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

Jack Butler
Amazon Web Services
jackbtlr@amazon.co.uk

Nikita Kozodoi
Amazon Web Services
kozodoi@amazon.de

Zainab Afolabi
Amazon Web Services
zafolabi@amazon.co.uk

Brian Tyacke
Zalando
brian.tyacke@zalando.de

Gaiar Baimuratov
Zalando
gaiar.baimuratov@zalando.de

Abstract

As Large Language Models (LLMs) continue to evolve, practitioners face increasing options for enhancing inference-time performance without model retraining, including budget tuning and multi-step techniques like self-reflection. While these methods improve output quality, they create complex trade-offs among accuracy, cost, and latency that remain poorly understood across different domains. This paper systematically compares self-reflection and budget tuning across mathematical reasoning and translation tasks. We evaluate prominent LLMs, including Anthropic Claude, Amazon Nova, and Mistral families, along with other models under varying reflection depths and compute budgets to derive Pareto optimal performance frontiers. Our analysis reveals substantial domain dependent variation in self-reflection effectiveness, with performance gains up to 220% in mathematical reasoning. We further investigate how reflection round depth and feedback mechanism quality influence performance across model families. To validate our findings in a real-world setting, we deploy a self-reflection enhanced marketing content localisation system at Lounge by Zalando, where it shows market-dependent effectiveness, reinforcing the importance of domain specific evaluation when deploying these techniques. Our results provide actionable guidance for selecting optimal inference strategies given specific domains and resource constraints. We open source our self-reflection implementation for reproducibility at https://github.com/aws-samples/sample-genai-reflection-for-bedrock.

1 Introduction

The deployment of large language models (LLMs) in production systems has created new challenges for practitioners seeking to optimise performance under real-world constraints. Recent advances in inference-time computation scaling Snell et al. (2025) offer promising solutions, allowing deployed systems to dynamically allocate computational resources based on task difficulty, available budget, and latency requirements without model retraining.

Two established approaches for adjusting inference-time resources are multi-step inference and budget tuning. Budget tuning, available for select LLMs such as Anthropic Claude 3.7 Sonnet and OpenAI o1, enables users to configure inference parameters like maximum thinking tokens or reasoning tiers (e.g., low or high), allocating greater computational effort to more challenging inputs. Multi-step approaches such as self-reflection Madaan et al. (2023) are model-agnostic and involve prompting a model to revise its responses through sequential follow-up calls to the model.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Efficient Reasoning.

Prior work demonstrates that self-reflection improves LLM performance in tasks with clearly defined evaluation criteria Madaan et al. (2023) and structured domains with informative feedback signals, such as programming or math Chen et al. (2024). For instance, executing generated code and providing the outputs back to the LLM as context creates concrete feedback signals for accurate self-reflection. However, most production applications such as translation or classification involve more ambiguous objectives and weaker feedback signals. The effectiveness of self-reflection on such tasks remains underexplored, making it unclear whether additional inference-time computation consistently yields performance gains across diverse domains.

This uncertainty is compounded by the rapidly evolving LLM landscape, where practitioners navigate dozens of model options across providers, each with different capabilities, pricing structures, and inference features. Production teams frequently face critical decisions: Should they deploy a smaller, cost-effective model with advanced inference strategies, or invest in larger models with simpler inference pipelines? How do these choices impact not just accuracy, but operational costs, latency requirements, and system reliability?

This paper addresses these practical deployment challenges through comprehensive evaluation of inference-time optimisation strategies across real-world applications. We make three primary contributions. First, we benchmark self-reflection and budget tuning across multiple LLMs and application domains including mathematical reasoning, text-to-SQL generation, sentiment classification, and translation. The derived Pareto-optimal frontiers illustrate accuracy-latency trade-offs of different strategies and provide actionable recommendations for selecting a suitable inference method based on domain-specific requirements, resource constraints, and the base model. Second, we analyse self-reflection trajectories across different LLMs and feedback mechanisms, revealing how reflection depth and feedback quality critically influence self-reflection performance. To the best of our knowledge, this work presents the first direct performance comparison between these two approaches. Third, we demonstrate our findings through a production deployment at Lounge by Zalando, where we implemented self-reflection for marketing content localisation across 17 European markets.

2 Related Work

2.1 Inference-Time Compute

LLMs such as Anthropic's Claude family Anthropic (2024) have been increasing in size over recent years, which has brought profound improvements in performance across a wide range of applications while simultaneously increasing training time and compute requirements Hoffmann et al. (2022). Inference-time compute optimisation techniques avoid modifying the pre-trained model and instead enable dynamic allocation of computational resources depending on the specific requirements of each input. This offers the ability to tune performance according to task demands, scaling performance at inference time Snell et al. (2025).

One of the prominent inference optimisation approaches is self-reflection Madaan et al. (2023), which performs sequential follow-up LLM calls, allowing it to revise initial responses. Other studies have explored drawing parallel samples from the LLM and implementing sampling and verification procedures such as tree-of-thoughts Yao et al. (2023) and graph-of-thoughts Besta et al. (2024). Recent work has also leveraged techniques such as temporary fine-tuning Akyürek et al. (2025) and nearest neighbour retrieval-based fine-tuning Hübotter et al. (2025) where model's parameters are temporarily updated during inference.

In this paper, we explore trade-offs between two established inference strategies: model-agnostic self-reflection Madaan et al. (2023) and built-in reasoning capabilities exposed through model provider APIs. Our goal is to provide insights for practitioners who lack resources to conduct large-scale per-sample fine-tuning or complex multi-step inference processes with feedback loops.

2.2 LLM Post-Training

Another research area aimed at LLM reasoning capabilities is the use of reinforcement learning on specialised reasoning datasets. These approaches implement reinforcement learning with access to either outcome supervision Trung et al. (2024); Xi et al. (2024), step-by-step process supervision Lightman et al. (2023); Setlur et al. (2025) or LLM-driven feedback mechanisms Lee et al. (2024).

Reasoning models are commonly deployed via API interfaces with configuration settings that allow users to adjust computational resources allocated per sample (e.g. Anthropic Claude 3.7 Sonnet thinking tokens budget). Crucially, the reasoning in these models happens as internal processing tokens, and discrete proposed solutions are not validated using external feedback during generation.

2.3 LLM Evaluation

As LLMs have become more capable and businesses increasingly incorporate them into their products and services, robust evaluation frameworks have become essential. Various evaluation platforms exist across different domains, such as HELM Liang et al. (2023) and Chatbot Arena Chiang et al. (2024), where models undergo evaluation and receive automated or human feedback scores depending on the task and domain.

