

Table 26: **Benchmark scores for language: Portuguese (pt)**. The highest and second-best scores are shown in **bold** and underlined, respectively.

	Model	Multi-IF	MLogiQA	INCLUDE	MMMLU	MT-AIME24	PolyMath	Average
<i>Thinking Mode</i>	Gemini2.5-Pro	80.5	73.8	83.9	88.9	73.3	52.2	75.4
	QwQ-32B	70.5	70.0	<u>80.4</u>	84.0	80.0	48.7	72.3
	Qwen3-235B-A22B	73.6	78.8	78.6	<u>86.2</u>	86.7	58.3	77.0
	Qwen3-32B	74.1	<u>76.2</u>	76.8	82.6	80.0	52.4	73.7
	Qwen3-30B-A3B	76.1	71.2	71.4	81.0	76.7	49.3	71.0
	Qwen3-14B	<u>77.3</u>	68.8	75.0	81.6	<u>83.3</u>	46.7	72.1
	Qwen3-8B	<u>73.9</u>	67.5	75.0	78.6	<u>56.7</u>	44.8	66.1
	Qwen3-4B	70.6	62.5	71.4	75.1	73.3	44.2	66.2
	Qwen3-1.7B	55.6	60.0	53.6	64.6	46.7	28.2	51.4
	Qwen3-0.6B	38.7	33.8	42.9	47.5	10.0	12.7	30.9
<i>Non-thinking Mode</i>	GPT-4o-2024-1120	66.8	57.5	<u>78.6</u>	80.7	10.0	15.0	51.4
	Gemma-3-27b-IT	72.9	55.0	75.0	77.1	33.3	20.9	55.7
	Qwen2.5-72B-Instruct	68.8	55.0	71.4	<u>82.2</u>	23.3	11.3	52.0
	Qwen3-235B-A22B	72.5	67.5	82.1	83.5	33.3	28.3	61.2
	Qwen3-32B	71.1	<u>61.3</u>	73.2	80.6	30.0	23.9	56.7
	Qwen3-30B-A3B	72.3	47.5	67.9	77.8	26.7	24.0	52.7
	Qwen3-14B	75.5	58.8	75.0	76.5	26.7	<u>25.8</u>	56.4
	Qwen3-8B	71.9	56.2	71.4	72.9	20.0	19.7	52.0
	Qwen3-4B	66.1	50.0	73.2	66.7	10.0	18.1	47.4
	Qwen3-1.7B	49.5	33.8	39.3	52.9	6.7	12.8	32.5
	Qwen3-0.6B	36.6	37.5	42.9	37.5	3.3	5.7	27.2

Table 27: **Benchmark scores for language: Italian (it)**. The highest and second-best scores are shown in **bold** and underlined, respectively.

	Model	Multi-IF	INCLUDE	MMMLU	MT-AIME24	PolyMath	Average
<i>Thinking Mode</i>	Gemini2.5-Pro	80.9	100.0	87.2	90.0	<u>54.1</u>	82.4
	QwQ-32B	71.2	<u>96.4</u>	84.9	76.7	49.3	75.7
	Qwen3-235B-A22B	73.7	<u>96.4</u>	<u>85.7</u>	<u>80.0</u>	57.4	78.6
	Qwen3-32B	76.6	<u>90.9</u>	<u>81.6</u>	<u>80.0</u>	49.7	<u>75.8</u>
	Qwen3-30B-A3B	75.9	94.5	81.9	<u>80.0</u>	48.1	76.1
	Qwen3-14B	<u>79.0</u>	94.5	80.2	70.0	47.0	74.1
	Qwen3-8B	<u>74.6</u>	89.1	77.5	76.7	46.1	72.8
	Qwen3-4B	69.8	83.6	74.4	76.7	44.5	69.8
	Qwen3-1.7B	54.6	74.5	64.2	53.3	29.6	55.2
	Qwen3-0.6B	37.8	45.5	45.9	6.7	13.3	29.8
<i>Non-thinking Mode</i>	GPT-4o-2024-1120	67.6	98.2	<u>80.7</u>	13.3	15.2	55.0
	Gemma-3-27b-IT	<u>74.6</u>	90.9	<u>78.4</u>	23.3	20.5	57.5
	Qwen2.5-72B-Instruct	67.2	<u>94.5</u>	<u>80.7</u>	16.7	16.7	55.2
	Qwen3-235B-A22B	72.9	92.7	82.6	33.3	28.6	62.0
	Qwen3-32B	71.4	92.7	79.5	<u>30.0</u>	23.0	59.3
	Qwen3-30B-A3B	73.9	87.3	<u>77.7</u>	33.3	24.8	<u>59.4</u>
	Qwen3-14B	75.8	89.1	75.7	26.7	<u>27.6</u>	59.0
	Qwen3-8B	72.1	85.5	72.9	13.3	<u>23.8</u>	53.5
	Qwen3-4B	63.0	78.2	67.8	23.3	19.3	50.3
	Qwen3-1.7B	46.1	70.9	53.4	6.7	11.9	37.8
	Qwen3-0.6B	35.1	43.6	39.0	0.0	4.5	24.4