



# The Same Only Different: On Information Modality for Configuration Performance Analysis

Hongyuan Liang<sup>1</sup>, Yue Huang<sup>1</sup>, Tao Chen<sup>2\*</sup>

<sup>1</sup> School of Computer Science and Engineering, University of Electronic Science and Technology of China, China

<sup>2</sup> IDEAS Lab, School of Computer Science, University of Birmingham, United Kingdom  
lianghy16@outlook.com, huangyue16@outlook.com, t.chen@bham.ac.uk

**Abstract**—Configuration in software systems helps to ensure efficient operation and meet diverse user needs. Yet, some, if not all, configuration options have profound implications for the system’s performance. Configuration performance analysis, wherein the key is to understand (or infer) the configuration options’ relations and their impacts on performance, is crucial. Two major modalities exist that serve as the source information in the analysis: either the manual or source code. However, it remains unclear what roles they play in configuration performance analysis. Much work that relies on manuals claims their benefits of information richness and naturalness; while work that trusts the source code more prefers the structural information provided therein and criticizes the timeliness of manuals.

To fill such a gap, in this paper, we conduct an extensive empirical study over 10 systems, covering 1,694 options, 106,798 words in the manual, and 22,859,552 lines-of-code for investigating the usefulness of manual and code in two important tasks of configuration performance analysis, namely performance-sensitive options identification and the associated dependencies extraction. We reveal several new findings and insights, such as it is beneficial to fuse the manual and code modalities for both tasks; the current automated tools that rely on a single modality are far from being practically useful and generally remain incomparable to human analysis. All those pave the way for further advancing configuration performance analysis.

**Index Terms**—Software configuration, performance analysis, manual, source code analysis, configuration dependency

## I. INTRODUCTION

Software systems are often designed with a daunting number of configuration options. For example, until now, MySQL has more than 600 configuration options that can be set by users [1]. Although this certainly indicates the great flexibility of the software, it also comes with a cost: it has been reported that 59% of the performance issues worldwide are caused by configuration [2]. Therefore, much work has been done to mitigate these issues, mainly related to the modeling [3]–[7], testing [8]–[10], and tuning [11]–[16] of configuration.

However, all those research topics rely on or can benefit from two prerequisite tasks: (1) since not all the options are performance-sensitive, it is essential to preselect which are the ones that should be considered. For example, options that set a directory path or a port often do not impact performance. In contrast, options that control the code logic or some resource-heavy operations are performance sensitive, such as

the `wait_timeout` in MySQL and `maxThreads` in TOM-CAT; (2) there might be a dependency between two options, for example, in NGINX, the option `TCP_NOPUSH` can be used only when option `sendfile` is enabled. As a result, it is fundamental to know these dependencies beforehand. Those prerequisite tasks are the key steps in configuration performance analysis [17]–[19]. Indeed, much work has been done to investigate and propose tools for inferring performance-sensitive options [18], [20] and their dependencies [19], [21], [22], but they leverage a single modality—either using manual (e.g., official guidelines or API docs) or the source code.

The motivation for using either modality in configuration performance analysis is arguable: on the one hand, researchers who are in favor of manuals state that these artifacts provide strong and intuitive domain knowledge with rich information to be extracted [20], [22]. On the other hand, work that relies on source code claims that it offers comprehensive structure and state of the systems [18], [19], while being robust to the timeliness issue of the manual, i.e., what has been changed in the codebase might not have been timely reflected in the manual [17], [23]. Yet, there has been a lack of understanding on which modality should be preferred or tends to be more useful in configuration performance analysis and why.

In this paper, we seek to fulfill this gap. To that end, we conduct an extensive empirical study over 10 systems that come with a total of 1,694 configuration options, 106,798 words of manual, and 2,859,552 lines-of-code, aiming to understand the role of manual and code for identifying performance-sensitive options and extracting their dependencies in configuration performance analysis. We do that in two phases: in the first phase, we carry on a human analysis involving all authors that independently parse and analyze the manual and code, from each of which the results are compared. In this second phase, we examine the effectiveness of existing automated tools that leverage manual or code individually, while comparing their outcomes against those from the human analyzers. Through those, we aim to answer several important research questions (RQs) related to configuration performance analysis:

- **RQ1:** How can manual or code be relatively useful in identifying performance-sensitive options?
- **RQ2:** Comparatively, to what extent does manual or code help to extract performance-sensitive dependencies?

\*Tao Chen is the corresponding author. Hongyuan Liang and Yue Huang are also supervised in the IDEAS Lab.

- **RQ3:** How do existing automated tools in configuration performance analysis perform compared with human analysis using either manual or code information?

Among others, our key findings/contributions are that:

- To identify performance-sensitive options, the manual leads to fewer false negatives while the code is more false-positive resistant, but fusing both modalities can maximize the benefit. They both cause false positives due to misleading information but incur different causes for false negatives: the vague description in the manual and the non-standard patterns of executions in code.
- While code is often more helpful for extracting dependencies than manual, the manual can still be useful for some edge cases. We found no false positives for both modalities, and the false negatives of finding dependencies from manual and code are mainly due to lack of information and excessive recursions of function calls, respectively.
- Existing automated tools for both tasks are far from being practically useful, and they are not comparable to human analysis, except for identifying performance-sensitive options from the manual.

Deriving from these, we articulate several insights for this direction of research, such as the necessity of fusing both manual and code modality in identifying performance-sensitive options and the extraction of dependencies therein should also consider performance-insensitive options.

As part of our results, we have also made the dataset, code, and tools used available at: <https://github.com/ideas-labo/cp-mod>. Compared with existing datasets [18], [19], [22], our labeled dataset is of a larger scale with more diverse systems, complementing the research for automated configuration performance analysis.

The reminder of this paper is as follows: Section II presents the background and motivation. Section III delineates our empirical methodology. Section IV presents and analyzes the results followed by the insights learned in Section V. Sections VI and VII present the threats to validity and related work, respectively. Section VIII concludes the work.

## II. BACKGROUND AND MOTIVATION

### A. Analyzing Configurable Systems

A configurable system can be configured by giving a vector of configuration values  $\mathbf{c} = \{x_1, x_2, \dots, x_n\}$ , where  $x_n$  denotes the value of the  $n$ th configuration option, which can be binary, enumerate, or numeric [24]. Configuration performance analysis studies the correlation between the configuration options and the system’s performance, e.g., runtime or throughput, along with the interrelationships between the options. The results are fundamental for many configuration-related tasks, such as configuration performance modeling [4], [25], configuration tuning/testing [8], [14], and self-adaptation [26]–[29].

### B. Performance-sensitive Options

Not all configuration options are performance-sensitive, therefore it is essential to identify which are performance-sensitive in configuration performance analysis. In essence,

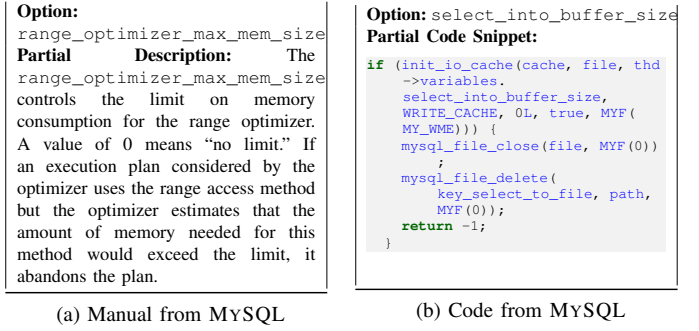


Fig. 1: Examples of manual texts and source code for revealing performance-sensitive options.

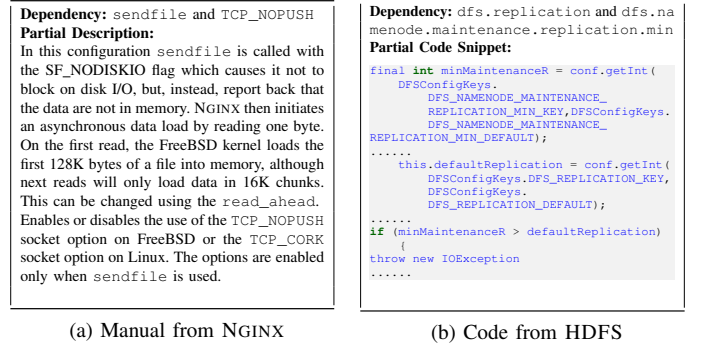


Fig. 2: Examples of manual texts and source code for revealing performance-sensitive dependencies.

two common modalities exist for finding performance-sensitive options: the documentation manual and source code. The former often provides clear information about the performance sensitivity. As can be seen in Figure 1a, the description states that the option can control the allocated memory, which would clearly be important for the performance.

The latter, in contrast, contains performance-sensitive information with the code execution logic. For example, in Figure 1b, the call logic shows that `select_into_buffer_size` affects the buffer size for initializing the cache—a key performance indicator.

### C. Performance-sensitive Dependencies

There can be a dependency between a pair of options. For example, in the HADOOP implementation of MAPREDUCE, any value of the option `mapreduce.jobhistory.loadedjobs.cache.size` would be ignored if the option `mapreduce.jobhistory.loaded.tasks.cache.size` is set to a positive value. Failure to comply with the dependency might result in, e.g., ineffective configuration or runtime exceptions.

Similar to the performance-sensitive options, extracting dependencies sensitive to the performance can also be achieved via analyzing the manual or code. Figure 2a shows an example of the manual, in which we see that the texts directly state that `TCP_NOPUSH` depends on the value of `sendfile`.

TABLE I: Configurable software and the manuals studied.

