpora, SeedAIchemy had the fastest triggering time in 55.0% (22/40) of bugs. Across the set of bugs reached by both corpora, SeedAIchemy was fastest at reaching 52.0% (39/75) of the bugs.



(a) Bugs Reached

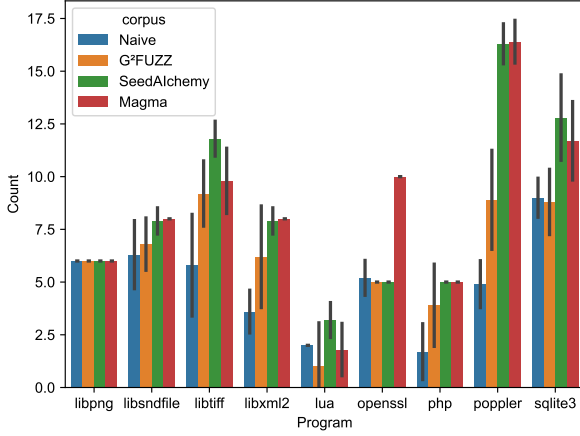


(b) Bugs Triggered

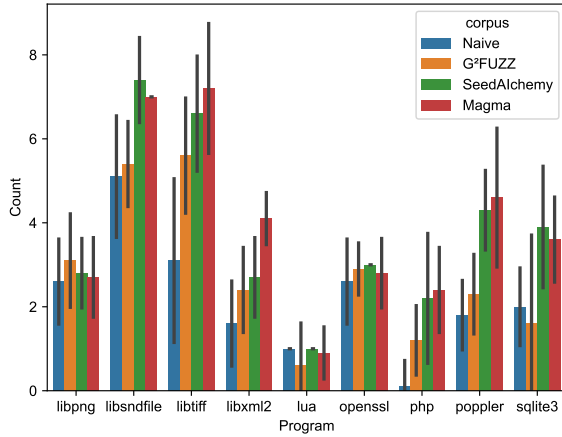


(c) Code Coverage (normalized)

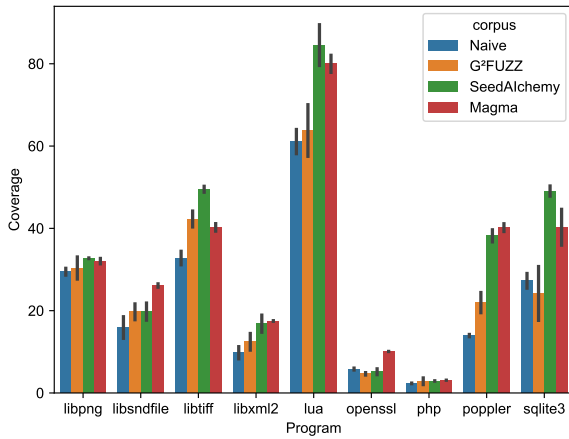
Fig. 3. Bugs reached, bugs triggered, and normalized code coverage averaged over 10 trials. Error bands show the 95% confidence interval. Coverage is normalized by the 24-hour coverage of each Magma fuzz target.



(a) Bugs Reached



(b) Bugs Triggered



(c) Code Coverage

Fig. 4. Bugs reached, bugs triggered, and code coverage for each target averaged over 10 trials. Error bars show the 95% confidence interval.

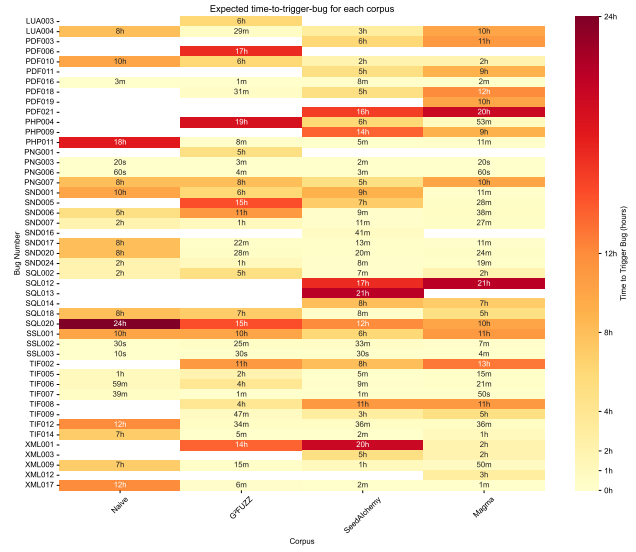


Fig. 5. Time to trigger bug averaged across 10 trials.