While these platforms provide standardised evaluation of different LLMs, including base models and fine-tuned or quantised variants, there is a lack of comparable evaluation frameworks for inference techniques. This gap makes it challenging for practitioners to understand and navigate trade-offs when combining LLMs with various inference-time compute methods. Throughout our experimentation, we demonstrate these trade-offs across model families, task domains, and inference budgets.

3 Experimental Setup

3.1 Datasets

The empirical study splits into two stages: testing on established benchmarks followed by evaluation on real-world deployment application. First, we perform experiments across four distinct domains using established benchmarks:

- **Flores-200 (Translation)** Costa-Jussà et al. (2022): Multilingual translation benchmark spanning 200 languages, allowing assessment of cross-lingual capabilities.
- **Math500 (Mathematical reasoning)** Lightman et al. (2024): Dataset with 500 problems across algebra, arithmetic and more, testing symbolic manipulation and logical reasoning.
- **Spider (Text-to-SQL)** Yu et al. (2018): Complex text-to-SQL task involving 200 databases with multiple tables, evaluating structured SQL query generation capabilities.
- **IMDB Reviews (Sentiment analysis)** Maas et al. (2011): Binary sentiment classification dataset on movie reviews, assessing natural language classification performance.

The diverse set of tasks allows us to evaluate structured mathematical and programming reasoning (Math500, Spider) and natural language understanding (Flores-200, IMDB). We use a subset of each dataset for evaluation. For Spider, we use 5 databases (voter_1, battle_death, museum_visits, employee_hire_evaluation, orchestra). For Flores-200, we sample 200 examples across 15 language pairs to ensure cross-linguistic variation (English to Arabic, German, Spanish, French, Hebrew, Hindi, Italian, Japanese, Korean, Dutch, Portuguese, Russian, Turkish, Chinese, Polish). For IMDB and Math500, we randomly sample 100 examples.

After validating approaches on benchmark data sets, we test and deploy the best model configurations on real-world data on marketing content localisation at Lounge by Zalando. Further details on the deployment are provided in Section 5.

3.2 Models and Inference Techniques

We compare performance across 10 LLMs:

- Amazon Nova (Premier, Pro, Micro, Lite Intelligence (2024))
- Anthropic Claude (Sonnet 3.7, Sonnet 3.5 v2, Haiku 3.5 Anthropic (2024))
- Llama 4 (Maverick 17B Meta (2025))
- Mistral (Small, Large Mistral (2024)).

For inference, we employ self-reflection with 0, 1, and 3 reflection rounds across all LLMs. Self-reflection is implemented through repeated LLM invocations, where at the end of each round we prompt the model to reflect on its response and update it if necessary. On the text-to-SQL task, we also compare 2 feedback mechanisms, which provide additional context before each self-reflection round. For Claude 3.7, we additionally make use of the built-in reasoning mode by defining 2 thinking budgets (4096 tokens and 1024 tokens). We refer to these thinking budgets as high and low in the rest of the paper.

We run experiments using Amazon Bedrock, maintaining default temperature and inference parameters associated with each model to ensure fair comparison. The token costs are recorded as of 02/05/2025, assuming on-demand pricing. Latency is measured as the total elapsed time between the input and the completion of the full response. Prompts used for each of the 4 domains, as well as for self-reflection and feedback mechanisms, are available in Appendix A.

3.3 Evaluation Metrics

We employ task-specific metrics for each dataset to evaluate LLM performance. For translation, we use METEOR Lavie and Agarwal (2007), which accounts for both precision and recall while handling synonyms and paraphrases. Sentiment analysis quality is assessed using classification accuracy on the binary prediction task. Spider and Math500 use additional verification procedures.

For Spider and Math500, we implement additional verification procedures to assess semantic equivalence of generated responses. For Math500, we evaluate accuracy through normalised comparison and symbolic verification. We use string matching on normalised and cleaned LaTeX expressions, followed by symbolic equivalence checking with SymPy Meurer et al. (2017) to identify equivalent answers even when expressed differently. For Spider, we assess both exact matches and functional equivalence by executing SQL queries, discarding failures, and comparing normalised result tables against ground truth. When exact row matches are not found, we calculate partial credit based on matching cell values.

For the content localisation deployment data, we use multiple technical metrics to evaluate the localisation quality, including METEOR, BLEU, and LLM-as-a-judge score. After technical evaluation, we also run human expert evaluation tests of the deployed system, where we compare the number of localisation mistakes raised by expert copywriters after analysing the generated localisations.

4 Results on Benchmarks

This section overviews empirical results. For each dataset, we illustrate the percentage change in accuracy relative to zero reflections for each model, highlighting improvements for each configuration. Next, we construct Pareto-optimal frontiers showing accuracy-latency trade-off when employing inference strategies and provide cost information for each model and strategy combination. In Section 4.5, we dive deeper on how performance of self-reflection is affected by different factors.

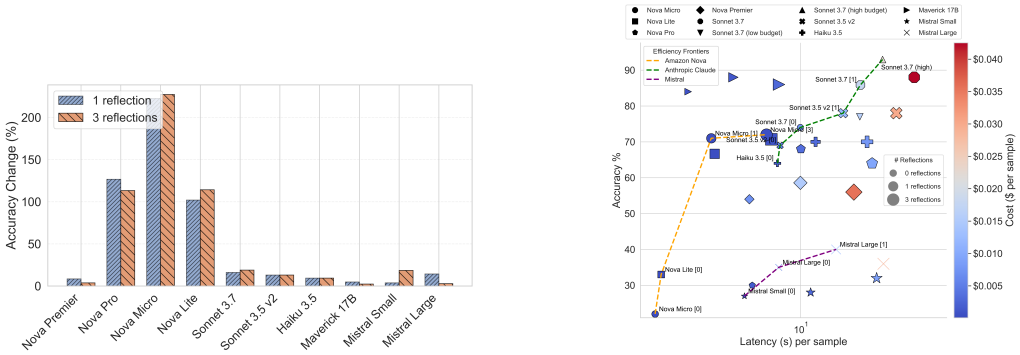
4.1 Mathematical Reasoning (Math500)

As depicted on Figure 1(a), all LLMs benefit from self-reflection. Nova Micro shows the largest gains, with accuracy improving by 220% with 1 reflection and maintaining this gain after 3 reflections. This suggests that Nova Micro's base mathematical reasoning capabilities are significantly enhanced through iterative self-correction. Similarly, Nova Lite and Pro show substantial improvements of approximately 100-130% with reflection, indicating that smaller models in the Nova family particularly benefit from reflection in mathematical reasoning.

In contrast, Nova Premier and Claude LLMs exhibit more modest but consistent improvements from reflection. Sonnet 3.7 shows approximately a 16% increase in accuracy with 1 and 20% with 3 reflections. Sonnet 3.5 v2 and Haiku 3.5 demonstrate similar patterns with gains of 13% and 9%, respectively. These results suggest that while Claude models benefit from reflection in mathematical tasks, their initial performance is already strong, resulting in less dramatic relative improvements.

