I. Detailed results of RQ1

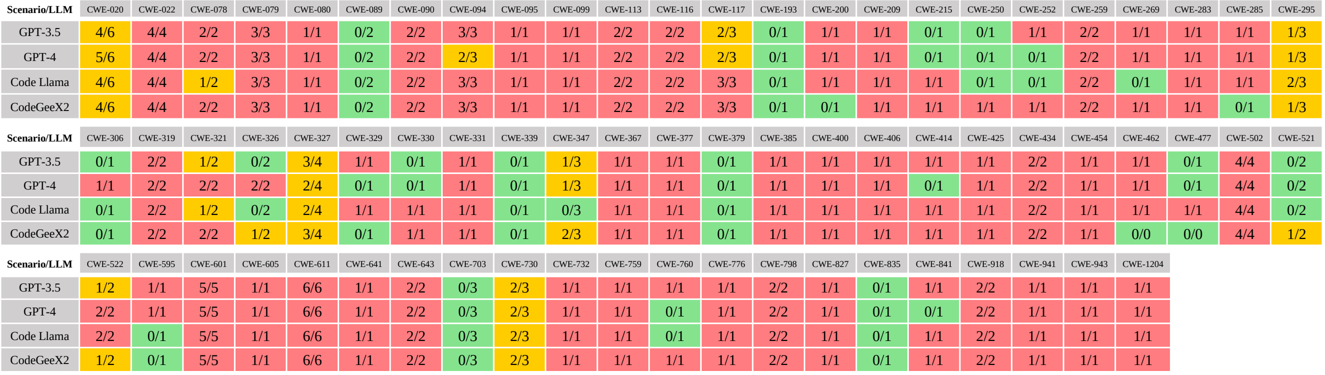
Fig. 1 shows the detailed performance of GPT-3.5, GPT-4, Code Llama, and CodeGeeX2 on SecurityEval, represented as the ratio of insecure code to total code generation tasks for each CWE ($\frac{number\ of\ insecure\ generated\ code\ pieces}{number\ of\ generation\ tasks}$). A red, green, or yellow cell represents, respectively, that all code generated by that LLM is vulnerable, secure, or partly vulnerable to the specified CWE.

II. Detailed results of RQ2

Fig. 2 and Fig. 3 respectively show the detailed results of GPT-3.5 and GPT-4 in detecting vulnerabilities in code generated by the four studied large language models. The results in the colored cells are presented as $\frac{number\ of\ correct\ identification}{number\ of\ identification\ tasks}$ for each CWE. For example, the first cell of Fig. 3 indicates that GPT-4 correctly identified whether the code is vulnerable to CWE-20 (Improper Input Validation) for 3 out of 6 pieces of code generated by GPT-3.5.

III. Detailed results of RQ3

Fig. 4 and Fig. 5 respectively show the detailed results of GPT-3.5 and GPT-4 in fixing vulnerabilities in code generated by the four studied large language models. The results in the colored cells are presented as $\frac{number\ of\ successful\ repair}{number\ of\ repair\ tasks}$ for each CWE. For example, the first cell of Fig. 5 indicates that GPT-4 successfully fixed all the 3 vulnerable pieces of code generated by GPT-3.5 with the vulnerability of CWE-20. A gray cell indicates that no code with that CWE weakness was generated in the first place.

