VQM Dataset Analysis

THIS IS FOR THE **BugFix Dataset**

TABLE - 1
WITH CWE-TAG

Dataset	Total Samples	Duplicate samples	Unique samples
Train	21246	17303	3943
Validation	2361	666	1695

TABLE - 2
WITHOUT CWE-TAG

Dataset	Total Samples	Duplicate samples	Unique samples
Train	21246	17303	3943
Validation	2361	666	1695

NOW, WE WILL REMOVE THE <S2SV_StartBug> and <S2SV_EndBug> AND THEN COMPARE

TABLE - 3
WITHOUT CWE-TAG

Dataset	Total Samples	Duplicate samples	Unique samples
Train	21246	18622	2624
Validation	2361	782	1579

We found **1319** (3943 - 2624) TRAIN and **116** (1695 - 1579) VALIDATION IN-FILE samples.

WHEN I DID CROSS FILE BETWEEN THE TWO FILE.

```
Found 1579 matching entries.
We have 0 unique entries written to ./vrepair_non_domain_data/final/no_val_in_train.csv.
Here are the full rows:
Empty DataFrame
Columns: [cwe_id, source, target]
Index: []
```

THIS IS FOR THE **VULNERABILITY DATASET**

TABLE - 3
WITHOUT CWE-TAG

Dataset	Total Samples	Duplicate samples	Unique samples
Train	3790	1	3789
Validation	536	0	536
Test	1090	0	1090

CrossFile between Vulnerability Dataset and BugFix dataset

- **511 matching entries out of 1090 from the vulnerability test file are present in the Bugfix dataset. We are left with 579 unique entries.**

```
Found 511 matching entries.
We have 579 unique entries written to ./cve_fixes_and_big_vul/final/test_removed_non_domain.csv.
```

- **We have 243 matching entries out of 536 from the vulnerability validation file in the BugFix dataset. We are left with 293 unique entries.**

```
Found 243 matching entries.
We have 293 unique entries written to ./cve_fixes_and_big_vul/final/val_removed_non_domain.csv.
```

- **We have 1747 matching entries out of 3789 from the vulnerability validation file in the BugFix dataset. We are left with 2044 unique entries.**

```
Found 1747 matching entries.
We have 2044 unique entries written to ./cve_fixes_and_big_vul/final/train_removed_non_domain.csv.
```

TABLE - 4
WITHOUT CWE-TAG AND <S2SV_StartBug> and <S2SV_EndBug>

Dataset	Total Samples	Duplicate samples	Unique samples
Train	3790	1747	2044
Validation	536	243	293
Test	1090	511	579