Table 1: Number of instances with compiler errors in unconstrained (Standard), idealized syntax-only (Syntax), and our proposed type-aware (Types) constraining. Type-aware constraining reduces compiler errors by 75.3% and 52.1% in the synthesis of HumanEval and MBPP problems respectively, compared to only 9.0% and 4.9% ideal improvement through syntax-only constraining on the two datasets respectively.

		Synthesis			Translation		
	Model	Standard	Syntax	Types	Standard	Syntax	Types
HumanEval	Gemma 2 2B	103	$92_{\downarrow 10.7\%}$	$\mathbf{44}_{\downarrow 57.3\%}$	177	$149_{\downarrow 15.8\%}$	$\mathbf{80}_{\downarrow 54.8\%}$
	Gemma 2 9B	45	$41_{\downarrow 8.9\%}$	$\mathbf{13}_{\downarrow 71.1\%}$	75	$63_{\downarrow 16.0\%}$	$\mathbf{16}_{\downarrow 78.7\%}$
	Gemma 2 27B	15	$13_{\downarrow 13.3\%}$	$\mathbf{2}_{\downarrow 86.7\%}$	20	$20_{\downarrow 0.0\%}$	$\mathbf{3}_{\downarrow 85.0\%}$
	DeepSeek Coder 33B	26	$25_{\downarrow 3.8\%}$	$\mathbf{5}_{\downarrow 80.8\%}$	18	$17_{\downarrow 5.6\%}$	$\mathbf{7}_{\downarrow 61.1\%}$
	CodeLlama 34B	86	$71_{\downarrow 17.4\%}$	$\mathbf{28}_{\downarrow 67.4\%}$	158	$124_{\downarrow 21.5\%}$	$\mathbf{59}_{\downarrow 62.7\%}$
	Qwen2.5 32B	17	$17_{\downarrow 0.0\%}$	$\mathbf{2}_{\downarrow 88.2\%}$	24	$21_{\downarrow 12.5\%}$	$\mathbf{5}_{\downarrow 79.2\%}$
MBPP	Gemma 2 2B	67	$64_{\downarrow 4.5\%}$	$\mathbf{27}_{\downarrow 59.7\%}$	126	$111_{\downarrow 11.9\%}$	$\mathbf{79}_{\downarrow 37.3\%}$
	Gemma 2 9B	30	$29_{\downarrow 3.3\%}$	$\mathbf{10}_{\downarrow 66.7\%}$	67	$61_{\downarrow 9.0\%}$	$\mathbf{33}_{\downarrow 50.7\%}$
	Gemma 2 27B	20	$19_{\downarrow 5.0\%}$	$\mathbf{7}_{\downarrow 65.0\%}$	37	$36_{\downarrow 2.7\%}$	$\mathbf{22}_{\downarrow 40.5\%}$
	DeepSeek Coder 33B	32	$32_{\downarrow 0.0\%}$	$\mathbf{19}_{\downarrow 40.6\%}$	29	$27_{\downarrow 6.9\%}$	$\mathbf{13}_{\downarrow 55.2\%}$
	CodeLlama 34B	80	$71_{\downarrow 11.2\%}$	$\mathbf{41}_{\downarrow 48.8\%}$	126	$114_{\downarrow 9.5\%}$	$\mathbf{54}_{\downarrow 57.1\%}$
	Qwen2.5 32B	19	$18_{\downarrow 5.3\%}$	$\mathbf{13}_{\downarrow 31.6\%}$	22	$22_{\downarrow 0.0\%}$	$\mathbf{16}_{\downarrow 27.3\%}$

Table 2: Trying to repair non-compiling instances generated with unconstrained synthesis using unconstrained decoding (Standard) and type-aware constraining (Types). The results are the number of non-repaired instances. Constrained generation boosts repair by on average 50.1%.

