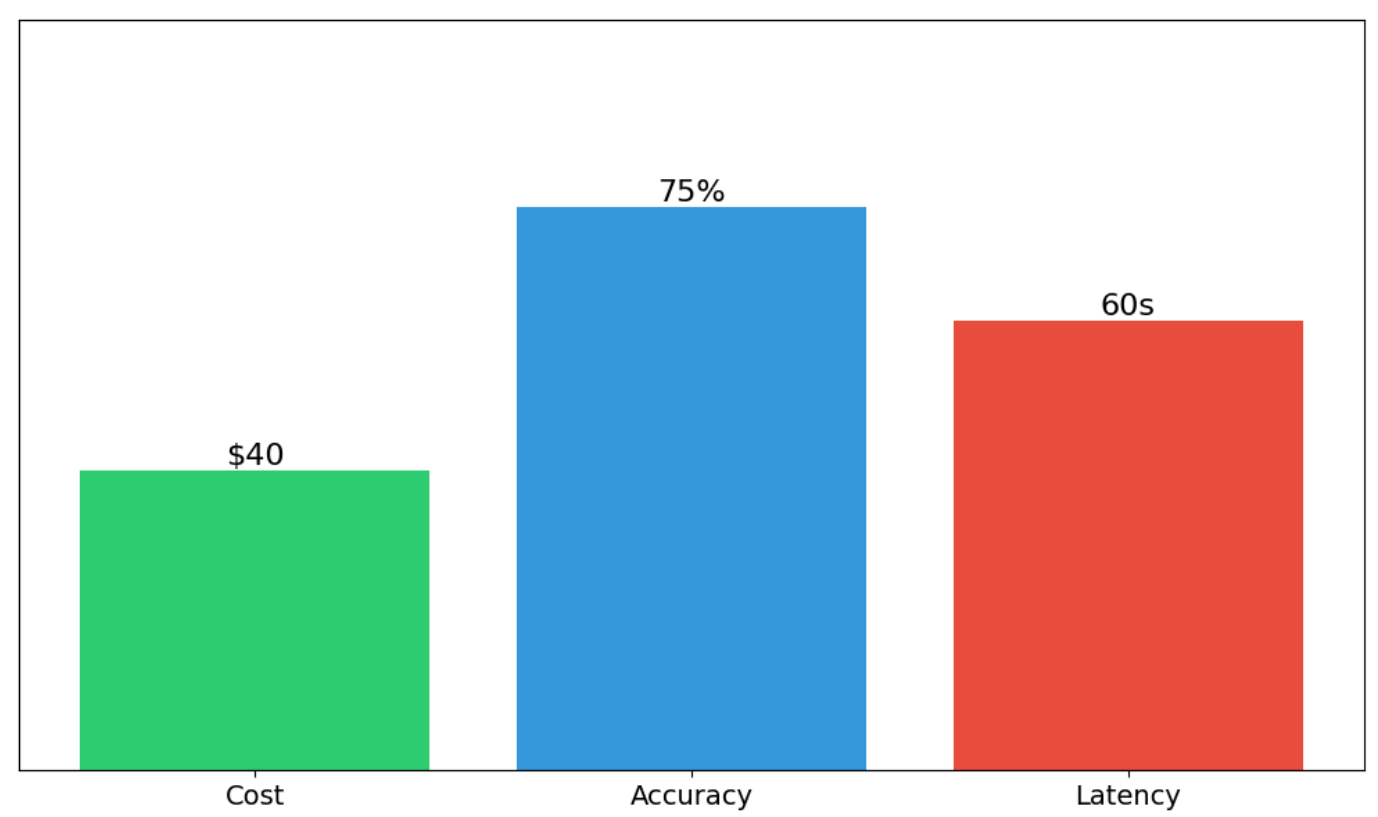
Finding the Sweet Spot: Trading Quality, Cost, and Speed During Inference-Time LLM Reflection

Agentic & GenAI Evaluation KDD 2025

Motivation

- Practitioners often need to choose the best combination of desired task accuracy, latency budget and cost for their use case.



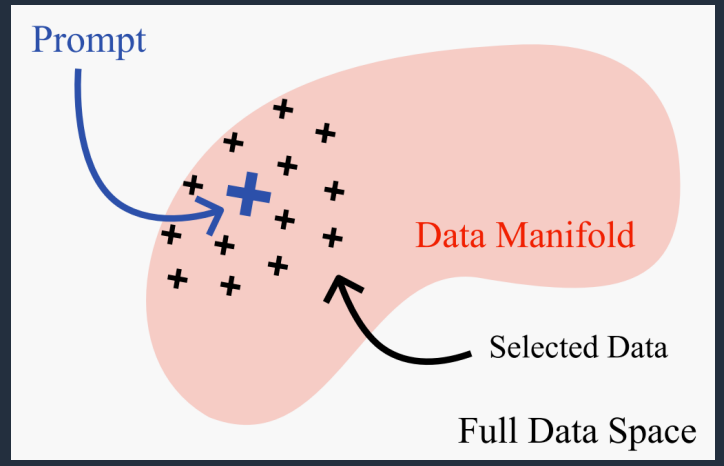
Motivation

➢ *Inference Time Compute* allows a direct trade-off of accuracy / cost / latency.

Training-based

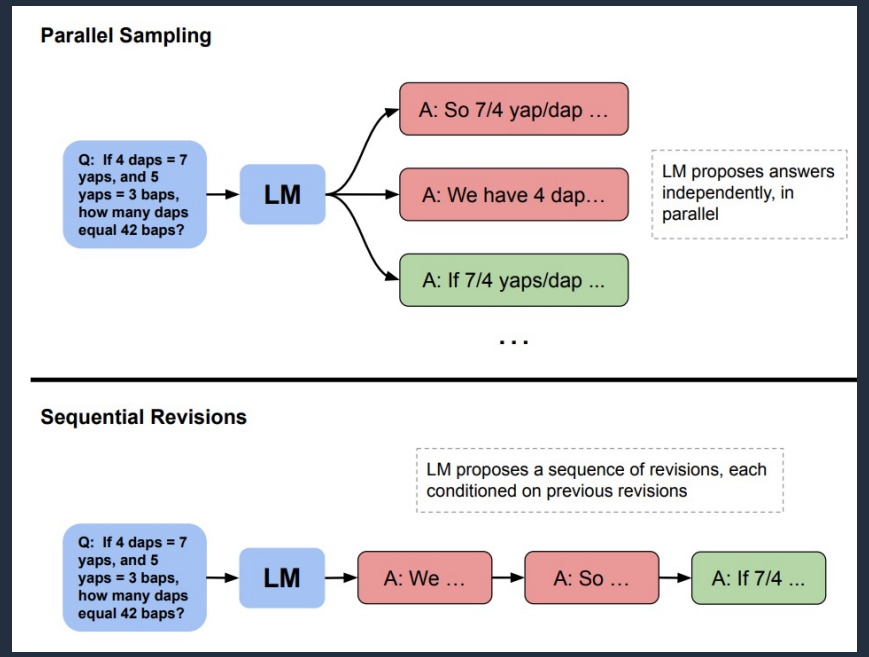
The Surprising Effectiveness of Test-Time Training for Abstract Reasoning

OpenAI o1



[1]