Figure 1(b) provides accuracy measurements across model configurations. Overall, Sonnet 3.7 demonstrates superior mathematical reasoning capabilities, starting with a baseline accuracy of 74% without reflection and improving to 86% with 1 and 88% with 3 reflections. Nova Micro starts with

at just 22% accuracy without reflection (omitted from the plot) but jumps to 71% accuracy with 1 and 72% with 3 reflections. This pattern is similar for other Nova, Llama Maverick, Mistral and Claude LLMs and suggests that for mathematical reasoning, a single well-implemented reflection round captures most of the potential performance benefit, with diminishing returns for further rounds.



(a) Relative Self-Reflection Gains

(b) Accuracy-Latency Pareto Frontiers

Figure 1: Inference-Time Performance (Math500)

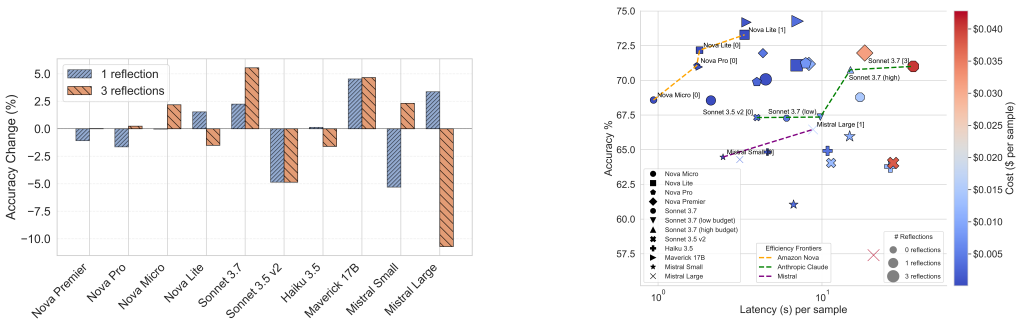
The Pareto frontier for the Claude family offers a rich selection in the accuracy-latency space, ranging from Haiku 3.5 with no reflection offering 64% accuracy at $0.0015 per example and 7.5 sec. latency, up to Sonnet 3.7 with a high thinking budget, which reaches 93% accuracy at $0.0224 and 27.9 sec. latency. At the same time, Sonnet 3.7 with a low thinking budget is dominated by Sonnet 3.7 with 1 self-reflection, which reaches a higher accuracy at the same latency. Considering the Amazon Nova family, Nova Micro with 1 and 3 reflections dominate Haiku 3.5 and Sonnet 3.5 in low-latency space but can not reach the same accuracy as higher-end Claude models even with self-reflection. This implies that practitioners should consider Nova Micro with self-reflection under strict cost/latency constraints and switch to Sonnet 3.7 with high reasoning for the best performance. Llama Maverick provides superior accuracy across the Nova Frontier for budgets larger than Nova Lite or Micro without reflections, and also match the performance of Sonnet 3.7 with 1 reflection.

To validate statistical significance of the results, we draw 100 bootstrap samples of individual examples and run pairwise t-tests comparing mean accuracies per model configuration. All models on the efficiency frontiers show significant accuracy differences at 1% level, which are also confirmed by pairwise Nemenyi tests.

4.2 Text-to-SQL (Spider)

In contrast to Math500, in text-to-SQL generation Sonnet 3.7 is the only model to show consistent and limited improvements with additional reflections, gaining 2.3% accuracy with 1 round and a 5.6% gain with 3 rounds. Most other LLMs show mixed or negative responses to self-reflection. Sonnet 3.5 v2 demonstrates the most pronounced quality degradation, with accuracy declining by approximately 4.8% with 1 and reflection rounds. Similarly, Nova Pro and Haiku 3.5 show noticeable performance decreases with added reflection rounds.

Nova Micro, Sonnet 3.7 and Llama Maverick are the only models benefiting from additional reflections. Nova Micro maintains neutral performance with 1 round and achieves a 2.2% accuracy improvement with 3 reflections. Nova Lite displays an inconsistent pattern, with a slight improvement (1.5%) at 1 reflection but declining by 1.5% with 3 rounds. Overall, these results suggest that self-reflection is less useful in domains like text-to-SQL, where revising the generated query without any additional context may mislead LLMs to change previously correct SQL queries. Mistral Small shows benefits with three rounds of reflection but decreases in accuracy with 1 reflection whereas the Large variant demonstrates the opposite behaviour.



(a) Relative Self-Reflection Gains

(b) Accuracy-Latency Pareto Frontiers

Figure 2: Inference-Time Performance (Spider)

The Amazon Nova Pareto frontier on Figure 2(b) represents optimal configurations for lower-latency applications. Overall, Amazon Nova models consistently outperform Claude LLMs, with 2 Nova Lite variants dominating all Claude counterparts. Nova Lite with 1 reflection achieves the highest absolute accuracy (74%) with approximately 3-second latency, while Nova Micro with 0 reflections offers the fastest and cheapest option to reach 68% accuracy. The Claude frontier represents a different set of trade-offs, with Sonnet 3.7 using 3 reflections achieving 71% accuracy but at a substantially higher latency (>35 seconds) and cost. Interestingly, built-in reasoning modes with both budget sizes fall behind the model variant with 3 reflections in terms of the accuracy, but are available at a lower latency and cost. Mistral frontier highlights that Mistral Small is performant for this use case but to achieve better performance one can leverage Mistral Large with 1 reflection. Llama Maverick provides the highest accuracy and also deliver this at a equivalent latency and cost the Nova Lite.

These results highlight the importance of model-specific optimisation strategies for SQL tasks, with Amazon Nova models generally performing best with minimal reflection, while Claude Sonnet 3.7 uniquely benefits from both reflection and built-in reasoning despite the increased latency and cost.

