



Efficient coding agent.

Build software efficiently with multiple strategies to save AI cost.

Three separate comparisons: long-context, short coding, and multi-file coding.

In each comparison, both sides run the same tasks on the same model.

01 Long-context coding tasks

02 Short coding tasks

03 Multi-file coding tasks

Three comparisons

Each row is a separate comparison. The same model is used on both sides.

Comparison	Completed	Input token savings (%)	Output token savings (%)	Model
Long-context coding tasks vs baseline agent	3 / 3 vs 3 / 3	31.1%	17.4%	qwen/qwen3.8-27b
Short coding tasks vs general agent	10 / 10 vs 10 / 10	58.0%	71.1%	qwen/qwen3.8-27b
Multi-file coding tasks vs general agent	15 / 15 vs 11 / 15	54.9%	18.8%	gemini-2.5-flash

Long-context coding tasks: 31.1% overall input-token savings (21,259 vs. 30,845) and 17.4% output-token savings (1,408 vs. 1,704).

Short coding tasks: 58.0% overall input-token savings (51,987 vs. 123,639) and 71.1% output-token savings (3,683 vs. 12,724).

Multi-file coding tasks: 54.9% input-token savings on matched runs and 18.8% overall output-token savings (18,527 vs. 22,815).

Long-context coding tasks

Baseline agent versus Wiyomi. Both completed every task and received a score of 5.

Task	Validated outcome	Baseline input	Wiyomi input	Baseline output	Wiyomi output	Input token savings (%)	Output token savings (%)
Landing page	Title, \$49, form, email, and phone	11,395	10,300	917	791	9.6%	13.7%
API client	POST /v1/leads and the required field	13,010	5,769	533	254	55.7%	52.3%
Status page	Three incidents displayed at the correct times	6,440	5,190	254	363	19.4%	-42.9%

Overall input-token savings: 31.1% (21,259 vs. 30,845).

Overall output-token savings: 17.4% (1,408 vs. 1,704). On the status-page task, Wiyomi used 363 vs. 254 output tokens (42.9% more).

Short coding tasks

General coding agent versus Wiyomi. Every code check passed.

Task	Validated outcome	Wiyomi input	Agent input	Wiyomi output	Agent output	Input token savings (%)	Output token savings (%)
Landing page	Title, \$49, form, email, and phone	5,790	16,135	1,051	6,015	64.1%	82.5%
API client	POST /v1/leads and the required field	6,482	14,151	403	1,013	54.2%	60.2%
Status page	Three incidents displayed at the correct times	4,472	13,057	289	1,275	65.8%	77.3%
Tax rate	tax(100) returns 7	5,726	12,649	293	842	54.7%	65.2%
Rename	rectangleArea(3, 4) returns 12	5,423	11,923	506	1,098	54.5%	53.9%
Prime test	All three bun tests pass	4,588	10,326	277	459	55.6%	39.7%
Clamp	Values remain between 0 and 100	5,576	12,235	360	886	54.4%	59.4%
Listen port	Server port is 4318	4,675	11,031	208	386	57.6%	46.1%
Crash line	Failure at store.ts:88	4,430	11,220	161	432	60.5%	62.7%
Catalog item	SKU-ORANGE-19 priced at 19	4,825	10,912	135	318	55.8%	57.5%

Overall input-token savings: 58.0% (51,987 vs. 123,639). Overall output-token savings: 71.1% (3,683 vs. 12,724).

Multi-file coding tasks, completion

Completed runs out of three. A run is complete when its code check passes.

