

Peer Review

Review of: "R+R: Security Vulnerability Dataset Quality Is Critical"

Evgenii Arsentev¹

1. Independent researcher

1. What the work claims

The authors take the VulRepair dataset (the Hugging Face merge of Big-Vul and CVEfixes used by Fu et al. and later by Zhang et al.) and ask how much of the published vulnerability-repair performance survives once the data are cleaned. They first replicate four models (VulRepair/CodeT5, CodeBERT, GraphCodeBERT, UniXcoder) as the mean of six seeds, then remove exact duplicates (Table 2: 1,593 in-set duplicates in train, 91 in test, 796 cross-set), then remove "CWE-duplicates", i.e., identical code carrying different CWE tags (Table 3: 1,858, 111, and 923). Each cleaning step is run twice, once dropping the cross-set copies from the training side (RQ2A, RQ3A) and once from the test side (RQ2B, RQ3B). The headline is that VulRepair falls from a replicated 40.42% perfect prediction (Table 5) to 8.91% (Table 7, RQ2A) and 13.17% (Table 9, RQ2B), with the other three models falling further in relative terms.

Two further contributions sit on top of this. A manual audit of the top-10 CWE test samples that remain after deduplication (Table 15) finds that only 30 of 68 are accurately labelled, 38 of 68 are complete, and 21 of 68 are both. And a transfer-learning experiment pre-trains on a deduplicated version of the Chen et al. bug-fix corpus before fine-tuning on RQ3B, lifting all four models to 18.67–26.07% at beam 50 (Table 17, column TL). The authors also show that the bug-fix corpus used for VQM pre-training overlaps heavily with the VQM fine-tuning splits (§6). Scripts are released on GitHub.

2. Whether the conclusion holds on the data

I think the central claim holds, and holds by a wide margin. This is the kind of paper the field needs more of: a careful, unglamorous replication that reports per-seed numbers for every configuration (Appendix Tables 18–35), shows the arithmetic of every dataset split (Tables 1–3, 6, 8, 10, 12), and states its own threats honestly in §5. I checked the split arithmetic, and it reconciles: $5,183 - 796 = 4,387$ (Table 6), 1,615

- 796 = 819 (Table 8), $4,918 - 923 = 3,995$ (Table 10), $1,595 - 923 = 672$ (Table 12). I also recomputed several appendix averages from the raw seed values; for example, Table 23, row VR, RQ1 gives $(22+18+18+12+17+17)/(6 \times 53) = 104/318 = 32.7\%$, matching Table 14. The top-10 column totals in Table 14 (449, 209, 175) also add up. A drop from roughly 40% to roughly 9–13% is far outside the seed-to-seed spread the authors themselves report, so no statistical test is needed to accept the direction of the effect.

Where I would ask for more care is in the secondary claims, which in places say more than the tables show.

The A/B comparison mixes two changes. §4.2 attributes the RQ2A→RQ2B improvement to "the increase in training samples from 4,387 to 5,183". But the test set also changes, from 1,615 (Table 6) to 819 (Table 8). RQ2B evaluates on a different, smaller population, so the gain cannot be assigned to training size alone. The same applies to RQ3A versus RQ3B (test 1,595 versus 672). The text also says the gain is "about 50% over RQ2A in all cases". From Tables 7 and 9, the ratios are $13.17/8.91 = 1.48$ (VulRepair), $8.83/5.58 = 1.58$ (CodeBERT), $9.22/5.31 = 1.74$ (GraphCodeBERT), and $9.10/4.82 = 1.89$ (UniXcoder). Only VulRepair is close to 50%; the other three range from 58% to 89%. A cleaner design would evaluate both A and B models on the common 819-sample (or 672-sample) test set, which the authors already have.

The bug-fix-only baseline is described in the wrong direction for three models. After Table 17, the text says pre-training alone achieves "results nearly as good as models trained on the de-duplicated VulRepair dataset". In Table 17, column BF versus column VR, that is true only of VulRepair (6.55% vs 10.27%). For CodeBERT (11.76% vs 5.38%), GraphCodeBERT (11.76% vs 6.25%), and UniXcoder (11.31% vs 6.18%), the bug-fix-only model is roughly twice as good as the model fine-tuned on vulnerability data. That is actually a more striking finding than the one stated, and the Conclusion (§7, paragraph 3) should reflect it.

The beam comparison sentence does not quite match Table 16. The paper says TL at beam 1 is "slightly better" than RQ3B at beam 50 and TL at beam 5 is "more than twice that of RQ3B with a beam size of 5". RQ3B is reported only at beam 50, so I assume the second comparison is also against beam 50. For VulRepair, Table 16 TL beam 5 is 20.2% against 10.27%, a ratio of 1.97, so "more than twice" holds for the three encoder models but not for VulRepair. For CodeBERT, TL beam 1 (12.5%) against RQ3B (5.38%) is more than double, which is not "slightly".

The manual audit is small and does not cover the largest categories. The abstract generalizes Table 15 to "56% of the samples had incorrect labels and 44% comprised incomplete samples". The percentages are right ($1 - 30/68 = 56\%$, $1 - 38/68 = 44\%$, $21/68 = 30.9\%$), but the denominator is 68 samples, of which

62 come from just two CWEs (CWE-787: 33, CWE-416: 29). The two biggest top-10 categories in RQ2B, CWE-125 (68 samples) and CWE-20 (73 samples, Table 14), were not audited, CWE-79 has no RQ2B samples, and CWE-434 none at all, so "8 of the top 10" is in practice six categories with data. The authors say as much in §4.4.1 ("the statistical importance of the results becomes much less apparent"); the abstract should carry the same caveat.

