

Answers to Reviewer Questions

A. Answer to R1Q3: Why didn't you try your approach on an AI coding assistant like CoPilot or Cursor?

Motivation&Approach. To address the reviewers' suggestions, we additionally evaluate CompMT on Commercial development tool: GitHub Copilot. As Copilot does not provide an API, we manually collect its code generation results. Due to time constraints, we evaluate only our method and two main baselines (PPM-V and PPM-T) on Copilot on the more complex HumanEval dataset, and focused on the most critical metric, Pass@1.

TABLE I
RESULTS ON COPILOT

Methods	Pass@1
PPM-T	0.52
PPM-V	0.45
MR1	0.68
MR2	0.37
MR3	0.47

Results. As shown in Table I, our method achieves the lowest Pass@1 score with MR2 (0.37), significantly lower than both baselines PPM-T (0.52) and PPM-V (0.45). This demonstrates that MR2 imposes stricter testing conditions and is more effective at revealing failures, highlighting the stronger rigor of our CompMT approach on GitHub Copilot.

Answer to **R1Q3**. Overall, MR2 achieves the lowest Pass@1 score on Copilot, confirming the superior rigor of our CompMT method compared to existing baselines.

B. Answer to R2Q3: Try on modern and widely adopted models like GPT, LLaMA

Motivation&Approach. In response to the reviewers' suggestions, we further evaluate the effectiveness of our method on two representative models: (1) **Pre-trained open-source model (Llama)**: We select the latest Llama-3.2-1B as the evaluation target. (2) **Online closed-source model (Chat-GPT)**: We use the OpenAI gpt-3.5-turbo-0125 API to generate code. Due to the limited rebuttal period, we conduct experiments only on the more complex HumanEval dataset.

Results. The experimental results are summarized in Table II. Since a lower Pass@k indicates a stricter and more effective test, the smaller the value, the more capable the method is at identifying model weaknesses. We observe the following:

(1) For all models, the lowest Pass@k values consistently come from MR3, indicating that CompMT reliably offers the most rigorous testing among all methods.

(2) Regarding the most practically relevant setting—Pass@1 on real-world code generation models—CompMT demonstrates consistent advantages. For Llama, all three MRs achieve lower Pass@1 scores than any baseline. For ChatGPT, although MR1 is slightly higher than PPM-V, the other two MRs still outperform all baselines.

These results confirm that CompMT delivers more effective testing, especially under the strictest metric, Pass@1.

Answer to **R2Q3**. We further evaluate CompMT on two models, and the results show that our MRs consistently produce the lowest Pass@k scores, especially outperforming baselines on the most practical metric, Pass@1.

C. Answer to R2Q4: The paper should present detailed statistical results for metrics such as Pass@k

Results. The specific value results of Pass@k are shown in Tables III and Table IV. Note that these results correspond exactly to the data visualized in Figure 6 of the main paper.

TABLE II
RESULTS ON LLAMA AND CHATGPT

Methods	Llama3.2-1b					Chatgpt				
	Pass@1	Pass@3	Pass@5	Pass@7	Pass@10	Pass@1	Pass@3	Pass@5	Pass@7	Pass@10
Base	0.14	0.22	0.26	0.29	0.32	0.59	0.78	0.83	0.86	0.87
Insert_line	0.14	0.23	0.28	0.31	0.35	0.63	0.79	0.83	0.84	0.86
Comment	0.12	0.19	0.23	0.25	0.27	0.75	0.84	0.87	0.88	0.89
PPM-T	0.03	0.05	0.06	0.07	0.08	0.26	0.41	0.46	0.49	0.52
PPM-V	0.01	0.03	0.04	0.05	0.06	0.15	0.26	0.31	0.34	0.37
MR1	0.01	0.04	0.07	0.09	0.13	0.16	0.30	0.37	0.42	0.47
MR2	0.01	0.03	0.06	0.08	0.11	0.10	0.22	0.27	0.31	0.33
MR3	0.01	0.02	0.03	0.04	0.06	0.02	0.05	0.08	0.10	0.12

TABLE III
THE EFFECTIVENESS EVALUATION ON THE HUMANEVAL DATASET (VALUES IN PARENTHESES INDICATE THE PERCENTAGE DECREASE COMPARED TO THE BASE METHOD)

