Table 11: **Zero-shot pass@1 (%) performance on the HUMANEVALPACK (synthesize) benchmark.** The baseline results are partly from OCTOPACK (Muennighoff et al., 2023)

Model	Params	Programming Language						
		Python	JavaScript	Java	Go	C++	Rust	Avg.
Proprietary models								
GPT-4	-	86.6	82.9	81.7	72.6	78.7	67.1	78.3
Open-source models								
InstructCodeT5+	16B	37.0	18.9	17.4	9.5	19.8	0.3	17.1
StarChat-β	15B	33.5	31.4	26.7	25.5	26.6	14.0	26.3
StarCoder	15B	33.6	30.8	30.2	17.6	31.6	21.8	27.6
CodeGeeX2	6B	35.9	32.2	30.8	22.5	29.3	18.1	28.1
OCTOGEEX	6B	44.7	33.8	36.9	21.9	32.3	15.7	30.9
OCTOCODER	15B	46.2	39.2	38.2	30.4	35.6	23.4	35.5
WizardCoder	15B	59.8	49.5	36.1	36.4	40.9	20.2	40.5
QWEN-CHAT	7B	37.2	23.2	32.9	20.7	22.0	9.1	24.2
	14B	43.9	38.4	42.7	34.1	24.4	18.9	33.7
CODE-QWEN	7B	40.2	40.4	40.2	26.2	20.7	15.8	30.6
	14B	45.1	51.8	57.3	39.6	18.2	20.7	38.8
CODE-QWEN-CHAT	7B	43.3	41.5	49.4	29.3	32.9	20.1	36.1
	14B	**66.4**	**58.5**	**56.1**	**47.6**	**54.2**	**28.7**	**51.9**

Table 12: **Results of models on mathematical reasoning**. We report the accuracy of QWEN for all benchmarks using greedy decoding. For MATH, we are reporting QWEN's performances on the test set from Lightman et al. (2023).

Model	Params	GSM8K	MATH	Math401	Math23K
Proprietary models					
GPT-4	-	**92.0**	**42.5**	83.5	74.0
GPT-3.5	-	80.8	34.1	75.1	60.0
Minerva	8B	16.2	14.1	-	-
	62B	52.4	27.6	-	-
	540B	58.8	33.6	-	-
Open-source models					
LLaMA-1 RFT	7B	46.5	5.2	-	-
	13B	52.1	5.1	-	-
WizardMath	7B	54.9	10.7	-	-
	13B	63.9	14.0	-	-
	70B	81.6	22.7	-	-
GAIRMath-Abel	7B	59.7	13.0	-	-
	13B	66.4	17.3	-	-
	70B	83.6	28.3	-	-
QWEN-CHAT	7B	50.3	6.8	57.4	51.2
	14B	60.1	18.4	70.1	67.0
MATH-QWEN-CHAT	7B	62.5	17.2	80.8	75.4
	14B	69.8	24.2	**85.0**	**78.4**