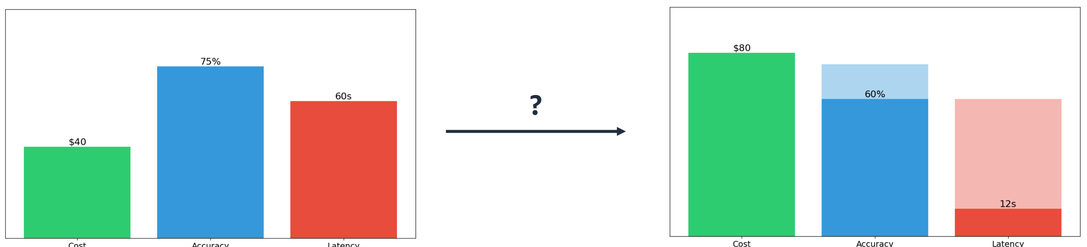
Finding the Sweet Spot: Trading Quality, Cost & Speed During Inference-Time LLM Reflection

Jack Butler, AWS, The United Kingdom
Nikita Kozodoi, AWS, Germany
Zainab Afolabi, AWS, The United Kingdom
Brian Tyacke, Zalando, Germany
Gaiar Baimuratov, Zalando, Germany

Helping customers accelerate MORE
Generative AI Innovation Center

Motivation

- Frequently we want to choose **the best combination** of desired task accuracy, latency budget and cost
- However, it **not always intuitive** how to allocate inference budget, especially across different domains
- **Inference-time compute** allows a direct trade-off of accuracy / cost / latency

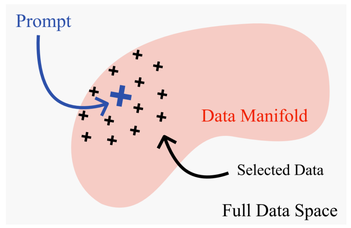


Inference-Time Compute

Training-based

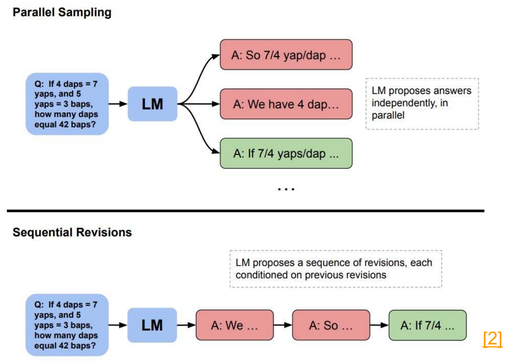
The Surprising Effectiveness of Test-Time Training for Abstract Reasoning

OpenAI o1



[1]

Sampling-based



[2]

Methods & Data

Language models:

amazon Nova	ANTHROPIC CLAUDE
Premier	Sonnet 3.7
Pro	Sonnet 3.5 v2
Micro	Haiku 3.5
Lite	

Accessed via Amazon Bedrock

Our **open-source library** implementing self-reflection

Datasets:

Flores-200

*200 samples across 15 language pairs

IMDB Reviews

*100 random samples

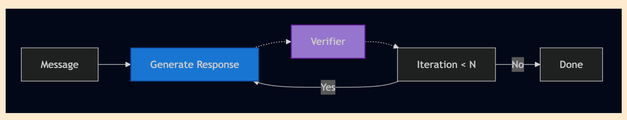
Spider 1.0

*5 sampled DBs

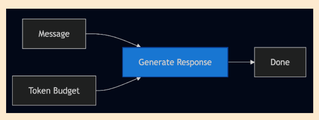
Math500

*100 random samples

Inference strategies:



Model-agnostic self-reflection

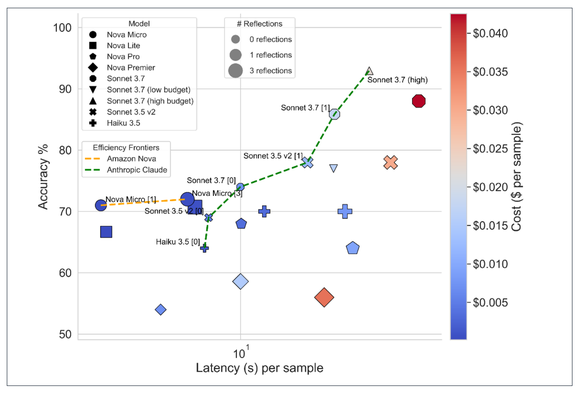


Claude's budget tuning

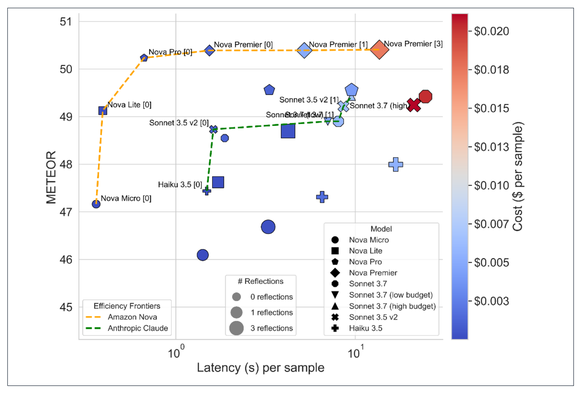
Overall Results

Task	Best-Performing Model	
	🎯 Pure Accuracy	📈 Largest Improvement
Math	**Claude 3.7 Sonnet:** 88% *(3 reflections)*	**Nova Micro:** 22% => 71% *(~220%, 1 reflection)*
Text-to-SQL	**Nova Lite:** 74% *(1 reflection)*	**Claude 3.7 Sonnet:** 67.5% => 71% *(~5%, 3 reflections)*
Translation	**Nova Premier:** 50.4 METEOR *(0 reflections)*	**Claude 3.7 Sonnet:** 48.5 => 49.4 METEOR *(~2%, 3 reflections)*
Sentiment Classification	**Claude 3.7 Sonnet:** 97% *(1 reflection)*	**Nova Micro:** 85% => 95% *(~12%, 1 reflection)*