Methods	Incoder-1B					CodeGen-2B				
	Pass@1	Pass@3	Pass@5	Pass@7	Pass@10	Pass@1	Pass@3	Pass@5	Pass@7	Pass@10
Base	0.06 (0.00%)	0.11 (0.00%)	0.13 (0.00%)	0.14 (0.00%)	0.16 (0.00%)	0.1 (0.00%)	0.17 (0.00%)	0.21 (0.00%)	0.23 (0.00%)	0.25 (0.00%)
Insert_line	0.05 (-16.67%)	0.09 (-18.18%)	0.12 (-7.69%)	0.13 (-7.14%)	0.16 (0.00%)	0.11 (10.00%)	0.19 (11.76%)	0.23 (9.52%)	0.26 (13.04%)	0.29 (16.00%)
Comment	0.03 (-50.00%)	0.07 (-36.36%)	0.09 (-30.77%)	0.11 (-21.43%)	0.12 (-25.00%)	0.09 (-10.00%)	0.15 (-11.76%)	0.19 (-9.52%)	0.21 (-8.70%)	0.23 (-8.00%)
PPM-T	0.01 (-83.33%)	0.02 (-81.82%)	0.02 (-84.62%)	0.02 (-85.71%)	0.02 (-87.50%)	0.01 (-90.00%)	0.02 (-88.24%)	0.03 (-85.71%)	0.03 (-86.96%)	0.04 (-84.00%)
PPM-V	0.01 (-83.33%)	0.01 (-90.91%)	0.02 (-84.62%)	0.02 (-85.71%)	0.03 (-81.25%)	0.01 (-90.00%)	0.02 (-88.24%)	0.03 (-85.71%)	0.03 (-86.96%)	0.03 (-88.00%)
MR1	0 (-100.00%)	0.01 (-90.91%)	0.02 (-84.62%)	0.03 (-78.57%)	0.05 (-68.75%)	0.01 (-90.00%)	0.03 (-82.35%)	0.05 (-76.19%)	0.07 (-69.57%)	0.09 (-64.00%)
MR2	0 (-100.00%)	0.01 (-90.91%)	0.02 (-84.62%)	0.03 (-78.57%)	0.04 (-75.00%)	0.02 (-80.00%)	0.06 (-64.71%)	0.09 (-57.14%)	0.13 (-43.48%)	0.19 (-24.00%)
MR3	0.01 (-83.33%)	0.02 (-81.82%)	0.03 (-76.92%)	0.04 (-71.43%)	0.06 (-62.50%)	0 (-100.00%)	0 (-100.00%)	0 (-100.00%)	0 (-100.00%)	0 (-100.00%)
Methods	CodeGen2-1B					Santacoder				
	Pass@1	Pass@3	Pass@5	Pass@7	Pass@10	Pass@1	Pass@3	Pass@5	Pass@7	Pass@10
Base	0.06 (0.00%)	0.1 (0.00%)	0.12 (0.00%)	0.13 (0.00%)	0.14 (0.00%)	0.15 (0.00%)	0.22 (0.00%)	0.26 (0.00%)	0.28 (0.00%)	0.3 (0.00%)
Insert_line	0.07 (16.67%)	0.1 (0.00%)	0.11 (-8.33%)	0.12 (-7.69%)	0.13 (-7.14%)	0.13 (-13.33%)	0.19 (-13.64%)	0.22 (-15.38%)	0.24 (-14.29%)	0.25 (-16.67%)
Comment	0.04 (-60.00%)	0.09 (-10.00%)	0.1 (-16.67%)	0.11 (-15.38%)	0.11 (-21.43%)	0.11 (-26.67%)	0.18 (-18.18%)	0.21 (-19.23%)	0.23 (-17.86%)	0.25 (-16.67%)
PPM-T	0.01 (-83.33%)	0.02 (-80.00%)	0.03 (-75.00%)	0.03 (-76.92%)	0.04 (-71.43%)	0.02 (-86.67%)	0.04 (-81.82%)	0.05 (-80.77%)	0.06 (-78.57%)	0.06 (-80.00%)
PPM-V	0 (-100.00%)	0.01 (-90.00%)	0.01 (-91.67%)	0.02 (-84.62%)	0.02 (-85.71%)	0.01 (-93.33%)	0.02 (-90.91%)	0.03 (-88.46%)	0.04 (-85.71%)	0.04 (-86.67%)
MR1	0 (-100.00%)	0.01 (-90.00%)	0.02 (-83.33%)	0.02 (-84.62%)	0.03 (-78.57%)	0.03 (-80.00%)	0.07 (-68.18%)	0.1 (-61.54%)	0.13 (-53.57%)	0.16 (-46.67%)
MR2	0 (-100.00%)	0.01 (-90.00%)	0.02 (-83.33%)	0.03 (-76.92%)	0.04 (-71.43%)	0.01 (-93.33%)	0.04 (-81.82%)	0.07 (-73.08%)	0.09 (-67.86%)	0.11 (-63.33%)
MR3	0 (-100.00%)	0 (-100.00%)	0 (-100.00%)	0 (-100.00%)	0 (-100.00%)	0.01 (-93.33%)	0.04 (-81.82%)	0.06 (-76.92%)	0.08 (-71.43%)	0.12 (-60.00%)

TABLE IV
THE EFFECTIVENESS EVALUATION ON THE MBPP DATASET (VALUES IN PARENTHESES INDICATE THE PERCENTAGE DECREASE COMPARED TO THE BASE METHOD)