4.3 Sentiment Classification (IMDB)

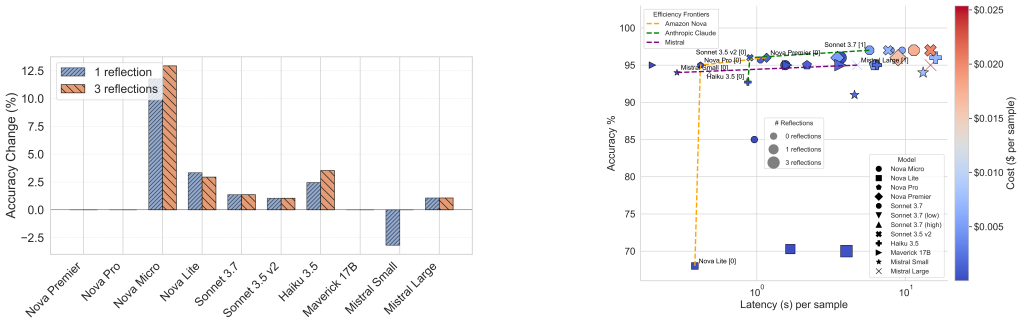
On sentiment analysis, Figure 3(a) clearly illustrates the positive impact of self-reflection on the accuracy across the LLMs. For the most models, adding reflection rounds improves accuracy, though with diminishing returns after the 1st reflection. Nova Micro shows the highest relative improvement from 0 to 1 reflections (85% to 95% accuracy), while maintaining similar latency to 0 reflections of Sonnet 3.7 (1.56 vs 1.06) and similar resulting accuracy (95% vs 95.7%) at 1/18th of the cost. Nova Pro, Premier and Llama Maverick are the only models whose accuracy is not affected by reflection. There is an outlier, Mistral Small, which decreases in accuracy with reflections.

As depicted on Figure 3(b), for applications requiring the highest possible accuracy, Sonnet 3.5 without reflection or Sonnet 3.7 with 1 reflection round offer the best performance. Built-in reasoning in Claude 3.7 performs similar to 1 round of self-reflection regardless of the thinking budget, but introduces higher latency and cost, making these configurations less attractive. For cost-sensitive deployments with moderate latency requirements, Nova Premier with 0 reflections presents a good compromise. Interestingly, Nova Micro with 3 reflections is able to reach a higher accuracy compared to Nova Premier, but results in a substantially higher overall latency and a marginally higher cost.

The results indicate that despite the ambiguity of the sentiment analysis task, LLMs consistently benefit from self-reflection in this domain. At the same time, the average gains are one order of magnitude smaller compared to mathematical reasoning, which makes inference-time techniques less attractive in terms of the cost-latency implications that may outweigh the accuracy gains.

4.4 Translation (Flores-200)

Figure 4(a) highlights distinct divergence in translation performance across LLM families. Claude models generally demonstrate enhanced performance after reflection. In contrast, all Amazon Nova models except Nova Premier exhibit an inverse trend, where incorporating 1 reflection diminishes translation accuracy. Despite some recovery when going from 1 to 3 reflections, Nova Micro,

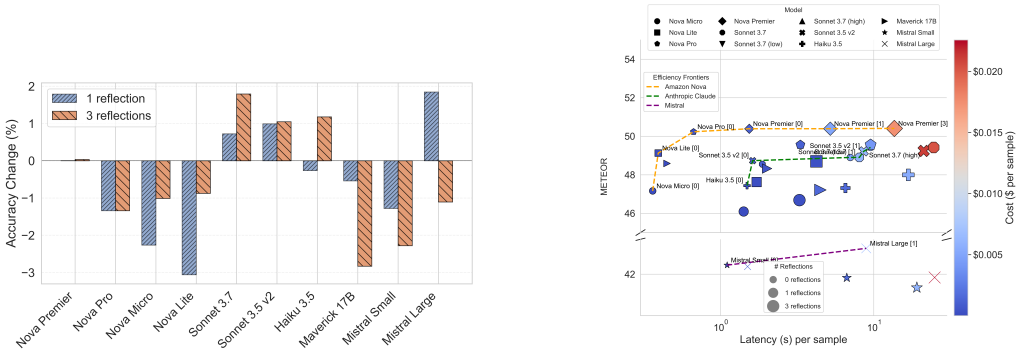


(a) Relative Self-Reflection Gains

(b) Accuracy-Latency Pareto Frontiers

Figure 3: Inference-Time Performance (IMDB)

Lite and Pro still under-perform compared to their baseline with no reflection. Mistral Small and Llama Maverick also show negative performance with 1 reflection but unlike Amazon Nova, there is no recovery after 3. Mistral Large shows initial improvement after 1 reflection round but after 3 reflections we see a degradation of similar scale to the Nova models. his implies that for Mistral Large only 1 reflection would be recommended where as using self-reflection for Amazon Nova LLMs, Mistral Small and Llama Maverick in translation tasks is not recommended.



(a) Relative Self-Reflection Gains

(b) Accuracy-Latency Pareto Frontiers

Figure 4: Inference-Time Performance (Flores-200)

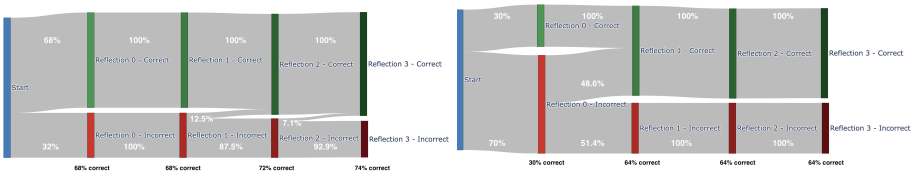
Figure 4(b) suggests that Amazon Nova models dominate all considered Claude models in the latency-accuracy space. Nova Pro reaches higher accuracy when all Claude variants at a lower latency compared to Haiku 3.5. Furthermore, Nova Premier with self-reflection provides a further marginal gain in translation accuracy, but brings a substantial latency increase. This suggests that Amazon Nova performs particularly well in translation tasks, with an important caveat that integrating self-reflection for smaller variants hurts their performance. Focusing on Claude family, we note that Sonnet 3.7 built-in reasoning with a high thinking budget achieves the best METEOR score among Claude models, outperforming low thinking budget, self-reflections, and other Claude variants.

4.5 Ablation Studies

Reflection Transitions

Figure 5 illustrates how LLM performance evolves throughout self-reflection rounds. We focus on mathematical reasoning, which benefits most from reflection and showcase results for 2 LLMs from model families with distinct performance patterns: Claude Sonnet 3.5 and Nova Micro.

Sonnet 3.5 v2 demonstrates superior initial accuracy at 68% compared to Nova Micro's 30%. Through 3 reflection stages, Sonnet 3.5 shows consistent improvements, ultimately reaching 74% accuracy. Interestingly, while the first reflection does not change Sonnet's accuracy, each subsequent round



(a) Claude Sonnet 3.5 v2 (b) Amazon Nova Micro

Figure 5: Errors Across Reflections (Math500)

Table 1: Impact of Feedback Mechanisms on Self-Reflection

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**

successfully corrects a portion of incorrect responses. In contrast, Nova Micro exhibits a dramatic improvement during the first reflection, jumping to 64% accuracy after successfully correcting 48.6% of initial errors. However, Nova's performance plateaus thereafter, showing no further improvement in subsequent reflection stages. Another notable pattern across both LLMs is their perfect preservation of initially correct responses throughout all reflection rounds. These findings suggest that smaller models like Nova Micro have capacity for initial self-correction, whereas more capable LLMs like Sonnet 3.5 have stronger foundational performance and potential for continuous improvement through iterative reflection rounds.

