

REPORT BENCH 20_08 — Replica 01/09/2026

OF V.9.11 vs OpenClaw 2026.8.1 · modello deepseek-v4-flash · stessa macchina (KittProxy), cache azzerate, installazioni vanilla

Configurazione

- **Piattaforme:** OF V.9.11 · OC 2026.8.1 (vanilla)
- **Modello:** deepseek-v4-flash — reasoning: OF F1 low / F4 high · OC high
- **Tariffe identiche (\$/M):** input miss 0,14 · output 0,28 · cache read 0,028
- **Token fatturati:** new_in = input totale – cacheRead (delta reale a prezzo pieno)
- **Metodo:** per ogni test OF poi OC (mai in parallelo); cache/sessioni azzerate dopo ogni test; T7 su copia di lavoro rinfrescata dal master (master mai toccato, md5 verificato)

Suite (7 test)

ID	Tipo	Richiesta
T1	Conoscenza	Il dio Giano per i romani
T2	Preparazione	Test di ammissione all'università
T3	Consulenza	34 anni, 2.300 € netti, mutuo
T4	Creazione	Web app gestione spesa casa
T5	Creazione	Negozio online di cappelli
T6	Creazione	Programma impresa edile artigiana
T7	Riparazione	gestionale.py (14 test)

Risultati

Test	Piattaforma	Tempo (s)	new_in	output	cache	Costo \$	Esito
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	atteso: <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	atteso: 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

Totali

Piattaforma	Tempo totale (s)	Costo totale \$	Esiti
OF	250,6	0,025531	7/7
OC	884,7	0,116422	4/7 (T1-T3, T7)

Conclusioni

- **OF 3,5× più veloce** (250,6s vs 884,7s) e **4,6× più economico** (0,0255\$ vs 0,1164\$).
- **T4-T6 (creazione):** OF consegna sempre; OC non consegna i file richiesti — stesso pattern del bench originale.
- **T7 (riparazione):** entrambe riparano la copia di lavoro (14/14 test verdi); **master intatto** (md5 verificato).
- Cache OC molto alta nei test falliti (fino a 1,37M token): contesto ripetuto nei retry, pagato a tariffa cache.

Replica del benchmark ufficiale del 20/08/2026 eseguita il 01/09/2026 con le versioni aggiornate: OF V.9.11 e OpenClaw 2026.8.1, stesso modello (deepseek-v4-flash), stessa macchina, cache azzerate, installazioni vanilla.