TABLE IV
WILCOXON SIGNED-RANK TESTS COMPARING SEEDAICHEMY TO OTHER CORPORA

Comparison	Bug Type	p-value
SeedAIchemy vs. Naive	Reached	0.00116
	Triggered	0.00110
SeedAIchemy vs. G ² FUZZ	Reached	0.00094
	Triggered	0.00789
SeedAIchemy vs. Magma	Reached	0.42525
	Triggered	0.46874

Based on Wilcoxon signed-rank tests¹, SeedAIchemy significantly outperformed the Naive and G²FUZZ corpora on bugs reached and triggered ($p < 0.01$ for all comparisons) while showing no significant difference from Magma corpora (reached: $p = 0.4253$; triggered: $p = 0.4687$). Table IV summarizes the Wilcoxon signed-rank test results comparing SeedAIchemy with other corpora.

B. Code coverage

We evaluated the average code coverage for each corpus as shown in Fig. 3c and 4c. To bring coverage scores into a common range, we normalized the data by dividing each coverage data point by the final 24 hour coverage of the associated Magma corpus coverage value so that values higher than 1 represent better coverage than the Magma corpus and values lower than 1 represent worse coverage. Overall, code coverage with the SeedAIchemy corpus was significantly better than the naive and G²FUZZ corpus, and not quite as good as the Magma corpus (Fig. 3c).

¹For each target, we computed the average number of bugs reached/triggered across 10 trials and paired these averages between corpora. Then, we applied the one-sided Wilcoxon signed-rank test to the pairs using the null hypothesis that $\text{median}(X) = \text{median}(Y)$, where X and Y denote the paired number of bugs reached/triggered of two corpora.

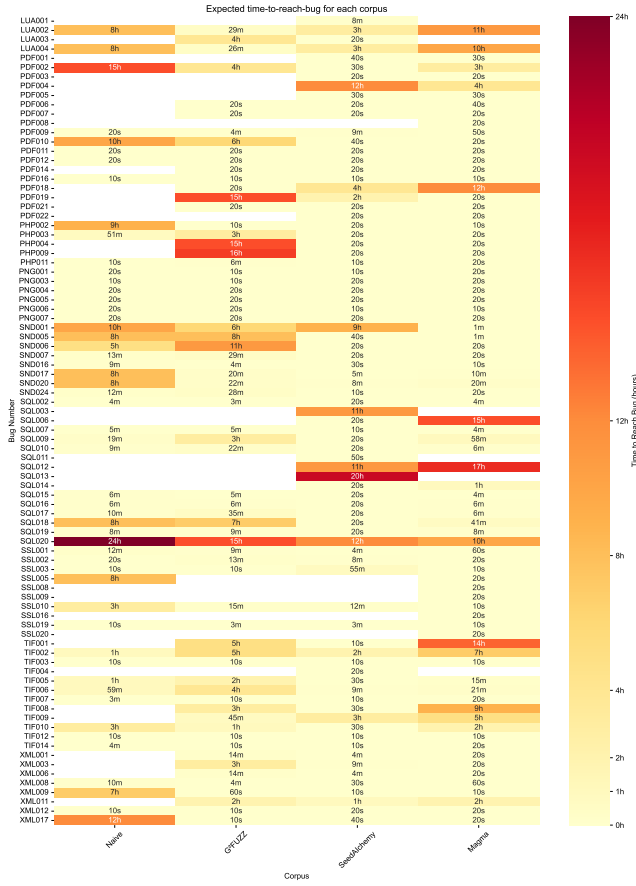


Fig. 6. Time to reach bug averaged across 10 trials.

Looking at individual targets in Fig. 4c, SeedAIchemy outperformed Magma’s corpus on 44% (4/9) of the targets and underperformed on 56% (5/9) of the targets. Compared to the Naive and G²FUZZ corpora, SeedAIchemy had higher or equivalent coverage on all targets except openssl, where the Naive corpus slightly outperformed the G²FUZZ and SeedAIchemy corpora.

C. Discussion

Our results show that SeedAIchemy generates high-quality corpora with fuzzing performance that is comparable to a manually constructed one. Corpora generated by SeedAIchemy significantly outperform Naive corpora and G²FUZZ corpora in bugs reached, bugs triggered, and code coverage. Our results demonstrate that SeedAIchemy is a suitable replacement for a manually constructed corpus on a wide variety of fuzz targets.

