

REPORT BENCH 20_08 — Replica of September 1, 2026

OF V.9.11 vs OpenClaw 2026.8.1 · model deepseek-v4-flash · same machine (KittProxy), caches cleared, vanilla installs

Configuration

- **Platforms:** OF V.9.11 · OC 2026.8.1 (vanilla)
- **Model:** deepseek-v4-flash — reasoning: OF F1 low / F4 high · OC high
- **Identical pricing (\$/M):** input miss 0.14 · output 0.28 · cache read 0.028
- **Billed tokens:** new_in = total input – cacheRead (real delta at full price)
- **Method:** each test run on OF first, then OC (never in parallel); caches/sessions cleared after each test; T7 run on a working copy refreshed from the master (master never touched, md5 verified)

Suite (7 tests)

ID	Type	Request
T1	Knowledge	The god Janus for the Romans
T2	Preparation	University entrance exam
T3	Advisory	34 years old, €2,300 net, mortgage
T4	Creation	Web app for household expense management
T5	Creation	Online hat shop
T6	Creation	Program for an artisan building firm
T7	Repair	gestionale.py (14 tests)

Results

Test	Platform	Time (s)	new_in	output	cache	Cost \$	Result
T1	OF	14.9	8,835	1,446	9,856	0.001918	
T1	OC	12.9	30,257	668	0	0.004423	
T2	OF	12.0	1,034	1,306	13,568	0.000890	
T2	OC	13.6	25,212	662	2,560	0.003787	
T3	OF	23.6	1,214	2,260	13,568	0.001183	
T3	OC	51.9	13,753	4,548	8,320	0.003432	
T4	OF	21.8	5,798	3,269	8,832	0.001974	
T4	OC	116.7	0	13,875	308,096	0.012512	expected: <html, total
T5	OF	49.9	20,319	7,699	896	0.005025	
T5	OC	305.8	0	40,004	182,656	0.016315	timeout
T6	OF	62.1	10,391	12,099	4,864	0.004979	

T6	OC	230.9	0	23,238	1,366,528	0.044769	expected: client/agenda
T7	OF	66.3	49,441	8,867	5,632	0.009562	14/14
T7	OC	152.9	0	17,495	938,752	0.031184	14/14

Totals

Platform	Total time (s)	Total cost \$	Results
OF	250.6	0.025531	7/7
OC	884.7	0.116422	4/7 (T1-T3, T7)

Conclusions

- OF **3.5× faster** (250.6s vs 884.7s) and **4.6× cheaper** (\$0.0255 vs \$0.1164).
- **T4-T6 (creation)**: OF always delivers; OC does not deliver the requested files — same pattern as the original benchmark.
- **T7 (repair)**: both fix the working copy (14/14 green tests); master untouched (md5 verified).
- **OC cache very high on failed tests** (up to 1.37M tokens): repeated context in retries, billed at cache rate.

Note: Replica of the official benchmark of August 20, 2026 run on September 1, 2026 with updated versions: OF V.9.11 and OpenClaw 2026.8.1, same model (deepseek-v4-flash), same machine, caches cleared, vanilla installs.