1. Pareto Frontiers

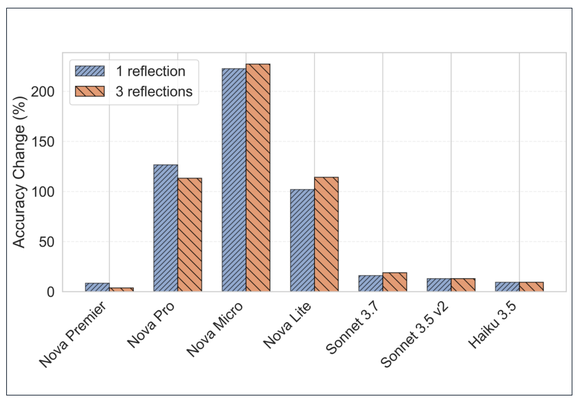


Math500

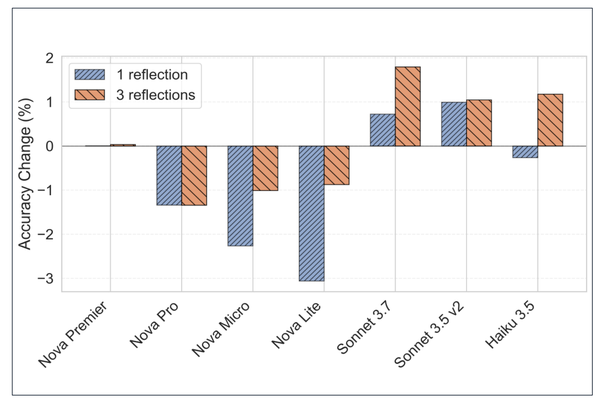


Flores-200

2. Self-Reflection Rounds



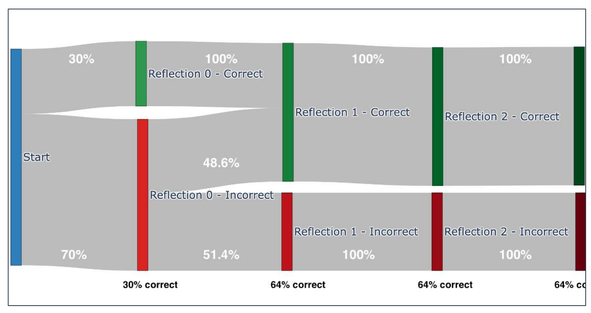
Math500



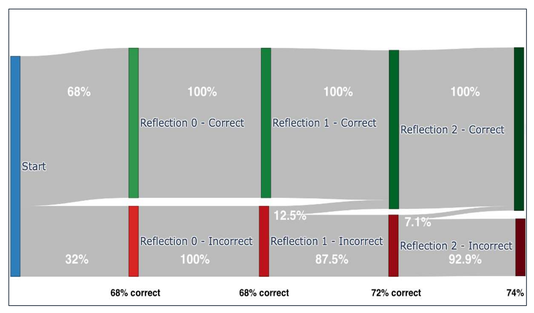
Flores-200

💡 **Different models benefit from different number of reflections, which varies across datasets**

3. Transition Dynamics



Nova Micro



Claude 3.5 Sonnet v2

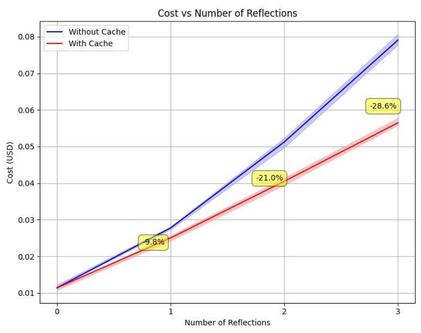
💡 **There are distinct correction patterns during consecutive self-reflection rounds depending on the LLM**

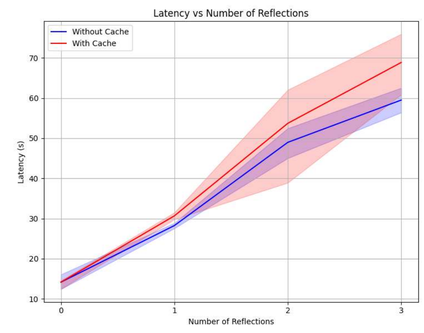
4. Feedback Mechanisms

Model	No feedback		LLM judge feedback		SQL execution feedback	
	1 round	3 rounds	1 round	3 rounds	1 round	3 rounds
Amazon Nova Premier	72.58	**74.98**	73.97	72.58	73.74	71.14
Amazon Nova Pro	71.75	**73.67**	71.71	66.96	68.62	73.50
Amazon Nova Lite	75.41	73.05	**79.57**	74.02	72.63	72.83
Amazon Nova Micro	70.73	72.14	**77.34**	75.77	73.15	70.41
Claude Sonnet 3.7	70.78	72.69	70.82	66.78	67.20	**73.32**
Claude Sonnet 3.5 v2	65.71	64.99	67.28	65.43	67.22	**67.33**
Claude Haiku 3.5	67.09	66.36	68.16	68.64	68.56	**72.58**

💡 **Providing feedback as context between reflection rounds helps to improve the accuracy**

5. Prompt Caching





💡 **Prompt caching uniquely benefits sampling-based reflection, caching past inference iterations**

Corresponding author: jackbtlr@amazon.co.uk