Malware-Domain Continued Pre-Training for Binary Malware Classification

Costin Teodorescu | C.Teodorescu-1@student.tudelft.nl
A leakage-aware study of code models on the SBAN corpus | BSc Computer Science and Engineering, TU Delft | Supervisors: Dr. Soham Chakraborty and Dr. Przemyslaw Pawelczak

Main Takeaway

Matched controls show no reliable gain from malware-domain continued pre-training.

1 Main Question

Question: Can continued pre-training on malware-related code make code language models more useful for malware classification than cheaper baselines and general-code training?

Answer: No reliable improvement is observed in the tested setting. The apparent +0.002 AUROC Qwen 0.5B mixed-domain difference is not supported by CodeBERT or the Qwen 1.5B scale check, so it is treated as insufficient evidence for malware-domain benefit.

2 Research Questions

- RQ1: How strong are cheap baselines?
- RQ2: Is source code or NLD more useful?
- RQ3: Does CPT beat the unadapted model?
- RQ4: Does malware data beat general-code CPT?
- RQ5: Does the pattern hold across models and zero-shot scoring?

The main claim is accepted only if it survives baseline, data-view, model-family, and operating-point controls.

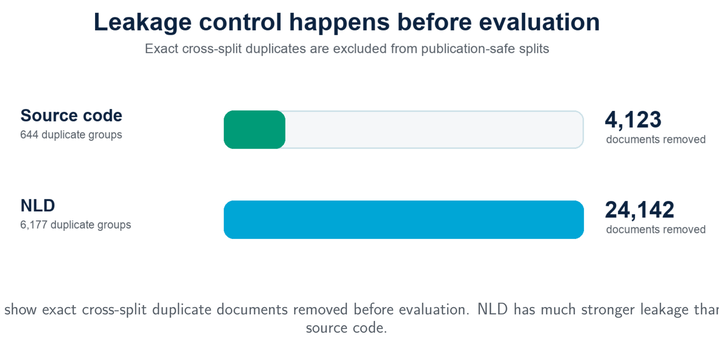
3 Leakage-Aware Benchmark

The reported benchmark is a paired source+NLD binary split built from BODMAS, Dike, MalwareBazzar, and Sorel20m. It keeps only explicit `BENIGN`/`MALWARE` labels; partial, unknown, and missing labels are excluded.

The benchmark fixes 165,519 paired samples before model comparisons: 118,450 train, 15,608 validation, and 31,461 test.

4 Split And Leakage Control

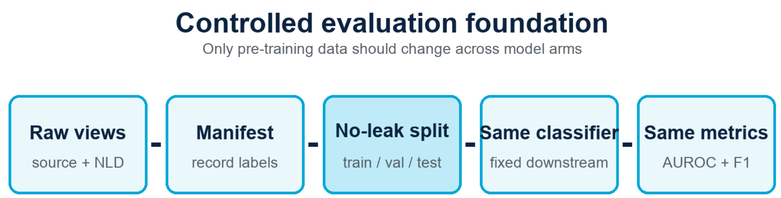
paired samples	benign	malware
165,519	**131,545**	**33,974**
source + NLD	binary labels	binary labels



Bars show exact cross-split duplicate documents removed before evaluation. NLD has much stronger leakage than source code.

Direct source-vs-NLD and model comparisons use the same paired sample IDs.

5 Controlled Comparison

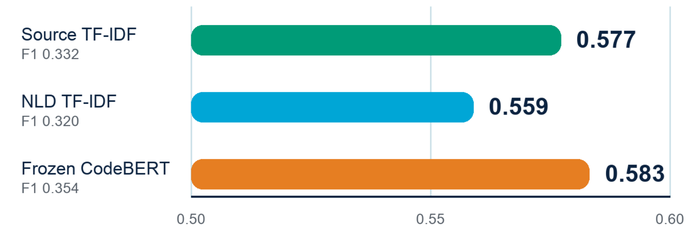


Process reads left to right. The downstream task stays fixed; only the continued-pre-training data changes between arms.

The arms separate no adaptation, extra general-code exposure, mixed malware-domain training, and malware-only training. Every supervised comparison uses the same paired source-code task, validation-threshold rule, and held-out test set.

A malware-domain gain must beat both the base model and the matched general-code control.

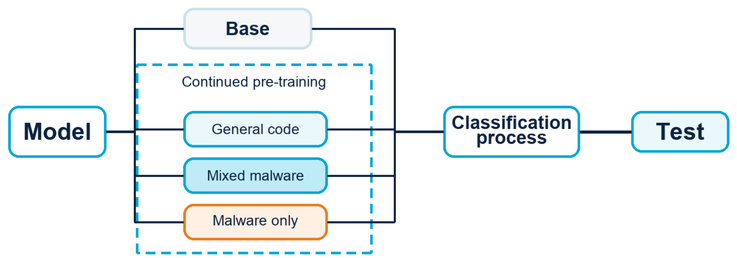
6 Baseline And Input-View Signal



Horizontal bars report test AUROC; F1 is shown next to each input or model row.

Source text is slightly stronger than NLD in this paired shallow setup; frozen CodeBERT is the strongest early baseline.

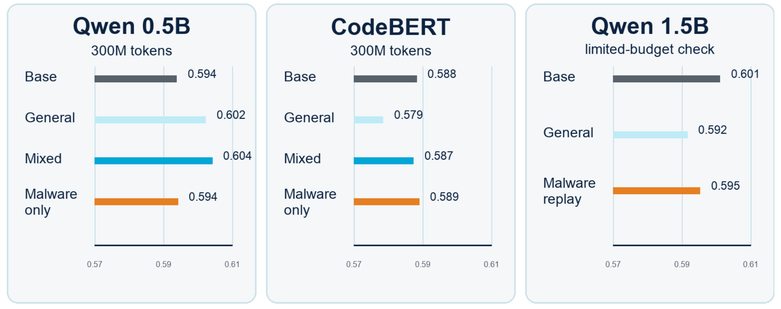
7 Training And Classification Procedure



Each arm enters the same downstream classification process and the same held-out test.

The model arm changes; the downstream classifier and test protocol stay fixed.

8 Model Trend



Each panel uses the same AUROC axis. The small differences do not form a reliable malware-domain CPT trend.

Across model families and controls, malware-domain CPT does not produce a reliable classification improvement.

9 Operating Point

Question: Does higher recall or zero-shot F1 imply a better malware detector?

Answer: No. Malware-only zero-shot scoring shifts predictions strongly toward malware: 98.9% of Qwen 0.5B test samples and 100.0% of CodeBERT test samples are predicted as malware.

AUROC, F1, precision, recall, and prediction rate must be read together. Validation-F1 thresholds often trade precision for recall, so zero-shot scores are diagnostics rather than deployment evidence.

The supervised matched comparisons remain the main evidence.

10 Final Answer And Limits

Final answer: the tested malware-domain CPT arms do not provide reliable evidence of better malware classification. The largest apparent margin is the Qwen 0.5B mixed-domain arm at +0.002 AUROC over the matched general-code control, which is too small to support the hypothesis.

The split removes exact duplicates, but near-duplicates may still remain. These scores use the paired result-protocol split, not the newer clustered rerun split, and each model arm is currently a single run without confidence intervals.

Under this protocol, malware-domain CPT is not enough to improve the classifier.