Reflection Feedback

Table 1 investigates if providing informative feedback to the LLM between self-reflection rounds facilitates stronger accuracy gains. We focus on text-to-SQL and compare 2 feedback mechanisms as LLM context: i) output of SQL query execution; ii) LLM-as-a-judge response with Nova Pro judge.

The results reveal no clearly dominating feedback strategy. On average, incorporating feedback mechanisms improves reflection quality in 61% of cases, confirming that additional feedback can be beneficial. However, model families respond differently to feedback types: Amazon Nova generally performs better with LLM-as-judge feedback or no feedback at all, while Claude shows higher accuracy with SQL execution feedback. This may be explained by the fact that Amazon Nova Pro judge is not able to provide stronger feedback to Claude models compared to their own reasoning, which may risk getting them off track. These findings emphasise the importance of identifying optimal configurations for specific business applications by accounting for resource constraints, the LLM being used, and the task domain. Throughout our experiments, we consistently find that no single inference optimisation strategy proves universally effective across the diverse range of scenarios we tested.

5 Results on Real-World Deployment

To validate our benchmark findings in a production environment, we present the real-world evaluation of a self-reflection-enhanced marketing content localisation system at Lounge by Zalando. Zalando is an online multi-brand fashion destination with more than 52 million active customers. Lounge by Zalando represents a shopping club, where customers can browse through a curated selection of fashion products. One of the critical business tasks at Lounge By Zalando is localising the marketing content for different European markets, including different distribution channels such as email newsletters, push notifications, and display.

Table 2: Self-Reflection Performance on Real-World Marketing Content Localisation Task

Language	No reflection			Self-reflection with LLM judge feedback		
	BLEU	METEOR	LLM judge score	BLEU	METEOR	LLM judge score
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**
German	0.32	0.61	0.38	**0.33**	**0.62**	**0.47**

Prior to our deployment, manual localisation process at Lounge by Zalando created significant bottlenecks, as copywriters required 2-3 days per campaign to localise content for major markets, limiting campaign agility and time-to-market. The core challenge was not merely translation into a local language, but sophisticated localisation incorporating: (i) market-specific tonality guidelines (e.g. using formal or informal pronouns when referring to a customer, (ii) local regulatory compliance (e.g. using correct terms for different sales types), and (iii) consistent brand voice adaptation.

5.1 Technical Performance Evaluation

To perform initial evaluation, we collect ground truth localised texts produced by copywriters independently from our system. We compare generated and human localisations using three metrics:

- BLEU score comparing the generated and ground truth localisations;
- METEOR score comparing the generated and ground truth localisations;
- LLM-as-a-judge score. The judge (Claude Sonnet 3.5 in our case) picks the best localisation out of the generated and ground truth versions. We then aggregate preferences across the dataset to calculate the judge's preferred generation or whethet there was a tie.

Table 2 illustrates the results on three markets. Considering the LLM-as-a-judge metric, we see that self-reflection improves the localisation quality on all languages, with the strongest gains observed for German. Here, the number of cases where the generated localisation is better than or the same quality as the human translation increases from 38% to 47%. On French and Spanish markets, the no-reflection version demonstrates very strong performance with 61% and 49%, diminishing the gains from self-reflection.

Considering the text similarity metrics BLEU and METEOR, we observe mixed results with consistent improvements from self-reflection on the German market, its negative impact on localisation quality on the French market and similar performance with and without reflection on the Spanish market. It is important to note that manual inspection of selected localisations indicated that LLM-as-a-judge provides a more reliable quality measure, as similarity metrics do not incorporate the tonality guidelines and do not account for multiple accepted alternatives of formulating a sentence.

5.2 Human Expert Evaluation

To complement our automated metrics, we conduct a blind A/B evaluation with domain expert copywriters. Three human experts, each with multiple years of experience in marketing localisation for European markets, were asked to evaluate localised content without knowledge of which leveraged self-reflection. The experts assessed localisations across the same five evaluation criteria and reported the list of issues violating the guidelines. The results provided in Table 3 reveal significant improvements from self-reflection.

Expert evaluation results demonstrate substantial improvements in content quality, particularly for French localisations where self-reflection reduced identified issues by 88% (from 384 to 46 issues). Spanish localisations showed a 39% reduction in issue, while for German 100% of original 15 issues were resolved. These results are in line with LLM-as-a-judge scores, which also favor the localisation variants with self-reflection. These findings underscore why business metrics are essential for evaluating production NLP systems. Technical metrics provide relevant development feedback, but real world metrics such as domain expert evaluations results in higher time-to-market acceleration, cost reductions, and compliance adherence.

Overall, the results indicate it is valuable to employ self-reflection on markets with more challenging localisation rules (such as German), whereas using it on markets where the base model already achieves high quality provides minimal quality gains that may not justify the additional LLM cost. While this confirms some of the findings from Sections 4 and 4.5, it also emphasises the importance of testing self-reflection performance on a specific dataset, as the gains may vary significantly depending on the use case.

Given that self-reflection enhances output quality at increased computational cost, we examine whether prompt caching Gim et al. (2024) can partly mitigate this cost overhead. Our analysis demonstrates that combining self-reflection with prompt caching achieves cost reductions of up to 28% when employing three reflection rounds with detailed cost-latency analysis presented in Appendix B.4.

Table 3: Human Expert Evaluation Results

Language	No reflection	Self-reflection
French	384 issues	**46 issues** (-88%)
Spanish	49 issues	**30 issues** (-39%)
German	15 issues	**0 issues** (-100%)

6 Conclusion

This paper presents a systematic analysis of inference optimisation techniques such as self-reflection and budget tuning across different domains, base models, and reflection parameters. We derive Pareto frontiers in accuracy-latency space, test self-reflection in a production deployment, and provide actionable recommendations to practitioners regarding suitable inference optimisation methods for real-world applications.

Our results reveal no universally dominant inference strategy, with both the magnitude and direction of performance impacts varying significantly across tasks. Self-reflection consistently improves performance in math (with gains up to 220%) and sentiment analysis, while showing mixed or negative effects in translation and text-to-SQL generation. This domain-specific variability highlights the importance of empirical evaluation before deploying inference optimisation techniques in production.

Several key patterns emerge from our analysis: (i) smaller LLMs often benefit more dramatically from reflection than larger ones; (ii) a single reflection round frequently captures most potential performance benefits; (iii) in several cases, smaller LLMs with reflection outperform larger models without it, offering potential cost savings; and (iv) Claude's built-in reasoning sometimes underperforms compared to self-reflection techniques despite its specialised design, and results in higher additional cost as it does not benefit from prompt caching.