The fact that SeedAIchemy’s corpora outperform those generated by G²FUZZ suggests that fuzzing performance is better when LLMs are prompted to generate search queries for seed files rather than generate seed files directly. Searching for files using search queries allows SeedAIchemy to take advantage of the diversity of files on the internet, which is hard for an LLM to replicate on its own. SeedAIchemy’s approach is also more cost effective because it only relies

on a few LLM queries to produce a fully functioning seed corpus. By automating the process of seed corpus generation, SeedAIchemy reduces the time and effort it takes to construct a high quality seed corpus for fuzzing.

One potential concern in evaluating on benchmarks like Magma constructed from past bugs is the possibility of data leakage. Before fuzzing on the generated corpus, we filtered out any seeds that directly cause a crash to the target program, which reduces the risk of inadvertent leakage. Nevertheless, we acknowledge that it is difficult to rule out entirely such “time-traveling” advantage for SeedAIchemy, since any corpus generation method that relies on existing seed repositories could include posterior inputs that trigger prior bugs.

VI. RELATED WORK

Fuzzing has emerged as one of the most effective automated techniques for discovering software vulnerabilities. It works by executing programs with large volumes of generated or mutated inputs and monitoring their behavior for crashes or anomalies [1], [23]. The quality of the seed corpus is a critical factor in fuzzing effectiveness [6], [10], [24].

A. Corpus Generation and Optimization

Traditionally, fuzzing has relied on domain experts manually collecting representative inputs [25]. Some studies evaluate fuzzing with a single seed file, which measures the effectiveness of fuzzing when little attention is put into corpus construction. Other studies use a corpus of files collected by fuzzing teams or crawling public repositories and the web [26].

Constructing a comprehensive corpus remains challenging and time-consuming for developers [27], [28]. Our work automates corpus construction by finding existing corpora from open-source projects and supplementing them with a wide variety of seed files scraped from the web. We apply corpus minimization, which improves fuzzing efficiency by removing redundant seed files from the corpus [6]. With this approach, we aim to lower the barrier to corpus generation [29], while improving the overall effectiveness of fuzzing.

B. AI in Corpus Generation

AI and machine learning have been previously applied to corpus and seed generation to address the limitations of manual approaches. Studies have used generative models to automatically produce syntactically valid C programs for fuzzing compilers [30] or augment corpora [31], [32]. However, these approaches are typically limited in their ability to support binary file formats and long-tail less widespread file formats.

Recent studies have focused on leveraging LLMs for seed corpus generation. [33] uses LLMs to generate and select corpora tailored to known vulnerabilities; [34] uses LLMs to generate a script that produces seeds and iteratively improves upon them based on the code coverage data; similarly, [15] uses LLMs to write code to generate seed files, making use of existing libraries for these file formats; [35] is an open source tool that uses LLMs to generate seed files for text files in widely used formats like JSON or XML; [36] uses

LLMs for target-specific fuzzing seed generation to identify vulnerabilities efficiently.

We explore a different approach: rather than using AI to generate valid input files, we collect input files from the web by using AI to construct intelligent web search terms that help us find a diverse set of seed files. Our strategy is less computationally intensive than training generative models, remains portable across domains without retraining, works for non-textual file formats, and allows practitioners to inspect and refine the generated queries for greater transparency and control [37].

VII. FUTURE WORK

Although SeedAIchemy improves the process of generating seed corpora, there are some areas for future improvement:

- **Testing multiple fuzzers:** Our evaluation was conducted exclusively with AFL++, which represents one of many widely used fuzzers. Future work could investigate the effectiveness of SeedAIchemy across different fuzzers such as libFuzzer [38] or Honggfuzz [39].
- **Testing individual module contributions:** We have not evaluated the contributions of each individual module on the performance of the overall corpus. Future work could investigate whether certain modules contribute more to bug discovery and code coverage than others.

VIII. CONCLUSION

In this paper, we present SeedAIchemy, an LLM-driven seed corpus generation tool for fuzzing. The corpora produced by SeedAIchemy perform significantly better than naive corpora and similarly to expert-curated corpora. SeedAIchemy also outperforms corpora generated by G²FUZZ, an existing LLM-based corpus generation tool that prompts LLMs to generate seed files directly. SeedAIchemy makes the fuzzing process cheaper and more accessible to developers who are trying to incorporate fuzzing into their development cycle by providing an automated tool for generating high-quality corpora.

ACKNOWLEDGMENT

This work was supported by the KACST-UC Berkeley Center of Excellence for Secure Computing and the NSF ACTION center through NSF grant 2229876.