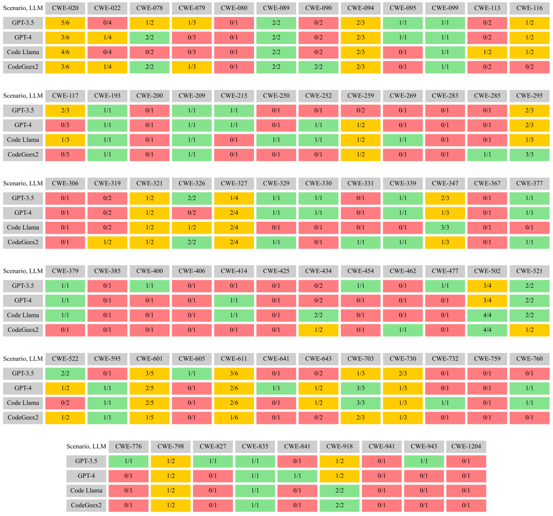


Scenario/LLM	CWE-020	CWE-022	CWE-078	CWE-079	CWE-080	CWE-089	CWE-090	CWE-094	CWE-095	CWE-099	CWE-113	CWE-116	CWE-117	CWE-193	CWE-200	CWE-209	CWE-215	CWE-250	CWE-252	CWE-259	CWE-269	CWE-283	CWE-285	CWE-295
GPT-3.5	4/6	4/4	2/2	3/3	1/1	0/2	2/2	3/3	1/1	1/1	2/2	2/2	2/3	0/1	1/1	1/1	0/1	0/1	1/1	2/2	1/1	1/1	1/1	1/3
GPT-4	5/6	4/4	2/2	3/3	1/1	0/2	2/2	2/3	1/1	1/1	2/2	2/2	2/3	0/1	1/1	1/1	0/1	0/1	0/1	2/2	1/1	1/1	1/1	1/3
Code Llama	4/6	4/4	1/2	3/3	1/1	0/2	2/2	3/3	1/1	1/1	2/2	2/2	3/3	0/1	1/1	1/1	1/1	0/1	0/1	2/2	0/1	1/1	1/1	2/3
CodeGeeX2	4/6	4/4	2/2	3/3	1/1	0/2	2/2	3/3	1/1	1/1	2/2	2/2	3/3	0/1	0/1	1/1	1/1	1/1	1/1	2/2	1/1	1/1	0/1	1/3

Scenario/LLM	CWE-306	CWE-319	CWE-321	CWE-326	CWE-327	CWE-329	CWE-330	CWE-331	CWE-339	CWE-347	CWE-367	CWE-377	CWE-379	CWE-385	CWE-400	CWE-406	CWE-414	CWE-425	CWE-434	CWE-454	CWE-462	CWE-477	CWE-502	CWE-521
GPT-3.5	0/1	2/2	1/2	0/2	3/4	1/1	0/1	1/1	0/1	1/3	1/1	1/1	0/1	1/1	1/1	1/1	1/1	1/1	2/2	1/1	1/1	0/1	4/4	0/2
GPT-4	1/1	2/2	2/2	2/2	2/4	0/1	0/1	1/1	0/1	1/3	1/1	1/1	0/1	1/1	1/1	1/1	0/1	1/1	2/2	1/1	1/1	0/1	4/4	0/2
Code Llama	0/1	2/2	1/2	0/2	2/4	1/1	1/1	1/1	0/1	0/3	1/1	1/1	0/1	1/1	1/1	1/1	1/1	1/1	2/2	1/1	1/1	1/1	4/4	0/2
CodeGeeX2	0/1	2/2	2/2	1/2	3/4	0/1	1/1	1/1	0/1	2/3	1/1	1/1	0/1	1/1	1/1	1/1	1/1	1/1	2/2	1/1	0/0	0/0	4/4	1/2

Scenario/LLM	CWE-522	CWE-595	CWE-601	CWE-605	CWE-611	CWE-641	CWE-643	CWE-703	CWE-730	CWE-732	CWE-759	CWE-760	CWE-776	CWE-798	CWE-827	CWE-835	CWE-841	CWE-918	CWE-941	CWE-943	CWE-1204
GPT-3.5	1/2	1/1	5/5	1/1	6/6	1/1	2/2	0/3	2/3	1/1	1/1	1/1	1/1	2/2	1/1	0/1	1/1	2/2	1/1	1/1	1/1
GPT-4	2/2	1/1	5/5	1/1	6/6	1/1	2/2	0/3	2/3	1/1	1/1	0/1	1/1	2/2	1/1	0/1	0/1	2/2	1/1	1/1	1/1
Code Llama	2/2	0/1	5/5	1/1	6/6	1/1	2/2	0/3	2/3	1/1	1/1	0/1	1/1	2/2	1/1	0/1	1/1	2/2	1/1	1/1	1/1
CodeGeeX2	1/2	0/1	5/5	1/1	6/6	1/1	2/2	0/3	2/3	1/1	1/1	1/1	1/1	2/2	1/1	0/1	1/1	2/2	1/1	1/1	1/1

