radium

METHOD
SAME-CLIENT, POINT-IN-TIME

STATUS
INTERNAL BENCHMARK

WORKLOAD
CODING ANALYSIS

TEST RUN
JUNE 23, 2026

Inference Benchmark Report

Radium Hal vs Claude Opus

Under identical test conditions, Radium Hal returned results at notably lower latency and higher sustained throughput than Claude Opus. Both providers achieved 100% success and quality parity, but the performance gap favors Radium decisively on speed.

The figures below are medians across a single coding-analysis run on June 23, 2026. Streaming was disabled, so time to first token was not measured. Every result is specific to this run and these test conditions.

MEDIAN LATENCY

LOWER IS BETTER

Radium Hal 1.0	9,865.83 ms
Claude Opus	38,310.27 ms

CLAUDE OPUS TOOK ABOUT 3.9× LONGER AT THE MEDIAN IN THIS RUN.

OUTPUT THROUGHPUT

HIGHER IS BETTER

Radium Hal 1.0	207.61 tok/s
Claude Opus	38.81 tok/s

RADIUM HAL SUSTAINED ABOUT 5.4× HIGHER THROUGHPUT IN THIS RUN.

EXACT-ANSWER QUALITY

HIGHER IS BETTER

Radium Hal 1.0	100%
Claude Opus	100%

PARITY ON THE BUILT-IN SANITY CHECK, WHICH IS NOT A FULL QUALITY EVALUATION.

SUCCESS RATE

HIGHER IS BETTER

Radium Hal 1.0	100%
Claude Opus	100%

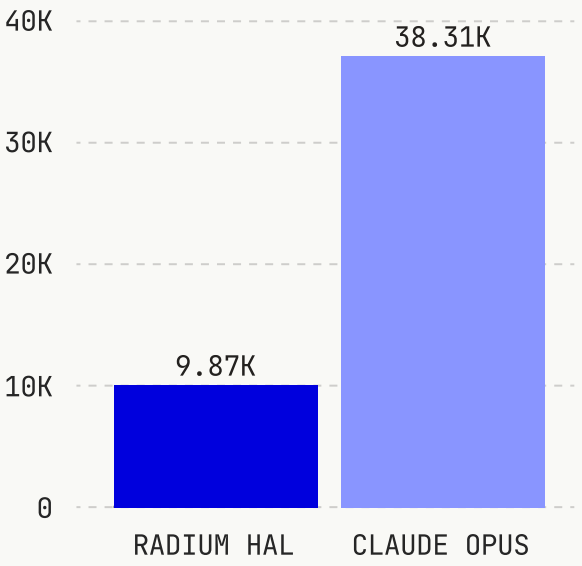
BOTH PROVIDERS COMPLETED THE WORKLOAD SUCCESSFULLY.

Both Radium Hal and Claude Opus were run through the same benchmark harness with the same prompt set, approximate input size, max token setting, and request format, so the comparison holds the client and the request fixed across providers.

SETTING	WHAT WAS HELD CONSTANT
Date	June 23, 2026
Client harness	The same benchmark harness for both providers
Prompt set	Identical prompt set across providers
Request parameters	Matched input size, max token setting, and request format
Streaming	Disabled, so time to first token was not measured
Providers	Radium Hal and Claude Opus, via their respective APIs
Reported values	Medians across the run

Median latency Milliseconds.

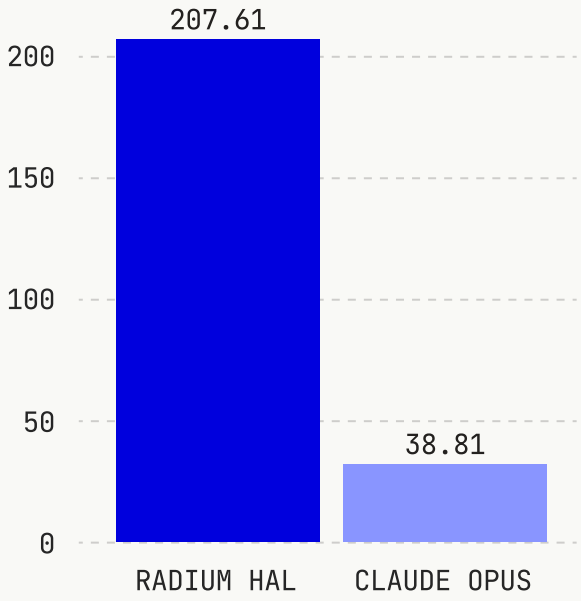
Lower is better.



