Table 10: Results of pass@1 (%) on HumanEval and MBPP. Most scores are retrieved from the papers of StarCoder (Li et al., 2023d), CodeT5+ (Wang et al., 2023d), WizardCoder (Luo et al., 2023b) and CODE LLAMA (Rozière et al., 2023).

Model	Params	HumanEval	MBPP
Proprietary models			
PaLM	540B	26.2	36.8
PaLM-Coder	540B	36.0	47.0
PaLM 2-S	-	37.6	50.0
Code-Cushman-001	-	33.5	45.9
Code-Davinci-002	-	47.0	58.1
GPT-3.5	-	73.2	-
GPT-4	-	86.6	-
Open-source models			
LLAMA 2	7B	12.2	20.8
	13B	20.1	27.6
	34B	22.6	33.8
	70B	30.5	45.4
CodeGen-Multi	16B	18.3	20.9
CodeGen-Mono	16B	29.3	35.3
CodeGeeX2	6B	35.9	-
StarCoder-Prompted	15B	40.8	49.5
CodeT5+	16B	30.9	-
InstructCodeT5+	16B	35.0	-
CODE LLAMA	7B	33.5	41.4
	13B	36.0	47.0
	34B	48.8	55.0
CODE LLAMA-INSTRUCT	7B	34.8	44.4
	13B	42.7	49.4
	34B	41.5	57.0
CODE LLAMA-PYTHON	7B	38.4	47.6
	13B	43.3	49.0
	34B	53.7	56.2
UNNATURAL CODE LLAMA	34B	62.2	61.2
WizardCoder-Python	13B	64.0	**55.6**
	34B	73.2	61.2
QWEN-CHAT	7B	37.2	35.8
	14B	43.9	46.4
CODE-QWEN	7B	40.2	41.8
	14B	45.1	51.4
CODE-QWEN-CHAT	7B	43.3	44.2
	14B	**66.4**	52.4