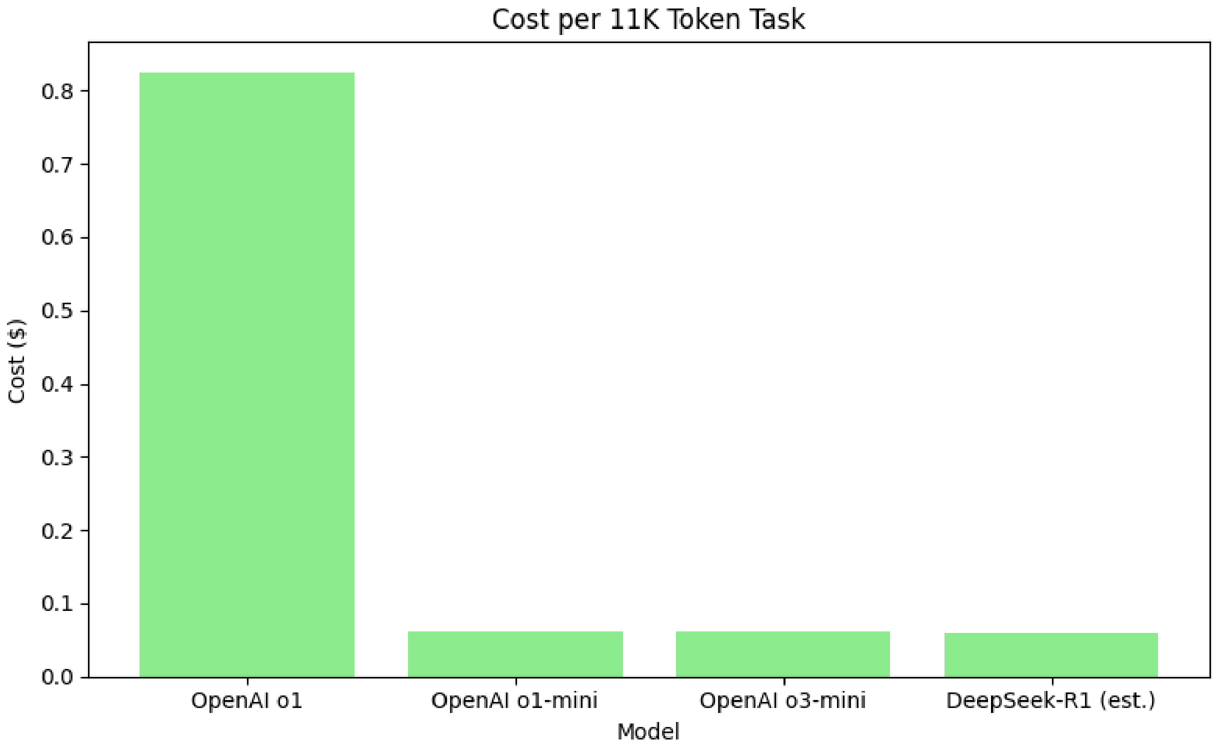
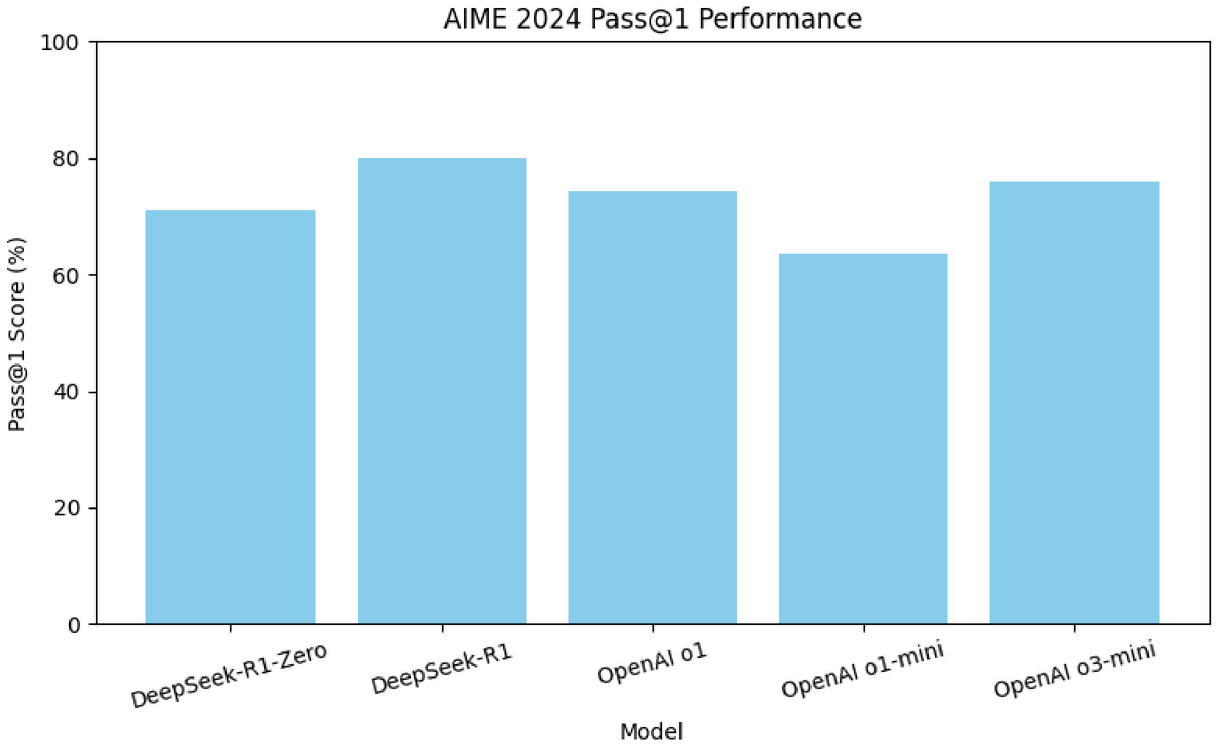
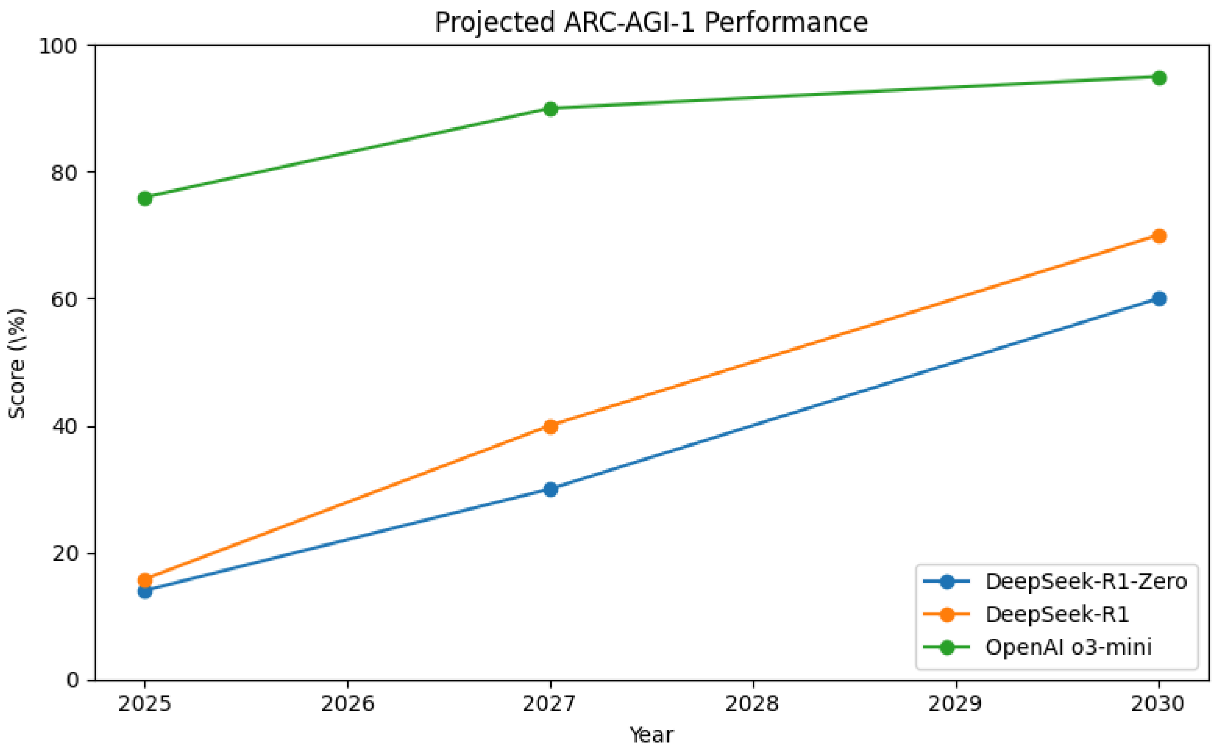