Software	Domain	Language	Words	LoC	Version	Used By
HTTPD	Web Server	C	6,161	634,144	2.4.57	[33] [20] [34]
LIGHTTPD	Web Server	C	686	252,192	1.4	[34]
NGINX	Web Server	C	4,741	413,976	1.24.0	[33] [20] [34]
REDIS	In-memory Database	C	14,614	437,076	6.2.12	[34] [35]
MYSQL	SQL Database	C/C++	36,424	10,332,544	8.0.33	[22] [21]
HDFS	File System	Java	7,974	2,344,750	3.3.5	[19] [36]
MAPREDUCE <sup>1</sup>	Distributed Computing	Java	7,677	708,280	3.3.5	[20] [19] [37]
HBASE	Distributed NoSQL Database	Java	9,666	3,960,100	2.5.5	[19] [36] [37]
YARN	Resource Management	Java	13,277	2,575,072	3.3.5	[19] [36]
TOMCAT	Application Server	Java	9,498	1,201,418	10.1.16	[38]

In contrast, the source code also provides such information through some joint implication to the control flow between the two options. For example, in Figure 2b, the two options control two variables independently, which would then be compared to determine if there is an exception, hence a dependency.

However, manually parsing/analyzing dependencies from the software manuals and code is a labor-intensive process (regardless whether they are performance-sensitive or not). This is because the description in the manual can be lengthy and hard to be interpreted; while the code of relevant options can involve complex call chains and code structure.

### D. Motivation

1) *Why Performance?*: It is worth noting that even valid options might not be performance-sensitive. Those non-performance-sensitive options overcomplicate tasks related to configuration performance, e.g., configuration tuning and debugging. Therefore, a common way in those works would be to select (either manually or automatically) only the performance-sensitive options to work on [20], [30]–[32]. As a result, it is crucial to specifically study performance-sensitive options and understand how their dependencies can impact the mutation of those options in the tuning/testing process, which is the key motivation of this work.

2) *Why Manual and Code?*: Indeed, much work relies on either manual or code individually. Those who use manuals claim that the manual contains a wide range of natural descriptions of the options and, hence, is more information-rich [20], [22]. Conversely, work that prefers code analysis states that the code is more structural and is less prone to outdated issues caused by rapid version updating [17]–[19], [23]. Despite the above, it remains unclear which modalities are more useful, particularly for identifying the performance-sensitive options and their dependencies. For example, SafeTune [20] is an automated tool that can be used for configuration performance analysis, which only relies on manual texts while completely ignoring code information without giving clear justification. This work seeks to bridge such a gap through an empirical study, providing insights into the roles of different modalities in configuration performance analysis.

## III. METHODOLOGY

### A. Software Systems

To collect data from widely-used configurable systems, we study those open-sourced ones that also come with well-

<sup>1</sup>We use the HADOOP implementation of MAPREDUCE.

TABLE II: Keywords for extracting performance-sensitive options from the manual (all are case-insensitive).

	Terms
Blacklist	<i>file, proxy, forward, path, port, address, location, updates, version compatibility, legacy, address, link, restore, plugin, dir, url, host, name, precision descriptor, principal, key, ui, profile, info, catalogue, password, pwd</i>
Whitelist	<i>debug, optimize, table, cpu, pct, interleave, block thread, worker, executor, error, time, depth, max, logger, range size, min, length, timeout, limit, cache, mode, log</i>

documented manuals. The list has been illustrated in Table I. We can clearly see that the systems studied come from diverse domains, with different languages and scales, while being widely used by the community. To ensure the validity of each system, we take the latest stable version and analyze its code and the corresponding manual by the date of analysis, i.e., Sep 2023, representing the most stable state of a system since the newest releases are likely to be error-prone.

### B. Identifying Performance-sensitive Options

1) *Manual Analysis*: Two authors who are experienced software engineers independently analyzed the manual to identify performance-sensitive options. This follows a mix of blacklist and whitelist approaches. The blacklist contains several keywords that, if detected in the manual’s description, would cause the corresponding option to be eliminated immediately. In contrast, if a description contains any of the terms in the whitelist, then the corresponding option is considered performance-sensitive. The procedure is as follows:

- Screening**: As from Table II, we conduct an initial screening by identifying the keywords for blacklist (e.g., names and paths) and whitelist (e.g., resource and thread).
- Blacklist Exclusion**: We eliminate any options which have descriptions that contain the keywords in the blacklist. The two authors exchanged their filtered list and any disagreement would be discussed or consulting the other author or external expert until a consensus can be made.
- Whitelist Inclusion**: In the remaining options, the description of which contains any of the terms in the whitelist is said as performance-sensitive. Extensive discussions were taken place to resolve disagreements.
- The two authors independently read the options collected and picked the performance-sensitive ones as understood from the description. A final list is shared and discussed until agreements are reached. The same process is repeated for the manual of every system studied.

For example, `mapreduce.cshuffle.port` is an option of MAPREDUCE that was filtered out using the blacklist since it merely specifies the port for connecting the component. Conversely, for MYSQL, the `wait_timeout` is an option that is included under the whitelist since the waiting time would have a dramatic impact on the system’s performance.

2) *Code Analysis*: To understand the performance-sensitive options from code, we leverage a semi-automated approach

that relies on taint analysis tools<sup>2</sup>. The same two leading authors took the lead, but all authors helped to resolve conflicts:

- (a) **Operation Categorization:** Any operation/function that belongs to the categories below are performance sensitive:
- Memory operations, including array or stack allocations, e.g., the `init_mem()` in MYSQL.
  - Cache operations, updates or synchronization, e.g., the `ngx_open_cached_file()` in NGINX.
  - I/O operations, e.g., those accessing the persistent storage, such as the `BufferedReader` for JAVA.
  - Multi-threading process, such as thread pooling, e.g., the `newCachedThreadPool` in MYSQL.
  - Loop operations, such as `for`, `while`, and `do-while` loops.
  - Network communication: such as the `http_range_header_filter()` in HTTPD.
  - Error control process: code such as `try-catch`, `assert`, `my_error()`.

Those categories are discussed among all authors until an agreement has been reached.

- (b) **Variable Identification:** It is easy to understand what the configuration options are by reading the configuration file, e.g., `yaml` files. Yet, since the configuration options in the code are codified in different ways depending on the systems, we use different approaches below to establish the mappings between configuration options from the manual and the corresponding variables in the code:
- **Unified Analysis:** For MYSQL, LIGHTTPD, NGINX, and REDIS, the variables for configuration options are centrally organized in a single file with identical names, and hence the mappings can be found directly.
  - **Segmented Analysis:** For HTTPD and TOMCAT, the variables are written in separate source code files but they are still centrally involved in a `main` thread. For those, we analyze the control flow to find the relevant code files and the corresponding variables.
  - **Scattered Analysis:** For HDFS, MAPREDUCE, HBASE, and YARN, the variables are distributed across the codebase and are only called on-demand. Yet, those variables are all accessible via the `setter()` and `getter()` function, which can be easily allocated.

Again, the two lead authors conduct the analysis independently and exchange the results to reach a consensus.

- (c) **Taint Analysis:** To extract those performance-sensitive options by analyzing the code, we leverage taint analysis, where we use variables identified serve as the sources and the performance-sensitive operations are the sink. For C/C++ systems, we use Clang [42], and for JAVA systems, we use JavaParser [43], in two ways:
- In the **top-down manner**, the taint analysis would procedure some taint flows that indicate which performance-sensitive operations would interact with

the variables identified. However, it is unclear whether those variables can impact the behavior of the operation. To confirm that, we further extract the corresponding code fragments from the Abstract Syntax Tree to investigate whether the path between the variables identified and performance-sensitive operations involve logical operators, e.g., `If/Else` or `While`. If it does, we find a performance-sensitive option/variable.

- We also follow a **bottom-up** approach starting from each of the performance-sensitive operations, based on which we identify the related parameters. We then manually traced back to the origin of those parameters to confirm their interactions with the variables. If an interaction exists, the corresponding variable is performance-sensitive. This helps to mitigate the false negative cases from the taint analysis.

3) *System Measurement (Ground Truth):* To identify the ground truth, we manually change each configuration option and examine whether such a change leads to significant performance deviation. This is achieved by running and profiling each system under 3-4 different benchmarks/workloads setting according to the commonly used ones from the literature [20], [30], [35], [44] and practical applications<sup>3</sup>. For example, we use SYSBENCH and TPC for MYSQL with different concurrent users, test duration, and table counts, etc.

For each option, we sample at least two possible values as the representatives: for the numeric options, we use the default, maximum, and minimum within the range documented. If no range is provided, we use the positive upper limit as the maximum value and 0 as the minimum value. For boolean options, we simply flip their `true` and `false`. For other enumerate options, we compare all pairs of the possible values in the eligible range. For all cases, we compare the performance achieved by setting the two values of an option, while keeping the other options with their default values.

We repeat the process five times to ensure the stability and reliability. A numeric or boolean option is said to be indeed performance-sensitive if, under any workload, the average performance of setting one value can have more than 5% deviation compared with that of the setting the said option with the other value; an enumerate option is performance-sensitive if the average performance of any pair of its value has deviation greater than 5% under any workload. The above is a typical setting being acknowledged as a de facto standard in the field [18], [45]. In total, we run over 17,000 performance tests, resulting in more than 1,500 CPU hours.

### C. Extracting Performance-sensitive Dependencies

Next, we extract the performance-sensitive dependencies, which involve at least one performance-sensitive option.

1) *Dependencies by Manual:* To do that by using manual, two authors independently perform the following:

- (a) The non-performance-sensitive options are filtered out.

<sup>2</sup>Our taint analysis is technically similar to prior work [39]–[41], but applied under different definitions of source and sink that fulfill our needs.

<sup>3</sup>Details: <https://github.com/ideas-labo/cp-mod/blob/main/dataset/perfsensitive-groundtruth/workloads.md>.