(a) Repair of HumanEval

Model	Standard	Types
Gemma 2 2B	$194_{\downarrow 33.6\%}$	$\mathbf{103}_{\downarrow 64.7\%}$
Gemma 2 9B	$113_{\downarrow 61.3\%}$	$\mathbf{52}_{\downarrow 82.2\%}$
Gemma 2 27B	$45_{\downarrow 84.6\%}$	$\mathbf{22}_{\downarrow 92.5\%}$
DeepSeek Coder 33B	$36_{\downarrow 87.7\%}$	$\mathbf{15}_{\downarrow 94.9\%}$
CodeLlama 34B	$153_{\downarrow 47.6\%}$	$\mathbf{48}_{\downarrow 83.6\%}$
Qwen2.5 32B	$36_{\downarrow 87.7\%}$	$\mathbf{13}_{\downarrow 95.5\%}$

(b) Repair of MBPP

Model	Standard	Types
Gemma 2 2B	$215_{\downarrow 20.7\%}$	$\mathbf{125}_{\downarrow 53.9\%}$
Gemma 2 9B	$151_{\downarrow 44.3\%}$	$\mathbf{82}_{\downarrow 69.7\%}$
Gemma 2 27B	$91_{\downarrow 66.4\%}$	$\mathbf{50}_{\downarrow 81.5\%}$
DeepSeek Coder 33B	$107_{\downarrow 60.5\%}$	$\mathbf{59}_{\downarrow 78.2\%}$
CodeLlama 34B	$178_{\downarrow 34.3\%}$	$\mathbf{95}_{\downarrow 64.9\%}$
Qwen2.5 32B	$72_{\downarrow 73.4\%}$	$\mathbf{45}_{\downarrow 83.4\%}$

Table 3: Pass@1 (in %) of unconstrained (Standard) and type-aware constrained (Types) generated code for the tasks of Synthesis, Repair, and Translation.

		Synthesis		Translation		Repair	
	Model	Standard	Types	Standard	Types	Standard	Types
HumanEval	Gemma 2 2B	29.1	**30.2**	50.2	**53.9**	11.6	**20.9**
	Gemma 2 9B	56.6	**58.3**	73.7	**78.3**	24.0	**34.9**
	Gemma 2 27B	69.5	**71.2**	86.6	**87.7**	38.4	**41.1**
	DS Coder 33B	68.9	**71.1**	88.7	**90.1**	47.6	**50.7**
	CodeLlama 34B	41.0	**43.4**	58.6	**63.5**	17.5	**27.4**
	Qwen2.5 32B	79.6	**81.8**	92.1	**93.9**	65.4	**71.2**
MBPP	Gemma 2 2B	40.4	**42.4**	52.3	**56.0**	11.1	**20.7**
	Gemma 2 9B	65.4	**67.4**	71.4	**75.8**	22.1	**29.2**
	Gemma 2 27B	70.6	**72.1**	83.1	**84.4**	35.8	**41.3**
	DS Coder 33B	65.4	**67.2**	85.9	**89.1**	32.1	**39.5**
	CodeLlama 34B	42.2	**45.6**	55.7	**63.3**	14.8	**24.7**
	Qwen2.5 32B	76.3	**76.6**	89.6	**90.4**	44.6	**50.2**

Model	HumanEval	MBPP
Gemma 2 2B	$6.7_{\uparrow 38.3\%}$	$6.3_{\uparrow 35.4\%}$
Gemma 2 9B	$8.3_{\uparrow 29.2\%}$	$9.5_{\uparrow 46.8\%}$
Gemma 2 27B	$11.7_{\uparrow 19.9\%}$	$11.7_{\uparrow 32.8\%}$
DeepSeek Coder 33B	$11.5_{\uparrow 36.2\%}$	$9.4_{\uparrow 59.5\%}$
CodeLlama 34B	$7.6_{\uparrow 40.8\%}$	$7.0_{\uparrow 37.6\%}$
Qwen2.5 32B	$7.3_{\uparrow 39.6\%}$	$4.9_{\uparrow 54.8\%}$

Table 4: Median time taken per instance in seconds.