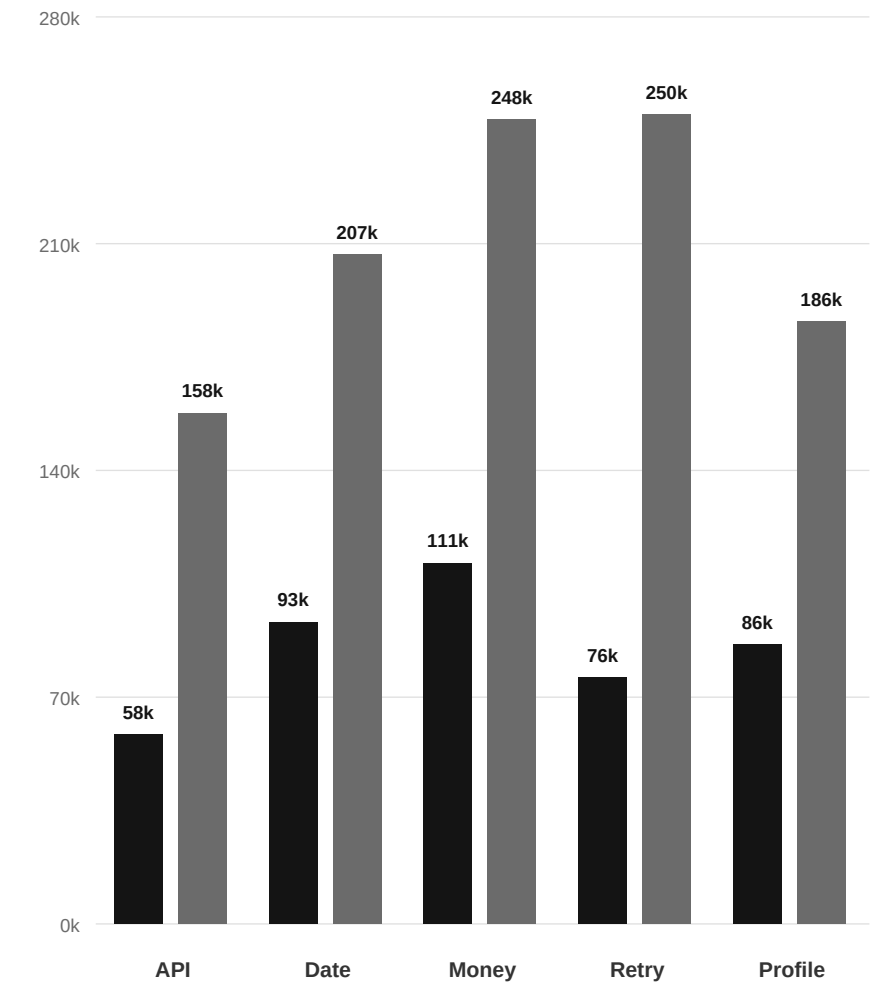
Wiyomi General agent



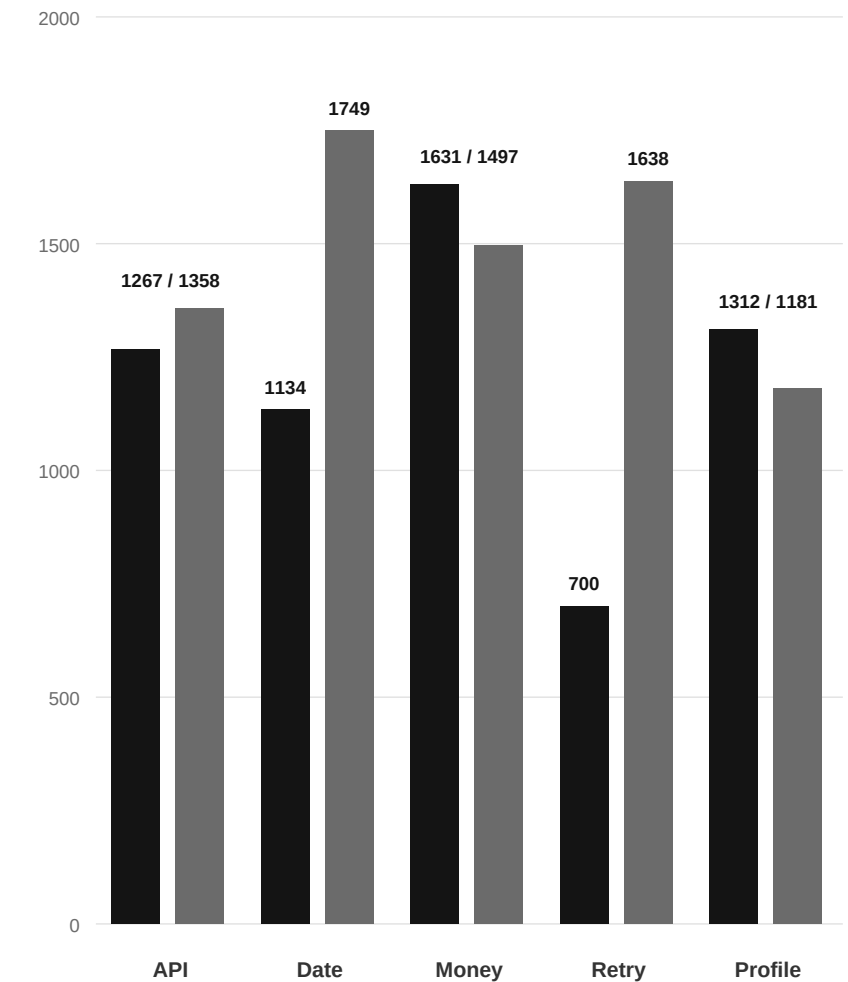
Multi-file coding tasks, tokens

Per-task medians. Black: Wiyomi. Gray: general coding agent.

Input



Output



Multi-file coding tasks, scorecard

Median input and output. Completion is shown as completed runs out of three.

Mission	Completed	Wiyomi input	Agent input	Wiyomi output	Agent output	Input token savings (%)	Output token savings (%)
API repair	3 / 3 vs 3 / 3	58,399	157,618	1,267	1,358	62.9%	6.7%
Date window	3 / 3 vs 2 / 3	93,130	206,589	1,134	1,749	39.7%	35.2%
Money rename	3 / 3 vs 1 / 3	111,291	248,315	1,631	1,497	41.2%	-9.0%
Retry policy	3 / 3 vs 2 / 3	75,999	249,819	700	1,638	58.6%	57.3%
Profile update	3 / 3 vs 3 / 3	86,165	185,942	1,312	1,181	56.6%	-11.1%

Input-token savings: 54.9% on matched runs. Overall output-token savings: 18.8% (18,527 vs. 22,815).

On money-rename and profile-update tasks, Wiyomi used 9.0% and 11.1% more output tokens, respectively.

WIYOMI

Efficient coding agent.

Build software efficiently with multiple strategies to save AI cost.