- (b) Each author analyzed the description of all options remaining. An option is said to have a dependency on the other if its description has mentioned the other option.
- (c) In the end, both lead authors combined the list of dependencies identified and discussed among all authors for cross-checking on, e.g., the accuracy, completeness, and relationships between descriptions and related options.
- (d) The above process was repeated by two rounds of rigorous review to mitigate bias.

The extracted dependencies represent the information that can be obtained by analyzing the manual. Those dependencies are then further verified by investigating if the two variables can indeed be involved in some logical operation in the relevant code. However, since the dependencies are initially extracted from the manual, they are said to be manual-related. Indeed, as we will show, manual analysis can reveal certain dependencies that are deeply hidden in purely code analysis.

2) *Dependencies by Code*: To extract dependency information from the code, we again focus on the performance-sensitive options only. We then apply taint analysis using the same aforementioned tools, such that the variable of a performance-sensitive option serves as the source while that of the other option is the sink. Within the results, we check whether there is a call path that involves both variables. If that is the case, we investigate the code segment and examine if both variables are involved in some form of logical operation, such as If/Else, Switch, or While loop. When such a logical relationship can be identified, we say the code indicates that there exists a dependency between the variables/options.

Two authors independently execute the above process, and the final results are discussed and agreed by all authors.

3) *Why not Testing Systems?*: For dependencies, in this work, we regard the outcomes from both manual and code analysis above as the ground truth without system testing, since the systems have mostly unclear responses when violating performance-sensitive dependencies. Indeed, Chen et al. [19] have shown that for up to 74.2% of the dependencies (we checked that these include many performance-sensitive ones), there are no/incomplete messages in the log from the system under their violations, hence we cannot verify if there is a dependency even when testing on the actual systems.

4) *Why not Extract all Dependencies First?*: We focus explicitly on the performance-sensitive options identified and extract their dependencies, as opposed to extracting all dependencies and then pruning dependencies not containing performance-sensitive options, for two reasons:

- In our work, extracting the dependencies only on the performance-sensitive options under the scale of our study, which constitutes merely 21% of the options, has already consumed several months of work per expert (we have two lead authors) merely for one round of analysis (we have two rounds), excluding the discussion time for resolving conflicts with all authors. Therefore, extracting the dependencies for all options and further pruning them down would require significantly increased efforts in the first place, since the number of initial dependencies

(pairs of options) to be analyzed can increase considerably along the number of options considered. Further, this would inevitably include many non-performance-sensitive dependencies which we are not interested.

- Since we are only interested in performance-sensitive dependencies, by which we mean those dependencies that involve at least one performance-sensitive option (which is known), focusing on the performance-sensitive options first and their dependencies would lead to similar results as to extracting all dependencies then prune them down. This is because the known performance-sensitive options are identical between the two processes.

## IV. EMPIRICAL STUDY RESULTS

### A. RQ1: Performance-sensitive Options

1) *Operationalization*: In comparison to the ground truth obtained by actual system measurement, we report on the true positive rate, false positive rate, and false negative rate of the results obtained by analyzing the manual and the code.

To study the discrepancy of correctly identifying performance-sensitive options by using the manual and the code individually, we use the following patterns observed:

- **Manual Only**: This refers to the true performance-sensitive options (against ground truth) that are found by studying the manual only. For example, the description form manual for `max_execution_time` option specifies that it determines MySQL's execution waiting time for SELECT statements. However, it does not involve any control flow related to performance-sensitive operations in the code.
- **Code Only**: This refers to the true performance-sensitive options (against ground truth) that are identified via analyzing the code but are not noticeable from the manual. For example, the `server.defer-accept` is merely described in the manual as the listening socket for a TCP protocol `TCP_DEFER_ACCEPT` ("on" or "off"). Yet, in the code, it is clear that `server.defer-accept` directly controls the function `setsockopt()`, which affects various behaviors of `LIGHTTPD`, e.g., network traffic, security, errors, and more.

To further understand why analyzing manual or code alone might lead to false positives, we distinguish some patterns:

- **FP<sub>1</sub>**: The configuration option is completely discarded. However, this might not be noticeable in the manual (since it might not be mentioned) or code (due to complicated call graphs). For example, `max_length_for_sort_data` in MySQL is a typical example of **FP<sub>1</sub>**. In the manual texts, it clearly indicates that the option is performance-sensitive but it has no impact on the system in the code. `tls-session-cache-size` from REDIS change the default number of TLS sessions cached, which is invoked within the function `tlsConfigure()` that has a clear impact on the system from the code. However, `tlsConfigure()` is never triggered.



TABLE III: Ground truth performance-sensitive options identified by measuring the systems.

Software	All Options	True Performance-sensitive Options	%
HTTPD	77	15	19%
LIGHTTPD	62	18	29%
NGINX	79	12	15%
REDIS	115	36	31%
MYSQL	103	26	25%
HDFS	223	58	26%
MAPREDUCE	218	56	26%
HBASE	227	35	15%
YARN	545	86	16%
TOMCAT	45	12	27%
<b>Total</b>	<b>1694</b>	<b>354</b>	<b>21%</b>

- **FP<sub>2</sub>**: The option has limited implication on the performance, i.e., the largest change is less than 1%. For example, the `LimitInternalRecursion` in HTTPD is identified as performance-sensitive options from both the manual and the code analysis. However, in actual profiling, it caused a maximal performance fluctuation of 0.76% across the workloads.
- **FP<sub>3</sub>**: The impact of the configuration option on performance is minimal. That is to say, there is indeed a performance fluctuation after option change, but the maximal magnitude is less than 5%. For example, the `KeepAlive` option in HTTPD was deemed to be performance-sensitive from both manual and code, but it merely has a maximum of 2.34% variation in performance on all workloads.

Similarly, for false negative cases, we draw on two patterns:

- **FN<sub>1</sub>**: There is indeed a significant performance implication. For example, the `Optimizer_prune_level` in MYSQL does not exhibit any control relationship with performance-sensitive operations in code. Yet, it significantly impacts the performance by 27% in the profiling, since its setting is passed to a third-party algorithm.
- **FN<sub>2</sub>**: The maximal performance implication of the options only marginally beyond 5%, e.g., by less than 1%, and it is possibly due to randomness in the execution. Yet, for the sake of completeness, we also classify these as performance-sensitive options. For example, the `mapreduce.job.am.node-label-expression` in MAPREDUCE determines whether range requests are permitted, and it is clearly non-performance-sensitive from the texts in the manual. However, it has non-trivial impacts on performance with a maximum of 5.41% change across all workloads.

We also discuss the main causes of the above false results.

2) *Findings*: From Table III, it is clear that the true performance-sensitive options constitute a small proportion of all options—merely 21% in general across all the systems.

**Finding 1:** Only a small set of options, between 15% and 31%, can non-trivially impact the performance.

For identifying performance-sensitive options from manual and code, as shown in Table IV, we see that it is easier to

identify more options by manual than by code (778 vs. 672), which makes sense since the naturalness of manual could easily lead to more inclusion. However, analyzing the manual does not lead to more true positive samples than using code (41% vs. 43%), meaning that the highly structural nature of code can be more beneficial. Taking the `tcp_nopush` in NGINX as an example, this option enables or disables the use of the `TCP_NOPUSH` socket option on FreeBSD or of the `TCP_CORK` socket option. It is difficult to obtain an accurate answer through manual texts, but in the code, this option is repeatedly invoked in recursive network operations, which strongly implies its performance sensitivity. On the other hand, analyzing the manual can lead to higher false positives while analyzing code is more likely to be false negative prone. However, the benefit of identifying more performance-sensitive options via manual is that it also allows us to find more ones that can only be detected therein, i.e., comparing the 12% of **manual only** to the 5% of **code only**. This suggests the good side of the richer naturalness in the manual: the more comprehensive summary of the interrelations between options can provide more understandable and intuitive information than analyzing code alone. For example, in the source code, there is no performance-sensitive flow for `optimizer_prune_level`, but we can intuitively understand from the manual description that this option controls the third-party heuristic algorithms applied during query optimization, pruning less promising parts of the plan from the optimizer’s search space in MYSQL.

**Finding 2:** Code is more helpful in finding the true performance-sensitive options and with fewer false positives than manual. Yet, the manual can still provide useful information for some that are difficult to identify in code and are less likely to be false negative prone.

As shown in Table V, clearly, the patterns of false positive for extracting performance-sensitive options from manual and code are of similar characteristics:  $\approx \frac{1}{3}$  are due to the fact that the options have limited implications on the performance;  $\approx \frac{2}{3}$  are due to noticeable but small implication; a very limited proportion is related to discarded, but have not been clearly marked options. We found that the fundamental cause is also similar: it is mainly due to both the manual and code giving misleading information about options’ performance sensitivity. For example, `useAsyncIO` in TOMCAT is a widely considered performance-sensitive option [46], [47], since it enables asynchronous IO while both the manual and code indicate such. However, upon the actual profiling, it only has up to 3.62% performance change over the workloads.

The patterns of false negatives when analyzing manual and code also differ marginally: ignoring the options that can largely impact performance has only a little higher proportion for code (70%) than manual (68%); while those suspiciously performance-sensitive options are merely a slightly more common pattern for manual (31%) than code (30%).

Yet, we found that the predominated root causes can differ:

TABLE IV: Identified performance-sensitive options (by manual and code) against the ground truth along with the patterns (TP%, FP%, and FN% are the true positive rate, false positive rate, and false negative rate against the ground truth, respectively).

