When Do You Repeat Yourself? Voices from the Trenches of Linux Kernel Maintainers on Code Duplication

Luan Arcanjo
luanicaro@usp.br
University of São Paulo
São Paulo, Brazil

David Tadokoro
davidbtadokoro@ime.usp.br
University of São Paulo
São Paulo, Brazil

Marcelo Spessoto
marcelomspessoto@usp.br
University of São Paulo
São Paulo, Brazil

Rafael Passos
rcpassos@ime.usp.br
University of São Paulo
São Paulo, Brazil

Paulo Meirelles
paulormm@ime.usp.br
University of São Paulo
São Paulo, Brazil

Abstract

The Don't Repeat Yourself (DRY) principle is central to software maintainability, but empirical studies challenge its rigid use, describing beneficial cases of duplication. However, these rely on retrospective analyses, leaving a gap in understanding real-time decision-making and socio-technical dynamics. This paper presents an ethnographic study on how the Linux kernel community manages duplication debt via deduplication contributions. Using a clone detection tool called ArKanjo, we conducted a multimethod ethnographic study: first as a complete participant submitting patches to AMDGPU, then as a participant-as-observer mentoring 23 newcomers contributing to AMDGPU and IIO. Analysis of patch reviews suggests maintainers can tolerate duplication to accommodate driver-forking (T0), prioritize readability (T1), reduce integration overhead (T2), lower cognitive load (T3), and preserve performance (T4). Our findings demonstrate that managing duplication in Linux is a nuanced process in which trade-offs among maintainability, clarity, and practicality outweigh dogmatic adherence to the DRY principle. This work also offers a novel ethnographic approach and illustrates how deduplication contributions can aid newcomer onboarding.

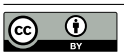
CCS Concepts

• **Software and its engineering → Open source model**; • **Human-centered computing → Open source software**.

Keywords

Linux kernel, Code Quality, Deduplication, Maintainers

ACM Reference Format:
Luan Arcanjo, David Tadokoro, Marcelo Spessoto, Rafael Passos, and Paulo Meirelles. 2026. When Do You Repeat Yourself? Voices from the Trenches of Linux Kernel Maintainers on Code Duplication. In *2026 IEEE/ACM 8th International Conference on Technical Debt (TechDebt '26), April 12–18, 2026, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 6 pages. https://doi.org/XX.XXXX/XXXXXXX.XXXXXXX



TechDebt '26, Rio de Janeiro, Brazil

ACM ISBN XXX-X-XXXX-XXXX-X/XXXX/XX
https://doi.org/XX.XXXX/XXXXXXX.XXXXXXX

1 Introduction

Having tens or hundreds of developers working together on a software product is a complex and demanding task. Similar or even exact copies are a frequent and often unavoidable outcome. Regarded as a bad smell [5] and a major source of technical debt, *code duplication* (or *code cloning*) can negatively impact a project's maintainability, as any change to a repeated code segment (to fix a bug, for instance) requires replication across all other copies. Developers must therefore remain constantly vigilant to synchronize different parts of the codebase, a seemingly simple yet highly error-prone responsibility [9]. Removing duplicated code (*deduplication*) demands great care to preserve the original behavior, with minor mistakes possibly leading to serious problems in the future.

A widespread dogma among developers is the *Don't Repeat Yourself* (DRY) principle [8], which asserts that to develop reliable and maintainable software, "***every piece of knowledge must have a single, unambiguous, authoritative representation within a system.***" Through static program analysis and mining software repository techniques, some empirical studies conducted on living real-world projects have challenged the DRY principle, showing that duplication can be advantageous depending on the context and evolution of the clones [7, 11, 14, 17]. Nevertheless, these types of analysis primarily operate in an archeological perspective; identifying, cataloguing, and drawing conclusions from the already made and final decisions regarding technical debt management. This, in turn, can obscure the reasons and the whole story behind accepting certain clone debts or actively blocking repayments in massive, complex projects.

