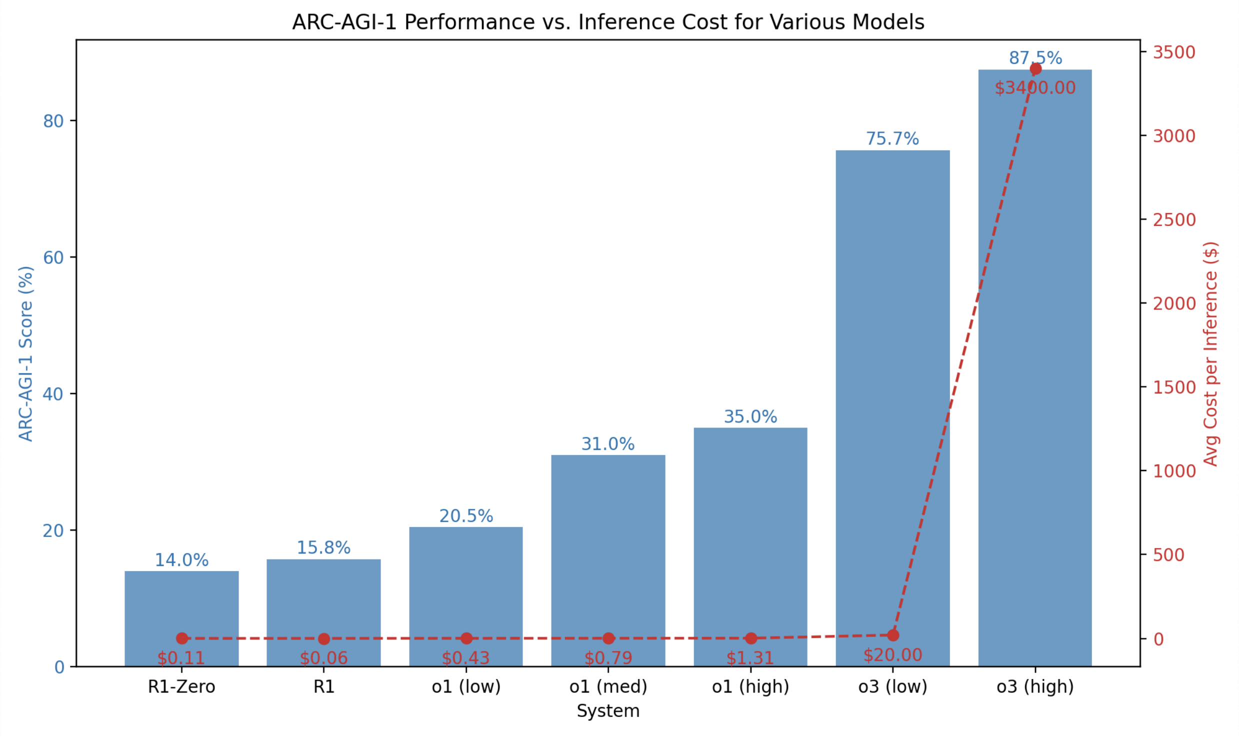
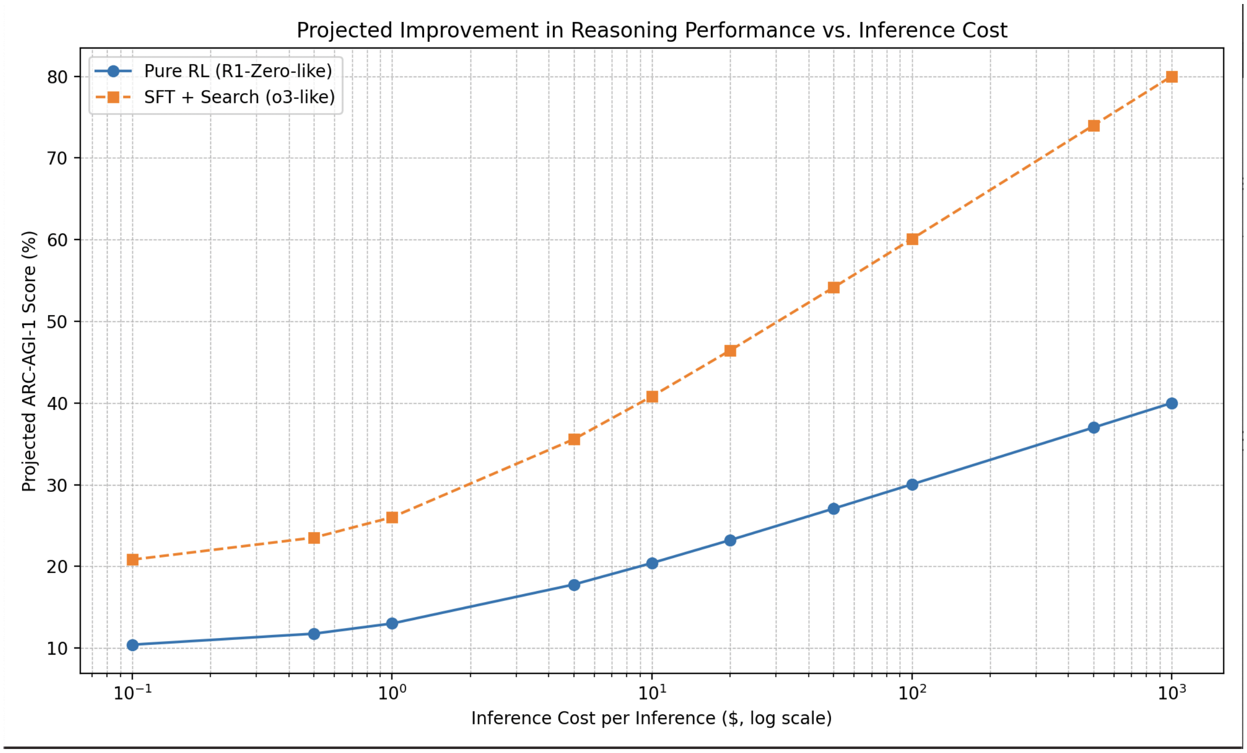
O3 Mini High





Based on our analysis, we propose a **hybrid deployment strategy**:

1. **For Cost-Sensitive Applications:**
    - **Deploy DeepSeek-R1 or R1-Zero:**
      These models are very economical and perform sufficiently well in domains with clear verification (e.g., math and coding tasks). They are especially attractive when the AI agent is used for preliminary reasoning, where the output can be later validated by a secondary system.
2. **For High-Stakes or Novel Problem Domains:**
    - **Adopt OpenAI o3-based Systems:**
      Although the cost per inference is higher, the dramatic improvement in performance (75–87% ARC-AGI-1) and robustness to novel queries makes these systems ideal for applications where reliability is paramount. For example, AI agents that must autonomously adapt in critical business workflows or support real-time decision making would benefit from this level of performance.
3. **Long-Term Strategy:**
    - **Monitor the Evolution of Pure RL Systems:**
      As research continues, future generations (e.g., a potential “R2-Zero”) may reduce the need for SFT while still achieving high performance. Turing Labs should continue to evaluate these developments as they could offer a breakthrough in lowering inference costs while maintaining high reliability.
    - **Hybrid Inference Pipelines:**
      Consider designing a system that uses a lightweight R1/R1-Zero model for initial inference and a fallback to an o3-like model for cases where additional reasoning and verification are required. This strategy could optimize costs while ensuring reliability.