Median end-to-end latency in milliseconds. Radium Hal completed the workload at a median of 9,865.83 ms against 38,310.27 ms for Claude Opus, about 3.9 times lower in this run.

Median output throughput

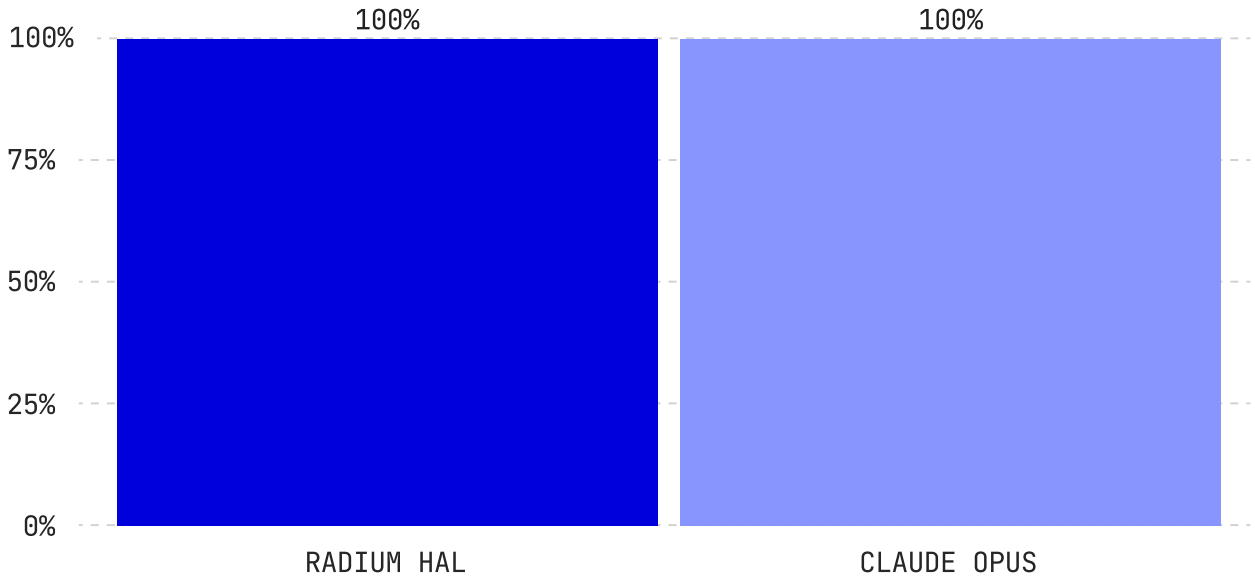
Output tokens / second. Higher is better.



Median output throughput in tokens per second. Radium Hal sustained 207.61 tokens per second against 38.81 for Claude Opus, about 5.4 times higher in this run.

Exact-answer quality check

Built-in coding and systems questions.
Higher is better.



Exact-answer match on the built-in coding and systems multiple-choice questions. Both providers scored 100 percent, which is a small sanity test rather than a substitute for human or workload-specific evaluation.

What this benchmark shows

This benchmark is a point-in-time, same-client comparison of API-level latency and output-token throughput on one coding-analysis workload. Both Radium Hal and Claude Opus were run through the same benchmark harness, with the same prompt set, approximate input size, max token setting, and request format. Under these test conditions, Radium Hal delivered lower latency and higher output-token throughput while both providers completed the workload successfully.

What this benchmark does not claim

This is not a comprehensive model-quality evaluation, independent benchmark, or guarantee of performance on every workload. Results may vary based on prompt type, input length, output length, network conditions, provider routing, region, system load, streaming settings, and model versions. The exact-answer quality check is a small sanity test, not a substitute for human evaluation or customer-specific testing. Customers should evaluate Radium on their own workloads.

Methodology note

Test run: June 23, 2026. Same prompt set and client harness used for both providers. Streaming was disabled, so TTFT was not measured. Results are from this specific benchmark run and may vary by workload and operating conditions.