For instance, the Linux kernel is a foundational *Free/Libre and Open Source Software* (FLOSS) project with more than 27 million lines of code, excluding non-programming code (verifiable through our replication package), and over 2,000 developers involved in version 6.17 [3]. Although it is known that some Linux device drivers deliberately duplicate code [11], it can be hard to determine why a specific clone exists, even when analyzing upstream commits in the mainline and subsystems. One could argue that the debt interest is too low and the debt principal is too high, or that there is a motive for accepting the clone unrelated to technical debt, or that simply no reasonable deduplication proposal has ever been submitted. We are left to speculate (in a well-informed manner) about the daily practices of contributors and maintainers, and the chain of thought behind each commit we analyze.

Leveraging a function-level clone detection tool we developed specifically for the context of Linux, called *ArKanjo* [1], we conducted a multimethod ethnographic study regarding code duplication. Ethnography is a well-established qualitative method for understanding people, their cultures, and work practices [4]. It provides insights into community members' values, beliefs, and practices [15]. In this study, we detected duplications and submitted deduplications (directly or indirectly) to two Linux subsystems.

Through interactions with maintainers, we gained insights into how the Linux community perceives code quality related to code clones in the project's "development trenches". Contrary to the DRY principle, our findings indicate that, in some contexts, maintainers can tolerate duplication to improve readability, enhance performance, or even avoid integration overhead. This empirical work also offers a novel approach to researching technical debt management practices related to code duplication in FLOSS projects, beyond indicating that deduplication contributions are a viable entryway for newcomers in the Linux kernel.

2 Related Work

The works by Kamiya et al. [10] and Li et al. [13] proposed techniques to identify duplications, using the Linux codebase as a case study. Kapser and Godfrey [11] discussed patterns of cloning used in software, including situations where this could be positive, and Rahman et al. [14] contested the notion of clones as a bad smell. Tornhill [17] claims that clones evolving independently might not be problematic, whilst Juergens et al. [9] argue that inconsistent changes to clones that should evolve together are what primarily lead to faults. Wang and Godfrey [18] explores the use of cloning as a development technique for the SCSI drivers in the Linux Kernel, where similar devices had drivers born from code cloning. Harder [7] acknowledges the harmful case of cloned bugs but provides empirical evidence contradicting popular beliefs regarding the adverse effects of clones. The work concludes without providing clear guidance for practitioners and leaving an open path for future research, such as this one.

Our study provides additional empirical evidence and a different methodological perspective for researchers and practitioners interested in technical debt, code quality, and socio-technical practices within the Linux kernel project. We obtained key insights by analyzing the reasoning of maintainers when they accepted or rejected submitted changes (*patches*). In this respect, we make a new contribution to a long-standing conversation among software engineering researchers about when and how duplicate code matters.

3 Research Strategy

First, we needed a tool for detecting code duplication suitable for the Linux context. After reviewing existing static analysis tools like *PMD*, *Duplo*, *Simian*, and *CodeClimate* (analysis of candidate tools is available in our replication package), we found practical limitations when applying them to the Linux kernel. Many were commercially licensed, used only token-based matching, or, like some FLOSS alternatives, were limited to file-scope comparisons, missing intra-file duplications. To address these constraints, we developed ArKanjo, a custom command-line tool for detecting duplicated functions in large C codebases. Released under LGPLv3, it uses TF-IDF vector embeddings and cosine similarity that identify the four clone types described in the literature (Type-1 to Type-4) [2], including within the same file. This lightweight CLI prioritizes maintainability, keeping our focus on code analysis and community interactions rather than tool development.

The Linux kernel is organized into subsystems like the process scheduler, memory management, and device drivers. Each subsystem usually has a dedicated maintainer or team overseeing its development and contributions. Our study focuses on two of these subsystems, selected to represent different contexts for contribution. We explore the *AMDGPU DRM drivers* [1], chosen for their importance within the hardware ecosystem, which is responsible for natively enabling AMD GPUs functionality in Linux. This subsystem is substantial, with over 391,000 lines of code across more than 1,089 files. We also investigate the *Industrial Input/Output* (IIO) subsystem, often considered an accessible entry point for new contributors. The IIO subsystem provides support for devices that perform analog-to-digital or digital-to-analog conversions and contains over 281,772 lines of code in more than 755 files.