Fig. 1. Detailed performance of GPT-3.5, GPT-4, Code Llama, and CodeGeeX2 on SecurityEval



Scenario, LLM	CWE-020	CWE-022	CWE-078	CWE-079	CWE-080	CWE-089	CWE-090	CWE-094	CWE-095	CWE-099	CWE-113	CWE-116
GPT-3.5	5/6	0/4	1/2	1/3	0/1	2/2	0/2	2/3	1/1	1/1	0/2	1/2
GPT-4	3/6	1/4	2/2	0/3	0/1	2/2	0/2	2/3	1/1	1/1	0/2	1/2
Code Llama	4/6	0/4	0/2	0/3	0/1	2/2	0/2	2/3	0/1	1/1	1/2	1/2
CodeGeex2	3/6	1/4	2/2	1/3	0/1	2/2	2/2	2/3	0/1	1/1	0/2	0/2

Scenario, LLM	CWE-117	CWE-193	CWE-200	CWE-209	CWE-215	CWE-250	CWE-252	CWE-259	CWE-269	CWE-283	CWE-285	CWE-295
GPT-3.5	2/3	1/1	0/1	1/1	1/1	0/1	0/1	0/2	0/1	0/1	0/1	2/3
GPT-4	0/3	1/1	0/1	1/1	1/1	0/1	1/1	1/2	0/1	0/1	0/1	2/3
Code Llama	1/3	1/1	0/1	1/1	0/1	1/1	1/1	1/2	1/1	0/1	0/1	1/3
CodeGeex2	0/3	1/1	0/1	1/1	0/1	0/1	0/1	1/2	0/1	0/1	1/1	3/3

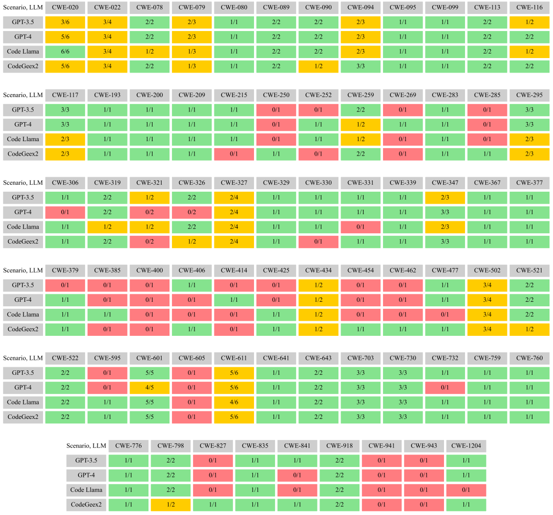
Scenario, LLM	CWE-306	CWE-319	CWE-321	CWE-326	CWE-327	CWE-329	CWE-330	CWE-331	CWE-339	CWE-347	CWE-367	CWE-377
GPT-3.5	0/1	0/2	1/2	2/2	1/4	1/1	1/1	0/1	1/1	2/3	0/1	1/1
GPT-4	0/1	0/2	1/2	0/2	2/4	1/1	1/1	0/1	1/1	1/3	0/1	1/1
Code Llama	0/1	0/2	1/2	1/2	2/4	0/1	0/1	0/1	0/1	3/3	0/1	0/1
CodeGeex2	0/1	1/2	1/2	2/2	2/4	1/1	0/1	1/1	1/1	1/3	0/1	1/1

Scenario, LLM	CWE-379	CWE-385	CWE-400	CWE-406	CWE-414	CWE-425	CWE-434	CWE-454	CWE-462	CWE-477	CWE-502	CWE-521
GPT-3.5	1/1	0/1	1/1	0/1	0/1	0/1	0/2	1/1	0/1	1/1	3/4	2/2
GPT-4	1/1	0/1	0/1	0/1	1/1	0/1	0/2	0/1	0/1	0/1	3/4	2/2
Code Llama	1/1	0/1	0/1	0/1	1/1	0/1	2/2	0/1	0/1	0/1	4/4	2/2
CodeGeex2	0/1	0/1	0/1	0/1	0/1	0/1	1/2	0/1	1/1	0/1	4/4	1/2