Software	Manual					Code				
	#Options	TP%	FP%	FN%	Manual Only	#Options	TP%	FP%	FN%	Code Only
HTTPD	37	35% (13/37)	39% (24/62)	13% (2/15)	20% (3/15)	36	31% (11/36)	40% (25/62)	27% (4/15)	7% (1/15)
LIGHTTPD	36	44% (16/36)	45% (20/44)	11% (2/18)	6% (1/18)	38	39% (15/38)	52% (23/44)	17% (3/18)	0% (0/18)
NGINX	52	19% (10/52)	63% (42/67)	17% (2/12)	17% (2/12)	47	21% (10/47)	55% (37/67)	17% (2/12)	17% (2/12)
REDIS	81	38% (31/81)	63% (50/79)	14% (5/36)	11% (4/36)	75	41% (31/75)	56% (44/79)	14% (5/36)	11% (4/36)
MYSQL	72	32% (23/72)	64% (49/77)	12% (3/26)	19% (5/26)	68	28% (19/68)	64% (49/77)	27% (7/26)	4% (1/26)
HDFS	112	44% (49/112)	38% (63/165)	16% (9/58)	7% (4/58)	84	57% (48/84)	22% (36/165)	17% (10/58)	5% (3/58)
MAPREDUCE	106	48% (51/106)	34% (55/162)	9% (5/56)	14% (8/56)	93	48% (45/93)	30% (48/162)	20% (11/56)	4% (2/56)
HBASE	74	43% (32/74)	22% (42/192)	9% (3/35)	11% (4/35)	64	45% (29/64)	18% (35/192)	17% (6/35)	3% (1/35)
YARN	176	45% (80/176)	21% (96/459)	7% (6/86)	10% (9/86)	141	51% (72/141)	15% (69/459)	16% (14/86)	1% (1/86)
TOMCAT	32	34% (11/32)	64% (21/33)	8% (1/12)	8% (1/12)	26	42% (11/26)	45% (15/33)	8% (1/12)	8% (1/12)
<b>Total</b>	778	41% (316/778)	34% (462/1340)	11% (38/354)	12% (41/354)	672	43% (291/672)	28% (381/1340)	18% (63/354)	5% (16/354)

TABLE V: Patterns for false positives/negatives on performance-sensitive options found by analyzing manual and code.

Software	Manual					Code				
	FP <sub>1</sub>	FP <sub>2</sub>	FP <sub>3</sub>	FN <sub>1</sub>	FN <sub>2</sub>	FP <sub>1</sub>	FP <sub>2</sub>	FP <sub>3</sub>	FN <sub>1</sub>	FN <sub>2</sub>
HTTPD	4% (1/24)	21% (5/24)	75% (18/24)	50% (1/2)	50% (1/2)	4% (1/25)	28% (7/25)	68% (17/25)	75% (3/4)	25% (1/4)
LIGHTTPD	0% (0/20)	5% (1/20)	95% (19/20)	50% (1/2)	50% (1/2)	0% (0/23)	17% (4/23)	63% (19/23)	67% (2/3)	33% (1/3)
NGINX	0% (0/42)	55% (23/42)	45% (19/42)	50% (1/2)	50% (1/2)	0% (0/37)	59% (22/37)	41% (15/37)	50% (1/2)	50% (1/2)
REDIS	4% (2/50)	34% (17/50)	62% (31/50)	80% (4/5)	20% (1/5)	7% (3/44)	27% (12/44)	66% (29/44)	80% (4/5)	20% (1/5)
MYSQL	4% (2/49)	29% (14/49)	67% (33/49)	67% (2/3)	33% (1/3)	4% (2/49)	29% (14/49)	67% (33/49)	57% (4/7)	43% (3/7)
HDFS	0% (0/63)	37% (23/63)	63% (40/63)	67% (6/9)	33% (3/9)	0% (0/36)	42% (15/36)	58% (21/36)	70% (7/10)	30% (3/10)
MAPREDUCE	2% (1/55)	36% (20/55)	62% (34/55)	80% (4/5)	20% (1/5)	2% (1/48)	38% (18/48)	60% (29/48)	82% (9/11)	18% (2/11)
HBASE	0% (0/42)	48% (20/42)	52% (22/42)	67% (2/3)	33% (1/3)	0% (0/35)	40% (14/35)	60% (21/35)	67% (4/6)	33% (2/6)
YARN	0% (0/96)	42% (40/96)	58% (56/96)	67% (4/6)	33% (2/6)	0% (0/69)	23% (16/69)	77% (53/69)	71% (10/14)	29% (4/14)
TOMCAT	0% (0/21)	52% (11/21)	48% (10/21)	100% (1/1)	0% (0/1)	0% (0/15)	53% (8/15)	47% (7/15)	0% (0/1)	100% (1/1)
<b>Total</b>	1% (6/462)	38% (174/462)	61% (282/462)	68% (26/38)	31% (12/38)	2% (7/381)	34% (130/381)	64% (244/381)	70% (44/63)	30% (19/63)

- **Cause of false negatives in manual analysis:** The texts in the manual have rather vague information related to performance. For instance, from the LIGHTTPD manual, the description of option `server.range-requests` is "defines whether range requests are allowed or not", which is highly imprecise. As such, we did not include that as a performance-sensitive option. Yet, in the actual profiling, it can considerably impact the performance.
- **Cause of false negatives in code analysis:** The code involves non-standard patterns of executions that affect the performance. For example, in MYSQL, `max_seeks_for_key` is only called in the function `find_cost_for_ref()`. Within this function, `max_seeks_for_key` is compared with several similar options and the highest value is returned as the budget, which does not involve our defined performance operations. However, MYSQL uses the cost estimate provided by `find_cost_for_ref()` to choose the most efficient query execution plan.

**Finding 3:** The false positives for extracting performance-sensitive options using manual and code are due to similar patterns and causes, but the causes for false negative samples vary even though their patterns are similar.

## B. RQ2: Performance-sensitive Dependencies

1) *Operationalization:* We found that all the performance-sensitive dependencies identified (via manual or code) are

indeed present, from which the unique dependencies for each system are reported, hence there is no false positive. However, there are indeed false negative cases. We observe the following discrepancy patterns when using manual or code:

- **Manual Only:** There exist performance-sensitive dependencies that are only clearly observable in the manual but deeply hidden in the code. For example, in our taint analysis results, there is no code segment where the taint flow between the options `read_buffer_size` and `select_into_buffer_size` in MYSQL. However, the manual provides detailed descriptions of the dependency relationship for them.
- **Code Only:** Similarly, certain performance-sensitive dependencies have not been mentioned in the manual at all but clearly noticeable in the code. For example, in the source code of LIGHTTPD, if the value of `server.max-connections` is greater than half of the value for `server.max-fds`, then the system will log an error. However, such a dependency between the two said options is not mentioned at all in the manual.

For a given dependency, we found two possible patterns depending on the options' type:

- **Mixed Options:** This means one of the options involved is not relevant to performance. For example, in MYSQL, the option `wait_timeout` is performance-sensitive while the option `interactive_timeout` is not, but there is a dependency that the former's value should be smaller than that of the latter.
- **Performance Only Options:** This refers to

TABLE VI: Performance-sensitive dependencies identified.  $|C|$  denotes the number of options involved in a dependency chain.

Software	All Dependencies	Mixed Options	Perf. Only Options	$ C  \geq 3$
HTTPD	4	50% (2/4)	50% (2/4)	75% (3/4)
LIGHTTPD	2	100% (2/2)	0% (0/2)	0% (0/2)
NGINX	6	33% (2/6)	66% (4/6)	50% (3/6)
REDIS	12	58% (7/12)	41% (5/12)	42% (5/12)
MYSQL	8	50% (4/8)	50% (4/8)	25% (2/8)
HDFS	13	76% (10/13)	23% (3/13)	15% (2/13)
MAPREDUCE	10	70% (7/10)	30% (3/10)	0% (0/10)
HBASE	7	42% (3/7)	57% (4/7)	0% (0/10)
YARN	20	60% (12/20)	40% (8/20)	65% (13/20)
TOMCAT	2	0% (0/2)	100% (2/2)	100% (2/2)
<b>Total</b>	<b>84</b>	<b>58% (49/84)</b>	<b>41% (35/84)</b>	<b>38% (32/84)</b>

the dependencies wherein both options are performance-sensitive, e.g., the dependency between `large_client_header_buffers` and `connection_pool_size` in NGINX.

In our analysis, a dependency chain involving multiple options is simply a concatenation of different dependencies with two options (providing that they have one option in common), hence we also analyze the number of options involved in a dependency chain, which is a direct reflection of the dependency of complexity for a system. Again, we showcase how those complex dependency chains, i.e., the number of options involved greater than 3, can be detected via manual and code analysis individually. For example, in REDIS, the `maxmemory` and `maxmemory-policy` options jointly determine when and which keys need to be deleted to free memory, hence the latter depends on the former. At the same time, option `lazyfree-lazy-eviction` depends on the policy of `maxmemory-policy`, and it determines whether the deletion operation will block the processing of commands.

Since some performance-sensitive dependencies can only be identified in manual not code (and vice versa), we also study why certain dependencies have been missed.

2) *Findings*: Table VI shows the general results and patterns for all the systems. We see that the performance-sensitive options can indeed involve dependencies. More than half of those are only be relevant to performance-sensitive options while the remaining half can involve non-performance-sensitive ones. In particular, there is a non-trivial proportion of the chain that covers three or more options (38%), meaning that the complexity of dependency can be considerably high.

**Finding 4:** Among the performance-sensitive dependencies, over half of these involve mixed options while there exist a non-trivial proportion of the chain that involves more than 3 options.

