# Supplementary Results for Measuring Clone Detection Model Alignment

February 27, 2024

## 1 Evaluation Results

Here we present the results from the second alternate code mutation strategy. For these mutations, a code pair is mutated using original label values of two human evaluators to obtain code segments for removal.

We perform similarity removals at two levels based on two mutation scopes; (1) removing code segments representing core similarities only for which a label value of +2 is assigned by both human evaluators and (2) removing code segments representing all core and non-core similarities for which a label value of either +1 or +2 is assigned by both the human evaluators. Note that the label values may be the same for both human evaluators (both assign +1 or both assign +2) or conflicting (at least one evaluator assigns +1 or +2).

Similarly, we perform differences removals at two levels based on two mutation scopes; (1) removing code representing core differences only for which a label value of -2 is assigned by both human evaluators and (2) removing code representing all core and non-core differences, for which a label value of either -1 or -2 is assigned by both the human evaluators. Note that the label values may be the same for both human evaluators (both assign -1 or both assign -2) or conflicting (at least one evaluator assigns -1 or -2).

| Model | Scope | $ACE_{\Delta s,TP}$ | $ACE_{\Delta s,FP}$ | $ACE_{\Delta d,TN}$ | $ACE_{\Delta d,FN}$ |
|---|---|---|---|---|---|
| CodeBERT | core | 0.024 | -0.033 | 0.556 | 0.046 |
| CodeBERT | any | 0.022 | 0.016 | 0.652 | 0.184 |
| CodeT5 | core | 0.294 | -0.000 | 0.043 | -0.004 |
| CodeT5 | any | 0.428 | 0.499 | 0.042 | 0.112 |
| CodeGraph4CCDetector | core | 0.422 | 0.136 | 0.043 | -0.057 |
| CodeGraph4CCDetector | any | 0.262 | 0.192 | 0.046 | -0.018 |
| GPT-Turbo-3.5 | core | 0.370 | - | 0.051 | 0.013 |
| GPT-Turbo-3.5 | any | 0.473 | 0.45 | 0.173 | 0.038 |

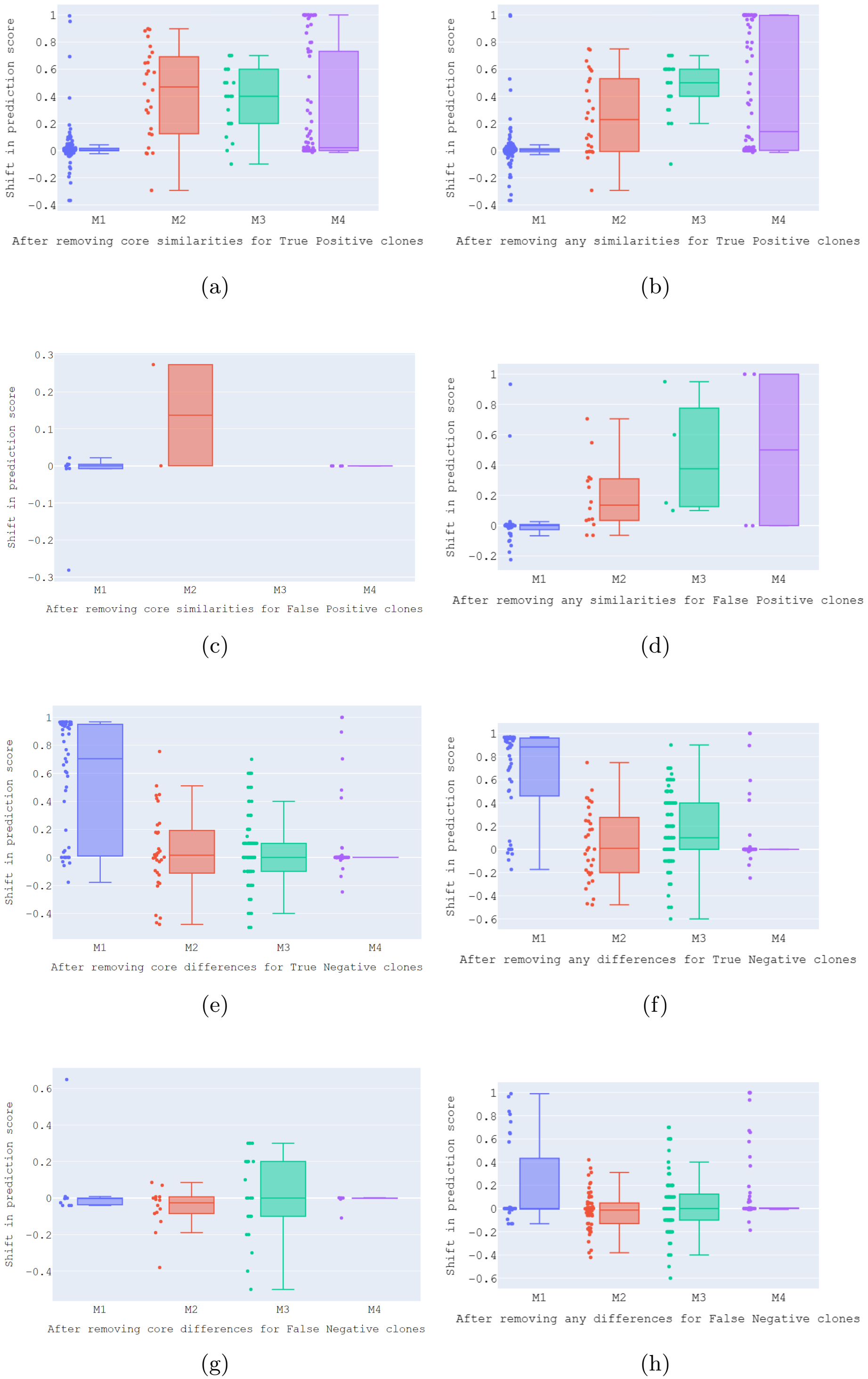
Table 1: Average Causal Effects (ACE) for models M across mutation scopes ‘core’ and ‘any’ for different mutation styles. Mutations are performed as value-based mutations using code label values to determine range of mutation scope.

