Table 1: The complete result of LLMs on three unit test tasks.

Model Name	Test Generation					Assertion Generation			Test Evolution		
	Correct	Passing	Failing	Build Error	Syntax Error	EM	BLEU	CodeBLEU	EM	BLEU	CodeBLEU
CodeBERT	5.56%	8.41%	5.99%	78.90%	6.70%	54.64%	74.31%	73.89%	6.15%	37.72%	46.42%
GraphCodeBERT	11.77%	14.02%	11.19%	65.89%	8.89%	58.32%	74.69%	76.02%	12.69%	63.49%	67.37%
UniXcoder	6.42%	8.12%	10.48%	69.02%	12.37%	55.12%	68.92%	73.14%	10.58%	58.03%	63.12%
CodeT5_small	17.93%	20.22%	22.54%	53.39%	3.86%	56.60%	83.30%	86.11%	11.54%	72.56%	72.56%
CodeT5_base	15.91%	18.76%	14.90%	61.36%	4.98%	60.26%	85.12%	87.32%	12.88%	78.31%	78.31%
CodeT5_large	16.06%	17.88%	22.63%	57.15%	2.34%	58.86%	84.23%	86.57%	13.85%	78.78%	79.39%
CodeT5p_220m	17.75%	19.79%	27.07%	50.26%	2.88%	61.76%	85.89%	88.25%	17.12%	81.54%	81.54%
CodeT5p_770m	14.13%	17.48%	22.30%	52.23%	7.99%	60.29%	85.04%	87.08%	21.92%	77.08%	79.31%
PLBART_base	12.41%	13.67%	13.01%	66.06%	7.26%	53.92%	82.09%	84.86%	12.88%	77.88%	78.45%
PLBART_large	17.50%	20.25%	20.63%	52.96%	6.16%	55.10%	82.52%	85.15%	15.19%	77.80%	78.67%
CodeGPT	7.66%	9.64%	9.81%	39.67%	40.88%	51.30%	61.28%	69.12%	15.38%	79.77%	81.58%
InCoder_1b	17.50%	19.32%	21.56%	50.73%	8.39%	62.24%	77.17%	75.50%	26.15%	79.95%	83.41%
CodeGen_350m	13.65%	16.51%	9.19%	34.63%	39.67%	59.23%	72.57%	73.96%	25.77%	80.80%	80.76%
CodeGen_2b	15.42%	17.37%	10.54%	42.18%	29.91%	64.23%	76.49%	78.09%	25.96%	80.93%	83.93%
CodeGen_6b	21.32%	23.83%	16.02%	34.91%	25.23%	64.14%	77.72%	79.09%	27.69%	82.25%	83.15%
CodeGen_16b	-	-	-	-	-	-	-	-	32.31%	83.89%	84.82%
CodeGen2_1b	-	-	-	-	-	61.24%	72.31%	75.95%	29.04%	82.89%	84.55%
CodeGen2_3b	-	-	-	-	-	64.23%	76.49%	78.09%	28.08%	83.73%	84.89%
StarCodeBase_1b	26.08%	29.56%	14.62%	35.36%	20.46%	62.34%	71.23%	76.66%	26.35%	82.40%	83.90%
StarCodeBase_3b	28.51%	30.74%	13.82%	34.41%	21.04%	65.97%	77.61%	79.88%	26.15%	78.84%	82.68%
StarCodeBase_7b	28.88%	32.76%	15.52%	34.29%	17.43%	67.40%	80.65%	81.52%	28.27%	81.37%	83.73%
StarCodeBase_15b	-	-	-	-	-	-	-	-	33.46%	84.33%	84.97%
StarCoder_15b	-	-	-	-	-	-	-	-	33.46%	84.01%	84.93%
StarCoder2_3b	21.66%	24.37%	19.04%	42.85%	13.74%	63.29%	72.50%	77.60%	21.73%	79.55%	81.97%
StarCoder2_7b	20.78%	23.10%	23.16%	41.45%	12.30%	64.77%	78.29%	78.71%	22.88%	80.49%	82.80%
StarCoder2_15b	-	-	-	-	-	-	-	-	29.04%	83.45%	84.71%
CodeLlama_7b	29.50%	32.14%	7.86%	30.87%	29.13%	71.42%	83.92%	83.34%	34.62%	81.22%	84.52%
CodeLlama_13b	-	-	-	-	-	-	-	-	34.62%	85.20%	86.13%
CodeGemma_2b	22.84%	25.50%	15.03%	50.37%	9.10%	61.06%	74.26%	77.79%	27.69%	84.43%	85.20%
CodeGemma_7b	-	-	-	-	-	-	-	-	28.08%	84.24%	84.91%
Phi_1	15.74%	22.56%	14.75%	51.40%	11.29%	54.93%	68.03%	72.71%	17.69%	79.49%	78.32%
Phi_2	13.80%	16.45%	9.32%	56.29%	17.93%	55.81%	67.53%	71.96%	19.42%	80.09%	82.25%
DeciCoder_1b	18.93%	23.98%	9.87%	42.47%	23.68%	61.23%	77.60%	75.90%	31.15%	83.15%	84.12%
DeepSeek-Coder_1b	17.58%	22.37%	13.05%	36.20%	28.38%	65.64%	79.92%	79.80%	33.65%	83.90%	85.74%
DeepSeek-Coder_6b	33.68%	36.02%	9.00%	21.36%	33.62%	70.57%	82.99%	82.66%	35.58%	84.48%	86.35%
SantaCoder	11.55%	15.44%	7.38%	39.33%	37.85%	57.60%	73.44%	74.66%	27.12%	82.46%	83.85%
CodeShell_7b	-	-	-	-	-	-	-	-	33.46%	82.90%	85.26%
GPT-3.5	49.16%	51.03%	11.33%	34.44%	3.20%	2.97%	32.41%	25.73%	15.96%	86.20%	83.61%
llama3.1-8b	30.34%	35.04%	10.22%	53.99%	0.75%	0.22%	20.93%	19.46%	4.23%	59.64%	65.25%
llama3.1-70b	31.02%	33.38%	7.99%	58.09%	0.54%	3.89%	31.58%	27.46%	5.39%	58.40%	63.37%
llama3-8b	36.88%	38.73%	16.98%	43.60%	0.69%	0.14%	11.49%	18.10%	5.96%	71.16%	76.29%
qwen2-7b	13.72%	18.89%	15.44%	49.34%	16.32%	0.91%	10.17%	22.49%	0.19%	42.70%	51.07%
qwen2-72b	29.48%	31.13%	21.81%	45.60%	1.46%	4.89%	27.31%	27.46%	3.46%	54.67%	62.18%