Table VII summarizes the statistics of dependencies that can be identified from manual and code. As can be seen, the code identifies 73 dependencies against the 53 from the manual; since no false positives are involved and there are 84 dependencies in total, this means that the information from the code only misses 11 ones while that of the manual can cause 31 misses. Further, it is much more likely to find dependencies from code that cannot be found in the manual (42%) than the

opposed way (20%). The key reason is that, for systems like LIGHTTPD, there is almost no information about dependencies in the manual. We believe that when the dependencies do not cause serious issues, e.g., crashes, the developers tend to omit their descriptions from the manual.

The same can be reflected in identifying dependencies over the patterns of option types. What we found is that for mixed options, the manual is more likely to miss out on their dependencies. For example, in LIGHTTPD, `server.errorlog-use-syslog` is performance-sensitive while `server.syslog-facility` is not, and the manual has no information of their dependency. Yet, they clearly have an “enabling” dependency in code.

Code also helps to identify most complicated dependency chains (87%) while manual can only provide clear information on 71% of those. For example, in HTTPD, the `KeepAlive` option has a dependency on the `Timeout` option, which influences the data flow of the `KeepAliveTimeout` option. This dependency chain with three options is not found in the manual but is clearly visible in the code.

**Finding 5:** Code often offers more useful information than manual in identifying performance-sensitive dependencies across the patterns and dependency length.

Since we found no false positive, we focus on the reason that causes false negatives using manual and code individually:

- **Cause of false negatives in manual analysis:** We found that the lack of manual information has led to the 31 false negative cases. Developers tend to omit option dependencies that do not have direct relationships or serious consequences when maintaining the manual, to avoid causing too much trouble for the users.
- **Cause of false negative in code analysis:** For the 11 edge cases where the code analysis has resulted in false negatives, the key reason is due to the depth limitations of taint analysis, which prevent locating all dependencies within the code. When configuration options are called through multiple recursions of function calls, the taint analysis often fails to trace the code, hence leading to most of the false negatives.

**Finding 6:** Although code is generally more useful in identifying performance-sensitive dependencies, the manual can still help to resolve some edge cases.

### C. RQ3: Usefulness of Current Automated Tools

1) *Operationalization*: We now seek to examine the effectiveness of existing automated tools for configuration performance analysis. We choose state-of-the-art tools which relies on the information modality of manual or code independently. Tools for performance-sensitive options inference are:

- **SafeTune [20]:** As a representative that relies on **manual**, SafeTune identifies performance-sensitive options by analyzing configuration documentation.



TABLE VII: Patterns between performance-sensitive dependencies identified via manual and code. Same formats as Table VI.

Software	Manual					Code				
	#Dependencies	Manual Only	Mixed Options	Perf. Only Options	$ C  \geq 3$	#Dependencies	Code Only	Mixed Options	Perf. Only Options	$ C  \geq 3$
HTTPD	1	0% (0/4)	0% (0/2)	50% (1/2)	33% (1/3)	4	75% (3/4)	100% (2/2)	100% (2/2)	100% (3/3)
LIGHTTPD	0	0% (0/2)	0% (0/2)	0% (0/0)	0% (0/0)	2	100% (2/2)	100% (2/2)	0% (0/0)	0% (0/0)
NGINX	4	33% (2/6)	100% (2/2)	50% (2/4)	100% (3/3)	4	33% (2/6)	50% (1/2)	75% (3/4)	66% (2/3)
REDIS	9	8% (1/12)	57% (4/7)	100% (5/5)	100% (5/5)	11	25% (3/12)	85% (6/7)	100% (5/5)	80% (4/5)
MYSQL	8	25% (2/8)	100% (4/4)	100% (4/4)	100% (2/2)	6	0% (0/8)	75% (3/4)	75% (3/4)	50% (1/2)
HDFS	5	7% (1/13)	40% (4/10)	33% (1/3)	0% (0/2)	12	61% (8/13)	90% (9/10)	100% (3/3)	100% (2/2)
MAPREDUCE	4	10% (1/10)	42% (3/7)	33% (1/3)	0% (0/0)	9	60% (6/10)	85% (6/7)	100% (3/3)	0% (0/0)
HBASE	5	28% (2/7)	66% (2/3)	75% (3/4)	0% (0/2)	5	28% (2/7)	66% (2/3)	75% (3/4)	100% (2/2)
YARN	15	10% (2/20)	66% (8/12)	87% (7/8)	76% (10/13)	18	25% (5/20)	100% (12/12)	75% (6/8)	92% (12/13)
TOMCAT	2	0% (0/2)	0% (0/0)	100% (2/2)	100% (2/2)	2	0% (0/2)	0% (0/0)	100% (2/2)	100% (2/2)
<b>Total</b>	<b>53</b>	<b>20% (11/53)</b>	<b>55% (27/49)</b>	<b>74% (26/35)</b>	<b>71% (23/32)</b>	<b>73</b>	<b>42% (31/73)</b>	<b>87% (43/49)</b>	<b>85% (30/35)</b>	<b>87% (28/32)</b>

- **DiagConfig [18]:** DiagConfig trains a Random Forest to predict whether the given configuration option and its associated code segments are performance-sensitive.

As such, it uses **code** as the main information modality.

Since we did not find readily available tools that can directly extract performance-sensitive dependencies, we use GPTuner and CDep, which are designed to extract any dependencies:

- **GPTuner [22]:** GPTuner leverages the understanding capabilities of Large Language Model (LLM) to infer dependencies by reading the texts from the **manual**.
- **CDep [19]:** In CDep, the dependencies are predicted by using rule mining on **code** based on predefined patterns.

Within the dependencies identified by those tools, we then further extract the performance-sensitive ones according to our ground truth of performance-sensitive options (i.e., at least one option in the pair of options is performance-sensitive). In the evaluation, we only consider those extracted samples against the performance-sensitive dependencies derived from the process in Section III-C. This mitigates the unfairness caused by the fact that those tools cover any dependency types.

Since all tools are cross-project, we use all the manual/code that exists in a target project as the testing data and pre-train them with the original dataset from their authors if possible. Notably, none of our studied systems were studied in DiagConfig hence we used all of their original datasets in training; there are 5 overlapping systems for SafeTune and hence we removed those systems and only used data from the remaining ones (8 systems) in their original datasets to train the model. These prevent data leakage. GPTuner is a special one since it relies on LLM/GPT3.5; as such we have no idea what data it has been using to pre-train the LLM. CDep is a rule-based tool that was designed using general understandings from the systems we studied and hence it is not subject to the data leakage issue in general machine learning-based tools.

We use recall, precision, and F1 score as the metrics as they are robust to data imbalance [48]. Note that DiagConfig and CDep only work for JAVA-based systems. For all cases, we follow the general “rule-of-thumb” that an F1 score of 0.7 or higher means a practically useful configuration predictor [49]. We repeat 10 runs of experiments for all stochastic tools.

2) *Findings:* For predicting performance-sensitive options, as from Figures 3a and 3b, we see that manual-based tools like SafeTune would likely have excellent recall but poor precision—this matches with the human analysis of manual in **RQ1**, where the false positive is more severe than false

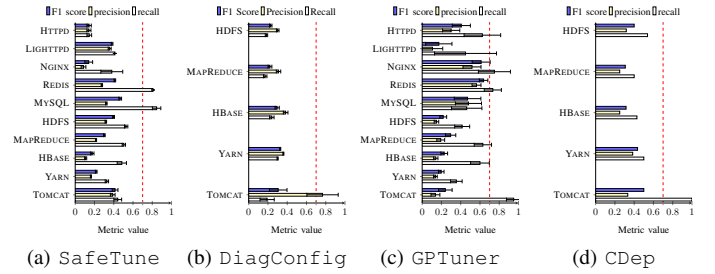


Fig. 3: Effectiveness (mean/deviation) of tools that predict performance-sensitive options (a and b) and configuration dependencies (c and d) over 10 runs. SafeTune and GPTuner are manual-based; DiagConfig and CDep are code-based.

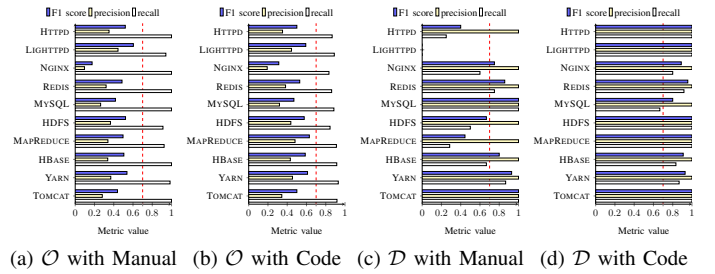


Fig. 4: Human analysis for performance-sensitive options, i.e.,  $\mathcal{O}$ , (a and b) and their dependencies, i.e.,  $\mathcal{D}$ , (c and d).

negatives. As for code-based tools like DiagConfig, the precision is slightly better than the recall. This deviates from the human results form **RQ1**, because its data-driven nature could parse code information that cannot be captured by human analysts. However, in both cases, we see that the F1 score is still far away from the threshold of 0.7, suggesting that those tools, using either manual or code, are still far from being practically useful in predicting performance-sensitive options.

Figures 3c and 3d illustrate the results for the tools that predict configuration dependencies using different modalities. As can be seen, there is generally poor precision but good recall, regardless of the modality they rely on. This means that false positives are a more serious issue therein, which again deviates from human analysis. As a result, the final F1 score can hardly reach 0.7 on any system. One possible explanation is that those tools have not been explicitly tailored to cater to the characteristics of performance-sensitive

dependencies. For example, CDep is mainly designed based upon the understanding of some manually derived patterns of dependencies, which are relevant to any dependencies type in general. Yet, some of those patterns might not apply to most performance-sensitive dependencies, albeit they can be common in other dependency types. For instance, we found that the *default value pattern*, which represents the dependency wherein the value of an option serves as the default of the other, does not apply to any performance-sensitive dependencies. Often, this is relevant to path-related options, e.g., the value of `dfs.namenode.name.dir` is the default for `dfs.namenode.edits.dir` in HDFS. The above can eventually amplify their issue of a high number of false positives for identifying performance-sensitive dependencies when correctly finding the dependencies of other types are no longer of concern, leading to poor precision and negatively impacting their overall performance.