Sampling-based



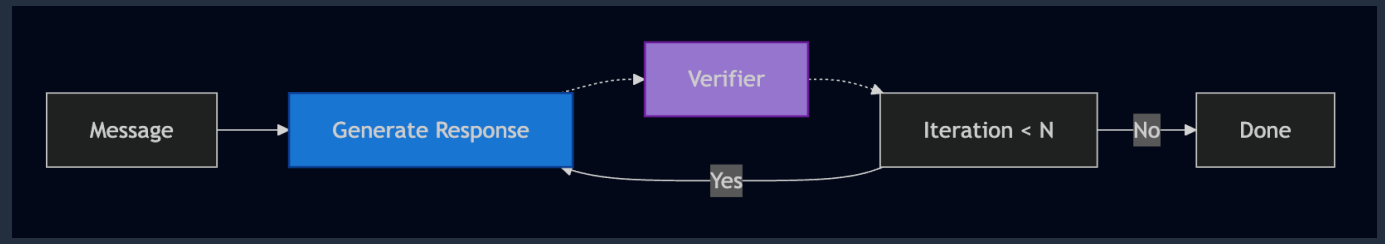
[2]

Experiments

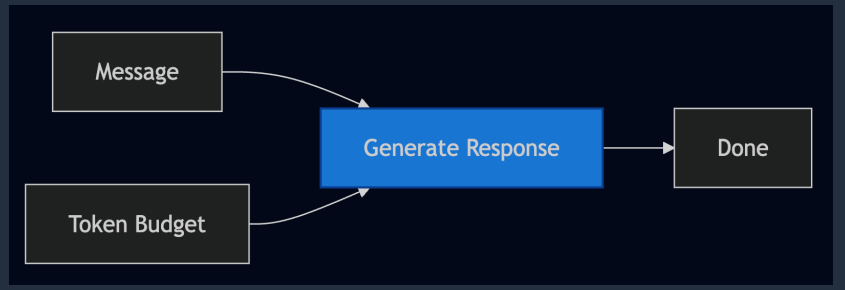
- *We compared the following models*

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

- *Leveraging self-reflection for all models and "budget tuning" for Claude 3.7 Sonnet*



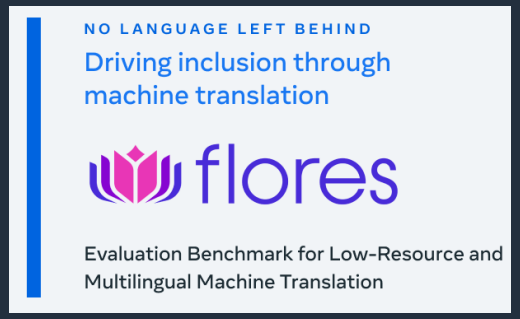
Self-reflection



Budget tuning

Datasets & Evaluation

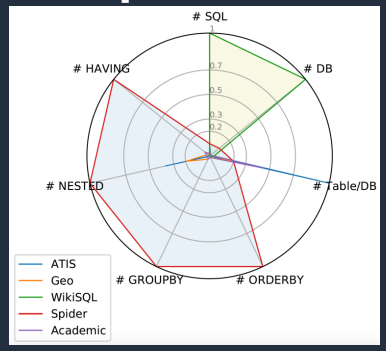
Flores 200



- **METEOR score**

*200 examples across 15 language pairs

Spider 1.0



- **Results accuracy** of normalised SQL results on row and cell level

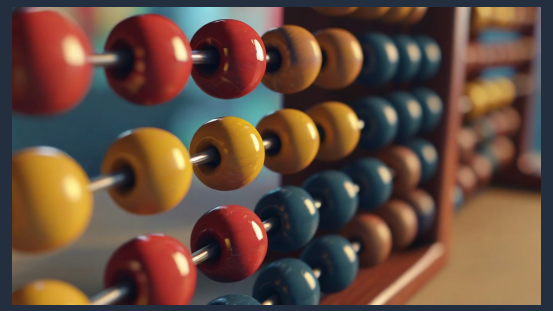
*5 sampled databases

IMDB Reviews

- **Classification accuracy**

*100 random samples

Math500



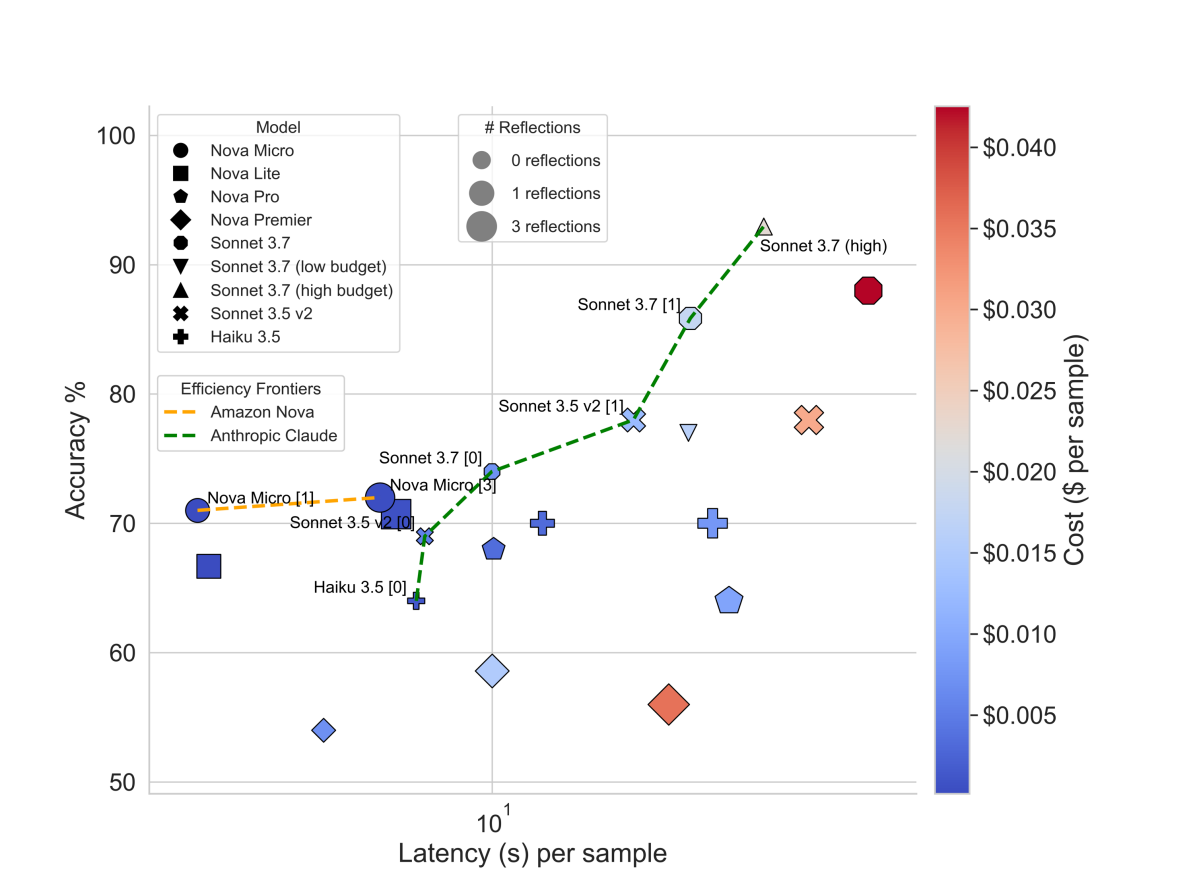
- **String matching** on cleaned LaTeX
- **Symbolic equivalence** checking with SymPy