For practitioners, these findings suggest task-specific optimisation strategies. For math, self-reflection is highly recommended, with Amazon Nova Micro offering an excellent cost-performance balance. For text-to-SQL, Amazon Nova generally outperform Claude variants, with minimal reflection recommended. For sentiment analysis, most models benefit from reflection, though gains may not justify increased costs. For translation, Claude generally benefits from reflection while Amazon Nova performs better without it. The case study on real-world marketing content localisation data confirms that self-reflection gains with Claude models may be higher on the tasks that are more challenging.

In future work, we aim to conduct a deeper interpretative analysis of the budget tuning methods, including providing transition analysis of the generated thinking tokens. We also wish to expand our analysis outside of the Amazon Nova, Mistral and Anthropic Claude model families to understand the influence of inference-time compute techniques on other leading model providers. We would also want to understand the benefits of combining complimentary techniques from inference-time compute such as parallel sampling, best-of-N majority voting and others.

References

Ekin Akyürek, Mehul Damani, Adam Zweiger, Linlu Qiu, Han Guo, Jyothish Pari, Yoon Kim, and Jacob Andreas. 2025. The Surprising Effectiveness of Test-Time Training for Few-Shot Learning. arXiv:2411.07279 [cs.AI] https://arxiv.org/abs/2411.07279

Anthropic. 2024. The Claude 3 Model Family: Opus, Sonnet, Haiku. (2024).

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. 2024. Graph of Thoughts: Solving Elaborate Problems with Large Language Models. *Proceedings of the AAAI Conference on Artificial Intelligence* 38, 16 (March 2024), 17682–17690. https://doi.org/10.1609/aaai.v38i16.29720

Xinyun Chen, Maxwell Lin, Nathanael Schärli, and Denny Zhou. 2024. Teaching Large Language Models to Self-Debug. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=KuPixIqPiq

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference. arXiv:2403.04132 [cs.AI] https://arxiv.org/abs/2403.04132

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672* (2022).

In Gim, Guojun Chen, Seung seob Lee, Nikhil Sarda, Anurag Khandelwal, and Lin Zhong. 2024. Prompt Cache: Modular Attention Reuse for Low-Latency Inference. arXiv:2311.04934 [cs.CL] https://arxiv.org/abs/2311.04934

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. Training Compute-Optimal Large Language Models. arXiv:2203.15556 [cs.CL] https://arxiv.org/abs/2203.15556

Jonas Hübotter, Sascha Bongni, Ido Hakimi, and Andreas Krause. 2025. Efficiently Learning at Test-Time: Active Fine-Tuning of LLMs. In *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=NS1G1Uhny3

Amazon Artificial General Intelligence. 2024. The Amazon Nova family of models: Technical report and model card. *Amazon Technical Reports* (2024). https://www.amazon.science/publications/the-amazon-nova-family-of-models-technical-report-and-model-card

Alon Lavie and Abhaya Agarwal. 2007. Meteor: an automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation* (Prague, Czech Republic) *(StatMT '07)*. Association for Computational Linguistics, USA, 228–231.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Ren Lu, Thomas Mesnard, Johan Ferret, Colton Bishop, Ethan Hall, Victor Carbune, and Abhinav Rastogi. 2024. RLAIF: Scaling Reinforcement Learning from Human Feedback with AI Feedback. https://openreview.net/forum?id=AAxIs3D2ZZ

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas

Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic Evaluation of Language Models. arXiv:2211.09110 [cs.CL] https://arxiv.org/abs/2211.09110

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's Verify Step by Step. arXiv:2305.20050 [cs.LG] https://arxiv.org/abs/2305.20050

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let's Verify Step by Step. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=v8L0pN6EOi

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning Word Vectors for Sentiment Analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Portland, Oregon, USA, 142–150. http://www.aclweb.org/anthology/P11-1015

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-Refine: Iterative Refinement with Self-Feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=S37hOerQLB

Meta. 2025. The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation. (2025).

Aaron Meurer, Christopher P. Smith, Mateusz Paprocki, Ondřej Čertík, Sergey B. Kirpichev, Matthew Rocklin, AMiT Kumar, Sergiu Ivanov, Jason K. Moore, Sartaj Singh, Thilina Rathnayake, Sean Vig, Brian E. Granger, Richard P. Muller, Francesco Bonazzi, Harsh Gupta, Shivam Vats, Fredrik Johansson, Fabian Pedregosa, Matthew J. Curry, Andy R. Terrel, Štěpán Roučka, Ashutosh Saboo, Isuru Fernando, Sumith Kulal, Robert Cimrman, and Anthony Scopatz. 2017. SymPy: symbolic computing in Python. *PeerJ Computer Science* 3 (Jan. 2017), e103. https://doi.org/10.7717/peerj-cs.103

Mistral. 2024. AI in abundance. (2024).

Amrith Setlur, Chirag Nagpal, Adam Fisch, Xinyang Geng, Jacob Eisenstein, Rishabh Agarwal, Alekh Agarwal, Jonathan Berant, and Aviral Kumar. 2025. Rewarding Progress: Scaling Automated Process Verifiers for LLM Reasoning. In *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=A6Y7AqlzLW

Charlie Victor Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2025. Scaling LLM Test-Time Compute Optimally Can be More Effective than Scaling Parameters for Reasoning. In *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=4FWAwZtd2n

Luong Trung, Xinbo Zhang, Zhanming Jie, Peng Sun, Xiaoran Jin, and Hang Li. 2024. ReFT: Reasoning with Reinforced Fine-Tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 7601–7614. https://doi.org/10.18653/v1/2024.acl-long.410

Zhiheng Xi, Wenxiang Chen, Boyang Hong, Senjie Jin, Rui Zheng, Wei He, Yiwen Ding, Shichun Liu, Xin Guo, Junzhe Wang, Honglin Guo, Wei Shen, Xiaoran Fan, Yuhao Zhou, Shihan Dou, Xiao Wang, Xinbo Zhang, Peng Sun, Tao Gui, Qi Zhang, and Xuanjing Huang. 2024. Training large language models for reasoning through reverse curriculum reinforcement learning. In *Proceedings of the 41st International Conference on Machine Learning* (Vienna, Austria) *(ICML'24)*. JMLR.org, Article 2217, 19 pages.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. arXiv:2305.10601 [cs.CL] https://arxiv.org/abs/2305.10601

Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, Zilin Zhang, and Dragomir Radev. 2018. Spider: A Large-Scale Human-Labeled Dataset for Complex and Cross-Domain Semantic Parsing and Text-to-SQL Task. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (Eds.). Association for Computational Linguistics, Brussels, Belgium, 3911–3921. https://doi.org/10.18653/v1/D18-1425

Appendix

A. Prompt Templates

This Appendix provides prompt templates used for each of the four predictions tasks considered in the paper. We also provide the prompt templates for the LLM-as-a-judge feedback mechanism and for the self-reflection iterations.