Using ArKanjo, we assessed the viability of reducing the identified duplications through a two-phase ethnographic study employing participant observation methodology [6]. In the first phase, adopting the role of *complete participant* [6], we engaged directly with the Linux kernel community by submitting deduplication patches to the AMDGPU drivers. In the second phase, shifting to the role of *participant-as-observer* [6], we introduced a similar activity in a university course, in which students were tasked with proposing patches for the AMDGPU or IIO subsystems. Throughout this process, we provided guidance and mentorship to the students as they interacted with the community.

3.1 Complete Participant Observation Study

To investigate whether we could use the tool to find ways to contribute to the Linux kernel, we first conducted a complete participant observation experiment, where the main author (a graduate student) acted as a first-time contributor to AMDGPU drivers, collecting artifacts of his experience in the process.

We ran the tool on the AMDGPU codebase and manually analyzed the largest duplications by line count, finding one pair we judged promising to try to deduplicate. We approached large duplications as proof of concept to gauge the tool's potential to enable significant impact. Given the duplicated function pair, we observed that the files of the functions contained multiple other clones. Thus, we proposed a simple systematic approach to mitigate all the duplicate functions in the context, not just the initial function pairs. After approaching this refactoring with a systematic strategy, we sent a patch to the AMDGPU drivers mailing list for the feedback of maintainers, while documenting the process, interactions, and impressions (available in our replication package).

In C programming, communication between source files is achieved by creating header files that specify libraries [12]. Since the Linux kernel is primarily written in C, our systematic strategy consisted of eliminating code clones by consolidating them into

[1] In particular, the *Display Core* component, which is the *hardware-abstraction layer* (HAL) for display functionality.

a single library, thereby replacing duplicate instances across the codebase. For each flagged function, we used the tool to locate all duplications of that function in the AMDGPU codebase, since it may have been replicated beyond the two occurrences initially detected. This strategy resulted in a collection of functions and corresponding duplication occurrences in code files.

To identify shared code more effectively, we extended this approach to search for other common functions across all collection files. We then applied specific refactoring methods to each shared function. If the functions were identical across files, we removed the duplicates and created a single function in the library. If modifications existed, we applied case-specific refactoring.

3.2 Participant-as-Observer Observation Study

To broaden our study and observations, we adopted a participant-as-observer approach, involving students as FLOSS project newcomers to make practical contributions to the Linux kernel. This approach formed the core activity of a university course, in which we guided students to remove code duplications in the Linux kernel as one possible pathway for contributing to the project in the first phase of the course.

The 2025 course offering had 37 students (25 undergraduate and 12 graduate), who were asked to form groups of two or three members, but graduate students could work individually. To assist students in this task, we (a professor and three more experienced graduate students) prepared several alternatives for contributing to the Linux kernel, including simpler contributions such as cleaning coding style issues. In this context, we specified two options for removing duplications in the Linux kernel.

Deduplication Option 1. We ran ArKanjo on the IIO subsystem to generate a list of duplicated function pairs, then curated it to highlight actionable cases for students. Specifically, we (i) kept only pairs within the same file, since cross-driver duplications could involve multiple maintainers or distinct interfaces; (ii) removed very short duplications; (iii) removed semantical Type-3 and Type-4 clones, for the higher complexity in refactoring them; and (iv) manually ranked the remaining pairs, adding annotations to guide students toward entries likely to be accepted. In this option, students were tasked with selecting a recommended entry, devising a way to remove the duplication, and submitting their patches to the IIO mailing list for review.

Deduplication Option 2. We offered an experience similar to our initial complete participant observation, where students managed the entire workflow: running the tool, analyzing results to identify potential deduplications, creating a patch, and submitting it to the driver's maintainers. This option provided more freedom of choice but made the students' task more complex.

Of the students who chose to work on the tool-related tasks, 23 (16 undergraduate and 7 graduate) formed 11 groups to pursue the first alternative. One graduate student opted for the second alternative. Notably, none of these students had prior experience contributing to the Linux kernel, making this their first attempt to submit a patch as newcomers to the project.

It is important to stress that, **no matter the chosen approach, we supported all groups by providing review cycles on their contributions before submission**, to prepare students for the interactions with the maintainers and alleviate rough, but natural, newcomers' mistakes. After the submission, we helped students understand the feedback from maintainers and gave directions on how to move forward. This close, yet indirect, participation through intimate mentoring characterizes the second phase as our participant-as-observer observation.