*100 random samples

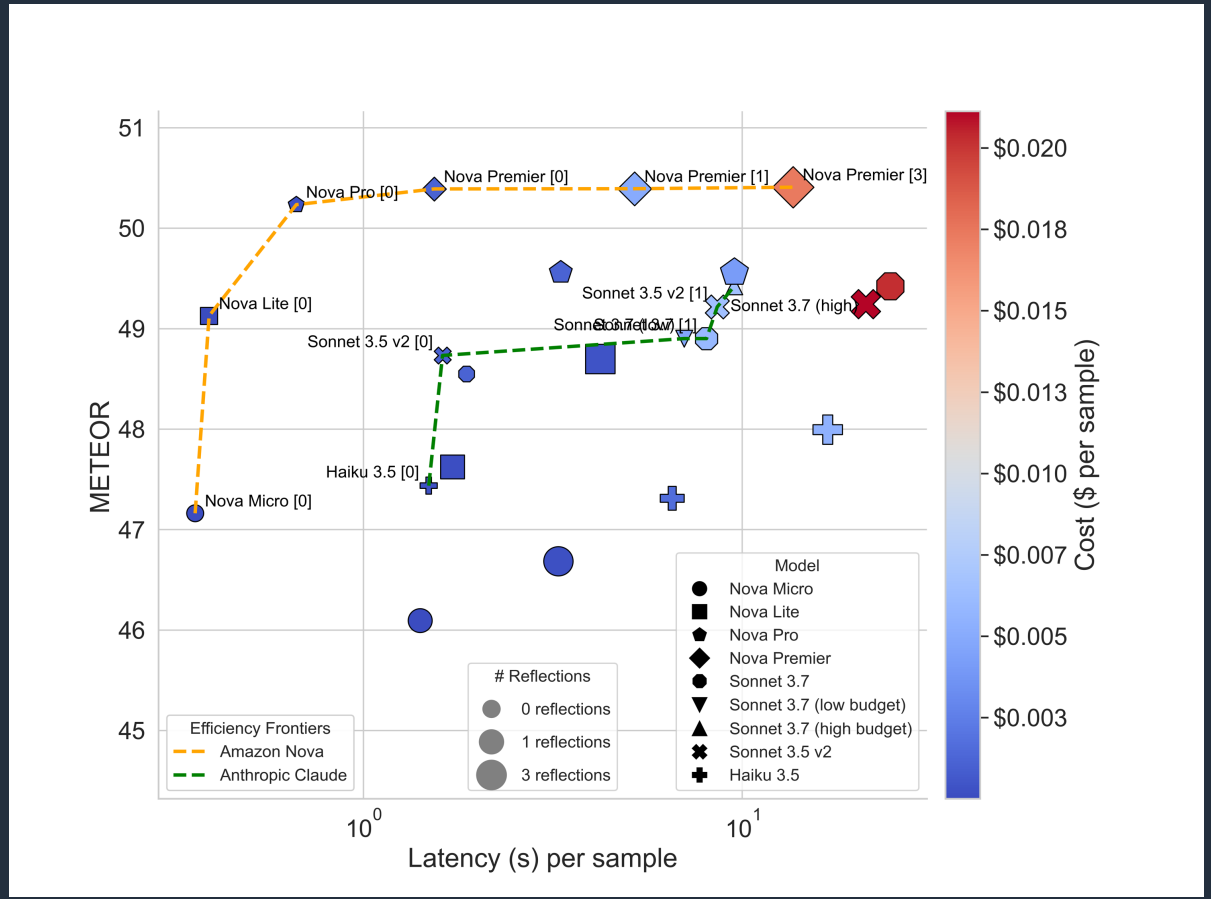
Results: Overall Performance

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)*

Results: Efficient Frontiers

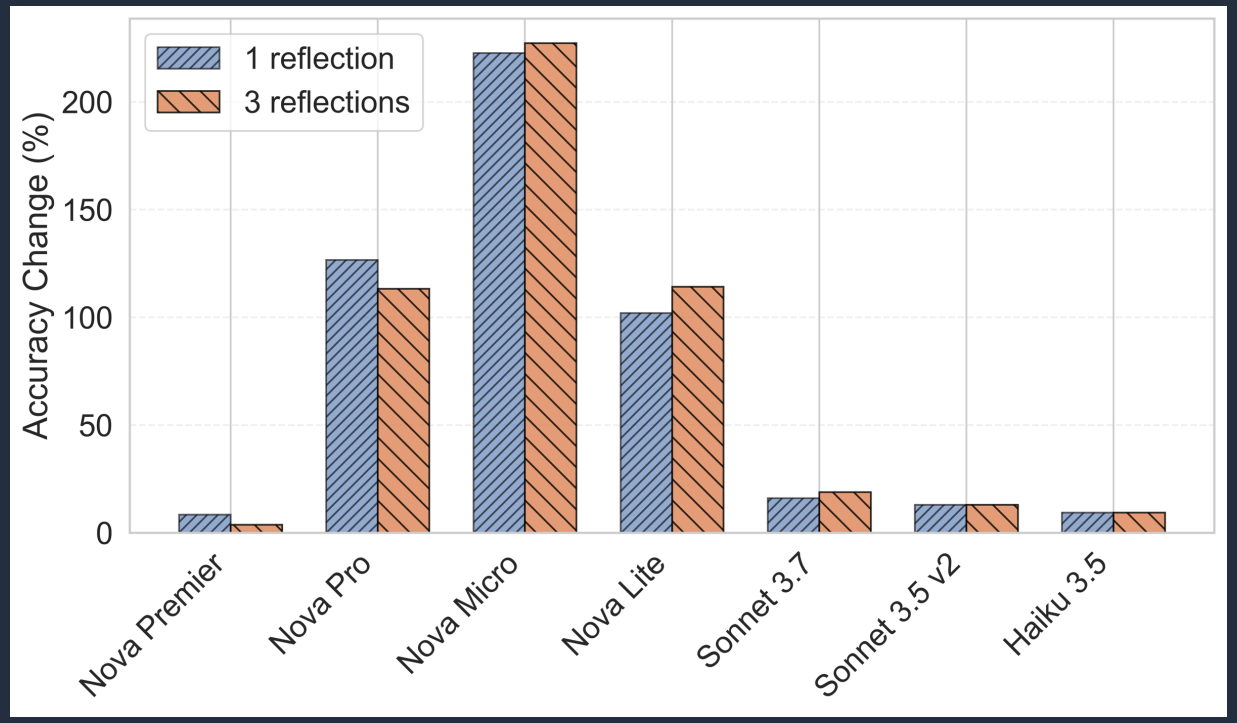


Math

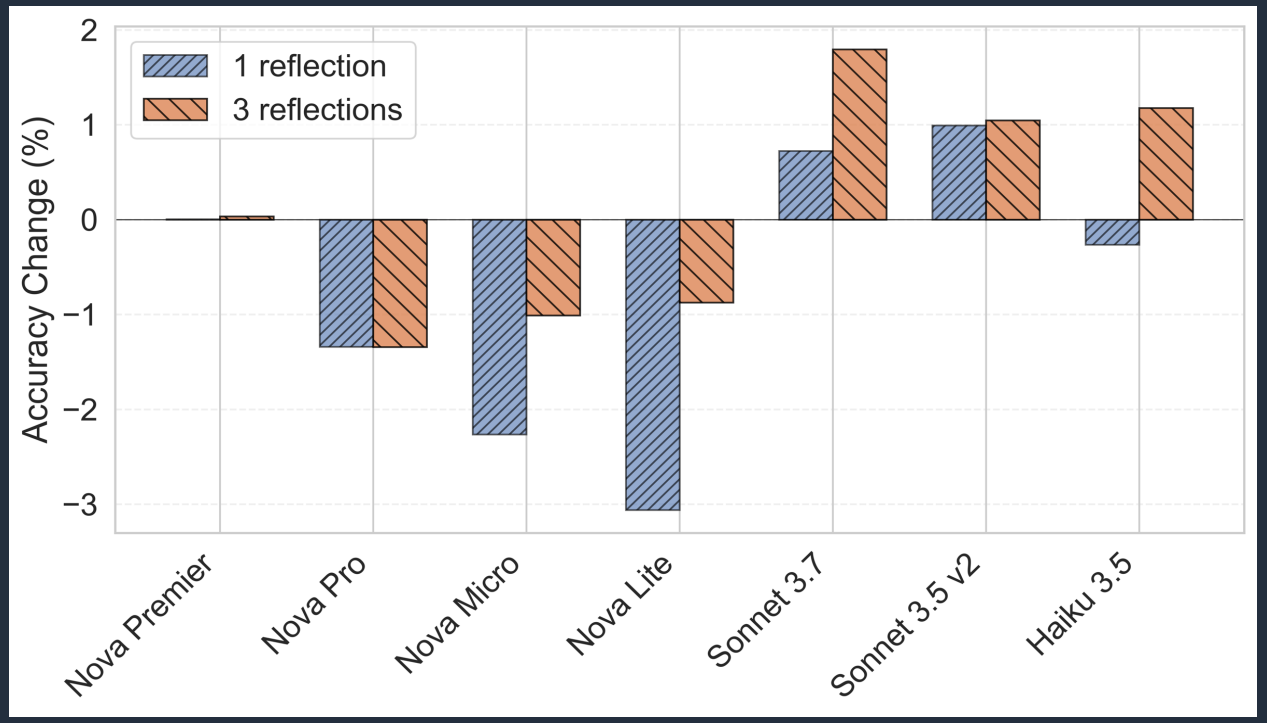


Translation

Results: Number of Self-Reflections



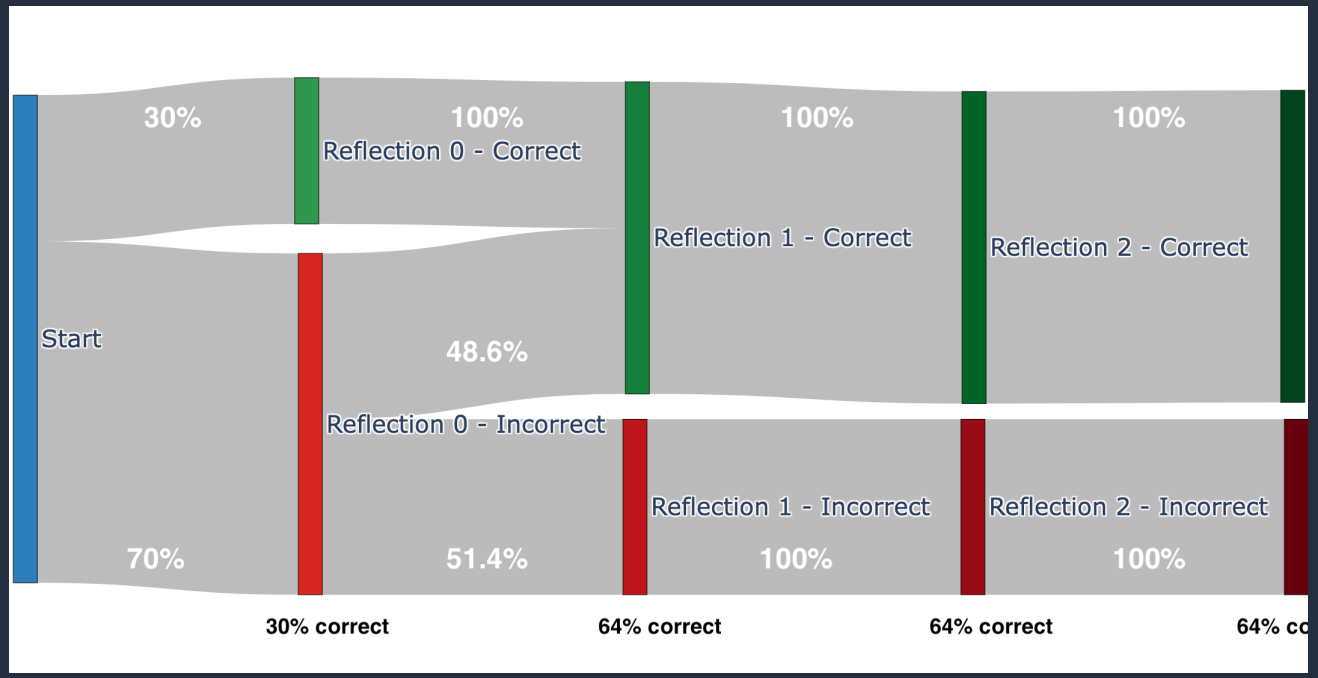
Math



Translation

Distinct reflection dynamics depending on the LLM and the domain

Ablation 1: Transitions Dynamics (Math)



Nova Micro

Claude 3.5 Sonnet v2

Distinct reflection dynamics depending on the LLM and the domain

Ablation 2: Feedback Mechanisms (Text-to-SQL)

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

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**				
Amazon Nova Pro	71.75	**73.67**				
Amazon Nova Lite	75.41	73.05				
Amazon Nova Micro	70.73	72.14				
Claude Sonnet 3.7	70.78	72.69				
Claude Sonnet 3.5 v2	65.71	64.99				
Claude Haiku 3.5	67.09	66.36				

Ablation 2: Feedback Mechanisms (Text-to-SQL)

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

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**			73.74	71.14
Amazon Nova Pro	71.75	**73.67**			68.62	73.50
Amazon Nova Lite	75.41	73.05			72.63	72.83
Amazon Nova Micro	70.73	72.14			73.15	70.41
Claude Sonnet 3.7	70.78	72.69			67.20	**73.32**
Claude Sonnet 3.5 v2	65.71	64.99			67.22	**67.33**
Claude Haiku 3.5	67.09	66.36			68.56	**72.58**

Ablation 2: Feedback Mechanisms (Text-to-SQL)

