Table 11: **Comparison among Qwen3-235B-A22B (Thinking) and other reasoning baselines. The highest and second-best scores are shown in bold and underlined, respectively.**

		OpenAI-o1	DeepSeek-R1	Grok-3-Beta (Think)	Gemini2.5-Pro	Qwen3-235B-A22B
	Architecture	-	MoE	-	-	MoE
	# Activated Params	-	37B	-	-	22B
	# Total Params	-	671B	-	-	235B
General Tasks	MMLU-Redux	92.8	92.9	-	**93.7**	92.7
	GPQA-Diamond	78.0	71.5	80.2	**84.0**	71.1
	C-Eval	85.5	**91.8**	-	82.9	89.6
	LiveBench 2024-11-25	75.7	71.6	-	**82.4**	77.1
Alignment Tasks	IFEval strict prompt	**92.6**	83.3	-	89.5	83.4
	Arena-Hard	92.1	92.3	-	**96.4**	95.6
	AlignBench v1.1	8.86	8.76	-	**9.03**	8.94
	Creative Writing v3	81.7	85.5	-	**86.0**	84.6
	WritingBench	7.69	7.71	-	**8.09**	8.03
Math & Text Reasoning	MATH-500	96.4	97.3		**98.8**	98.0
	AIME'24	74.3	79.8	83.9	**92.0**	85.7
	AIME'25	79.2	70.0	77.3	**86.7**	81.5
	ZebraLogic	81.0	78.7	-	**87.4**	80.3
	AutoLogi	79.8	86.1	-	85.4	**89.0**
Agent & Coding	BFCL v3	67.8	56.9	-	62.9	**70.8**
	LiveCodeBench v5	63.9	64.3	70.6	70.4	**70.7**
	CodeForces (Rating / Percentile)	1891 / 96.7%	2029 / 98.1%	-	2001 / 97.9%	**2056 / 98.2%**
Multilingual Tasks	Multi-IF	48.8	67.7	-	**77.8**	71.9
	INCLUDE	84.6	82.7	-	**85.1**	78.7
	MMMLU 14 languages	**88.4**	86.4	-	86.9	84.3
	MT-AIME2024	67.4	73.5	-	76.9	**80.8**
	PolyMath	38.9	47.1	-	52.2	**54.7**
	MLogiQA	75.5	73.8	-	75.6	**77.1**

Table 12: **Comparison among Qwen3-235B-A22B (Non-thinking) and other non-reasoning baselines. The highest and second-best scores are shown in bold and underlined, respectively.**

		GPT-4o -2024-11-20	DeepSeek-V3	Qwen2.5-72B -Instruct	LLaMA-4 -Maverick	Qwen3-235B-A22B
	Architecture	-	MoE	Dense	MoE	MoE
	# Activated Params	-	37B	72B	17B	22B
	# Total Params	-	671B	72B	402B	235B
General Tasks	MMLU-Redux	87.0	89.1	86.8	**91.8**	89.2
	GPQA-Diamond	46.0	59.1	49.0	**69.8**	62.9
	C-Eval	75.5	**86.5**	84.7	83.5	86.1
	LiveBench 2024-11-25	52.2	60.5	51.4	59.5	**62.5**
Alignment Tasks	IFEval strict prompt	86.5	86.1	84.1	**86.7**	83.2
	Arena-Hard	85.3	85.5	81.2	82.7	**96.1**
	AlignBench v1.1	8.42	8.64	7.89	7.97	**8.91**
	Creative Writing v3	**81.1**	74.0	61.8	61.3	80.4
	WritingBench	7.11	6.49	7.06	5.46	**7.70**
Math & Text Reasoning	MATH-500	77.2	90.2	83.6	90.6	**91.2**
	AIME'24	11.1	39.2	18.9	38.5	**40.1**
	AIME'25	7.6	**28.8**	15.0	15.9	24.7
	ZebraLogic	27.4	**42.1**	26.6	40.0	37.7
	AutoLogi	65.9	76.1	66.1	75.2	**83.3**
Agent & Coding	BFCL v3	**72.5**	57.6	63.4	52.9	68.0
	LiveCodeBench v5	32.7	33.1	30.7	**37.2**	35.3
	CodeForces (Rating / Percentile)	864 / 35.4%	1134 / 54.1%	859 / 35.0%	712 / 24.3%	**1387 / 75.7%**
Multilingual Tasks	Multi-IF	65.6	55.6	65.3	**75.5**	70.2
	INCLUDE	78.8	76.7	69.6	**80.9**	75.6
	MMMLU 14 languages	80.3	81.1	76.9	**82.5**	79.8
	MT-AIME2024	9.2	20.9	12.7	27.0	**32.4**
	PolyMath	13.7	20.4	16.9	26.1	**27.0**
	MLogiQA	57.4	58.9	59.3	59.9	**67.6**