Methods	Incoder-1B					CodeGen-2B				
	Pass@1	Pass@3	Pass@5	Pass@7	Pass@10	Pass@1	Pass@3	Pass@5	Pass@7	Pass@10
Base	0.1 (0.00%)	0.21 (0.00%)	0.27 (0.00%)	0.31 (0.00%)	0.35 (0.00%)	0.22 (0.00%)	0.38 (0.00%)	0.45 (0.00%)	0.49 (0.00%)	0.53 (0.00%)
Insert_line	0.1 (0.00%)	0.22 (4.76%)	0.28 (3.70%)	0.32 (3.23%)	0.37 (5.71%)	0.19 (-13.64%)	0.34 (-10.53%)	0.41 (-8.89%)	0.45 (-8.16%)	0.49 (-7.55%)
Comment	0.04 (-60.00%)	0.1 (-52.38%)	0.15 (-44.44%)	0.18 (-41.94%)	0.22 (-37.14%)	0.15 (-31.82%)	0.31 (-18.42%)	0.38 (-15.56%)	0.44 (-10.20%)	0.49 (-7.55%)
PPM-T	0.01 (-90.00%)	0.03 (-85.71%)	0.04 (-85.19%)	0.05 (-83.87%)	0.07 (-80.00%)	0.01 (-95.45%)	0.04 (-89.47%)	0.05 (-88.89%)	0.07 (-85.71%)	0.09 (-83.02%)
PPM-V	0.01 (-90.00%)	0.04 (-80.95%)	0.06 (-77.78%)	0.07 (-77.42%)	0.08 (-77.14%)	0.03 (-86.36%)	0.07 (-81.58%)	0.09 (-80.00%)	0.11 (-77.55%)	0.14 (-73.58%)
MR1	0.01 (-90.00%)	0.03 (-85.71%)	0.05 (-81.48%)	0.07 (-77.42%)	0.09 (-74.29%)	0.01 (-95.45%)	0.04 (-89.47%)	0.06 (-86.67%)	0.07 (-85.71%)	0.09 (-83.02%)
MR2	0.01 (-90.00%)	0.03 (-85.71%)	0.04 (-85.19%)	0.06 (-80.65%)	0.08 (-77.14%)	0.01 (-95.45%)	0.03 (-92.11%)	0.05 (-88.89%)	0.07 (-85.71%)	0.1 (-81.13%)
MR3	0 (-100.00%)	0 (-100.00%)	0 (-100.00%)	0 (-100.00%)	0 (-100.00%)	0 (-100.00%)	0.01 (-97.37%)	0.02 (-95.56%)	0.02 (-95.92%)	0.03 (-94.34%)
Methods	CodeGen2-1B					Santacoder				
	Pass@1	Pass@3	Pass@5	Pass@7	Pass@10	Pass@1	Pass@3	Pass@5	Pass@7	Pass@10
Base	0.11 (0.00%)	0.21 (0.00%)	0.28 (0.00%)	0.32 (0.00%)	0.37 (0.00%)	0.24 (0.00%)	0.41 (0.00%)	0.48 (0.00%)	0.52 (0.00%)	0.55 (0.00%)
Insert_line	0.1 (-9.09%)	0.2 (-4.76%)	0.26 (-7.14%)	0.29 (-9.38%)	0.33 (-10.81%)	0.18 (-25.00%)	0.34 (-17.07%)	0.41 (-14.58%)	0.46 (-11.54%)	0.5 (-9.09%)
Comment	0.06 (-45.45%)	0.14 (-33.33%)	0.19 (-32.14%)	0.22 (-31.25%)	0.26 (-29.73%)	0.23 (-4.17%)	0.4 (-2.44%)	0.48 (0.00%)	0.53 (1.92%)	0.58 (5.45%)
PPM-T	0.01 (-90.91%)	0.04 (-80.95%)	0.06 (-78.57%)	0.08 (-75.00%)	0.1 (-72.97%)	0.02 (-91.67%)	0.06 (-85.37%)	0.08 (-83.33%)	0.1 (-80.77%)	0.12 (-78.18%)
PPM-V	0.01 (-90.91%)	0.04 (-80.95%)	0.06 (-78.57%)	0.07 (-78.12%)	0.09 (-75.68%)	0.04 (-83.33%)	0.08 (-80.49%)	0.1 (-79.17%)	0.12 (-76.92%)	0.14 (-74.55%)
MR1	0 (-100.00%)	0.01 (-95.24%)	0.02 (-92.86%)	0.03 (-90.63%)	0.04 (-89.19%)	0.03 (-87.50%)	0.07 (-82.93%)	0.1 (-79.17%)	0.13 (-75.00%)	0.16 (-70.91%)
MR2	0 (-100.00%)	0.01 (-95.24%)	0.02 (-92.86%)	0.03 (-90.63%)	0.04 (-89.19%)	0.03 (-87.50%)	0.07 (-82.93%)	0.11 (-77.08%)	0.13 (-75.00%)	0.17 (-69.09%)
MR3	0 (-100.00%)	0 (-100.00%)	0 (-100.00%)	0 (-100.00%)	0 (-100.00%)	0.01 (-95.83%)	0.04 (-90.24%)	0.06 (-87.50%)	0.08 (-84.62%)	0.11 (-80.00%)