

## A EXPERIMENTAL DETAILS

Figure 5 shows the performance of the final solution code as it is iteratively refined by the Auto-RCA framework over five rounds. The results demonstrate a dramatic improvement in problem-solving capability compared to the baseline direct application of LLMs. Notably, when using Gemini-2.5-Pro as the reasoning engine within the framework, the solution’s F1-score is elevated from an initial baseline of 0.5899 to an impressive **0.9179** in Round 1, which stands as the best overall performance achieved by any model in the framework. This substantial increase highlights the framework’s ability to significantly enhance problem-solving accuracy.

The iterative progress is evident across the rounds for various models. A score of 0.0000 indicates a round where the generated code was syntactically invalid or performed worse than the previous best, leading the Orchestrator to reject the change—a key feature of the system’s robustness that prevents performance regression and ensures only improvements are adopted.

A critical factor contributing to Gemini-2.5-Pro’s superior performance within the Auto-RCA framework is its significantly larger Max Tokens capacity (1M). This extensive context window allows Gemini-2.5-Pro to process a greater volume of effective information, including detailed problem descriptions, historical interactions, and intermediate reasoning steps, without truncation. In contrast, other models with smaller context windows (e.g., Claude-Sonnet-4 and Qwen3-235B, both at 64K tokens) are more susceptible to information loss due to truncation, which can hinder their ability to generate optimal code modifications and lead to suboptimal performance. This suggests that for complex, iterative code generation tasks, the ability to maintain a comprehensive context is paramount.

To further analyze the framework’s output, we assessed the best-performing solution code (from Gemini-2.5-Pro, F1-score 0.9179) across different scenario difficulties, as detailed in Table 4.

- **Simple Scenarios:** The solution achieved a near-perfect F1-score of 0.9540, demonstrating mastery over foundational problems with clear causal links.
- **Difficult Scenarios:** Performance dropped to an F1-score of 0.9158, revealing the current limits of the solution when faced with complex causal chains or ambiguous signals. The higher Recall (0.9763) compared to Precision (0.8920) suggests the model tends to identify most true positives but also flags some false positives, adopting an “aggressive” diagnostic strategy in complex cases.
- **Mixed Scenarios:** The F1-score of 0.9179 aligns closely with the overall iterative performance, confirming the benchmark’s balanced composition and the solution’s real-world effectiveness.

In summary, the Auto-RCA framework is a highly effective paradigm for complex reasoning tasks. It achieves a final F1-score of 0.9179 (Gemini-2.5-Pro), a substantial increase over the best baseline F1-Score@1 of 0.6254 (Qwen3-235B in Mixed scenarios). This validates our hypothesis that an autonomous, self-optimizing agentic architecture can effectively master domain-specific challenges like those in TN-RCA.

