Table 17: **Comparison among Qwen3-8B / Qwen3-4B (Thinking) and other reasoning baselines. The highest and second-best scores are shown in bold and underlined, respectively.**

		DeepSeek-R1 -Distill-Qwen-14B	DeepSeek-R1 -Distill-Qwen-32B	Qwen3-4B	Qwen3-8B
	Architecture	Dense	Dense	Dense	Dense
	# Activated Params	14B	32B	4B	8B
	# Total Params	14B	32B	4B	8B
General Tasks	MMLU-Redux	84.1	**88.2**	83.7	87.5
	GPQA-Diamond	59.1	**62.1**	55.9	62.0
	C-Eval	78.1	82.2	77.5	**83.4**
	LiveBench 2024-11-25	52.3	45.6	63.6	**67.1**
Alignment Tasks	IFEval strict prompt	72.6	72.5	81.9	**85.0**
	Arena-Hard	48.0	60.8	76.6	**85.8**
	AlignBench v1.1	7.43	7.25	8.30	**8.46**
	Creative Writing v3	54.2	55.0	61.1	**75.0**
	WritingBench	6.03	6.13	7.35	**7.59**
Math & Text Reasoning	MATH-500	93.9	94.3	97.0	**97.4**
	AIME'24	69.7	72.6	73.8	**76.0**
	AIME'25	44.5	49.6	65.6	**67.3**
	ZebraLogic	59.1	69.6	81.0	**84.8**
	AutoLogi	78.6	74.6	87.9	**89.1**
Agent & Coding	BFCL v3	49.5	53.5	65.9	**68.1**
	LiveCodeBench v5	45.5	54.5	54.2	**57.5**
	CodeForces (Rating / Percentile)	1574 / 89.1%	1691 / 93.4%	1671 / 92.8%	**1785 / 95.6%**
Multilingual Tasks	Multi-IF	29.8	31.3	66.3	**71.2**
	INCLUDE	59.7	**68.0**	61.8	67.8
	MMMLU 14 languages	73.8	**78.6**	69.8	74.4
	MT-AIME2024	33.7	44.6	60.7	**65.4**
	PolyMath	28.6	35.1	40.0	**42.7**
	MLogiQA	53.6	63.3	65.9	**69.0**

Table 18: **Comparison among Qwen3-8B / Qwen3-4B (Non-thinking) and other non-reasoning baselines. The highest and second-best scores are shown in bold and underlined, respectively.**

		LLaMA-3.1-8B -Instruct	Gemma-3 -12B-IT	Qwen2.5-7B -Instruct	Qwen2.5-14B -Instruct	Qwen3-4B	Qwen3-8B
	Architecture	Dense	Dense	Dense	Dense	Dense	Dense
	# Activated Params	8B	12B	7B	14B	4B	8B
	# Total Params	8B	12B	7B	14B	4B	8B
General Tasks	MMLU-Redux	61.7	77.8	75.4	**80.0**	77.3	79.5
	GPQA-Diamond	32.8	40.9	36.4	**45.5**	41.7	39.3
	C-Eval	52.0	61.1	76.2	**78.0**	72.2	77.9
	LiveBench 2024-11-25	26.0	43.7	34.9	42.2	48.4	**53.5**
Alignment Tasks	IFEval strict prompt	75.0	80.2	71.2	81.0	81.2	**83.0**
	Arena-Hard	30.1	**82.6**	52.0	68.3	66.2	79.6
	AlignBench v1.1	6.01	7.77	7.27	7.67	8.10	**8.38**
	Creative Writing v3	52.8	**79.9**	49.8	55.8	53.6	64.5
	WritingBench	4.57	7.05	5.82	5.93	6.85	**7.15**
Math & Text Reasoning	MATH-500	54.8	85.6	77.6	83.4	84.8	**87.4**
	AIME'24	6.3	22.4	9.1	15.2	25.0	**29.1**
	AIME'25	2.7	18.8	12.1	13.6	19.1	**20.9**
	ZebraLogic	12.8	17.8	12.0	19.7	**35.2**	26.7
	AutoLogi	30.9	58.9	42.9	57.4	76.3	**76.5**
Agent & Coding	BFCL v3	49.6	50.6	55.8	58.7	57.6	**60.2**
	LiveCodeBench v5	10.8	**25.7**	14.4	21.9	21.3	22.8
	CodeForces (Rating / Percentile)	473 / 14.9%	462 / 14.7%	191 / 0.0%	904 / 38.3%	842 / 33.7%	**1110 / 52.4%**
Multilingual Tasks	Multi-IF	52.1	65.6	47.7	55.5	61.3	**69.2**
	INCLUDE	34.0	**65.3**	53.6	63.5	53.8	62.5
	MMMLU 14 languages	44.4	70.0	61.4	**70.3**	61.7	66.9
	MT-AIME2024	0.4	**16.7**	5.5	8.5	13.9	16.6
	PolyMath	5.8	17.6	11.9	15.0	16.6	**18.8**
	MLogiQA	41.9	**54.5**	49.5	51.3	49.9	51.4