REFERENCES

- [1] V. J. Manès, H. Han, C. Han, S. K. Cha, M. Egele, E. J. Schwartz, and M. Woo, "The art, science, and engineering of fuzzing: A survey," *IEEE Transactions on Software Engineering*, vol. 47, no. 11, pp. 2312–2331, 2019.
- [2] G. Klees, A. Ruef, B. Cooper, S. Wei, and M. Hicks, "Evaluating fuzz testing," in *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security (CCS '18)*. Toronto, ON, Canada: ACM, 2018.
- [3] Google, "OSS-Fuzz," <https://google.github.io/oss-fuzz/>, 2023, accessed: Aug. 16, 2025.
- [4] J. Bursey, A. Amiri Sani, and Z. Qian, "SyzRetrospector: A Large-Scale Retrospective Study of Syzbot," *arXiv preprint arXiv:2401.11642*, 2024. [Online]. Available: <https://arxiv.org/abs/2401.11642>
- [5] M. Jurczyk, "Effective file format fuzzing: Thoughts, techniques, and results," in *Black Hat Europe Conference*, London, UK, Nov. 2016, accessed: Aug. 26, 2025. [Online]. Available: <https://j00ru.vexillum.org/slides/2016/blackhat.pdf>
- [6] A. Herrera, H. Gunadi, S. Magrath, M. Norrish, M. Payer, and A. L. Hosking, "Seed selection for successful fuzzing," in *Proceedings of the 30th ACM SIGSOFT International Symposium on Software Testing and Analysis (ISSTA '21)*. Virtual, Denmark: ACM, 2021.
- [7] Scrapinghub, "Scrapy," <https://scrapy.org/>, 2025.
- [8] Qti3e, "extensions.json: File extension to MIME type mapping data," <https://gist.github.com/Qti3e/6341245314bf3513abb080677cd1c93b>, 2017.
- [9] J. Acharya and G. Ginde, "BugsRepo: A Comprehensive Curated Dataset of Bug Reports, Comments and Contributors Information from Bugzilla," 2025. [Online]. Available: <https://arxiv.org/abs/2504.18806>
- [10] A. Rebert, S. K. Cha, T. Avgerinos, J. Foote, D. Warren, G. Grieco, and D. Brumley, "Optimizing seed selection for fuzzing," in *23rd USENIX Security Symposium (USENIX Security 14)*, 2014, pp. 861–875.
- [11] G. Elbaz, "Common Crawl," <https://commoncrawl.org/>, 2025.
- [12] "AWS Open Data," <https://aws.amazon.com/opendata/>.
- [13] A. Herrera, H. Gunadi, L. Hayes, S. Magrath, F. Friedlander, M. Sebastian, M. Norrish, and A. L. Hosking, "Corpus distillation for effective fuzzing: A comparative evaluation," 2020. [Online]. Available: <https://arxiv.org/abs/1905.13055>
- [14] M. Zalewski, *AFL++ Documentation*, AFL++ Project, 2019, accessed: Oct. 5, 2025. [Online]. Available: https://afl-1.readthedocs.io/_/downloads/en/latest/pdf/
- [15] K. Zhang, Z. Li, D. Wu, S. Wang, and X. Xia, "Low-cost and comprehensive non-textual input fuzzing with LLM-synthesized input generators," in *Proceedings of the 34th USENIX Security Symposium (USENIX Security 25)*. Seattle, WA, USA: USENIX Association, Aug. 2025. [Online]. Available: <https://www.usenix.org/conference/usenixsecurity25/presentation/zhang-kunpeng>
- [16] M. Schloegel, N. Bars, N. Schiller, L. Bernhard, T. Scharnowski, A. Crump, A. Ale-Ebrahim, N. Bissantz, M. Muench, and T. Holz, "SoK: Prudent evaluation practices for fuzzing," in *2024 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2024, pp. 1974–1993.
- [17] M. Böhme, L. Szekeres, and J. Metzman, "On the reliability of coverage-based fuzzer benchmarking," in *Proceedings of the 44th International Conference on Software Engineering*, 2022, pp. 1621–1633.
- [18] AFL++ Team, "AFL++ technical details," https://aflplusplus.com/docs/technical_details/, 2025, accessed: Aug. 23, 2025.
- [19] A. Hazimeh, A. Herrera, and M. Payer, "Magma: A ground-truth fuzzing benchmark," *Proc. ACM Meas. Anal. Comput. Syst.*, vol. 4, no. 3, Dec. 2020. [Online]. Available: <https://doi.org/10.1145/3428334>
- [20] OpenAI, "Pricing," <https://platform.openai.com/docs/pricing>, 2025, accessed: Sept. 23, 2025.
- [21] "Amazon athena pricing," <https://aws.amazon.com/athena/pricing/>, accessed: Oct. 2, 2025.
- [22] "Amazon s3 pricing," <https://aws.amazon.com/s3/pricing/>, accessed: Oct. 2, 2025.
- [23] X. Zhu, S. Wen, S. Camtepe, and Y. Xiang, "Fuzzing: a survey for roadmap," *ACM Computing Surveys (CSUR)*, vol. 54, no. 11s, pp. 1–36, 2022.
- [24] J. Wang, B. Chen, L. Wei, and Y. Liu, "Skyfire: Data-driven seed generation for fuzzing," in *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2017, pp. 579–594.
- [25] R. Padhye, C. Lemieux, K. Sen, M. Papadakis, and Y. Le Traon, "Semantic fuzzing with zest," in *Proceedings of the 28th ACM SIGSOFT International Symposium on Software Testing and Analysis*, 2019, pp. 329–340.
- [26] P. Godefroid, H. Peleg, and R. Singh, "Learn&fuzz: Machine learning for input fuzzing," in *2017 32nd IEEE/ACM International Conference on Automated Software Engineering (ASE)*. IEEE, 2017, pp. 50–59.
- [27] J. Li, B. Zhao, and C. Zhang, "Fuzzing: a survey," *Cybersecurity*, vol. 1, no. 1, p. 6, 2018.
- [28] Y. Wang, P. Jia, L. Liu, C. Huang, and Z. Liu, "A systematic review of fuzzing based on machine learning techniques," *PloS one*, vol. 15, no. 8, p. e0237749, 2020.
- [29] O. Nourry, Y. Kashiwa, B. Lin, G. Bavota, M. Lanza, and Y. Kamei, "The human side of fuzzing: Challenges faced by developers during fuzzing activities," *ACM transactions on software engineering and methodology*, vol. 33, no. 1, pp. 1–26, 2023.
- [30] X. Liu, X. Li, R. Prajapati, and D. Wu, "DeepFuzz: automatic generation of syntax valid C programs for fuzz testing," in *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial*

