

FEMP-X

A Witness-Augmented Dataset for Candidate Functionally Equivalent Java Method Pairs

Yoshiki Higo

The University of Osaka

ICSME 2026 · Tool Demonstration and Data Showcase Track

16 September 2026 · Benevento, Italy

github.com/YoshikiHigo/FEMP-X

Do these two methods do the same thing?

```
Integer method1(String str) {  
    if (str == null) return null;  
    return Integer.decode(str);  
}
```

```
Integer method2(String value) {  
    if (value == null) return null;  
    return Integer.valueOf(value);  
}
```

Both take a String. Both return an Integer. Both handle null the same way.

They don't.

```
method1 -> Integer.decode(str)
method2 -> Integer.valueOf(value)
```

```
@Test
@BehavioralDifference(DifferenceType.RETURN_VALUE)
void methodsInterpretLeadingZeroDifferently() {
    assertEquals(Integer.valueOf(8), subject.method1("010"));
    assertEquals(Integer.valueOf(10), subject.method2("010"));
}
```

Not a label. A JUnit test that our workflow generated — and then actually ran.

The dataset answers with an executed test

```
method1 -> Integer.decode(str)
method2 -> Integer.valueOf(value)
```

```
@Test
@BehavioralDifference(DifferenceType.RETURN_VALUE)
void methodsInterpretLeadingZeroDifferently() {
    assertEquals(Integer.valueOf(8), subject.method1("010"));
    assertEquals(Integer.valueOf(10), subject.method2("010"));
}
```

"010" — **decode** reads the leading zero as octal and returns 8. **valueOf** is always decimal: 10.

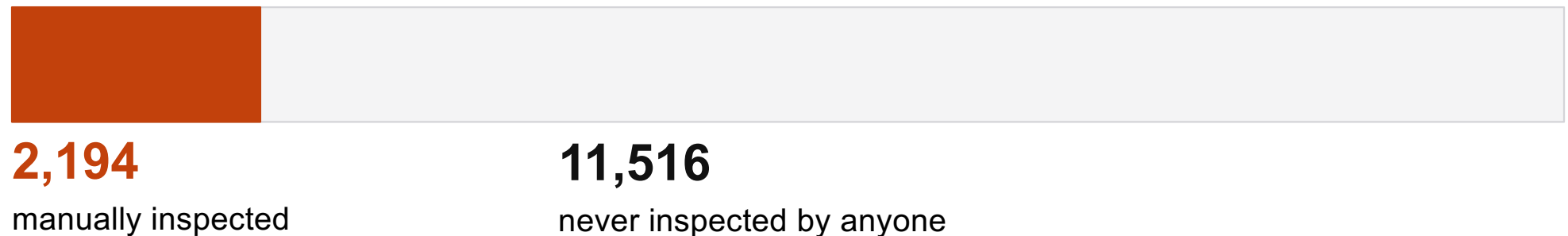
This is one record of FEMP-X. The dataset has **5,627** of them.

The bottleneck is curation, not collection

Datasets of functionally equivalent code can be used to evaluate clone detection detectors, semantic code search, and test generation.

Finding candidate pairs can be automated. Deciding whether a candidate really is equivalent cannot: equivalence is not provable in general, so datasets fall back on human inspection.

FEMPD candidate pool — **13,710** pairs, all compilable and executable



Y. Higo, "Dataset of Functionally Equivalent Java Methods and Its Application to Evaluating Clone Detection Tools", IEICE Transactions on Information and Systems, Vol.E107-D, No.6, pp.751--760, June 2024.

Falsify, don't prove

Prove equivalence

Must hold for every possible input.
Undecidable in general.