| Model | $ACE_{sim}$ | $ACE_{diff}$ | $ACE_{s \cup d}$ |
|---|---|---|---|
| CodeBERT | 0.020 | 0.468 | 0.169 |
| CodeT5 | 0.358 | 0.054 | 0.174 |
| CodeGraph4CCDetector | 0.305 | 0.007 | 0.107 |
| GPT-Turbo-3.5 | 0.425 | 0.083 | 0.128 |

Table 2: Aggregated Average Causal Effects of human-identified code similarities and differences on various models semantic code clone predictions. (Results based on code mutations based on unresolved code labels).

| M | $\lvert flips(\mu_1)\rvert$ | $\lvert\mu_1\rvert$ | $\lvert flips(\mu_2)\rvert$ | $\lvert\mu_2\rvert$ | $A_{sim}$ |
|---|---|---|---|---|---|
| CodeBERT | 5 | 117 | 5 | 128 | 0.041 |
| CodeT5 | 24 | 82 | 38 | 88 | 0.365 |
| CodeGraph4CCDetector | 19 | 26 | 13 | 26 | 0.615 |
| GPT-Turbo-3.5 | 20 | 22 | 23 | 23 | 0.955 |

Table 3: Human-model code similarity intuition alignment $A_{sim}$ for model types M over mutated sets of clones $\mu_1$ and $\mu_2$ (using original label values for value-based mutations)



(a) (b)

(c) (d)

(e) (f)

(g) (h)

Figure 1: Model prediction shifts on mutated clone pairs for core removals only (a,c,e,g) and for both core and non-core removals including conflicts (b,d,f,h). M1 = CodeBERT, M2 = CodeGraph4CCDetector, M3 = GPT-Turbo-3.5, M4 = CodeT5

| Model | Scope | $\lvert flips(\mu_1)\rvert$ | $\lvert\mu_1\rvert$ | $\lvert flips(\mu_2)\rvert$ | $\lvert\mu_2\rvert$ | $\lvert flips(\mu_3)\rvert$ | $\lvert\mu_3\rvert$ | $F$ |
|---|---|---|---|---|---|---|---|---|
| CodeBERT | core | 0 | 8 | 12 | 34 | 12 | 117 | 0.779 |
| CodeBERT | any | 2 | 37 | 8 | 31 | 123 | 128 | 0.678 |
| CodeT5 | core | 0 | 4 | 0 | 19 | 58 | 82 | 0.552 |
| CodeT5 | any | 3 | 8 | 7 | 62 | 50 | 88 | 0.38 |
| CodeGraph4CCDetector | core | 0 | 2 | 1 | 14 | 7 | 26 | 0.19 |
| CodeGraph4CCDetector | any | 7 | 14 | 4 | 56 | 13 | 26 | 0.25 |
| GPT-Turbo-3.5 | core | - | - | 1 | 23 | 2 | 22 | 0.066 |
| GPT-Turbo-3.5 | any | 2 | 4 | 6 | 93 | - | - | 0.08 |

Table 4: Confounding frequency $F$ for various models aggregated from frequency of prediction flips for different mutation configurations and prediction outcomes $\mu_1, \mu_2, \mu_3$. Where $\mu_1 = \mu(\Delta s_x, FP)$, $\mu_2 = \mu(\Delta d_x, FN)$, and $\mu_3 = \mu(\Delta s_x, TP)$. Mutations are value-based as the code label values are used directly to determine mutation scope.

| Model | $ACE_{s\cup d}$ | $A_{sim}$ | F | C | $A_M$ |
|---|---|---|---|---|---|
| CodeBERT | 0.169 | 0.041 | 0.728 | 1.00 | 0.12 |
| CodeT5 | 0.174 | 0.365 | 0.466 | 1.00 | 0.27 |
| CodeGraph4CCDetector | 0.107 | 0.615 | 0.220 | 1.00 | 0.37 |
| GPT-Turbo-3.5 | 0.128 | 0.955 | 0.073 | 0.75 | 0.44 |

Table 5: Model Alignment ($A_M$) of semantic code clone detection of models (Values based on mutations resulting from code label values without resolving labels)