- Intelligence*, ser. AAAI'19/IAAI'19/EAAI'19. AAAI Press, 2019. [Online]. Available: <https://doi.org/10.1609/aaai.v33i01.33011044>
- [31] N. Nichols, M. Raugas, R. Jasper, and N. Hilliard, "Faster fuzzing: Reinitialization with deep neural models," *arXiv preprint arXiv:1711.02807*, 2017.
 - [32] C. Lyu, S. Ji, Y. Li, J. Zhou, J. Chen, and J. Chen, "Smartseed: Smart seed generation for efficient fuzzing," *arXiv preprint arXiv:1807.02606*, 2018.
 - [33] Y. Xu, H. Jia, L. Chen, X. Wang, Z. Zeng, Y. Wang, Q. Gao, J. Wang, W. Ye, S. Zhang, and Z. Wu, "Isc4dggf: Enhancing directed grey-box fuzzing with llm-driven initial seed corpus generation," *arXiv preprint arXiv:2409.14329*, Sep. 2024. [Online]. Available: <https://arxiv.org/abs/2409.14329>
 - [34] W. Shi, Y. Zhang, X. Xing, and J. Xu, "Harnessing large language models for seed generation in greybox fuzzing," *arXiv preprint arXiv:2411.18143*, 2024.
 - [35] user1342, "Autocorpus: Automatic corpus generation tool using llms," <https://github.com/user1342/AutoCorpus>, 2023, accessed: 2025-09-10.
 - [36] Asmita, Y. Oliinyk, M. Scott, R. Tsang, C. Fang, and H. Homayoun, "Fuzzing busybox: Leveraging llm and crash reuse for embedded bug unearthing," in *Proceedings of the 33rd USENIX Security Symposium (USENIX Security 24)*. Philadelphia, PA, USA: USENIX Association, Aug. 2024. [Online]. Available: <https://www.usenix.org/conference/usenixsecurity24/presentation/asmita>
 - [37] G. J. Saavedra, K. N. Rodhouse, D. M. Dunlavy, and P. W. Kegelmeyer, "A review of machine learning applications in fuzzing," *arXiv preprint arXiv:1906.11133*, 2019.
 - [38] LLVM Project, "libFuzzer: A Library for Coverage-Guided Fuzz Testing," <https://llvm.org/docs/LibFuzzer.html>, 2025, accessed: Oct. 5, 2025.
 - [39] R. Swiecki, "Honggfuzz: General-purpose fuzzer," <https://honggfuzz.dev/>, 2018, accessed: Oct. 5, 2025.

