Model	AIME 2024 (Pass@1)	MATH-500 (Pass@1)	GPQA Diamond	MMLU
DeepSeek-R1	**79.8%**	**97.3%**	**71.5%**	90.8%
OpenAI-o1-1217	79.2%	96.4%	72.6%	**91.8%**
DeepSeek-R1-32B	72.6%	94.3%	59.1%	87.4%
OpenAI-o3-mini	80.0%	-	77.0%	-
OpenAI-o1-mini	63.6%	90.0%	60.0%	85.2%
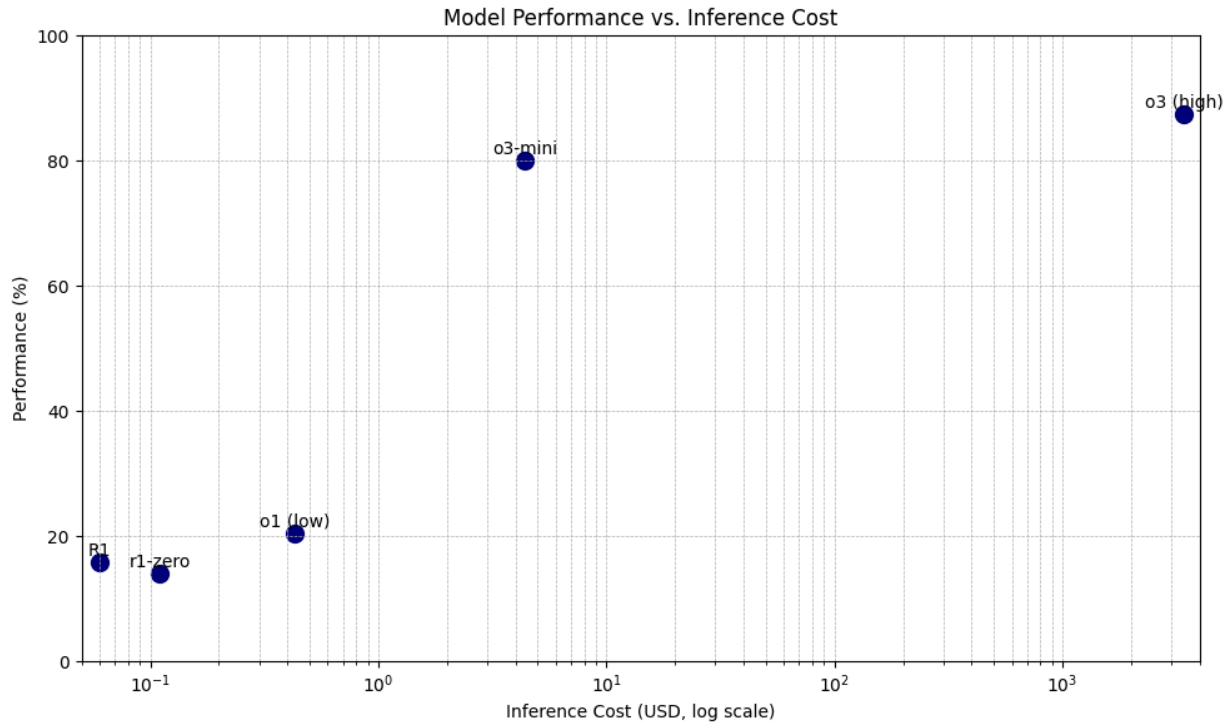
Recommendations for Turing Labs:

Given Turing Labs' business focus on deploying reliable AI agents for businesses:

- Deploy **DeepSeek-R1** as the main reasoning engine in use-cases focused purely on complex reasoning tasks, where language readability and safety concerns are secondary.
- For agent deployments involving sensitive or safety-critical applications, prefer **OpenAI-o3-mini**, acknowledging its superior risk mitigation and robustness in diverse scenarios.