Table 15: **Comparison among Qwen3-30B-A3B / Qwen3-14B (Thinking) and other reasoning baselines. The highest and second-best scores are shown in bold and underlined, respectively.**

		DeepSeek-R1 -Distill-Qwen-32B	**QwQ-32B**	**Qwen3-14B**	**Qwen3-30B-A3B**
	Architecture	Dense	Dense	Dense	MoE
	# Activated Params	32B	32B	14B	3B
	# Total Params	32B	32B	14B	30B
General Tasks	MMLU-Redux	88.2	**90.0**	88.6	89.5
	GPQA-Diamond	62.1	65.6	64.0	**65.8**
	C-Eval	82.2	**88.4**	86.2	86.6
	LiveBench 2024-11-25	45.6	72.0	71.3	**74.3**
Alignment Tasks	IFEval strict prompt	72.5	83.9	85.4	**86.5**
	Arena-Hard	60.8	89.5	**91.7**	91.0
	AlignBench v1.1	7.25	**8.70**	8.56	**8.70**
	Creative Writing v3	55.0	**82.4**	80.3	79.1
	WritingBench	6.13	**7.86**	7.80	7.70
Math & Text Reasoning	MATH-500	94.3	**98.0**	96.8	**98.0**
	AIME’24	72.6	79.5	79.3	**80.4**
	AIME’25	49.6	69.5	70.4	**70.9**
	ZebraLogic	69.6	76.8	88.5	**89.5**
	AutoLogi	74.6	88.1	**89.2**	88.7
Agent & Coding	BFCL v3	53.5	66.4	**70.4**	69.1
	LiveCodeBench v5	54.5	62.7	**63.5**	62.6
	CodeForces (Rating / Percentile)	1691 / 93.4%	**1982 / 97.7%**	1766 / 95.3%	1974 / 97.7%
Multilingual Tasks	Multi-IF	31.3	68.3	**74.8**	72.2
	INCLUDE	68.0	69.7	71.7	**71.9**
	MMMLU 14 languages	78.6	**80.9**	77.9	78.4
	MT-AIME2024	44.6	68.0	73.3	**73.9**
	PolyMath	35.1	45.9	45.8	**46.1**
	MLogiQA	63.3	**75.5**	71.1	70.1

Table 16: **Comparison among Qwen3-30B-A3B / Qwen3-14B (Non-thinking) and other non-reasoning baselines. The highest and second-best scores are shown in bold and underlined, respectively.**

		Phi-4	**Gemma-3 -27B-IT**	**Qwen2.5-32B -Instruct**	**Qwen3-14B**	**Qwen3-30B-A3B**
	Architecture	Dense	Dense	Dense	Dense	MoE
	# Activated Params	14B	27B	32B	14B	3B
	# Total Params	14B	27B	32B	14B	30B
General Tasks	MMLU-Redux	**85.3**	82.6	83.9	82.0	84.1
	GPQA-Diamond	**56.1**	42.4	49.5	54.8	54.8
	C-Eval	66.9	66.6	80.6	81.0	**82.9**
	LiveBench 2024-11-25	41.6	49.2	50.0	**59.6**	59.4
Alignment Tasks	IFEval strict prompt	62.1	80.6	79.5	**84.8**	83.7
	Arena-Hard	75.4	86.8	74.5	86.3	**88.0**
	AlignBench v1.1	7.61	7.80	7.71	8.52	**8.55**
	Creative Writing v3	51.2	**82.0**	54.6	73.1	68.1
	WritingBench	5.73	7.22	5.90	**7.24**	7.22
Math & Text Reasoning	MATH-500	80.8	**90.0**	84.6	**90.0**	89.8
	AIME’24	22.9	32.6	18.8	31.7	**32.8**
	AIME’25	17.3	**24.0**	12.8	23.3	21.6
	ZebraLogic	32.3	24.6	26.1	33.0	**33.2**
	AutoLogi	66.2	64.2	65.5	**82.0**	81.5
Agent & Coding	BFCL v3	47.0	59.1	**62.8**	61.5	58.6
	LiveCodeBench v5	25.2	26.9	26.4	29.0	**29.8**
	CodeForces (Rating / Percentile)	**1280 / 65.3%**	1063 / 49.3%	903 / 38.2%	1200 / 58.6%	1267 / 64.1%
Multilingual Tasks	Multi-IF	49.5	69.8	63.2	**72.9**	70.8
	INCLUDE	65.3	**71.4**	67.5	67.8	67.8
	MMMLU 14 languages	74.7	**76.1**	74.2	72.6	73.8
	MT-AIME2024	13.1	23.0	15.3	23.2	**24.6**
	PolyMath	17.4	20.3	18.3	22.0	**23.3**
	MLogiQA	53.1	58.5	58.0	**58.9**	53.3