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

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**

Learnings

1) No one-size-fits-all strategy across domains

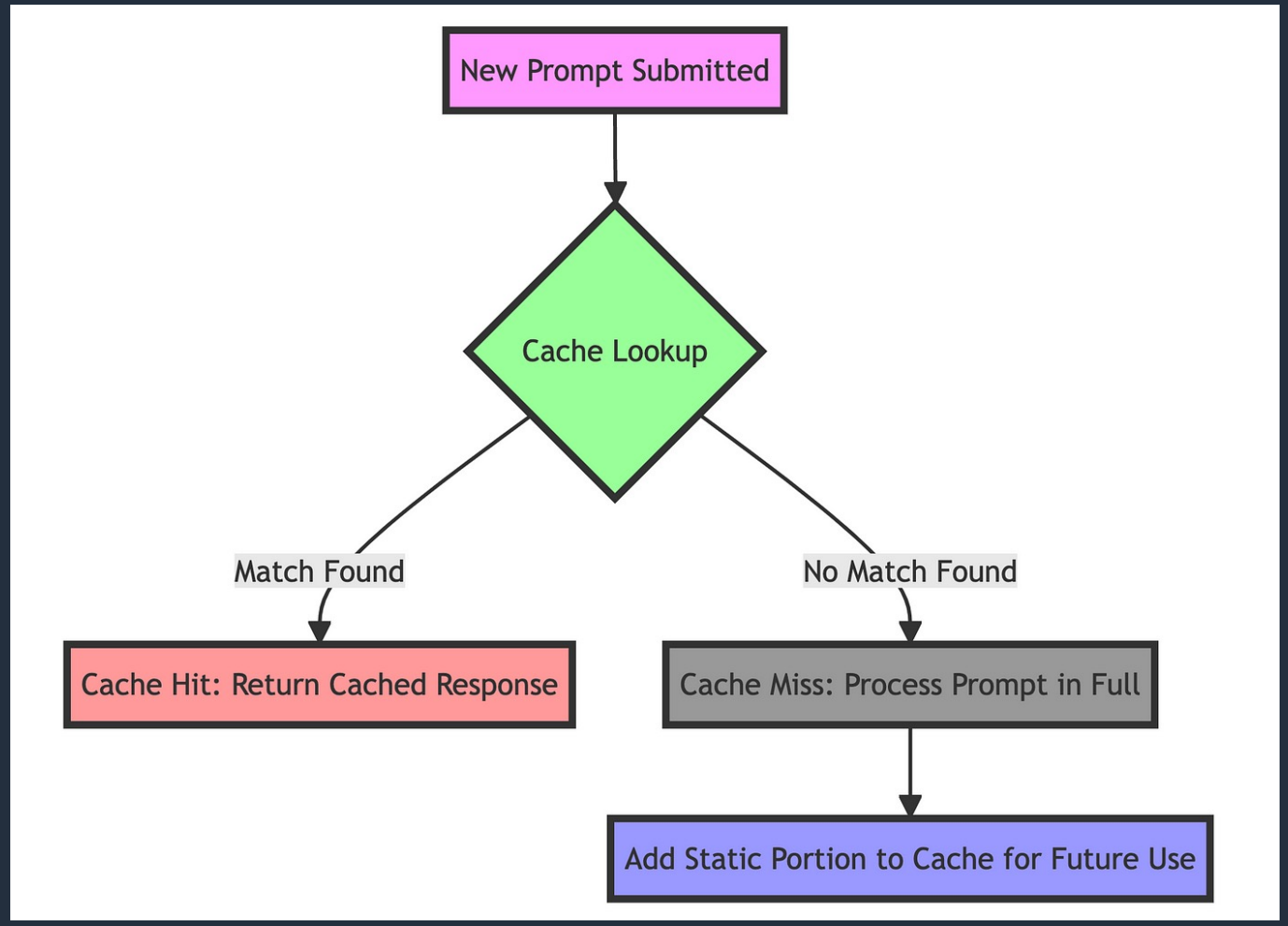
- Large gains in math (+220%) and sentiment analysis
- Mixed/negative results for translation and text-to-SQL

2) Smaller models benefit more and budget tuning < manual reflection

- Small models + reflection can outperform larger models, offering cost savings
- Claude’s built-in reasoning underperforms manual reflection and increases cost