**Finding 7:** Existing tools for predicting performance-sensitive options and their dependencies tend to be considerably affected by false positives in general, thus can hardly lead to practically useful results.

We also directly compare the results from tools with those of human analysis from **RQ1** and **RQ2**. For identifying performance-sensitive options, in Figure 3a and 4a, we see that the F1 score of using tool SafeTune is comparable to that of human analysis when analyzing manual; the latter merely slightly better since understanding natural language also shares similar difficulties as to machine interpretation. However, when comparing the results by analyzing code, we note that human analysis is much superior to the tool, i.e., DiagConfig (Figures 3b and 4b). This is because the code is of complex interrelationships and the learned model in DiagConfig has failed to retain the domain knowledge used in human analysis. Yet, in all cases, neither human analysis nor automated tools have reached the 0.7 F1 score threshold.

For dependencies extraction, clearly the human analysis (Figure 4c and 4d) is significantly better than GPTuner and CDep (Figures 3c and 3d) for the JAVA systems with either modality. Notably, the results of human analysis are mostly beyond the F1 score of 0.7 for all systems.

**Finding 8:** Existing tools are generally far from reaching the level of human analysis, except for identifying performance-sensitive options from manual.

## V. ACTIONABLE INSIGHTS

**Finding 1** clearly suggests that there is only a relatively small proportion of the configuration options that can non-trivially impact the performance. This means that:

**Insight 1:** It can be highly beneficial to extract performance-sensitive options within or before tasks such as configuration testing and configuration tuning.

**Finding 2** and **Finding 3** imply that, to better identify the performance-sensitive options, it is necessary to jointly investigate the information from both manual and code. Therefore:

**Insight 2:** When identifying performance-sensitive options (by hand or with tools), it is necessary to combine the information from both modalities of manual and code.

Yet, the patterns/causes of their false positives are similar but the causes for false negatives can differ:

**Insight 3:** When fusing the modality of code and manual for extracting performance-sensitive options, one should not 100% believe in the information from the manual and code, as it is possibly misleading. In particular, special attention needs to be paid to vague information (for manual) and code that does not follow standard performance-relevant operations.

The above insight challenges the existing studies (on either manual or code) in which the proposed approach fully trusts the information from the modality, implying the need for some potential confidence-driven approach or some reinforced methods in automated configuration performance learning (even when combining both manual and code).

The above also implies that key information might be missing in both modalities, therefore configuration should be better catered to at the software design level. This might include, e.g., standardizing the format between manual/comment and code; clearly annotating the property of configuration options designs, or, at the more fundamental level, ensuring performance regression test is conducted and the results are declared when adding any new configuration options.

**Finding 4** suggest that when finding performance-sensitive dependencies, 58% of the cases involve mixed options. This challenges the belief that dependencies used in, e.g., configuration tuning, are mainly considered for performance-sensitive options only [50]–[52]. That said:

**Insight 4:** Performance-sensitive dependencies identification should not be specific to performance-sensitive options, as mixed ones are highly possible.

Further, **Finding 5** and **Finding 6** suggest that:

**Insight 5:** Although it could be acceptable to only rely on the information from the code in identifying performance-sensitive dependencies in general, additionally incorporating manual information can still be useful for finding some edge cases. For manual analysis, the information regarding dependencies might be missing while for the source code, the depth of nested invocations/recursions is critical.

Finally, **Finding 7** and **Finding 8** examine how far are we at automatic performance-sensitive option identification and their dependencies extraction, the results imply that:

**Insight 6:** Relying on existing automated tools that leverage either manual or code individually is problematic, and human analysis remains generally far more reliable. Hence it is promising to investigate interactive tools.

Intuitively, there will be a trade-off between quality (more human analysis) and cost (more tool supports), which is highly case-dependent. Our view is that the inferior accuracy of the tool will have an impact on the downstream tasks benefiting from the configuration performance analysis. For example, if we seek to identify the performance-sensitive options and their dependencies for configuration tuning, then more inaccuracies can lead to unnecessary testing of some options and frequently violated dependencies. Since the configuration tuning is expensive too, in this case, the extra overhead caused by the worse accuracies of a tool might exceed its relative cost-saving.

For interactive tools, one possible solution is to incorporate output explainability (e.g., LIME [53] or SHAP [54]) into the tool fused with manual and code, based on which humans can then provide their inputs in various forms, e.g., a “Yes”/“NO” answer; some categories of confidence, or even weights.

## VI. THREATS TO VALIDITY

The human analysis using manual and code might pose threats to construct validity. To mitigate that, we have followed a systematic protocol involving all authors with multiple rounds of reviews. Yet, like any empirical study, human errors/oversights are always possible. When testing the systems for **RQ1**, a possible threat can come from the fact that the option is changed one at a time—the standard way that has been followed in many prior works [18]. Indeed, due to the finite resources and a large number of options/systems to test in our study, fully covering the option interactions is infeasible. There are also other forms of configuration beyond this study, e.g., alternative third-party libraries and versions, the understanding of which is promising for future research.

Threats to internal validity might be related to the parameter setting of the automated tools and the methodology, e.g., the threshold for determining performance sensitivity. In this work, we set the same parameter values and procedures, e.g., the training data for *DiaConfig*, as used in their prior work. While these serve as a convenient default setting, we cannot guarantee that they are the best status.

To minimize the threats to external validity, we ensure that our study covers a wide range of systems, with diverse manuals and complexity of code. We have also considered tools that leverage different modalities, as well as using various workloads when testing the systems. However, studying more systems, workloads, or tools might be more fruitful.

## VII. RELATED WORK

**Performance-sensitive Options Identification.** To identify performance-sensitive options, a common approach has been treating the configurable system as a black box, and hence leveraging data-driven methods in the prediction. For example, some studies have been using empirical observations,

e.g., certain patterns of how the performance-sensitive options influence the performance obtained via qualitative analysis [30], [55]. Recent work has followed a more white-box approach, where certain artifacts and modalities about configurable systems have been exploited. Among others, He et al. [20] propose *SafeTune*, a multi-intent-aware semi-supervised method that analyzes the manuals, aiming to infer the performance-sensitive options. *DiagConfig* [18] is an approach that also leverages machine learning (random forest). However, it exploits the code invocation chains as part of the features for predicting performance-sensitive options.

**Dependency Extraction.** Existing work on dependency extraction has relied on the systematic analysis of some modalities related to configurable systems. *GPTuner* [22] is an approach that is based on a large language model to parse the manuals for extracting information about dependencies between options. It proposes a prompt integration algorithm to unify the structured view of the refined knowledge. Apart from manual, the other modality that is commonly used is the source code. Zhou et al. [36] introduce a new tool for multi-configuration error diagnosis by analyzing the dependencies between options in the source code. Similarly, *Cdep* [19] detects dependencies between configuration parameters manifested in the code through pattern matching and taint analysis.

**Performance Bug Detection.** There are approaches that rely on either manual or code to detect performance bugs. Among others, *PracExtractor* [56] is a tool that uses natural language processing techniques to automatically extract best practice recommendations from manuals, which serve as an oracle in performance bug detection. *ECSTATIC* [57] and *TAINTMINI* [58] are tools that use taint analysis to parse the flow graph of the code, which detects the potential performance bugs. Yet, these approaches have not covered configuration-related performance.

**Related Empirical Studies.** Empirical studies have been conducted on different aspects of configurable systems [2], [24], [59]. Xu et al. [24] demonstrate the prevalence of configurability. Zhang et al. [59] investigate how configurations evolve among system versions. However, there has been no empirical study that compares configuration performance analysis using different modalities, i.e., manual or code.

## VIII. CONCLUSION

In this paper, we conduct an extensive empirical study that assesses the usefulness of two modalities, i.e., manual and code, for identifying performance-sensitive options and extracting their dependencies. Through analyzing 10 systems with 1,694 options, 106,798 words in the manual, and 2,859,552 lines-of-code, we reveal several insights that can impact the community of configuration performance analysis.

Our observations pave the way for vast future research directions, including but not limited to multiple-modal configuration performance analysis and interactive tools thereof.

## ACKNOWLEDGEMENT

This work was supported by a NSFC Grant (62372084) and a UKRI Grant (10054084).