For this second phase, our primary data source was the mailing list threads, where all development occurred (also available in our replication package). Using blog posts, the groups documented their experiences submitting a patch to the Linux kernel, and we also leveraged those to complement our observations.

4 Results and Discussion

In the **complete participant observation** study (first phase), we sent the first revision of the patch on August 9, 2024, and a resend followed on October 9, 2024, due to a lack of initial feedback. Nevertheless, for the first two versions, the maintainers' response requested minor changes for coding style, license use, and alignment with existing conventions. In the third version, the feedback was to move the new generic library functions into an existing file rather than creating a new one. Then, the fourth version was accepted and integrated into the kernel on February 25, 2025, removing 406 lines in total and impacting three generations of drivers.

Since we experienced a notable delay in the review process, we directly contacted a maintainer to understand the cause. From this conversation, we understood that it is commonplace for drivers to be cloned (sometimes entirely) to provide a solid base and enhance independence between them, allowing developers to make changes to a specific one without needing to check for regressions on others, which agrees with Tornhill [17]. Also, this practice is the hardware variation of the forking pattern defined by Kapser and Godfrey [11]. Even so, in the AMD Display Core, duplication of drivers is a *Self-Admitted Technical Debt* (SATD) [2]. In this framing, we can attribute the delay to the change impacting multiple drivers, which resulted in greater principal on the maintainer's side to ensure the repayment was stable. Prioritization of other repayments may also have contributed to the delay. With that in mind, this complete participant observation produced our initial takeaway:

> **Takeaway T0 (Driver Forking):** *Driver developers do not necessarily view duplicated code negatively, considering the interest low and the principal high, yet being open to repayments and even fostering them.*

What is notable is that this insight, grounded in the practice of an industry-caliber project such as the Linux kernel, corroborates the findings of related work, whilst adding nuance to technical debt management decisions.

To further explore, we examined other contributions on code duplication refactoring and analyzed how different maintainers responded to these patches. This broader analysis aimed to capture diverse maintainer perspectives and contextual factors influencing their decisions.

[2] https://www.kernel.org/doc/html/v6.17/gpu/amdgpu/display/display-contributing.html#reduce-code-duplication.

Table 1: Summary of newcomer deduplication contributions.

GID	SS	SML	Patch Status	Takeaway	Diff
0	AMD	100%	Accepted (v4)	T0	+114/-520
1	IIO	100%	Dropped (v2)	T1, T2	+7/-14
2	IIO	100%	Dropped (v1)	T1	+23/-27
3	IIO	100%	Accepted (v1)	None	+12/-30
4	IIO	100%	Accepted (v3)	T1	+2/-7
5	IIO	90%	Accepted (v1)	None	+17/-70
6	IIO	90%	Dropped (v1)	T3	+95/-92
7	IIO	90%	Accepted (v3)	None	+20/-36
8	IIO	90%	Accepted (v4)	T4	+20/-16
9	IIO	90%	Accepted (v6)	T4	+149/-243
10	IIO	90%	Dropped (v1)	T1, T4	+32/-58
11	IIO	100%	Dropped (v1)	None	+2/-12
	AMD	90%	Accepted (v2)	None	+90/-489

Table 1 summarizes all the contributions sent by the main author and student groups (GID) [3]. It contains the information about the chosen subsystem (SS) and the similarity threshold (SML). Each entry also has the patch status, takeaways that may have surfaced from the interaction, and a code differential corresponding to the final patch versions before it was dropped or accepted into the code base. The code differential for a set of patches corresponds to the sum of added and removed lines of each patch.

We derived four takeaways from our **participant-as-observer observation**. We substantiate these takeaways with quotes from the maintainers that originated during the review process[4].