A.1. Prediction Tasks

Flores-200

Translate the following text into {language}. Please output only the translated text with no prefix or introduction and put in in <translation></translation> XML tags.
Text to be translated: {source}

Math500

What is the answer to the following math problem: {problem}. Make sure to always state your final answer in <answer> </answer> tags.

Spider

You are a data scientist sqlite expert. Your job is to take user questions and translate them into SQL queries. For reference, today's date is 16/04/2025.
{table_name_and_schema}
<instruction> Only fetch the relevant columns for example partition is not generally required. </instruction>
The user question is provided inside <question></question> XML tags. Aim to generate a valid sqlite query for the user question using the table above. Always provide your thinking in <reasoning></reasoning> tags and then output the SQL statement in <SQL></SQL> tags.
Here is the question:{question}

IMDB Reviews

Read the following movie review. Classify the review sentiment as either positive or negative. Do not add any other words. Please output only the sentiment in <sentiment></sentiment> XML tags. Review to be classified: {review}

A.2. Self-Reflections and Feedback Mechanisms

Self-Reflection

Please reiterate your answer by thinking step by step, making sure to state your answer at the end of the response.
{feedback_mechanism_output}
As a reminder, the original question is {first_user_message}

LLM-as-a-Judge Feedback

You are evaluating the accuracy of a response to a question. Review the following context containing both a question and answer.
For your evaluation:

- Determine if the answer is factually correct and fully addresses the question
- Make a binary judgment: CORRECT or INCORRECT
- Provide a brief justification (1-2 sentences)

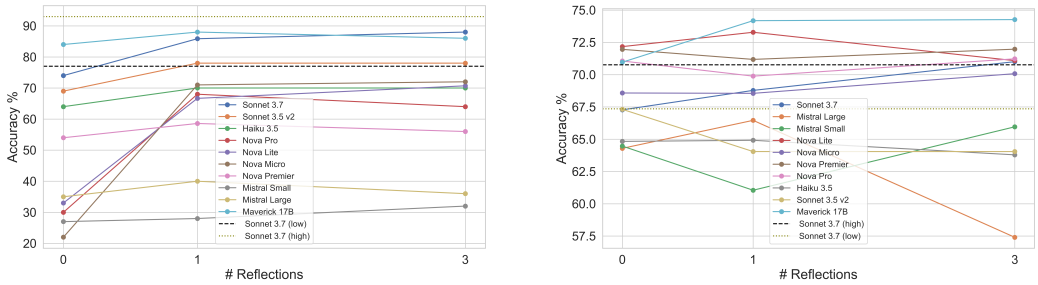
- If you don't have enough information to make a judgment, say so

User question: {user_query}
Provided response: {context}

B. Extended Results

This Appendix provides extended empirical results, including the plots depicting the impact of number of self-reflection rounds on the LLM accuracy, as well as additional Sankey diagrams revealing the transition dynamics during the self-reflection rounds on Math500.

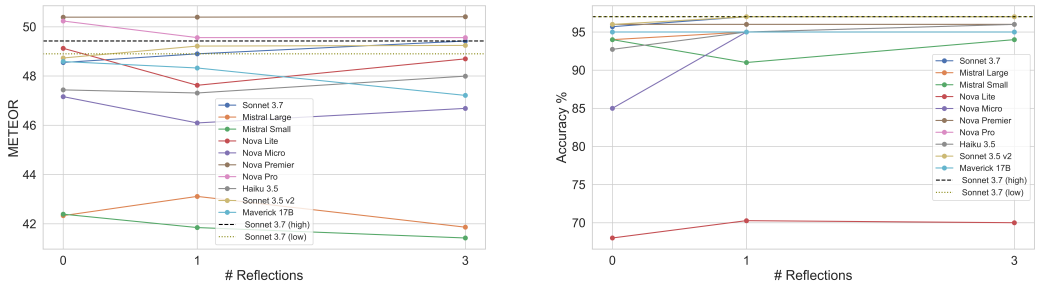
B.1. Impact of the Number of Reflections on Accuracy



(a) Math500, Reflection Impact

(b) Text-to-SQL (Spider), Reflection Impact

Figure 6: Number of Reflections (Math500 and Spider)



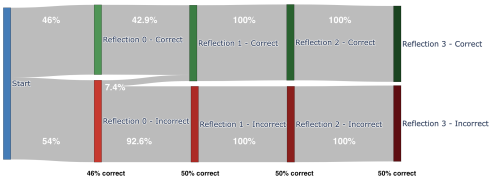
(a) Translation, Reflection Impact

(b) Sentiment Classification, Reflection Impact

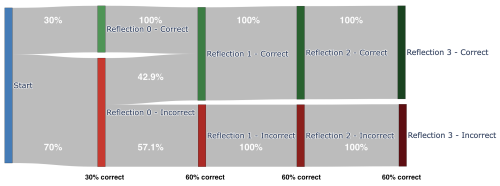
Figure 7: Number of Reflections (Flores-200 and IMDB)

B.2. Self-Reflection Transitions

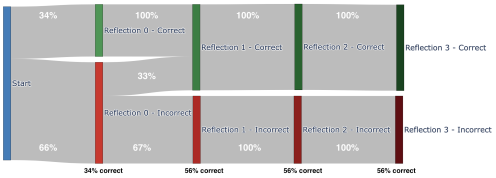
The Sankey diagrams (Figure 8 a-e) provide detailed visualisation of reflection pathways for additional models beyond Claude Sonnet 3.5 v2 and Amazon Nova Micro discussed in the main text. These diagrams reveal consistent patterns across model families while highlighting unique characteristics. Models with varying initial accuracy (34%-70%) all demonstrate perfect retention of correct answers through subsequent reflection stages; a pattern consistent across all tested LLMs. For models with moderate initial performance (46%-50%), we observe that the first reflection stage provides the most substantial correction opportunity, with 42.9%-67% of initially incorrect responses remaining incorrect after Reflection 0, while subsequent reflections yield minimal improvements. This mirrors Amazon Nova Micro's behavior described in the main text. In contrast, models with higher initial accuracy (70%) show more nuanced improvement patterns, with 13.3% of initially incorrect responses being corrected at Reflection 0 and accuracy stabilising at 74%—similar to Claude Sonnet 3.5's incremental improvement pattern. These findings reinforce our main conclusion that smaller models primarily benefit from initial self-correction, while more capable models can leverage both strong foundational performance and iterative improvement through extended reflection processes.