APPENDIX

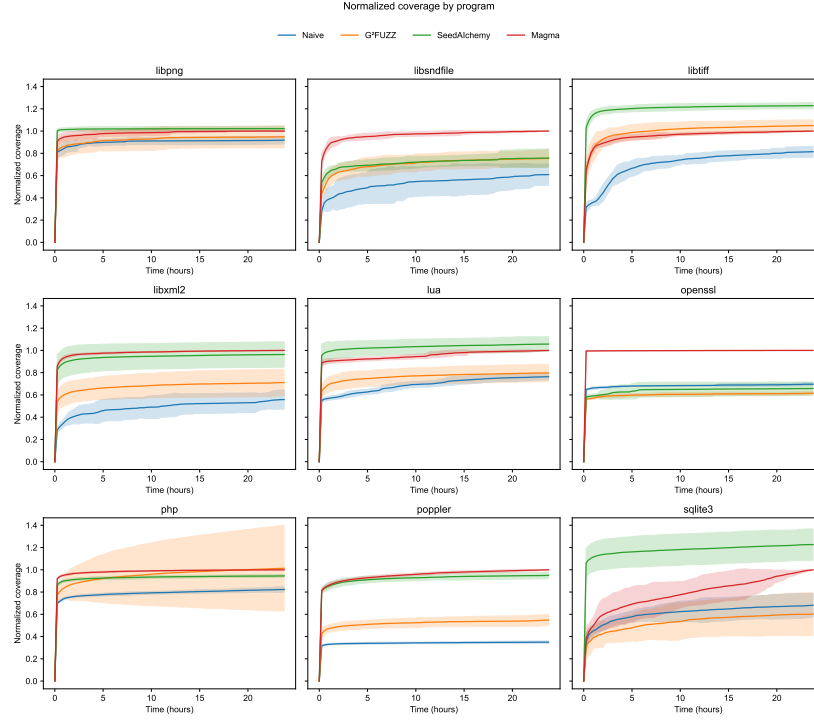


Fig. 7. Normalized coverage for each Magma target

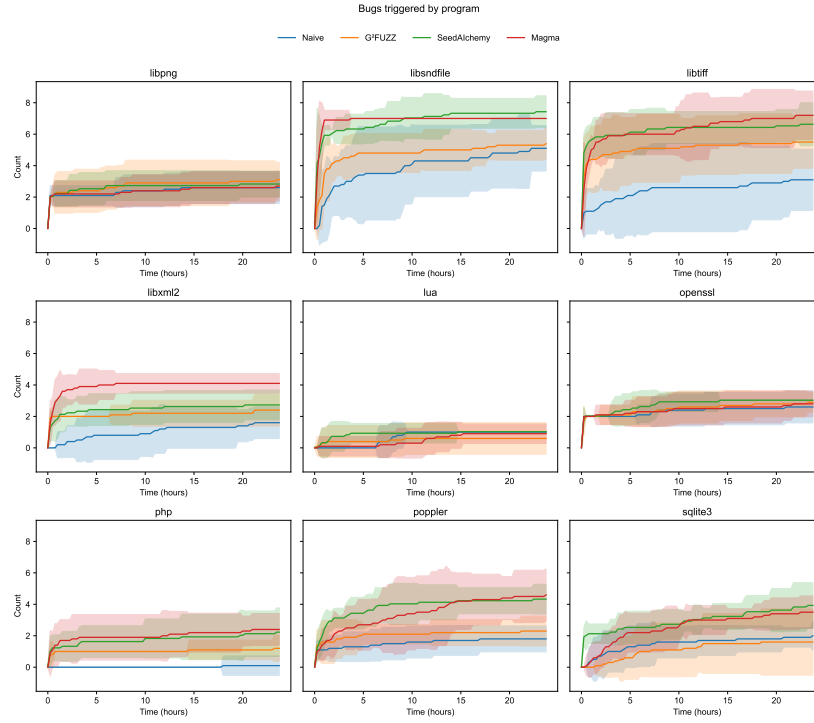


Fig. 8. Bugs triggered for each Magma target