Cost per Interaction (1000 Input + 1000 Output Tokens)

Performance on AIME Benchmark

3. Recommendations for Turing Labs

Model Selection

Based on the analysis, I recommend a tiered approach for Turing Labs' AI agent product:

- **Primary Recommendation: o3-mini**
  - **Justification**: Offers a strong balance of performance (80% AIME, 2036 Codeforces) and cost ($0.0055 per interaction). Its coding prowess and multilingual capabilities (e.g., 0.8070 MMLU Arabic) suit businesses needing automation, software development, or global support. Availability is assumed based on its pricing listing (o3-mini-2025-01-31).
- **Cost-Effective Option: GPT-4o-mini**
  - **Justification**: At $0.00075 per interaction, it's ideal for simple tasks (Q&A, basic automation) where high reasoning isn't required, despite lower performance (10% AIME).
- **High-End Option: o1**
  - **Justification**: For premium clients needing top-tier reasoning (78% AIME), though its $0.075 cost limits scalability unless justified by significant value-add.
- **Exploratory Option: DeepSeek R1**
  - **Justification**: If Turing Labs can deploy R1 internally, its 79.8% AIME and 2029 Codeforces performance rival o3-mini. However, lack of full open-source weights and support infrastructure makes it riskier.