(a) Math500, Amazon Nova Premier Reflection Transitions

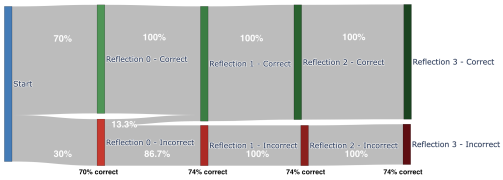


(b) Math500, Amazon Nova Pro Reflection Transitions



(c) Math500, Amazon Nova Lite Reflection Transitions

(d) Math500, Anthropic Claude 3.5 Haiku Reflection Transitions



(e) Math500, Anthropic Claude 3.7 Sonnet Reflection Transitions

Figure 8: Reflections Transitions (Math500)

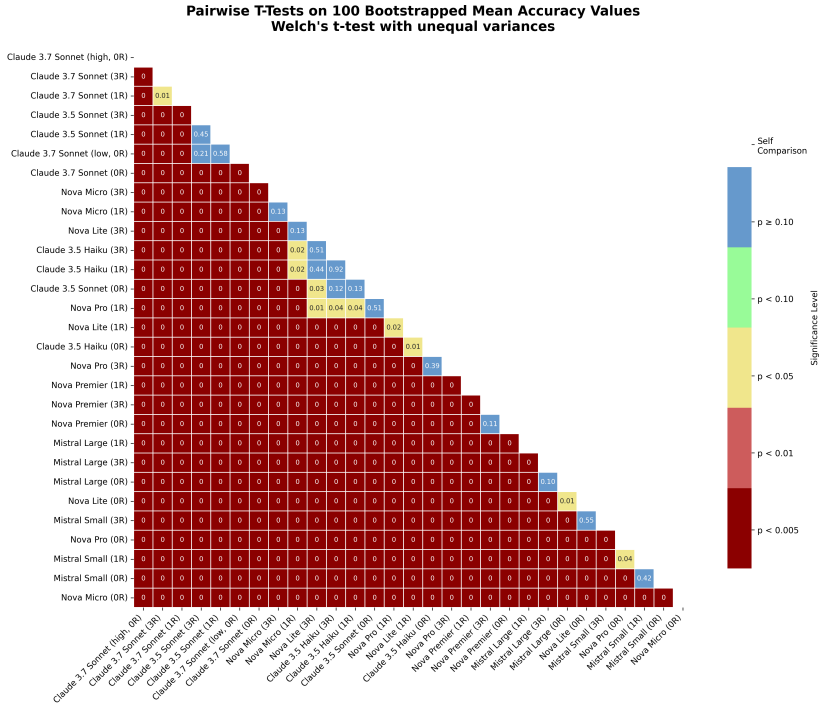


Figure 9: Math500, Pairwise T-Test P-Values

B.3. Statistical Significance Tests

To validate the statistical significance of the accuracy differences observed in Section 4, we draw 100 bootstrap samples of the individual LLM generations for each model and inference-time configuration. Next, we calculate accuracy scores per configuration on each of the bootstrap samples, which yields a distribution of 100 accuracy scores for each configuration. Finally, we run pairwise statistical tests comparing the accuracy distributions.

We use pairwise Welch's t-tests with unequal variances to compare mean accuracy values of different configurations. Figure 9 shows test p-values across all model configurations on Math500, sorted by the average accuracy. Out of 496 configuration pairs, only 26 accuracy differences are not significant at 1% level, including different variations of Claude 3.5 Haiku compared to each other and Nova Pro with 1 self-reflection.

Beyond the t-tests, significance of the accuracy differences is also confirmed by the Friedman test, which indicates there are significant differences between at least some of the 32 model configurations and rejects the null hypothesis that all models perform equally. Nemenyi post-hoc tests indicate that 71% of the pairwise accuracy differences are statistically significant, including the models on the efficiency frontiers.

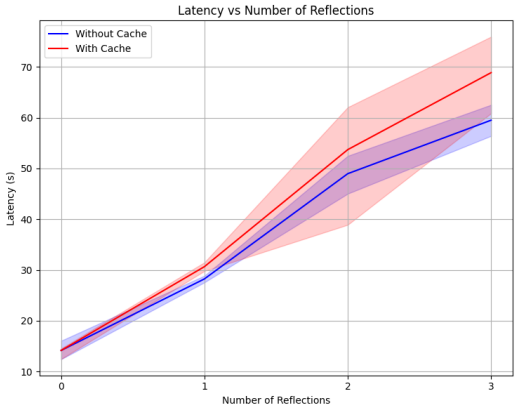
B.4. Prompt Caching

Prompt caching, as described generally in Gim et al. (2024), is a set of techniques for caching computed model states so they can be re-used over future invocations of an LLM. Amazon Bedrock has released a prompt caching feature which allows users to set cached checkpoints during their conversation history and then save on the cost of recomputing these past messages. This capability is often used to cache very long initial system prompts or initial context which is used across multiple interactions with the LLM. Additionally, our results such as Figure 3 have shown that while self-reflection can bring improved performance, the additional cost and latency can hurt the feasibility of integrating these techniques.

Self-reflection, as we have defined in the earlier sections, has the potential to benefit from the prompt caching approach as we are frequently asking the model to reflect on past messages and revise the response. This is different to reasoning models such as Claude Sonnet 3.7, as their thought process is typically contained within the internal thinking tokens and not explicitly defined as sequences of messages in a conversation chain, preventing the use of prompt caching features available in Amazon Bedrock.

To analyse this trade-off further, we explore the differences in cost and latency across multiple rounds of self-reflection with and without leveraging the prompt caching feature in Amazon Bedrock. Figure 10 shows the cost and latency for a typical sequence of self-reflection rounds, with the model prompted to solve a Text-to-SQL question using an initial prompt size of approximately 1,000 tokens. Interestingly, Figure 10a shows that prompt caching combined with self-reflection has minimal benefits in terms of reducing the latency. We hypothesise that this could be due to the additional overhead of reading from cache databases being approximately equal to the latency required to generate the relatively minimal 100's of tokens. However, Figure 10b demonstrates that integrating self-reflection with prompt caching can being up to 28% reduction in cost when sampling over 3 rounds of reflection.

This method allows for more cost effective, linear scaling of self-reflection where only the incremental cost of additional output tokens is expensed with each round of reflection. For practitioners, it implies that leveraging self-reflection techniques can be more valuable with the LLMs and model providers that support prompt caching, as it offsets a significant part of additional costs on reflection rounds. We see potential for these techniques to grow in impact as more model providers enable improved prompt caching mechanisms in the future.



(a) Latency Trade-off for Prompt Caching and Self-reflection

(b) Cost Trade-off for Prompt Caching and Self-reflection. The percentage difference is the difference in mean cost.

Figure 10: Prompt Caching cost ($) and latency trade-off results for a sampled Text-to-SQL prompt, repeated over 3 distinct rounds of generation with the mean and variance shown as $\mu \pm \sigma$