

Table 1: Baseline accuracy (%) of Qwen3 models on math competition benchmarks. pass@1 uses greedy decoding (temperature = 0); avg@16 uses temperature = 1.2, top_p = 0.95, 16 samples per problem.

Model	AIME 2024		AIME 2025		HMMT 2025		AMO-Bench	
	pass@1	avg@16	pass@1	avg@16	pass@1	avg@16	pass@1	avg@16
Qwen3-1.7B	—	—	—	—	—	—	—	—
Qwen3-4B	—	—	—	—	—	—	—	—
Qwen3-8B	—	—	—	—	—	—	—	—