Groups 1 and 4 worked on similar duplications using the extract method on predicate functions (functions that return a boolean) that, essentially, were the negation of one another. Group 4 initially used macros, but, after feedback from maintainers, both solutions converged toward keeping one of the functions while modifying the other to return the negation of the former. Both cases suffered substantial mutations, even though they were correct in their first iterations, and these changes were primarily driven by the feedback of maintainers regarding readability. For Group 1, a maintainer said that *"[...] the naming as* `_reg_check()` *is not helpful as it doesn't indicate anything specific is being checked."*, while for Group 4, the feedback explicitly highlighted the precedence of readability over deduplications: *"I think the old code is more readable than hiding the values in a macro even if it is duplicating a few lines of code."*

Even though the deduplication of Group 2 was not similar to that of Group 1 and Group 4 in terms of implementation (it used the parametized method), it also brought forth the readability factor: *"In my view this isn't a significant enough reduction to justify the more complex code."*

These experiences revealed a recurring concern among maintainers beyond functional correctness or redundancy reduction. Instead, it reflected a preference for code clarity and comprehension. From these consistent insights that originated from multiple review interactions, we formulated our first takeaway:

[3]Since the first author was also a newcomer, the first phase experience was included as Group 0.

[4]We **do not claim** these quotes fully represent the maintainers' stance on code duplication or any technical debt management matter.

> **Takeaway T1 (Readability):** *Maintainers prioritized readability over removing duplications in many different contexts. Accepting debt from duplication may be worth it if it results in easier-to-read, more understandable code.*

Still focusing on Group 1, maintainers questioned the merit of the proposed patch. While the previous interactions highlighted how maintainers often favor readability over strict code deduplication, this case added a new dimension to our understanding: the *cost-benefit* evaluation of proposed changes. A maintainer pointed out that the effort required to propagate such modifications after acceptance (i.e., during the upstreaming process) could outweigh the perceived benefits: "*[...] such patches might not worth it since the proposed improvement is very small (and questionable) while the upstreaming process still requires some effort.*" From this feedback, we crafted our second takeaway:

> **Takeaway T2 (Integration Overhead):** *Maintainers may drop a duplication repayment if it is not considered impactful enough, because just integrating it into the Linux upstream incurs considerable principal.*

Groups 3, 6, and 10 used the parameterized method by introducing a new struct used as a parameter to a generic function. Notwithstanding that the same maintainer reviewed all three contributions, only Group 3 had its patch accepted. Our understanding was that this change, in particular, acted on a simple iterative algorithm that (colaterally) also fixed a bug, which favored its straightforward acceptance.

Although removing duplications does not necessarily mean reducing lines of code, the patch from Group 6 resulted in a positive net change in lines of code, which the maintainer pointed out as a drawback. Still, the real reason for the rejection was that the deduplication scattered parts of a complex code that resided close together before: *"Wrapping this up doesn't provide any real advantage, requiring as it does the reviewer to look at this function AND where the value is set rather than seeing them in one place."* Needing to jump to a function definition (to see its behavior) or a function call (to see its concrete parameters) to understand code is a cognitively demanding task for developers [16]. Merging these informations, we derived our third takeaway:

> **Takeaway T3 (Cognitive Load):** *Maintainers can justify duplications if they avoid having to jump back and forth in the code when parsing through the file, as some repayments that do not hinder readability directly, can also incur debt.*

The reason for rejecting the patch from Group 10 involved the deduplication lowering readability, as the maintainer was *"[...] not sure the code reduction is sufficiently to cover the resulting loss of readability."*, reinforcing the takeaway T1.

In the interaction of Group 10, there was also a hint that performance could be considered when evaluating the worth of a deduplication, whilst reaffirming that readability was a priority in that context: *"I'd be slightly interested to see the optimized output*

of the two approaches, but this is far from a high performance path so we care a lot more about readability here." Nonetheless, this notion surfaced aggressively with Group 10, where the contribution could completely remove both duplicated functions using the inline method by leveraging helpers; however, one of the functions had a call that, if eliminated, would result in a significant performance impact: *"Note that is not an appropriate change for the large reads though [...] This is the one case were* `spi_write_then_read()` *is probably not appropriate due to the large buffers that are potentially involved."* These interactions, coupled with the low-level development context of Linux and driver development, which is usually related to devices with limited resources, produced our fourth and final takeaway:

> **Takeaway T4 (Performance):** *Interactions with maintainers hinted that duplications can be forgiven if they improve performance in contexts where resources are limited. In these scenarios, avoiding performance debt is prioritized.*

Groups 7 and 8 used the extract method using helpers to remove duplicated code in the function's body. Group 7 refined its patch from a trivial maintainers' feedback, resulting in a smooth acceptance process that did not produce any insights. Group 8 followed a similar course, but generated an interaction reinforcing the takeaway T4. In the feedback of version 1, a maintainer claimed that *"Even though there is less code repetition, we now have an extra comparison [...]"*, with another maintainer corroborating this in version three by saying: *"There is always a balance/trade-off between modularity and execution speed. I agree with* `anonymous-maintainer`*'s reply in the first patch [...]"*, regardless of the fact that the patch was eventually accepted in version four.

