Test-Time Compute Scaling: More Thinking = Better Answers

Accuracy on Hard Math Problems (%) vs Inference Tokens (log scale)

Extended reasoning enables harder problems

Diminishing returns

- GPT-4 (fixed reasoning)
- o1-mini
- o1