Table 19: **Comparison among Qwen3-1.7B / Qwen3-0.6B (Thinking) and other reasoning baselines. The highest and second-best scores are shown in bold and underlined, respectively.**

		DeepSeek-R1 -Distill-Qwen-1.5B	**DeepSeek-R1 -Distill-Llama-8B**	**Qwen3-0.6B**	**Qwen3-1.7B**
	Architecture	Dense	Dense	Dense	Dense
	# Activated Params	1.5B	8B	0.6B	1.7B
	# Total Params	1.5B	8B	0.6B	1.7B
General Tasks	MMLU-Redux	45.4	66.4	55.6	**73.9**
	GPQA-Diamond	33.8	**49.0**	27.9	40.1
	C-Eval	27.1	50.4	50.4	**68.1**
	LiveBench 2024-11-25	24.9	40.6	30.3	**51.1**
Alignment Tasks	IFEval strict prompt	39.9	59.0	59.2	**72.5**
	Arena-Hard	4.5	17.6	8.5	**43.1**
	AlignBench v1.1	5.00	6.24	6.10	**7.60**
	Creative Writing v3	16.4	**51.1**	30.6	48.0
	WritingBench	4.03	5.42	5.61	**7.02**
Math & Text Reasoning	MATH-500	83.9	89.1	77.6	**93.4**
	AIME'24	28.9	**50.4**	10.7	48.3
	AIME'25	22.8	27.8	15.1	**36.8**
	ZebraLogic	4.9	37.1	30.3	**63.2**
	AutoLogi	19.1	63.4	61.6	**83.2**
Agent & Coding	BFCL v3	14.0	21.5	46.4	**56.6**
	LiveCodeBench v5	13.2	**42.5**	12.3	33.2
Multilingual Tasks	Multi-IF	13.3	27.0	36.1	**51.2**
	INCLUDE	21.9	34.5	35.9	**51.8**
	MMMLU 14 languages	27.3	40.1	43.1	**59.1**
	MT-AIME2024	12.4	13.2	7.8	**36.1**
	PolyMath	14.5	10.8	11.4	**25.2**
	MLogiQA	29.0	32.8	40.9	**56.0**

Table 20: **Comparison among Qwen3-1.7B / Qwen3-0.6B (Non-thinking) and other non-reasoning baselines. The highest and second-best scores are shown in bold and underlined, respectively.**

		Gemma-3 -1B-IT	**Phi-4-mini**	**Qwen2.5-1.5B -Instruct**	**Qwen2.5-3B -Instruct**	**Qwen3-0.6B**	**Qwen3-1.7B**
	Architecture	Dense	Dense	Dense	Dense	Dense	Dense
	# Activated Params	1.0B	3.8B	1.5B	3.1B	0.6B	1.7B
	# Total Params	1.0B	3.8B	1.5B	3.1B	0.6B	1.7B
General Tasks	MMLU-Redux	33.3	**67.9**	50.7	64.4	44.6	64.4
	GPQA-Diamond	19.2	25.2	29.8	**30.3**	22.9	28.6
	C-Eval	28.5	40.0	53.3	**68.2**	42.6	61.0
	LiveBench 2024-11-25	14.4	25.3	18.0	23.8	21.8	**35.6**
Alignment Tasks	IFEval strict prompt	54.5	**68.6**	42.5	58.2	54.5	68.2
	Arena-Hard	17.8	32.8	9.0	23.7	6.5	**36.9**
	AlignBench v1.1	5.3	6.00	5.60	6.49	5.60	**7.20**
	Creative Writing v3	**52.8**	10.3	31.5	42.8	28.4	43.6
	WritingBench	5.18	4.05	4.67	5.55	5.13	**6.54**
Math & Text Reasoning	MATH-500	46.4	67.6	55.0	67.2	55.2	**73.0**
	AIME'24	0.9	8.1	0.9	6.7	3.4	**13.4**
	AIME'25	0.8	5.3	0.4	4.2	2.6	**9.8**
	ZebraLogic	1.9	2.7	3.4	4.8	4.2	**12.8**
	AutoLogi	16.4	28.8	22.5	29.9	37.4	**59.8**
Agent & Coding	BFCL v3	16.3	31.3	47.8	50.4	44.1	**52.2**
	LiveCodeBench v5	1.8	10.4	5.3	9.2	3.6	**11.6**
Multilingual Tasks	Multi-IF	32.8	40.5	20.2	32.3	33.3	**44.7**
	INCLUDE	32.7	**43.8**	33.1	**43.8**	34.4	42.6
	MMMLU 14 languages	32.5	51.4	40.4	**51.8**	37.1	48.3
	MT-AIME2024	0.2	0.9	0.7	1.6	1.5	**4.9**
	PolyMath	3.5	6.7	5.0	7.3	4.6	**10.3**
	MLogiQA	31.8	39.5	40.9	39.5	37.3	**41.1**