Find one input where they differ

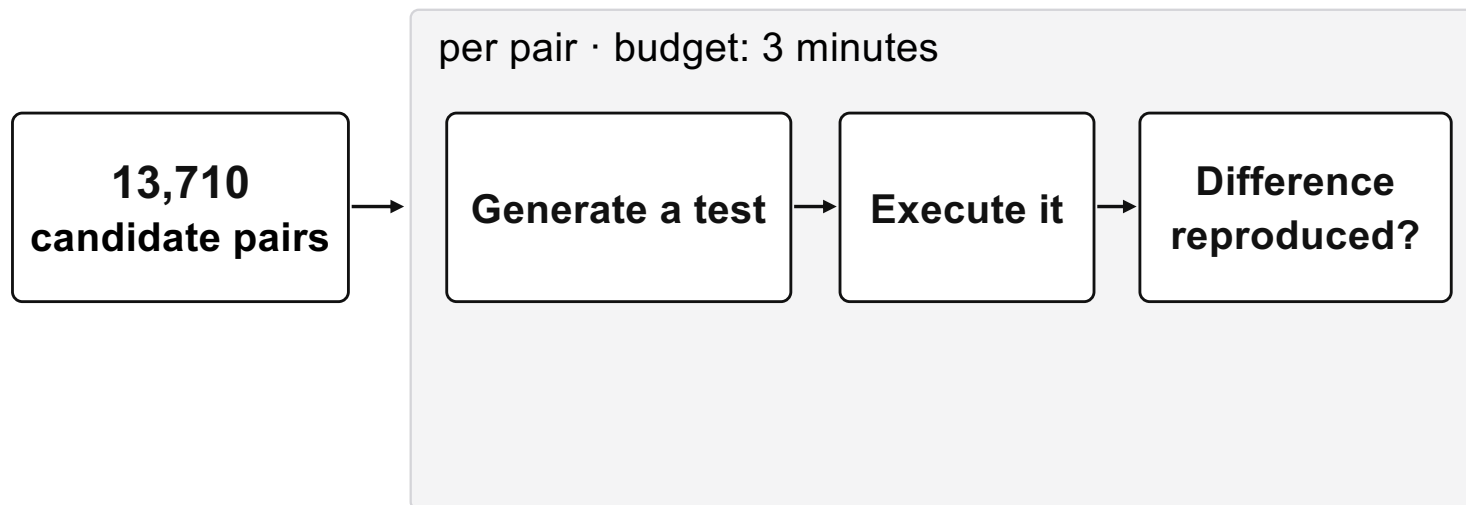
One execution settles it.
Differential testing, one pair at a time.

The asymmetry that matters

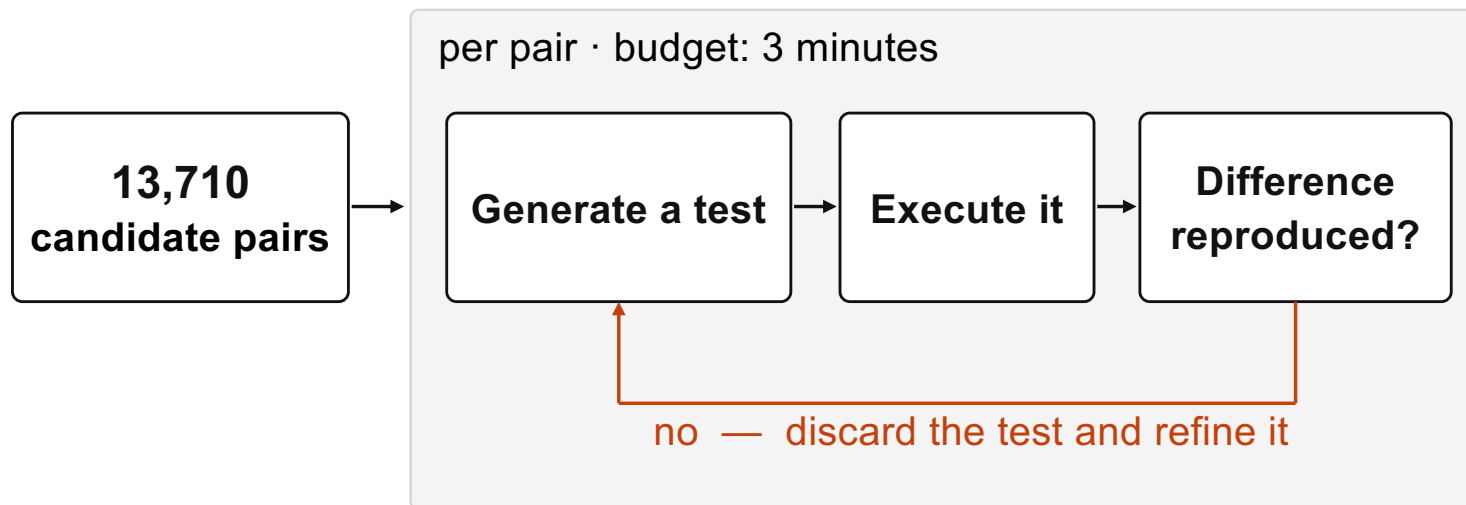
witness found $\Rightarrow$ the pair is inequivalent, and you can re-run the evidence yourself.

no witness found $\Rightarrow$ **not** proven equivalent. It only survived our attempt to break it.