Scenario, LLM	CWE-522	CWE-595	CWE-601	CWE-605	CWE-611	CWE-641	CWE-643	CWE-703	CWE-730	CWE-732	CWE-759	CWE-760
GPT-3.5	2/2	0/1	3/5	1/1	3/6	0/1	0/2	1/3	2/3	0/1	0/1	0/1
GPT-4	1/2	1/1	2/5	0/1	2/6	1/1	1/2	3/3	1/3	0/1	0/1	1/1
Code Llama	0/2	1/1	2/5	0/1	2/6	0/1	1/2	3/3	1/3	1/1	0/1	1/1
CodeGeex2	1/2	1/1	1/5	0/1	1/6	0/1	0/2	2/3	1/3	0/1	0/1	0/1

Scenario, LLM	CWE-776	CWE-798	CWE-827	CWE-835	CWE-841	CWE-918	CWE-941	CWE-943	CWE-1204
GPT-3.5	1/1	1/2	1/1	1/1	0/1	1/2	0/1	1/1	0/1
GPT-4	0/1	1/2	0/1	1/1	1/1	1/2	0/1	0/1	0/1
Code Llama	0/1	1/2	0/1	1/1	0/1	2/2	0/1	0/1	0/1
CodeGeex2	0/1	1/2	0/1	1/1	0/1	2/2	0/1	0/1	0/1

Fig. 2. Detailed results of GPT-3.5 in detecting vulnerabilities in code generated by GPT-3.5, GPT-4, Code Llama, and CodeGeeX2



Scenario, LLM	CWE-020	CWE-022	CWE-078	CWE-079	CWE-080	CWE-089	CWE-090	CWE-094	CWE-095	CWE-099	CWE-113	CWE-116
GPT-3.5	3/6	3/4	2/2	2/3	1/1	2/2	2/2	2/3	1/1	1/1	2/2	1/2
GPT-4	5/6	3/4	2/2	2/3	1/1	2/2	2/2	2/3	1/1	1/1	2/2	2/2
Code Llama	6/6	3/4	1/2	1/3	1/1	2/2	2/2	2/3	1/1	1/1	2/2	1/2
CodeGeex2	5/6	3/4	2/2	1/3	1/1	2/2	1/2	3/3	1/1	1/1	2/2	2/2

Scenario, LLM	CWE-117	CWE-193	CWE-200	CWE-209	CWE-215	CWE-250	CWE-252	CWE-259	CWE-269	CWE-283	CWE-285	CWE-295
GPT-3.5	3/3	1/1	1/1	1/1	1/1	0/1	0/1	2/2	0/1	1/1	0/1	3/3
GPT-4	3/3	1/1	1/1	1/1	1/1	0/1	1/1	1/2	1/1	1/1	0/1	3/3
Code Llama	2/3	1/1	1/1	1/1	1/1	0/1	1/1	1/2	0/1	1/1	0/1	2/3
CodeGeex2	2/3	1/1	1/1	1/1	0/1	1/1	0/1	2/2	0/1	1/1	1/1	2/3

