Table 7: The proportion of code generated by QWEN that is executable on the in-house evaluation benchmark for Code Interpreter. This benchmark examines QWEN's coding proficiency in math problem solving, data visualization, and general purposes. CODE LLAMA underperforms on visualization tasks because it hallucinates non-existent columns solely based on CSV file names (see Figure 5).

Model	**Params**	**Category**			
		Math (%)	Visualization (%)	General (%)	All (%)
GPT-4	-	91.9	85.9	82.8	86.8
GPT-3.5	-	89.2	65.0	74.1	72.9
LLAMA 2-CHAT	7B	41.9	33.1	24.1	33.6
	13B	50.0	40.5	48.3	44.4
CODE LLAMA-INSTRUCT	7B	85.1	54.0	70.7	65.1
	13B	93.2	55.8	74.1	68.8
InternLM-Chat	7B v1.1	78.4	44.2	62.1	56.3
	20B	70.3	44.2	65.5	54.9
QWEN-CHAT	1.8B	33.8	30.1	8.6	26.8
	7B	82.4	64.4	67.2	70.2
	14B	89.2	84.1	65.5	81.7

Table 8: Correctness of the final response on the in-house evaluation benchmark for Code Interpreter. Visualization-Hard tasks involve planning multiple steps, while Visualization-Easy tasks do not. Visualization-All measures both types of tasks. CODE LLAMA excels in performing Visualization-Easy tasks but tends to underperform in Visualization-Hard tasks, due to its inclination to hallucinate non-existent columns based on the name of a CSV file (see Figure 5).

Model	**Params**	**Category**			
		Math (%)	Vis.-Hard (%)	Vis.-Easy (%)	Vis.-All (%)
GPT-4	-	82.8	66.7	60.8	63.8
GPT-3.5	-	47.3	33.3	55.7	44.2
LLAMA 2-CHAT	7B	3.9	14.3	39.2	26.4
	13B	8.3	8.3	40.5	23.9
CODE LLAMA-INSTRUCT	7B	14.3	26.2	60.8	42.9
	13B	28.2	27.4	62.0	44.2
InternLM-Chat	7B v1.1	28.5	4.8	40.5	22.1
	20B	34.6	21.4	45.6	33.1
QWEN-CHAT	1.8B	14.7	3.6	20.3	11.7
	7B	41.9	40.5	54.4	47.2
	14B	58.4	53.6	59.5	56.4