Table 13: **Comparison among Qwen3-32B (Thinking) and other reasoning baselines. The highest and second-best scores are shown in bold and underlined, respectively.**

		DeepSeek-R1 -Distill-Llama-70B	QwQ-32B	OpenAI-o3-mini (medium)	Qwen3-32B
	Architecture	Dense	Dense	-	Dense
	# Activated Params	70B	32B	-	32B
	# Total Params	70B	32B	-	32B
General Tasks	MMLU-Redux	89.3	90.0	90.0	**90.9**
	GPQA-Diamond	65.2	65.6	**76.8**	68.4
	C-Eval	71.8	**88.4**	75.1	87.3
	LiveBench 2024-11-25	54.5	72.0	70.0	**74.9**
Alignment Tasks	IFEval strict prompt	79.3	83.9	**91.5**	85.0
	Arena-Hard	60.6	89.5	89.0	**93.8**
	AlignBench v1.1	6.74	8.70	8.38	**8.72**
	Creative Writing v3	62.1	**82.4**	74.8	81.0
	WritingBench	6.08	7.86	7.52	**7.90**
Math & Text Reasoning	MATH-500	94.5	**98.0**	**98.0**	97.2
	AIME'24	70.0	79.5	79.6	**81.4**
	AIME'25	56.3	69.5	**74.8**	72.9
	ZebraLogic	71.3	76.8	**88.9**	88.8
	AutoLogi	83.5	**88.1**	86.3	87.3
Agent & Coding	BFCL v3	49.3	66.4	64.6	**70.3**
	LiveCodeBench v5	54.5	62.7	**66.3**	65.7
	CodeForces (Rating / Percentile)	1633 / 91.4%	1982 / 97.7%	**2036 / 98.1%**	1977 / 97.7%
Multilingual Tasks	Multi-IF	57.6	68.3	48.4	**73.0**
	INCLUDE	62.1	69.7	73.1	**73.7**
	MMMLU 14 languages	69.6	**80.9**	79.3	80.6
	MT-AIME2024	29.3	68.0	73.9	**75.0**
	PolyMath	29.4	45.9	38.6	**47.4**
	MLogiQA	60.3	75.5	71.1	**76.3**

Table 14: **Comparison among Qwen3-32B (Non-thinking) and other non-reasoning baselines. The highest and second-best scores are shown in bold and underlined, respectively.**

		GPT-4o-mini -2024-07-18	LLaMA-4 -Scout	Qwen2.5-72B -Instruct	Qwen3-32B
	Architecture	-	MoE	Dense	Dense
	# Activated Params	-	17B	72B	32B
	# Total Params	-	109B	72B	32B
General Tasks	MMLU-Redux	81.5	86.3	**86.8**	85.7
	GPQA-Diamond	40.2	**57.2**	49.0	54.6
	C-Eval	66.3	78.2	**84.7**	83.3
	LiveBench 2024-11-25	41.3	47.6	51.4	**59.8**
Alignment Tasks	IFEval strict prompt	80.4	**84.7**	84.1	83.2
	Arena-Hard	74.9	70.5	81.2	**92.8**
	AlignBench v1.1	7.81	7.49	7.89	**8.58**
	Creative Writing v3	70.3	55.0	61.8	**78.3**
	WritingBench	5.98	5.49	7.06	**7.54**
Math & Text Reasoning	MATH-500	78.2	82.6	83.6	**88.6**
	AIME'24	8.1	28.6	18.9	**31.0**
	AIME'25	8.8	10.0	15.0	**20.2**
	ZebraLogic	20.1	24.2	26.6	**29.2**
	AutoLogi	52.6	56.8	66.1	**78.5**
Agent & Coding	BFCL v3	**64.0**	45.4	63.4	63.0
	LiveCodeBench v5	27.9	29.8	30.7	**31.3**
	CodeForces (Rating / Percentile)	1113 / 52.6%	981 / 43.7%	859 / 35.0%	**1353 / 71.0%**
Multilingual Tasks	Multi-IF	62.4	64.2	65.3	**70.7**
	INCLUDE	66.0	**74.1**	69.6	70.9
	MMMLU 14 languages	72.1	**77.5**	76.9	76.5
	MT-AIME2024	6.0	19.1	12.7	**24.1**
	PolyMath	12.0	20.9	16.9	**22.5**
	MLogiQA	42.6	53.9	59.3	**62.9**