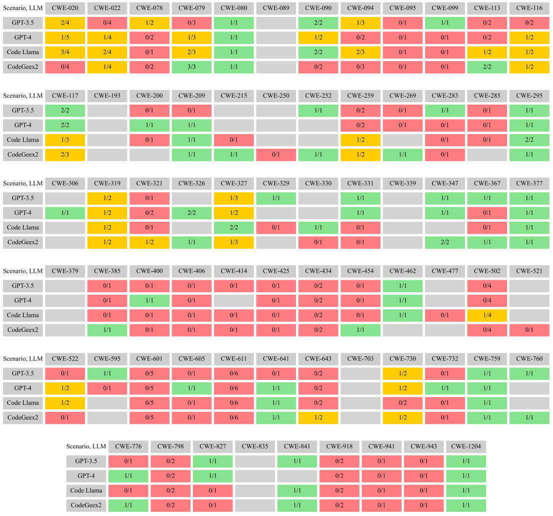
Scenario, LLM	CWE-306	CWE-319	CWE-321	CWE-326	CWE-327	CWE-329	CWE-330	CWE-331	CWE-339	CWE-347	CWE-367	CWE-377
GPT-3.5	1/1	2/2	1/2	2/2	2/4	1/1	1/1	1/1	1/1	2/3	1/1	1/1
GPT-4	0/1	2/2	0/2	0/2	2/4	1/1	1/1	1/1	1/1	3/3	1/1	1/1
Code Llama	1/1	1/2	1/2	2/2	2/4	1/1	1/1	0/1	1/1	2/3	1/1	1/1
CodeGeex2	1/1	2/2	0/2	1/2	2/4	1/1	0/1	1/1	1/1	3/3	1/1	1/1

Scenario, LLM	CWE-379	CWE-385	CWE-400	CWE-406	CWE-414	CWE-425	CWE-434	CWE-454	CWE-462	CWE-477	CWE-502	CWE-521
GPT-3.5	0/1	0/1	0/1	1/1	0/1	0/1	1/2	0/1	0/1	1/1	3/4	2/2
GPT-4	1/1	0/1	0/1	0/1	1/1	0/1	1/2	0/1	0/1	1/1	3/4	2/2
Code Llama	1/1	1/1	0/1	0/1	0/1	1/1	1/2	0/1	0/1	0/1	3/4	2/2
CodeGeex2	1/1	0/1	0/1	1/1	0/1	1/1	1/2	1/1	1/1	1/1	3/4	1/2

Scenario, LLM	CWE-522	CWE-595	CWE-601	CWE-605	CWE-611	CWE-641	CWE-643	CWE-703	CWE-730	CWE-732	CWE-759	CWE-760
GPT-3.5	2/2	0/1	5/5	0/1	5/6	1/1	2/2	3/3	3/3	1/1	1/1	1/1
GPT-4	2/2	0/1	4/5	0/1	5/6	1/1	2/2	3/3	3/3	0/1	1/1	1/1
Code Llama	2/2	1/1	5/5	0/1	4/6	1/1	2/2	3/3	3/3	1/1	1/1	1/1
CodeGeex2	2/2	1/1	5/5	0/1	5/6	1/1	2/2	3/3	3/3	1/1	1/1	1/1

Scenario, LLM	CWE-776	CWE-798	CWE-827	CWE-835	CWE-841	CWE-918	CWE-941	CWE-943	CWE-1204
GPT-3.5	1/1	2/2	0/1	1/1	1/1	2/2	0/1	0/1	1/1
GPT-4	1/1	2/2	0/1	1/1	0/1	2/2	0/1	0/1	1/1
Code Llama	1/1	2/2	0/1	1/1	0/1	2/2	0/1	0/1	0/1
CodeGeex2	1/1	1/2	1/1	1/1	1/1	2/2	0/1	0/1	1/1

Fig. 3. Detailed results of GPT-4 in detecting vulnerabilities in code generated by GPT-3.5, GPT-4, Code Llama, and CodeGeeX2



Scenario, LLM	CWE-020	CWE-022	CWE-078	CWE-079	CWE-080	CWE-089	CWE-090	CWE-094	CWE-095	CWE-099	CWE-113	CWE-116
GPT-3.5	2/4	0/4	1/2	0/3	1/1		2/2	1/3	0/1	1/1	0/2	0/2
GPT-4	1/5	1/4	0/2	1/3	1/1		1/2	0/2	0/1	0/1	0/2	1/2
Code Llama	3/4	2/4	0/1	2/3	1/1		2/2	2/3	0/1	0/1	1/2	1/2
CodeGeex2	0/4	1/4	0/2	3/3	1/1		0/2	0/3	0/1	0/1	2/2	1/2