One more point I could not resolve from the text: whether the Chen et al. bug-fix corpus was checked for overlap with the RQ3B test set. The paper does exactly this check for the VQM bug-fix corpus in §6 and finds 511 of 1,090 test samples leaked. If the same check was not done for the corpus used in RQ5, the transfer-learning gains in Tables 16 and 17 could include some of the leakage the paper warns about. I suspect the overlap is small (the corpus is tagged CWE-000 and is generic), but it should be shown.

3. Reproducibility

This is better than most papers in the area. The GitHub repository is linked in §1, the six seeds are named in every appendix table, hardware is given (A100-SXM4-80GB, §3.4.3), and the pre-training hyperparameters are listed in §3.4.1 (Adam, lr 2e-5, sequence lengths 512/256, 75 epochs with early stopping after three epochs, batch 128). The deduplication method is simple and exactly specified (Pandas `drop_duplicates()`, keep first; CWE tag stripped for the consistency pass).

What I could not find:

- The exact pretrained checkpoints used for each model (for example, which CodeT5 size and Hugging Face identifier) and the commit hashes of the upstream repositories that were modified (§5 lists the changes but not the versions).
- The fine-tuning batch size and whether early stopping was used in fine-tuning; §3.4.2 gives 75 epochs but no stopping rule.
- Which data served as the validation split during fine-tuning. §3.1.4 says the dataset "consists of a train and test set" and in the next sentence refers to "both the train and validation set".
- Whether exact match is computed on raw token strings or after whitespace normalization.
- For the manual audit, the per-sample decisions are said to be on GitHub, but the paper gives no agreement measure. Only samples that "raised questions" were seen by the second author (§5).

Table 31 notes that pre-training was done once, on seed 26312. The BF column is therefore a single run, while every other number in the paper is a six-seed mean. Given the authors' own observation that

results can vary by more than 1.5× across seeds (§3.5; for CodeBERT in Table 18 the range is 0.2327 to 0.3992), I would flag this asymmetry in Table 17.

4. *What specifically to fix*

1. §4.2, RQ2B result paragraph: replace "about 50% over RQ2A in all cases" with the actual ratios (1.48, 1.58, 1.74, 1.89), and either evaluate RQ2A and RQ2B models on the shared 819-sample test set or state that the test sets differ, so the gain is not attributed to training size alone. Same for RQ3A/RQ3B in §4.3.
2. Paragraph after Table 17 and §7, paragraph 3: rewrite the BF-versus-VR comparison. Three of four models do better on bug-fix pre-training alone than on RQ3B fine-tuning alone.
3. Same paragraph: check the "slightly better" and "more than twice" claims against Table 16. As written, "RQ3B with a beam size of 5" refers to a number the paper never reports.
4. Table 13, row VulRepair: the denominator is printed as 40.24; the replicated value in Table 5 is 40.42. With 40.42 the ratio is $10.27/40.42 = 25.4\%$, not 25.5%.
5. Table 7, row GraphCodeBERT: $5.31/38.51 = 0.1379$, so 13.8% rather than 13.7%.
6. Bug-fix corpus size: §1 and §6 say 534,848 training samples; §3.1.3 and Table 1 say 534,858. One of them is a typo.
7. §3.1.4 gives 1,707 test samples; Tables 2, 3, and 4 give 1,706.
8. §6, VQM analysis: $2,362 - 666 = 1,696$, not 1,695, and $2,362 - 782 = 1,580$, not 1,579. Probably an off-by-one in how the first occurrence is counted, but worth checking, since the "all of the validation set" claim rests on 1,579.
9. Add a cross-corpus overlap check between the deduplicated Chen bug-fix corpus and the RQ3B test set (as done for VQM in §6), and report the count before Table 16.
10. Abstract: say that the label audit covers 68 test samples from six populated top-10 categories, and that CWE-125 and CWE-20 were not audited.
11. Main result tables (5, 7, 9, 11, 13, 17) would read better with a standard deviation or min-max across seeds next to each mean. The data are already in the appendix.
12. Table 5, column "Change": these are percentage points, not percent.
13. Small editorial items. §3.1.2 refers to Sections 4.2 and 4.3 for uniqueness and consistency; those definitions are in §2.1 and §2.2. §4.4.1 and §4.4.2 cite "Table XV" for Table 15. Tables 24–30 label GraphCodeBERT as "GDB", Table 23 as "GCB".

14. §2.5: I would soften the sentence about the ACM badge audit. The point stands without the inference about what the auditors did.

5. Recommendation

Minor revisions. The main finding is well supported, carefully replicated over six seeds, and documented down to per-seed counts, which makes it genuinely useful to anyone still reporting results on this dataset. The corrections I list are mostly about bringing the prose in line with the tables: the RQ2A/RQ2B attribution, the BF-versus-VR sentence, a handful of arithmetic and typographic inconsistencies, and a leakage check for the RQ5 pre-training corpus. None of them changes the headline, and all can be done with data the authors already hold.

Declarations

Potential competing interests: No potential competing interests to declare.