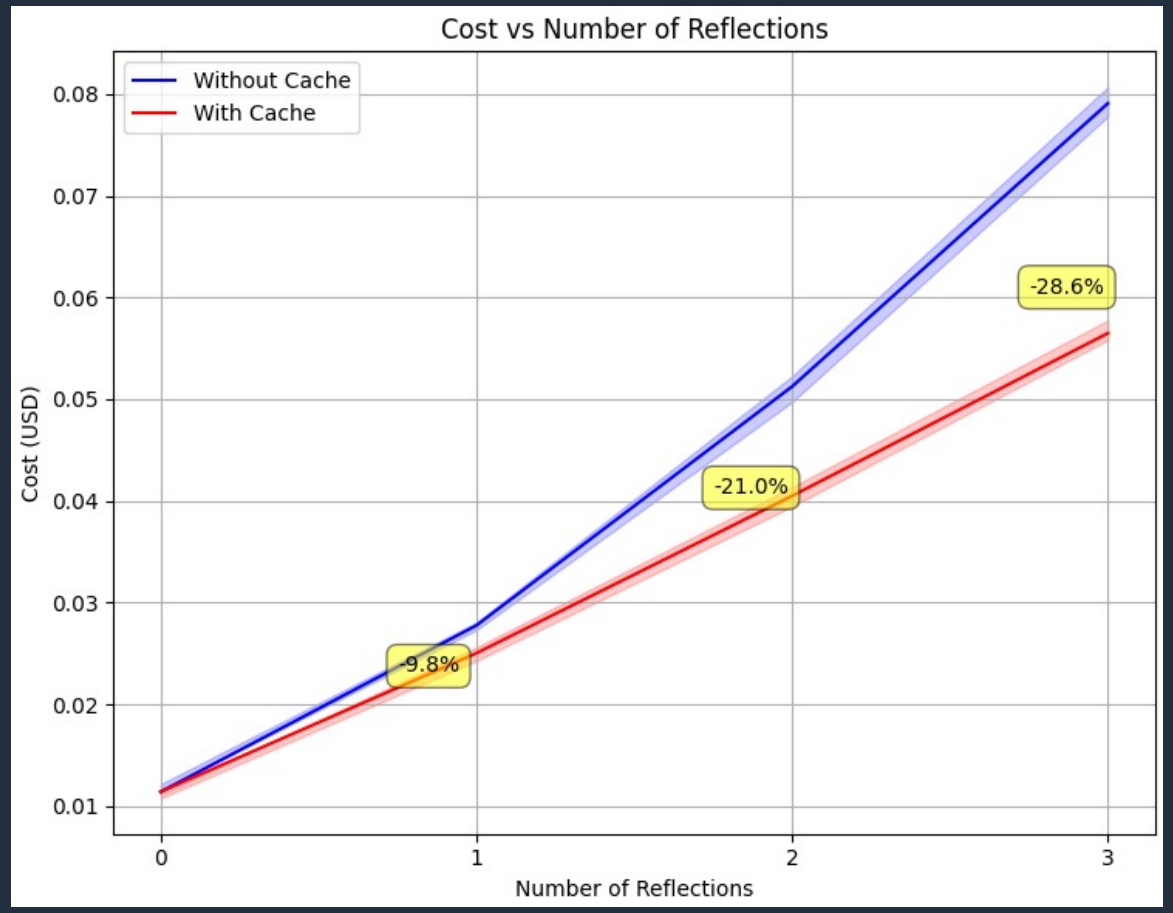
Appendix

Ablation 3: Prompt Caching

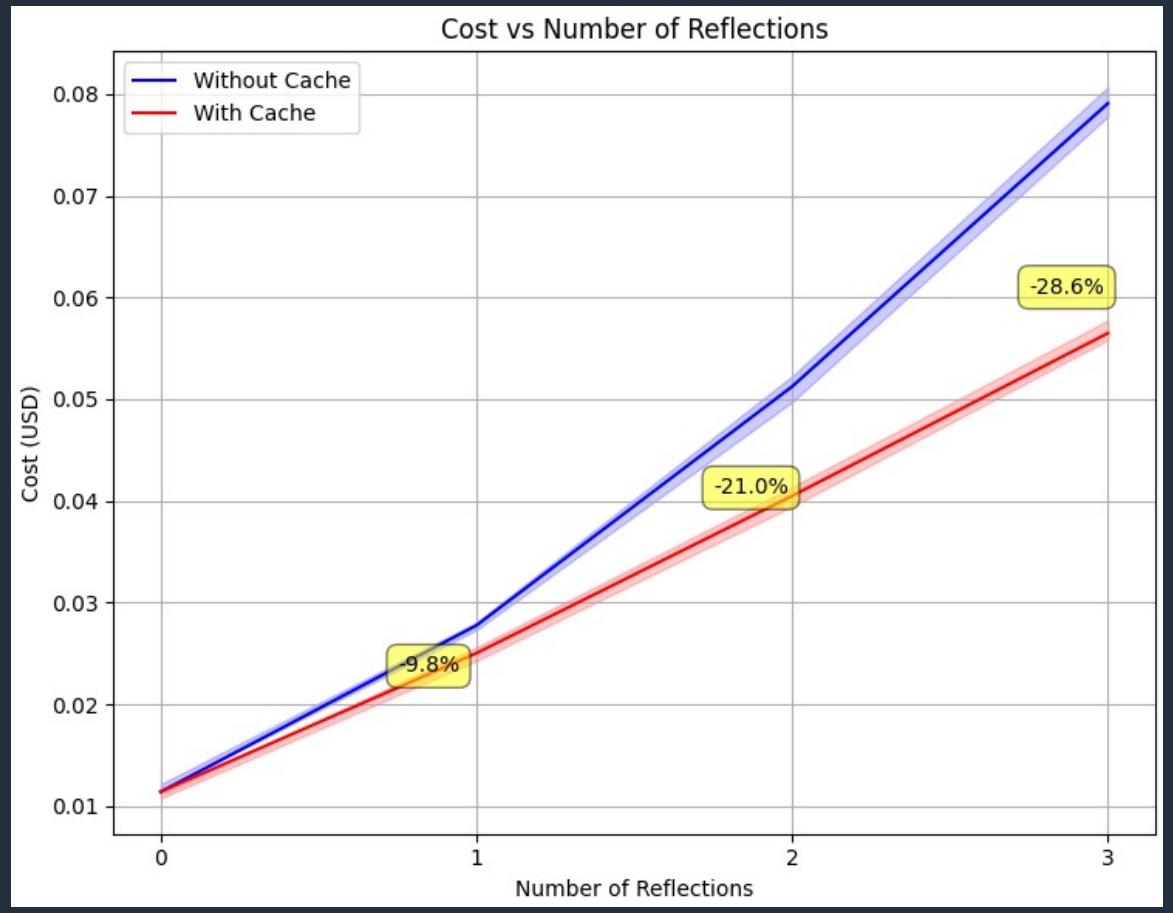


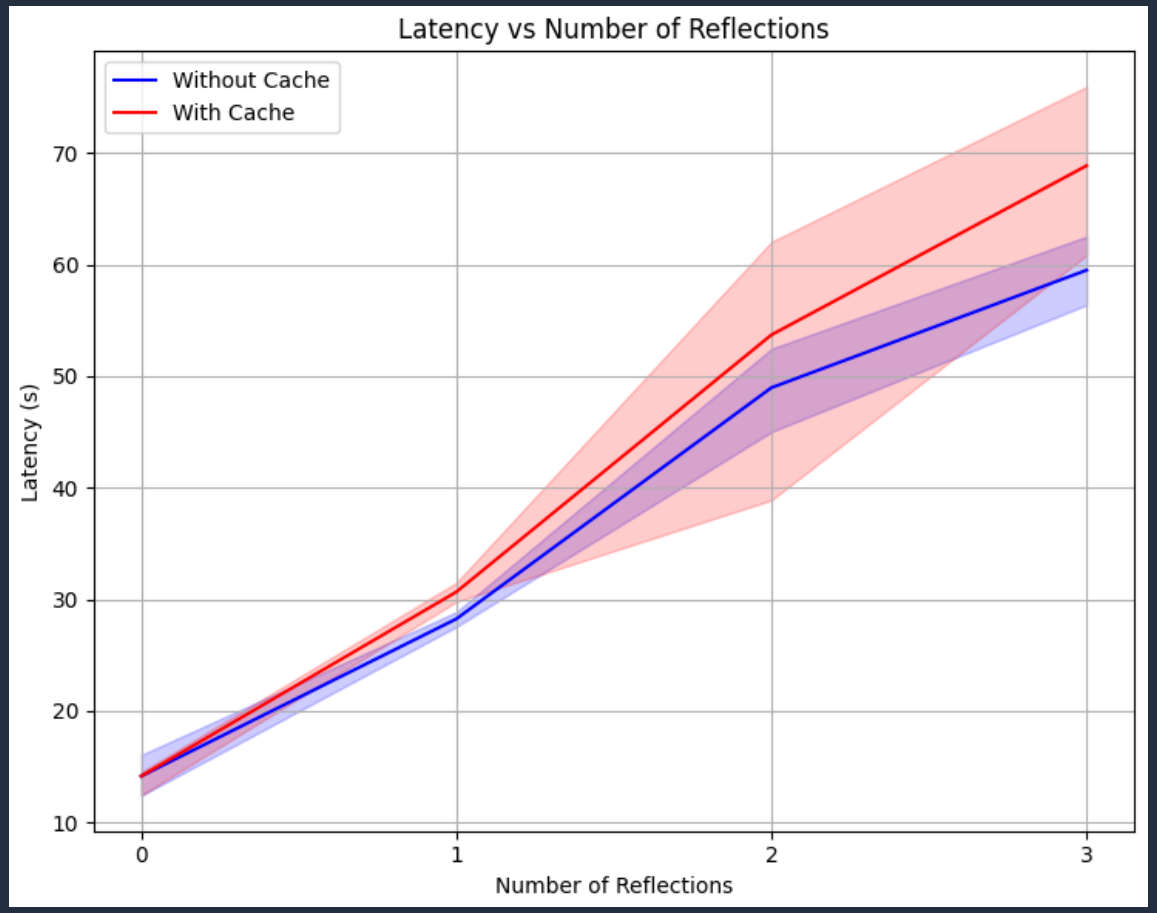
- Iterative self-reflection allows caching the previous part of the prompt
- This brings latency & cost improvements to manual self-reflection
- Built-in reasoning does not benefit from prompt caching