We did not discuss in detail the contributions of Group 5 and Group 11, as they did not result in (or influence) any nuanced insights or corroborations. However, they were still meaningful, as they made impactful contributions that reduced significant duplication, especially in Group 11, where a single student executed the entire workflow we did in the first phase and removed 400 net lines of code.

In this sense, we briefly examine another aspect of importance stemming from this work: using ArKanjo to create deduplication opportunities in the Linux kernel can be an excellent introduction for newcomers.

To illustrate this point, we display some statistics concerning the 13 contributions contemplated in this paper: (i) Collectively, the patches proposed the insertion of 585 lines and the removal of 1626 lines; (ii) 8 out of 13 contributions were accepted, resulting in a success rate of 62%; (iii) for accepted contributions, the code diff value was +255/-1152, averaging +31.9/-144.0 additions/removals; (iv) For rejected contributions, the code diff value was +181/-231, averaging +36.2/-46.2 additions/removals.

4.1 Threats to Validity

We outline limitations using standard validity categories.

Internal Validity. Selection bias may affect findings. Targeting large duplications in AMDGPU likely captured complex, conservative cases, while curating simpler duplications for participants may have oversimplified challenges and favored easier patches. Results may not generalize to medium-complexity or inter-file duplications.

External Validity. Findings are based on two Linux kernel subsystems (AMDGPU and IIO) and 23 student contributors. The kernel is a unique FLOSS project with specific norms, so, maintainer perspectives may vary. Participants' lack of previous kernel experience also limits generalizability to experienced contributors.

Construct Validity. Only Type-1 and Type-2 duplications were studied, excluding more complex clones, although ArKanjo also detects Type-3 and Type-4 clones. This restricts insights into maintainers' attitudes toward semantically similar duplications.

Conclusion Validity. The ethnographic approach provides qualitative insight but not statistical generalization. Conclusions are drawn from 13 patch interactions. In this sense, a larger sample would strengthen robustness.

5 Concluding Remarks

This ethnographic study, combining complete participant and participant-as-observer observations, provided a realistic view of the opportunities and challenges in repaying duplication debts in the Linux kernel. It shows that first-time Linux contributors could effectively use the ArKanjo tool to identify duplicated functions and submit meaningful patches.

More interestingly, the study uses a novel approach to corroborate and nuance consolidated conclusions in the technical debt literature that run counter to the DRY principle. As evidenced by our takeaways, desirable repayments can be slow due to prioritization and high principal from driver forking (T0). Readability can be regarded as a greater debt than duplication (T1), while the effort to propagate changes adds considerably to the principal (T2), and cognitive overhead is also considered (T3). Finally, preserving performance in resource-constrained environments (T4) emerged as an additional justification for tolerating duplication, suggesting more specific debt management decisions for Linux.

Without disregarding its limitations, this work shows that the Linux community has particular technical debt management practices, carefully weighing the pros and cons of accepting versus repaying duplication debts. Maintainability, clarity, and practical trade-offs shape maintainers' judgments, suggesting that consciously asking *"When do you repeat yourself?"* can be better than mindlessly following the DRY principle.

6 Replication Package

All anonymized review threads are available in this Zenodo repository: doi.org/10.5281/zenodo.17503684. Replication resources include the ArKanjo repository (github.com/arkanjo-tool/arkanjo), its documentation (arkanjo-tool.github.io/), and an analysis of candidate tools (github.com/arkanjo-tool/arkanjo/wiki/Evaluation-of-other-duplication-tools). The container for the Linux v6.17 LoC statistic from section 1 is also provided (github.com/linux-duks/linux-programming-loc).

Acknowledgments

This study is funded by the São Paulo Research Foundation (FAPESP) and the São Paulo State Data Analysis System Foundation (SEADE), under grants 2023/18026-8 and 2025/05395-0.