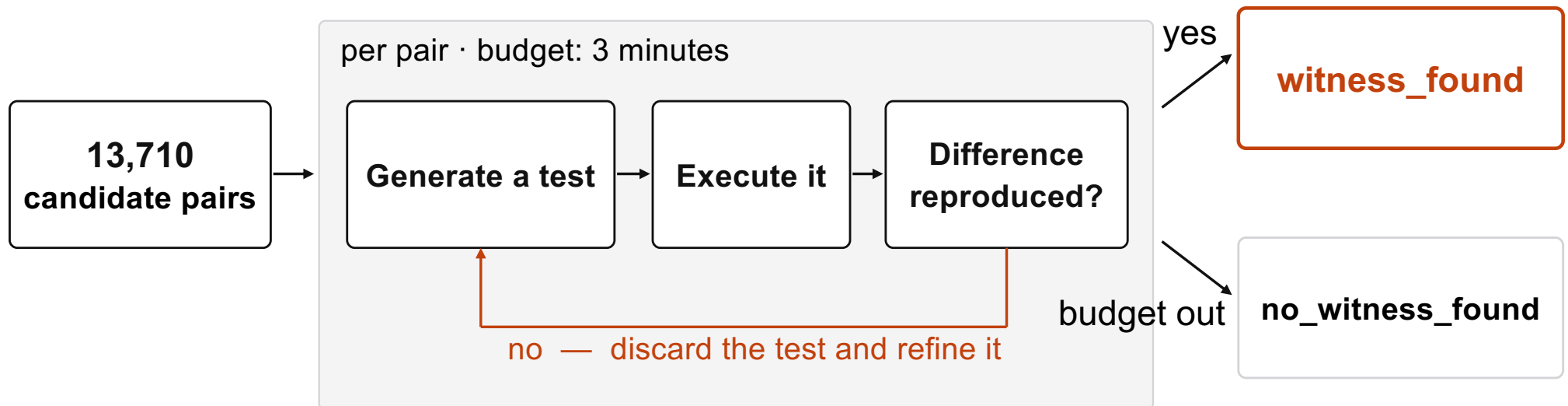
One agentic loop, 13,710 times



One agentic loop, 13,710 times



One agentic loop, 13,710 times



Codex (GPT-5.4, GPT-5.5) is a counterexample generator, not a labeler — execution decides the verdict.

Difference types recorded: return value · exception · non-termination · side effect

The model was given no human labels, no distinguishing tests, and no reasons from the original dataset.

Can it rediscover what humans already found?

Gold reference: the **852** pairs humans judged inequivalent — with reasons and distinguishing tests.
The model was shown none of that. It had to find its own counterexamples.

Can it rediscover what humans already found?

Gold reference: the **852** pairs humans judged inequivalent — with reasons and distinguishing tests.
The model was shown none of that. It had to find its own counterexamples.

848 / 852 **99.5% rediscovered,**
every one backed by a test that was actually executed

Can it rediscover what humans already found?

Gold reference: the **852** pairs humans judged inequivalent — with reasons and distinguishing tests.
The model was shown none of that. It had to find its own counterexamples.

848 / 852 **99.5% rediscovered,**
every one backed by a test that was actually executed

+153

of the **1,342** pairs humans judged **equivalent** also
produced a witness — all 153 manually confirmed valid.
The pair I opened with is one of them.

+4,626

witness-backed pairs from the **11,516** that had
never been inspected.
A random audit of 100 found all of them valid.

FEMP-X is not a bigger dataset.

It is a dataset that ships its evidence.

5,627 execution-validated witnesses over 13,710 candidate Java method pairs.

99.5% of the known inequivalent pairs rediscovered without access to the original judgments.

GitHub

github.com/YoshikiHigo/FEMP-X

Zenodo

doi.org/10.5281/zenodo.21149373



github.com/YoshikiHigo/FEMP-X