Ablation 3: Prompt Caching



Ablation 3: Prompt Caching

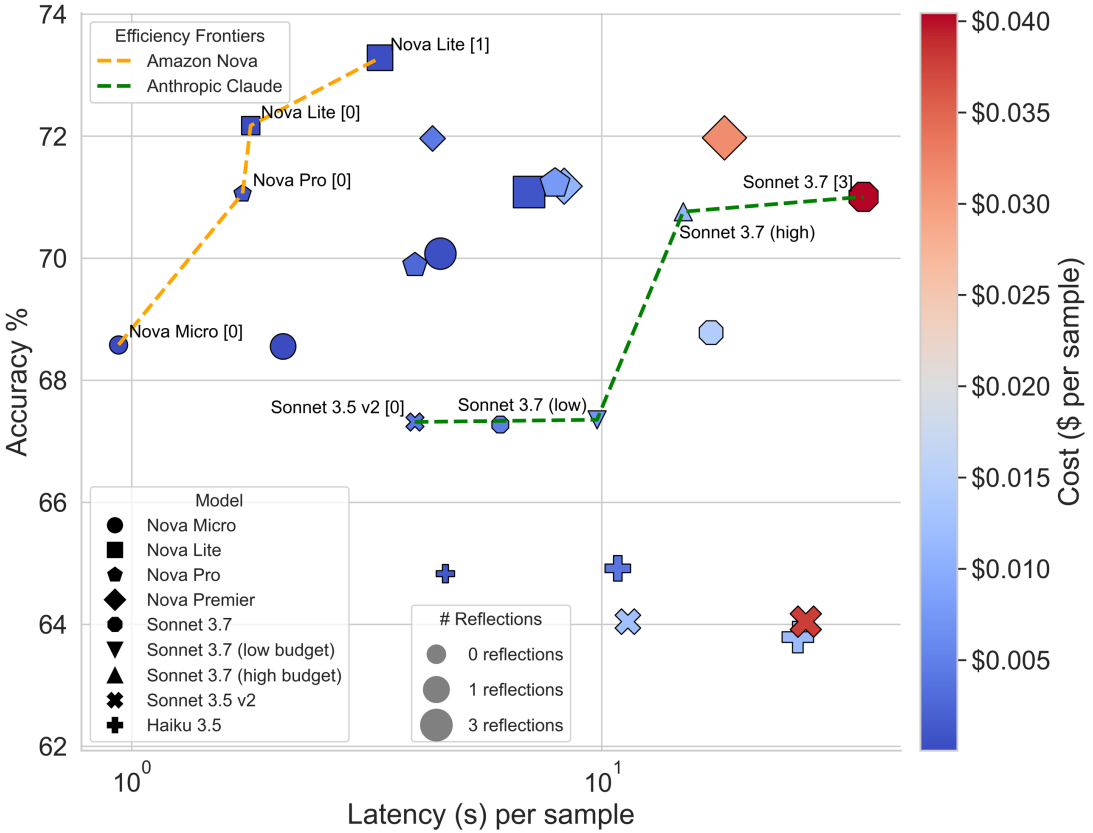




Results: Marketing content localization case study

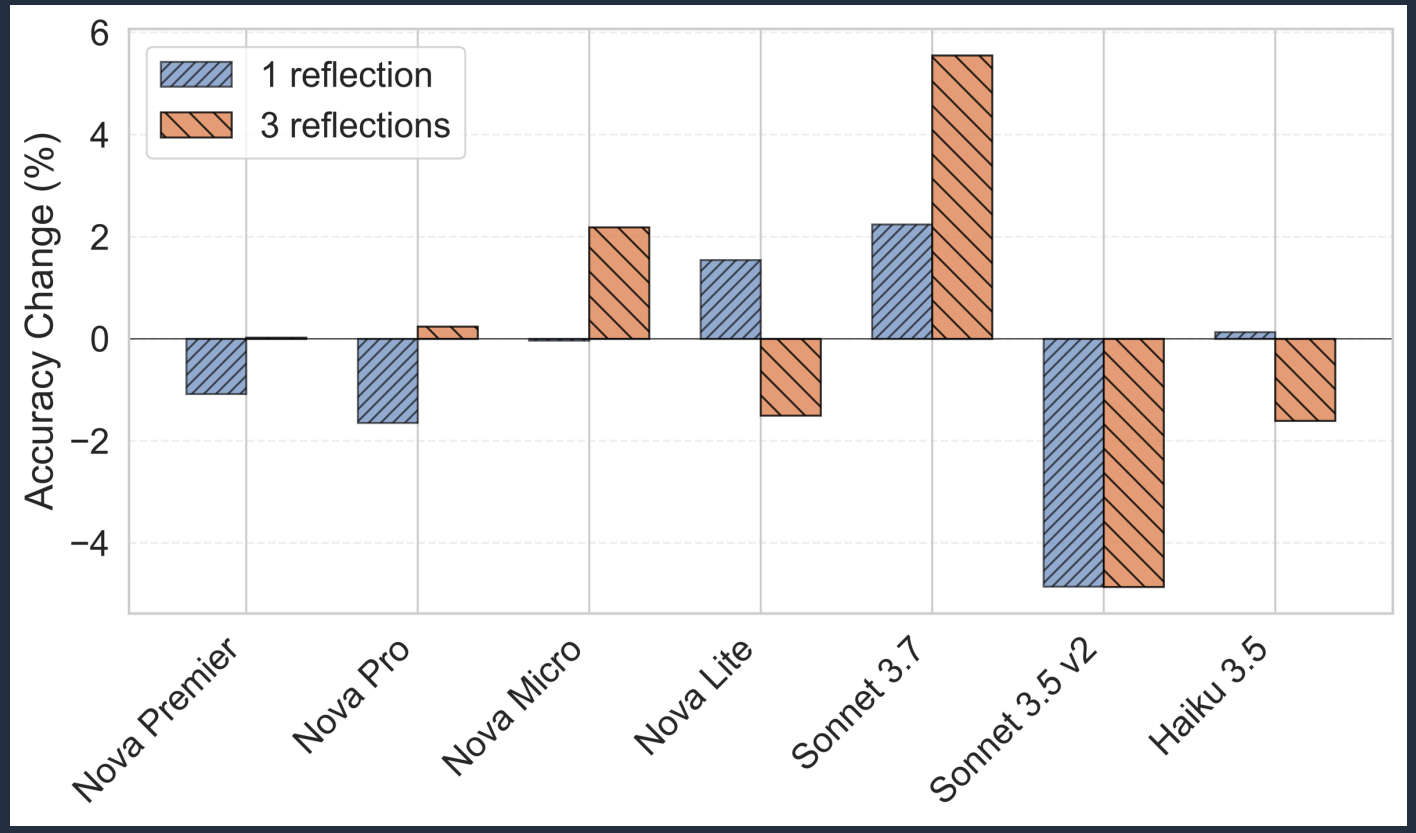
Language	No Reflection			Self-Reflection with LLM Judge Feedback		
	BLEU	METEOR	LLM Judge Score	BLEU	METEOR	LLM Judge Score
German	0.32	0.61	0.38	**0.33**	**0.62**	**0.47**
French	**0.16**	**0.47**	0.61	0.14	0.42	**0.62**
Spanish	0.29	**0.61**	0.49	**0.29**	0.59	**0.50**

Results – texttosql



- Amazon Nova Lite Outperforms all other models
- Nova models are the optimal configuration with respect to latency, cost and accuracy
- …