References

[1] Luan Arcanjo, David Tadokoro, and Paulo Meirelles. 2025. ArKanjo: a tool for detecting function-level Code Duplication in the Linux Kernel. In *DebConf25*, Olivier Barais and Bastien Roucariès (Eds.). IRISA, Brest, France, 3. https://hal.science/hal-05335545

[2] Chang-Feng Chen, Azlan Zain, and Kai-Qing Zhou. 2022. Definition, approaches, and analysis of code duplication detection (2006–2020): a critical review. *Neural Computing and Applications* 34 (08 2022), 1–31. doi:10.1007/s00521-022-07707-2

[3] Jonathan Corbet. 2025. *Development statistics for 6.17.* Linux Weekly News. https://lwn.net/Articles/1038358/ [Online; Last accessed on Jan. 25th, 2026].

[4] W Alex Edmonds and Thomas D Kennedy. 2016. *An applied guide to research designs: Quantitative, qualitative, and mixed methods.* Sage Publications.

[5] Martin Fowler. 2018. *Refactoring: improving the design of existing code* (2 ed.). Addison-Wesley Professional.

[6] Raymond L. Gold. 1958. Roles in Sociological Field Observation. *Social Forces* 36 (1958), 217–223. doi:10.2307/2573808

[7] Jan Harder. 2017. *Software Clones-Guilty Until Proven Innocent?* Logos Verlag Berlin GmbH.

[8] Andrew Hunt and David Thomas. 2000. *The pragmatic programmer: from journeyman to master.* Addison-Wesley Longman Publishing Co., Inc., USA.

[9] Elmar Juergens, Florian Deissenboeck, Benjamin Hummel, and Stefan Wagner. 2009. Do code clones matter?. In *2009 IEEE 31st International Conference on Software Engineering*. 485–495. doi:10.1109/ICSE.2009.5070547

[10] T. Kamiya, S. Kusumoto, and K. Inoue. 2002. CCFinder: a multilinguistic token-based code clone detection system for large scale source code. *IEEE Transactions on Software Engineering* 28, 7 (July 2002), 654–670. doi:10.1109/TSE.2002.1019480

[11] Cory J. Kapser and Michael W. Godfrey. 2008. "Cloning considered harmful" considered harmful: patterns of cloning in software. *Empirical Software Engineering* 13, 6 (Dec. 2008), 645–692. doi:10.1007/s10664-008-9076-6

[12] Brian W Kernighan and Dennis M Ritchie. 1988. *The C programming language.* prentice-Hall.

[13] Z. Li, S. Lu, S. Myagmar, and Y. Zhou. 2006. CP-Miner: finding copy-paste and related bugs in large-scale software code. *IEEE Transactions on Software Engineering* 32, 3 (March 2006), 176–192. doi:10.1109/TSE.2006.28

[14] Foyzur Rahman, Christian Bird, and Premkumar Devanbu. 2012. Clones: what is that smell? *Empirical Software Engineering* 17, 4 (01 Aug 2012), 503–530. doi:10.1007/s10664-011-9195-3

[15] Helen Sharp, Yvonne Dittrich, and Cleidson R. B. de Souza. 2016. The Role of Ethnographic Studies in Empirical Software Engineering. *IEEE Transactions on Software Engineering* 42, 8 (2016), 786–804. doi:10.1109/TSE.2016.2519887

[16] Jonathan Sillito, Gail C. Murphy, and Kris De Volder. 2006. Questions programmers ask during software evolution tasks. In *Proceedings of the 14th ACM SIGSOFT International Symposium on Foundations of Software Engineering* (Portland, Oregon, USA) *(SIGSOFT '06/FSE-14)*. Association for Computing Machinery, New York, NY, USA, 23–34. doi:10.1145/1181775.1181779

[17] Adam Tornhill. 2018. *Software Design X-Rays: Fix Technical Debt with Behavioral Code Analysis* (1st ed ed.). The Pragmatic Programmers, LLC.

[18] Wei Wang and Michael W. Godfrey. 2011. A Study of Cloning in the Linux SCSI Drivers. In *2011 IEEE 11th International Working Conference on Source Code Analysis and Manipulation*. 95–104. doi:10.1109/SCAM.2011.17