Scenario, LLM	CWE-117	CWE-193	CWE-200	CWE-209	CWE-215	CWE-250	CWE-252	CWE-259	CWE-269	CWE-283	CWE-285	CWE-295
GPT-3.5	2/2		0/1	0/1			1/1	0/2	0/1	1/1	0/1	1/1
GPT-4	2/2		1/1	1/1				0/2	0/1	0/1	0/1	1/1
Code Llama	1/3		0/1	1/1	0/1			1/2		0/1	0/1	2/2
CodeGeex2	2/3			1/1	1/1	0/1	1/1	1/2	1/1	0/1		1/1

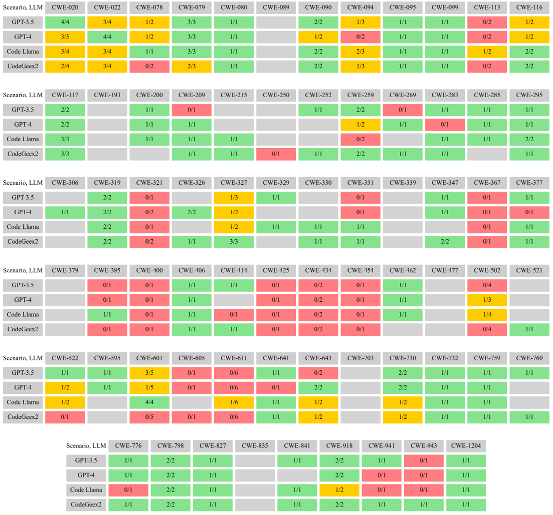
Scenario, LLM	CWE-306	CWE-319	CWE-321	CWE-326	CWE-327	CWE-329	CWE-330	CWE-331	CWE-339	CWE-347	CWE-367	CWE-377
GPT-3.5		1/2	0/1		1/3	1/1		1/1		1/1	1/1	1/1
GPT-4	1/1	1/2	0/2	2/2	1/2			1/1		1/1	0/1	1/1
Code Llama		1/2	0/1		2/2	0/1	1/1	0/1			0/1	1/1
CodeGeex2		1/2	1/2	1/1	1/3		0/1	0/1		2/2	1/1	1/1

Scenario, LLM	CWE-379	CWE-385	CWE-400	CWE-406	CWE-414	CWE-425	CWE-434	CWE-454	CWE-462	CWE-477	CWE-502	CWE-521
GPT-3.5		0/1	0/1	0/1	0/1	0/1	0/2	0/1	1/1		0/4	
GPT-4		0/1	1/1	0/1		0/1	0/2	0/1	1/1		0/4	
Code Llama		0/1	0/1	0/1	0/1	0/1	0/2	0/1	1/1	0/1	1/4	
CodeGeex2		1/1	0/1	0/1	0/1	0/1	0/2	1/1			0/4	0/1

Scenario, LLM	CWE-522	CWE-595	CWE-601	CWE-605	CWE-611	CWE-641	CWE-643	CWE-703	CWE-730	CWE-732	CWE-759	CWE-760
GPT-3.5	0/1	1/1	0/5	0/1	0/6	0/1	0/2		1/2	0/1	1/1	1/1
GPT-4	1/2	0/1	0/5	1/1	0/6	1/1	0/2		1/2	1/1	1/1	
Code Llama	1/2		0/5	0/1	0/6	1/1	0/2		0/2	0/1	1/1	
CodeGeex2	0/1		0/5	0/1	0/6	1/1	1/2		1/2	0/1	1/1	1/1

Scenario, LLM	CWE-776	CWE-798	CWE-827	CWE-835	CWE-841	CWE-918	CWE-941	CWE-943	CWE-1204
GPT-3.5	0/1	0/2	1/1		1/1	0/2	0/1	0/1	1/1
GPT-4	1/1	0/2	1/1			0/2	0/1	0/1	1/1
Code Llama	0/1	0/2	0/1		1/1	0/2	0/1	0/1	1/1
CodeGeex2	1/1	0/2	0/1		1/1	0/2	0/1	0/1	1/1