Cost per 11K Token Task
Cost ($)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
OpenAI o1
OpenAI o1-mini
OpenAI o3-mini
DeepSeek-R1 (est.)
Model


AIME 2024 Pass@1 Performance
Pass@1 Score (%)
0
20
40
60
80
100
DeepSeek-R1-Zero
DeepSeek-R1
OpenAI o1
OpenAI o1-mini
OpenAI o3-mini
Model



3. Recommendations for Turing Labs

For our AI agent product, we prioritize **cost-effectiveness**, **reliability**, and **business applicability** (e.g., task automation, customer support, data analysis).

1. **Primary Recommendation: OpenAI o3-mini**
   - **Justification**: Balances high reasoning performance (76% on ARC-AGI-1), reliability (agentic task success), and cost ($0.0605/task). Its safety mitigations and multilingual capabilities suit diverse business clients. The low cost aligns with scaling inference demand noted in ARC Prize.
   - **Use Case**: Complex automation (e.g., SWE-Bench tasks, 61% success with tools) and customer-facing QA.
2. **Secondary Recommendation: DeepSeek-R1**
   - **Justification**: Open-source flexibility and near-o1 performance (79.8% AIME) make it ideal for customization. Lower hosting costs offset initial setup effort. Suitable for businesses needing tailored solutions.
   - **Use Case**: Custom workflows and domains with verifiable outcomes (math, coding).
3. **Avoid: DeepSeek-R1-Zero and OpenAI o1**
   - **R1-Zero**: Poor readability limits client-facing use despite RL potential.

- **o1**: High cost ($0.825/task) outweighs marginal performance gains over o3-mini.