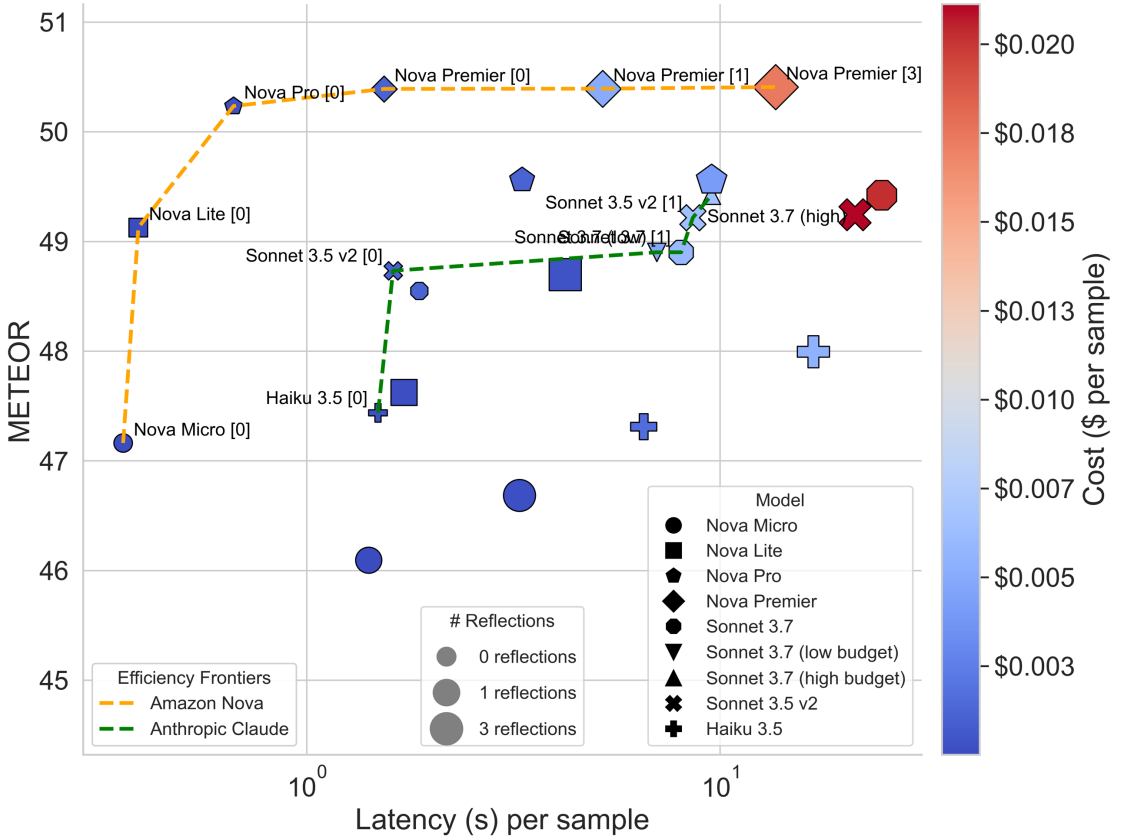
Results – texttosql



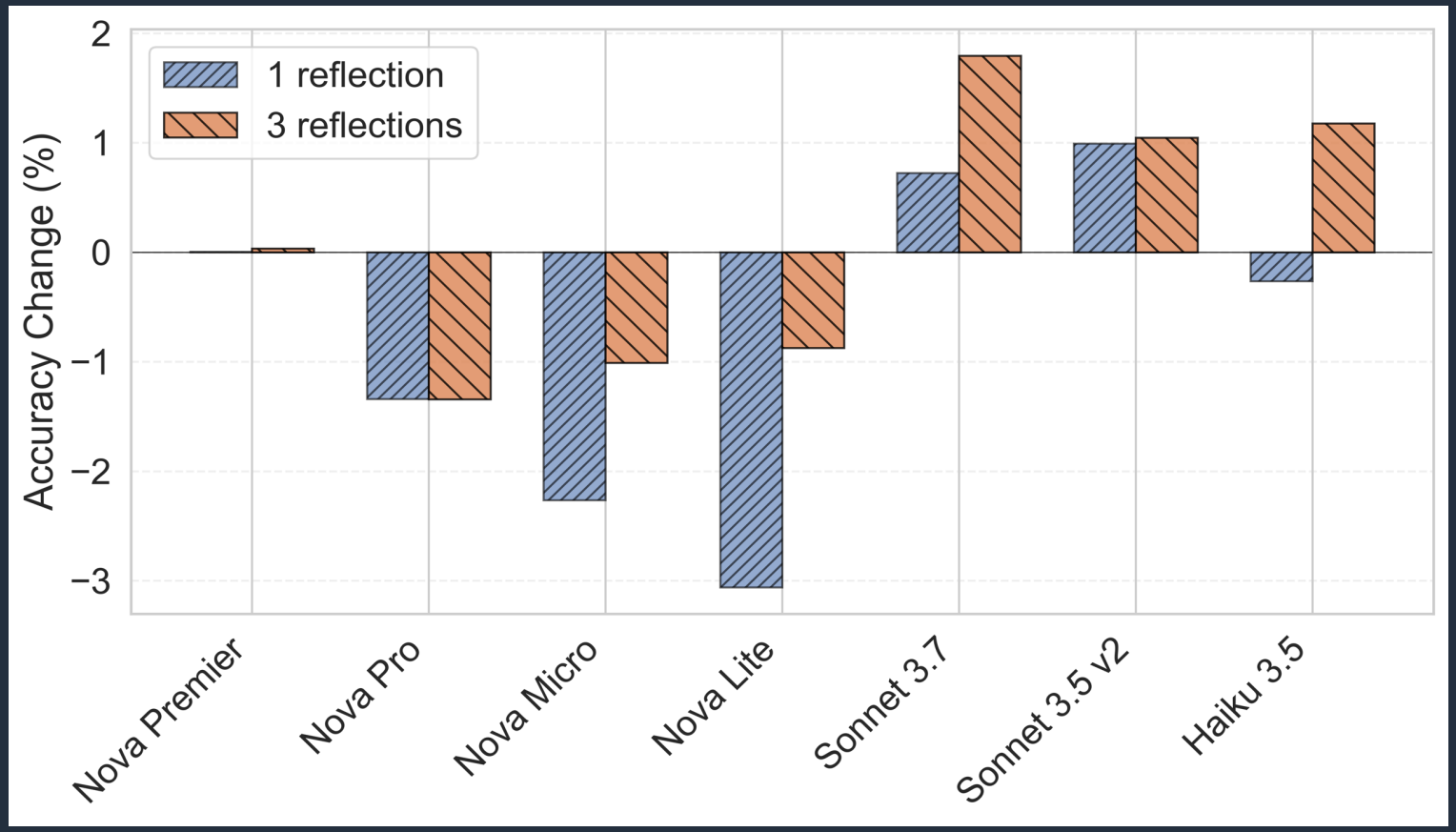
Only Claude 3.7 Shows consistent improvement

No pattern across families of models

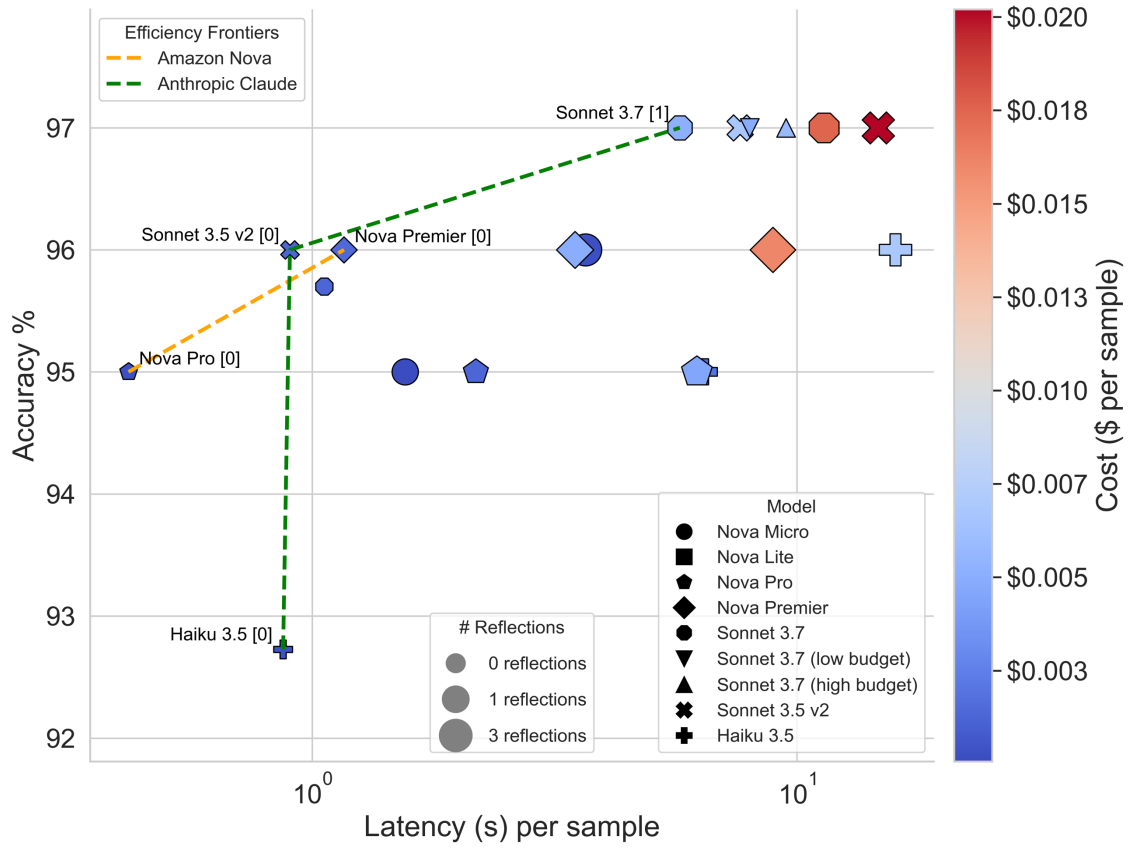
Results – translation



Results – translation



Results – sentiment



Results – sentiment