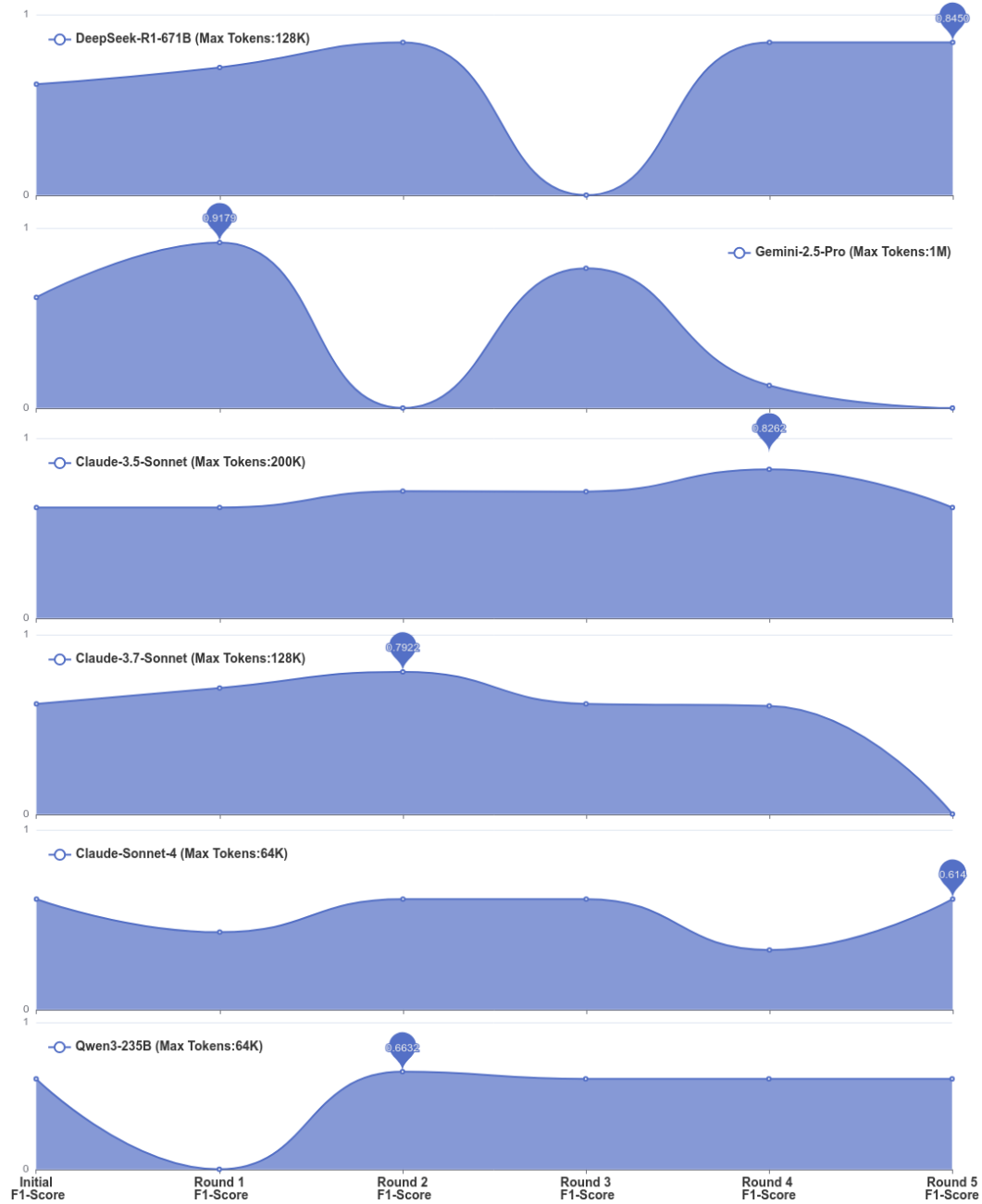


Figure 5: Performance Improvement on TN-RCA using the Auto-RCA Framework. The line chart shows the F1-score evolution of the best solution code across five refinement iterations. Markpoints indicate the highest F1-score achieved by each model on the mixed dataset. A score of 0.0000 signifies that a proposed code modification was rejected due to a performance decrease or an error, ensuring monotonic improvement.

Table 4: Performance Breakdown of Best Solution Code by Scenario Difficulty. The best-performing value in each column is highlighted in **bold**.

Model	Precision@1			Recall@1			F1-Score@1		
	Simple	Difficult	Mixed	Simple	Difficult	Mixed	Simple	Difficult	Mixed
DeepSeek-R1-671B	0.7832	0.8386	0.8115	0.8409	0.9627	0.9288	0.7997	0.8746	0.8450
Gemini-2.5-Pro	<b>0.9483</b>	<b>0.8920</b>	<b>0.8951</b>	<b>0.9828</b>	0.9763	0.9767	<b>0.9540</b>	<b>0.9158</b>	<b>0.9179</b>
Claude-3.5-Sonnet	0.9234	0.8547	0.8097	0.7453	0.9113	0.8633	0.8250	0.8721	0.8262
Claude-3.7-Sonnet	0.9310	0.8500	0.8544	0.7586	0.7660	0.7656	0.8161	0.7908	0.7922
Claude-Sonnet-4	<b>0.9483</b>	0.4407	0.4685	<b>0.9828</b>	<b>0.9823</b>	<b>0.9823</b>	<b>0.9540</b>	0.5943	0.6140
Qwen3-235B	0.9109	0.6950	0.6682	0.7839	0.6930	0.6626	0.8422	0.6923	0.6632