## REFERENCES

- [1] “Mysql manual,” <https://dev.mysql.com/doc/refman/8.4/en/server-option-variable-reference.html>, 2024, retrieved on Aug 01, 2024.
- [2] X. Han and T. Yu, “An empirical study on performance bugs for highly configurable software systems,” in *Proceedings of the 10th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, ESEM 2016, Ciudad Real, Spain, September 8-9, 2016*. ACM, 2016, pp. 23:1–23:10. [Online]. Available: <https://doi.org/10.1145/2961111.2962602>
- [3] J. Gong and T. Chen, “Deep configuration performance learning: A systematic survey and taxonomy,” *ACM Trans. Softw. Eng. Methodol.*, vol. 34, no. 1, Dec. 2024. [Online]. Available: <https://doi.org/10.1145/3702986>
- [4] H. Ha and H. Zhang, “Deepperf: performance prediction for configurable software with deep sparse neural network,” in *Proceedings of the 41st International Conference on Software Engineering, ICSE 2019, Montreal, QC, Canada, May 25-31, 2019*. IEEE / ACM, 2019, pp. 1095–1106. [Online]. Available: <https://doi.org/10.1109/ICSE.2019.00113>
- [5] J. Gong and T. Chen, “Predicting configuration performance in multiple environments with sequential meta-learning,” *Proc. ACM Softw. Eng.*, vol. 1, no. FSE, pp. 359–382, 2024. [Online]. Available: <https://doi.org/10.1145/3643743>
- [6] —, “Predicting software performance with divide-and-learn,” in *Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering, ESEC/FSE 2023, San Francisco, CA, USA, December 3-9, 2023*. ACM, 2023, pp. 858–870. [Online]. Available: <https://doi.org/10.1145/3611643.3616334>
- [7] J. Gong, T. Chen, and R. Bahsoon, “Dividable configuration performance learning,” *IEEE Trans. Software Eng.*, vol. 51, no. 1, pp. 106–134, 2025.
- [8] H. He, Z. Jia, S. Li, E. Xu, T. Yu, Y. Yu, J. Wang, and X. Liao, “Cp-detector: Using configuration-related performance properties to expose performance bugs,” in *35th IEEE/ACM International Conference on Automated Software Engineering, ASE 2020, Melbourne, Australia, September 21-25, 2020*. IEEE, 2020, pp. 623–634. [Online]. Available: <https://doi.org/10.1145/3324884.3416531>
- [9] Y. Ma, T. Chen, and K. Li, “Faster configuration performance bug testing with neural dual-level prioritization,” in *47th IEEE/ACM International Conference on Software Engineering (ICSE)*. IEEE, 2025.
- [10] R. Cheng, L. Zhang, D. Marinov, and T. Xu, “Test-case prioritization for configuration testing,” in *ISSTA ’21: 30th ACM SIGSOFT International Symposium on Software Testing and Analysis, Virtual Event, Denmark, July 11-17, 2021*. ACM, 2021, pp. 452–465. [Online]. Available: <https://doi.org/10.1145/3460319.3464810>
- [11] T. Chen and M. Li, “Multi-objectivizing software configuration tuning,” in *ESEC/FSE ’21: 29th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering, Athens, Greece, August 23-28, 2021*. ACM, 2021, pp. 453–465.
- [12] V. Nair, Z. Yu, T. Menzies, N. Siegmund, and S. Apel, “Finding faster configurations using FLASH,” *IEEE Trans. Software Eng.*, vol. 46, no. 7, pp. 794–811, 2020. [Online]. Available: <https://doi.org/10.1109/TSE.2018.2870895>
- [13] T. Chen and M. Li, “Adapting multi-objectivized software configuration tuning,” *Proc. ACM Softw. Eng.*, vol. 1, no. FSE, pp. 539–561, 2024. [Online]. Available: <https://doi.org/10.1145/3643751>
- [14] P. Chen, T. Chen, and M. Li, “MMO: meta multi-objectivization for software configuration tuning,” *IEEE Trans. Software Eng.*, vol. 50, no. 6, pp. 1478–1504, 2024. [Online]. Available: <https://doi.org/10.1109/TSE.2024.3388910>
- [15] Y. Ye, T. Chen, and M. Li, “Distilled lifelong self-adaptation for configurable systems,” in *47th IEEE/ACM International Conference on Software Engineering (ICSE)*. IEEE, 2025.
- [16] T. Chen and M. Li, “Do performance aspirations matter for guiding software configuration tuning? an empirical investigation under dual performance objectives,” *ACM Trans. Softw. Eng. Methodol.*, vol. 32, no. 3, pp. 68:1–68:41, 2023. [Online]. Available: <https://doi.org/10.1145/3571853>
- [17] M. Sayagh, N. Kerzazi, B. Adams, and F. Petrillo, “Software configuration engineering in practice interviews, survey, and systematic literature review,” *IEEE Trans. Software Eng.*, vol. 46, no. 6, pp. 646–673, 2020. [Online]. Available: <https://doi.org/10.1109/TSE.2018.2867847>
- [18] Z. Chen, P. Chen, P. Wang, G. Yu, Z. He, and G. Mai, “Diagconfig: Configuration diagnosis of performance violations in configurable software systems,” in *Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering, ESEC/FSE 2023, San Francisco, CA, USA, December 3-9, 2023*. ACM, 2023, pp. 566–578. [Online]. Available: <https://doi.org/10.1145/3611643.3616300>
- [19] Q. Chen, T. Wang, O. Legunsen, S. Li, and T. Xu, “Understanding and discovering software configuration dependencies in cloud and datacenter systems,” in *ESEC/FSE ’20: 28th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering, Virtual Event, USA, November 8-13, 2020*. ACM, 2020, pp. 362–374. [Online]. Available: <https://doi.org/10.1145/3368089.3409727>
- [20] H. He, Z. Jia, S. Li, Y. Yu, C. Zhou, Q. Liao, J. Wang, and X. Liao, “Multi-intention-aware configuration selection for performance tuning,” in *44th IEEE/ACM 44th International Conference on Software Engineering, ICSE 2022, Pittsburgh, PA, USA, May 25-27, 2022*. ACM, 2022, pp. 1431–1442. [Online]. Available: <https://doi.org/10.1145/3510003.3510094>
- [21] T. Xu, J. Zhang, P. Huang, J. Zheng, T. Sheng, D. Yuan, Y. Zhou, and S. Pasupathy, “Do not blame users for misconfigurations,” in *ACM SIGOPS 24th Symposium on Operating Systems Principles, SOSP ’13, Farmington, PA, USA, November 3-6, 2013*. ACM, 2013, pp. 244–259. [Online]. Available: <https://doi.org/10.1145/2517349.2522727>
- [22] J. Lao, Y. Wang, Y. Li, J. Wang, Y. Zhang, Z. Cheng, W. Chen, M. Tang, and J. Wang, “Gptuner: A manual-reading database tuning system via gpt-guided bayesian optimization,” *Proc. VLDB Endow.*, vol. 17, no. 8, pp. 1939–1952, 2024. [Online]. Available: <https://www.vldb.org/pvldb/vol17/p1939-tang.pdf>
- [23] T. Xu and Y. Zhou, “Systems approaches to tackling configuration errors: A survey,” *ACM Comput. Surv.*, vol. 47, no. 4, pp. 70:1–70:41, 2015. [Online]. Available: <https://doi.org/10.1145/2791577>
- [24] T. Xu, L. Jin, X. Fan, Y. Zhou, S. Pasupathy, and R. Talwadker, “Hey, you have given me too many knobs!: understanding and dealing with over-designed configuration in system software,” in *Proceedings of the 2015 10th Joint Meeting on Foundations of Software Engineering, ESEC/FSE 2015, Bergamo, Italy, August 30 - September 4, 2015*. ACM, 2015, pp. 307–319. [Online]. Available: <https://doi.org/10.1145/2786805.2786852>
- [25] J. Gong and T. Chen, “Does configuration encoding matter in learning software performance? an empirical study on encoding schemes,” in *19th IEEE/ACM International Conference on Mining Software Repositories, MSR 2022, Pittsburgh, PA, USA, May 23-24, 2022*. ACM, 2022, pp. 482–494. [Online]. Available: <https://doi.org/10.1145/3524842.3528431>
- [26] T. Chen, “Lifelong dynamic optimization for self-adaptive systems: Fact or fiction?” in *IEEE International Conference on Software Analysis, Evolution and Reengineering, SANER 2022, Honolulu, HI, USA, March 15-18, 2022*. IEEE, 2022, pp. 78–89. [Online]. Available: <https://doi.org/10.1109/SANER53432.2022.00022>
- [27] S. Kumar, T. Chen, R. Bahsoon, and R. Buyya, “DATESSO: self-adapting service composition with debt-aware two levels constraint reasoning,” in *SEAMS ’20: IEEE/ACM 15th International Symposium on Software Engineering for Adaptive and Self-Managing Systems, Seoul, Republic of Korea, 29 June - 3 July, 2020*. ACM, 2020, pp. 96–107. [Online]. Available: <https://doi.org/10.1145/3387939.3391604>
- [28] T. Chen, “Planning landscape analysis for self-adaptive systems,” in *International Symposium on Software Engineering for Adaptive and Self-Managing Systems, SEAMS 2022, Pittsburgh, PA, USA, May 22-24, 2022*. ACM/IEEE, 2022, pp. 84–90. [Online]. Available: <https://doi.org/10.1145/3524844.3528060>
- [29] T. Chen and R. Bahsoon, “Toward a smarter cloud: Self-aware autoscaling of cloud configurations and resources,” *Computer*, vol. 48, no. 9, pp. 93–96, 2015. [Online]. Available: <https://doi.org/10.1109/MC.2015.278>
- [30] K. Kanellis, R. Alagappan, and S. Venkataraman, “Too many knobs to tune? towards faster database tuning by pre-selecting important knobs,” in *12th USENIX Workshop on Hot Topics in Storage and File Systems, HotStorage 2020, July 13-14, 2020*. USENIX Association, 2020. [Online]. Available: <https://www.usenix.org/conference/hotstorage20/presentation/kanellis>