Fig. 4. Detailed results of GPT-3.5 in repairing vulnerabilities in code generated by GPT-3.5, GPT-4, Code Llama, and CodeGeeX2



Scenario, LLM	CWE-020	CWE-022	CWE-078	CWE-079	CWE-080	CWE-089	CWE-090	CWE-094	CWE-095	CWE-099	CWE-113	CWE-116
GPT-3.5	4/4	3/4	1/2	3/3	1/1		2/2	1/3	1/1	1/1	0/2	1/2
GPT-4	3/5	4/4	1/2	3/3	1/1		1/2	0/2	1/1	1/1	0/2	1/2
Code Llama	3/4	3/4	1/1	3/3	1/1		2/2	2/3	1/1	1/1	1/2	2/2
CodeGeex2	2/4	3/4	0/2	2/3	1/1		2/2	1/3	1/1	1/1	0/2	2/2

Scenario, LLM	CWE-117	CWE-193	CWE-200	CWE-209	CWE-215	CWE-250	CWE-252	CWE-259	CWE-269	CWE-283	CWE-285	CWE-295
GPT-3.5	2/2		1/1	0/1			1/1	2/2	0/1	1/1	1/1	1/1
GPT-4	2/2		1/1	1/1				1/2	1/1	0/1	1/1	1/1
Code Llama	3/3		1/1	1/1	1/1			0/2		1/1	1/1	2/2
CodeGeex2	3/3			1/1	1/1	0/1	1/1	2/2	1/1	1/1		1/1

Scenario, LLM	CWE-306	CWE-319	CWE-321	CWE-326	CWE-327	CWE-329	CWE-330	CWE-331	CWE-339	CWE-347	CWE-367	CWE-377
GPT-3.5		2/2	0/1		1/3	1/1		0/1		1/1	0/1	1/1
GPT-4	1/1	2/2	0/2	2/2	1/2			0/1		1/1	0/1	0/1
Code Llama		2/2	0/1		1/2	1/1	1/1	1/1			0/1	1/1
CodeGeex2		2/2	0/2	1/1	3/3		1/1	1/1		2/2	0/1	1/1

Scenario, LLM	CWE-379	CWE-385	CWE-400	CWE-406	CWE-414	CWE-425	CWE-434	CWE-454	CWE-462	CWE-477	CWE-502	CWE-521
GPT-3.5		0/1	0/1	1/1	1/1	0/1	0/2	0/1	1/1		0/4	
GPT-4		0/1	0/1	1/1		0/1	0/2	0/1	1/1		1/3	
Code Llama		1/1	0/1	1/1	0/1	0/1	0/2	0/1	1/1		1/4	
CodeGeex2		0/1	0/1	1/1	1/1	0/1	0/2	0/1			0/4	1/1

Scenario, LLM	CWE-522	CWE-595	CWE-601	CWE-605	CWE-611	CWE-641	CWE-643	CWE-703	CWE-730	CWE-732	CWE-759	CWE-760
GPT-3.5	1/1	1/1	3/5	0/1	0/6	1/1	0/2		2/2	1/1	1/1	1/1
GPT-4	1/2	1/1	1/5	0/1	0/6	0/1	2/2		2/2	1/1	1/1	
Code Llama	1/2		4/4		1/6	1/1	1/2		1/2	1/1	1/1	
CodeGeex2	0/1		0/5	0/1	0/6	1/1	1/2		1/2	1/1	1/1	1/1

Scenario, LLM	CWE-776	CWE-798	CWE-827	CWE-835	CWE-841	CWE-918	CWE-941	CWE-943	CWE-1204
GPT-3.5	1/1	2/2	1/1		1/1	2/2	1/1	0/1	1/1
GPT-4	1/1	2/2	1/1			2/2	0/1	0/1	1/1
Code Llama	0/1	2/2	1/1		1/1	1/2	0/1	0/1	1/1
CodeGeex2	1/1	2/2	1/1		1/1	2/2	1/1	1/1	1/1

Fig. 5. Detailed results of GPT-4 in repairing vulnerabilities in code generated by GPT-3.5, GPT-4, Code Llama, and CodeGeeX2