- [31] M. Zhu and D. Hao, "Compiler auto-tuning via critical flag selection," in *38th IEEE/ACM International Conference on Automated Software Engineering, ASE 2023, Luxembourg, September 11-15, 2023*. IEEE, 2023, pp. 1000–1011. [Online]. Available: <https://doi.org/10.1109/ASE56229.2023.00209>
- [32] P. Chen, J. Gong, and T. Chen, "Accuracy can lie: On the impact of surrogate model in configuration tuning," *IEEE Trans. Software Eng.*, 2025.
- [33] T. Wang, H. He, X. Liu, S. Li, Z. Jia, Y. Jiang, Q. Liao, and W. Li, "Conftainter: Static taint analysis for configuration options," in *38th IEEE/ACM International Conference on Automated Software Engineering, ASE 2023, Luxembourg, September 11-15, 2023*. IEEE, 2023, pp. 1640–1651. [Online]. Available: <https://doi.org/10.1109/ASE56229.2023.00067>
- [34] S. Zhou, X. Liu, S. Li, W. Dong, X. Liao, and Y. Xiong, "Confmapper: Automated variable finding for configuration items in source code," in *2016 IEEE International Conference on Software Quality, Reliability and Security, QRS 2016, Companion, Vienna, Austria, August 1-3, 2016*. IEEE, 2016, pp. 228–235. [Online]. Available: <https://doi.org/10.1109/QRS-C.2016.35>
- [35] R. Cao, L. Bao, C. Q. Wu, P. Zhangsun, Y. Li, and Z. Zhang, "CM-CASL: comparison-based performance modeling of software systems via collaborative active and semisupervised learning," *J. Syst. Softw.*, vol. 201, p. 111686, 2023. [Online]. Available: <https://doi.org/10.1016/j.jss.2023.111686>
- [36] Y. Zhou, X. Hu, S. Xu, Y. Jia, Y. Liu, J. Wang, G. Xu, W. Wang, S. Liu, and T. Baker, "Multi-misconfiguration diagnosis via identifying correlated configuration parameters," *IEEE Trans. Software Eng.*, vol. 49, no. 10, pp. 4624–4638, 2023. [Online]. Available: <https://doi.org/10.1109/TSE.2023.3308755>
- [37] C. Li, S. Wang, H. Hoffmann, and S. Lu, "Statically inferring performance properties of software configurations," in *EuroSys '20: Fifteenth EuroSys Conference 2020, Heraklion, Greece, April 27-30, 2020*. ACM, 2020, pp. 10:1–10:16. [Online]. Available: <https://doi.org/10.1145/3342195.3387520>
- [38] Y. Zhu, J. Liu, M. Guo, Y. Bao, W. Ma, Z. Liu, K. Song, and Y. Yang, "Bestconfig: tapping the performance potential of systems via automatic configuration tuning," in *Proceedings of the 2017 Symposium on Cloud Computing, SoCC 2017, Santa Clara, CA, USA, September 24-27, 2017*. ACM, 2017, pp. 338–350. [Online]. Available: <https://doi.org/10.1145/3127479.3128605>
- [39] M. Velez, P. Jamshidi, N. Siegmund, S. Apel, and C. Kästner, "White-box analysis over machine learning: Modeling performance of configurable systems," in *43rd IEEE/ACM International Conference on Software Engineering, ICSE 2021, Madrid, Spain, 22-30 May 2021*. IEEE, 2021, pp. 1072–1084. [Online]. Available: <https://doi.org/10.1109/ICSE43902.2021.00100>
- [40] M. Velez, P. Jamshidi, F. Sattler, N. Siegmund, S. Apel, and C. Kästner, "Configcrusher: towards white-box performance analysis for configurable systems," *Autom. Softw. Eng.*, vol. 27, no. 3, pp. 265–300, 2020. [Online]. Available: <https://doi.org/10.1007/s10515-020-00273-8>
- [41] M. Copik, A. Calotou, T. Grosser, N. Wicki, F. Wolf, and T. Hoefler, "Extracting clean performance models from tainted programs," in *PPoPP '21: 26th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming, Virtual Event, Republic of Korea, February 27- March 3, 2021*. ACM, 2021, pp. 403–417. [Online]. Available: <https://doi.org/10.1145/3437801.3441613>
- [42] <https://clang.llvm.org/docs/LibASTMatchers.html>.
- [43] <https://javaparser.org/>.
- [44] J. Shi, J. Zou, J. Lu, Z. Cao, S. Li, and C. Wang, "Mrtuner: A toolkit to enable holistic optimization for mapreduce jobs," *Proc. VLDB Endow.*, vol. 7, no. 13, pp. 1319–1330, 2014. [Online]. Available: <http://www.vldb.org/pvldb/vol7/p1319-shi.pdf>
- [45] S. Mühlbauer, F. Sattler, C. Kaltenecker, J. Dorn, S. Apel, and N. Siegmund, "Analysing the impact of workloads on modeling the performance of configurable software systems," in *45th IEEE/ACM International Conference on Software Engineering, ICSE 2023, Melbourne, Australia, May 14-20, 2023*. IEEE, 2023, pp. 2085–2097. [Online]. Available: <https://doi.org/10.1109/ICSE48619.2023.00176>
- [46] A. Karthikeyan, N. Natarajan, G. Somashekar, L. Zhao, R. Bhagwan, R. Fonseca, T. Racheva, and Y. Bansal, "Selftune: Tuning cluster managers," in *20th USENIX Symposium on Networked Systems Design and Implementation, NSDI 2023, Boston, MA, April 17-19, 2023*. USENIX Association, 2023, pp. 1097–1114. [Online]. Available: <https://www.usenix.org/conference/nsdi23/presentation/karthikeyan>
- [47] T. Chiba, R. Nakazawa, H. Horii, S. Suneja, and S. Seelam, "Confadvisor: A performance-centric configuration tuning framework for containers on kubernetes," in *IEEE International Conference on Cloud Engineering, IC2E 2019, Prague, Czech Republic, June 24-27, 2019*. IEEE, 2019, pp. 168–178. [Online]. Available: <https://doi.org/10.1109/IC2E.2019.00031>
- [48] S. Bej, N. Davtyan, M. Wolfien, M. Nassar, and O. Wolkenhauer, "Loras: an oversampling approach for imbalanced datasets," *Mach. Learn.*, vol. 110, no. 2, pp. 279–301, 2021. [Online]. Available: <https://doi.org/10.1007/s10994-020-05913-4>
- [49] L. K. Shar, A. Goknil, E. J. Husom, S. Sen, Y. N. Tun, and K. Kim, "Autoconf: Automated configuration of unsupervised learning systems using metamorphic testing and bayesian optimization," in *38th IEEE/ACM International Conference on Automated Software Engineering, ASE 2023, Luxembourg, September 11-15, 2023*. IEEE, 2023, pp. 1326–1338. [Online]. Available: <https://doi.org/10.1109/ASE56229.2023.00094>
- [50] N. Siegmund, A. Grebahn, S. Apel, and C. Kästner, "Performance-influence models for highly configurable systems," in *Proceedings of the 2015 10th Joint Meeting on Foundations of Software Engineering, ESEC/FSE 2015, Bergamo, Italy, August 30 - September 4, 2015*. ACM, 2015, pp. 284–294. [Online]. Available: <https://doi.org/10.1145/2786805.2786845>
- [51] M. Weckesser, R. Kluge, M. Pfannemüller, M. Matthé, A. Schürr, and C. Becker, "Optimal reconfiguration of dynamic software product lines based on performance-influence models," in *Proceedings of the 22nd International Systems and Software Product Line Conference - Volume 1, SPLC 2018, Gothenburg, Sweden, September 10-14, 2018*. ACM, 2018, pp. 98–109. [Online]. Available: <https://doi.org/10.1145/3233027.3233030>
- [52] T. Chen, K. Li, R. Bahsoon, and X. Yao, "FEMOSAA: feature-guided and knee-driven multi-objective optimization for self-adaptive software," *ACM Trans. Softw. Eng. Methodol.*, vol. 27, no. 2, pp. 5:1–5:50, 2018. [Online]. Available: <https://doi.org/10.1145/3204459>
- [53] M. T. Ribeiro, S. Singh, and C. Guestrin, "“why should I trust you?”: Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*. ACM, 2016, pp. 1135–1144. [Online]. Available: <https://doi.org/10.1145/2939672.2939778>
- [54] S. M. Lundberg and S. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, 2017*, pp. 4765–4774. [Online]. Available: <https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html>
- [55] Z. Cao, G. Kuenning, and E. Zadok, "Carver: Finding important parameters for storage system tuning," in *18th USENIX Conference on File and Storage Technologies, FAST 2020, Santa Clara, CA, USA, February 24-27, 2020, 2020*, pp. 43–57. [Online]. Available: <https://www.usenix.org/conference/fast20/presentation/cao-zhen>
- [56] C. Xiang, H. Huang, A. Yoo, Y. Zhou, and S. Pasupathy, "Pracextractor: Extracting configuration good practices from manuals to detect server misconfigurations," in *2020 USENIX Annual Technical Conference, USENIX ATC 2020, July 15-17, 2020, 2020*, pp. 265–280. [Online]. Available: <https://www.usenix.org/conference/atc20/presentation/xiang>
- [57] A. Mordahl, Z. Zhang, D. Soles, and S. Wei, "ECSTATIC: an extensible framework for testing and debugging configurable static analysis," in *45th IEEE/ACM International Conference on Software Engineering, ICSE 2023, Melbourne, Australia, May 14-20, 2023*. IEEE, 2023, pp. 550–562. [Online]. Available: <https://doi.org/10.1109/ICSE48619.2023.00056>
- [58] C. Wang, R. Ko, Y. Zhang, Y. Yang, and Z. Lin, "Taintmini: Detecting flow of sensitive data in mini-programs with static taint analysis," in *45th IEEE/ACM International Conference on Software Engineering, ICSE 2023, Melbourne, Australia, May 14-20, 2023*. IEEE, 2023, pp. 932–944. [Online]. Available: <https://doi.org/10.1109/ICSE48619.2023.00086>
- [59] Y. Zhang, H. He, O. Legunsen, S. Li, W. Dong, and T. Xu, "An evolutionary study of configuration design and implementation in cloud systems," in *43rd IEEE/ACM International Conference on Software Engineering, ICSE 2021, Madrid, Spain, 22-30 May 2021*. IEEE, 2021, pp. 188–200. [Online]. Available: <https://doi.org/10.1109/ICSE43902.2021.00029>