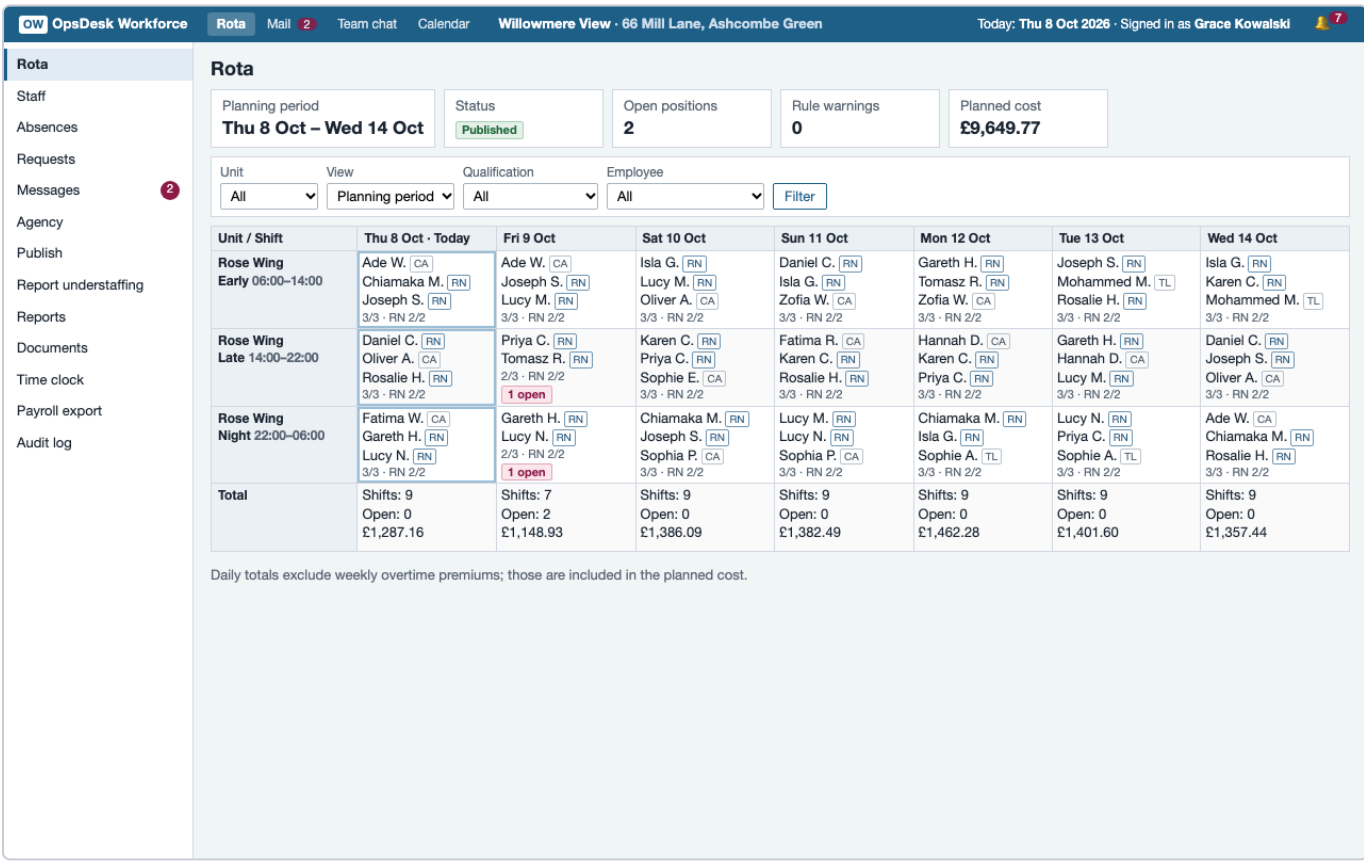
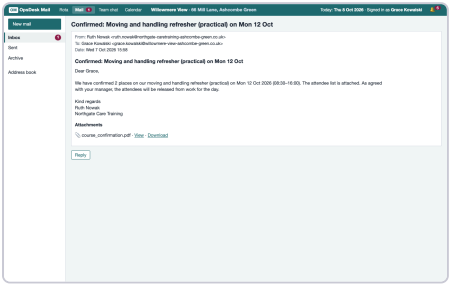
OpsDesk Workforce

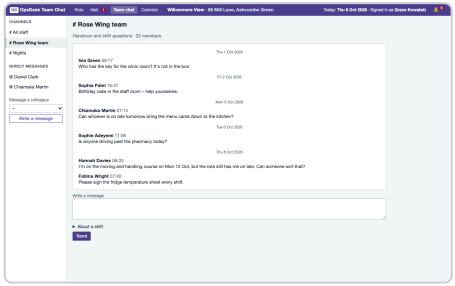
A self-hosted replica of care-home rota and staffing software, with a mail client, a team chat and a calendar on the same clock. Computer-use and tool-use agents get natural requests from the home manager and work across the apps; a deterministic verifier scores the final state. No LLM judge.



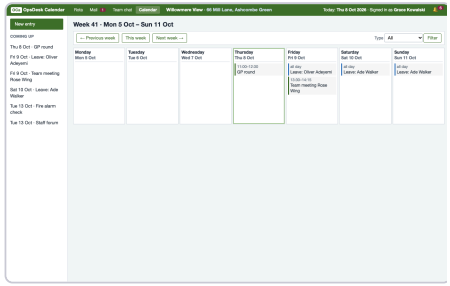
The rota of a UK home (English edition). German homes run the German edition with € and the Arbeitszeitgesetz.



Mail: a training provider confirms a course, attendee list as PDF.



Team chat: a sick call and a question nobody has recorded yet.



Calendar: course days, meetings and inspections.

What the agent does

- Builds a rota from scratch, or brings its cost down
- Repairs it after sick calls, covers extra demand, books agency cover within a budget
- Decides swap requests against the house rules and asks colleagues in the chat
- Answers the payroll office's query from the time clock and their CSV
- Fixes a works-council rejection (German homes); resolves conflicting documents
- Plans course days from the provider's mail into the calendar and the rota
- Covers what is legal in a period that cannot be fully staffed and reports the rest

A workspace that answers back

- **4 apps, one state:** rota (14 screens), mail, team chat, calendar. Information for a request is spread across them.
- **Colleagues reply by fixed rules:** staff, home manager, agencies and the payroll office answer and decide; a swap partner confirms in the staff app when asked in the chat. No model at run time.
- **Things change mid-task:** a second sick call after the first gap is filled, a swap request withdrawn.
- **GUI and tools:** browser for computer use; HTTP and MCP for tool use; both produce identical states.

A request, as the agent gets it

> We had sick calls overnight – they're in the messages. Could you get those shifts covered again? See Northgate Care Training's mail: two carers are on a course on Mon 12 Oct. Please add it to the calendar with everyone who's going and take them off the rota for the day. Let anyone whose shift changes know. Everything has to stay within the rules. Don't reshuffle more than you have to.

How every result is checked

Check	Rule
Gates	Valid final state; no hard-rule breach (UK Working Time Regulations or German Arbeitszeitgesetz, house rules, qualifications, absences); no collateral damage to unrelated records; coverage complete, or the proven maximum where full cover is impossible. Any failed gate scores 0.
Cost	Against a frozen CP-SAT reference: full credit at or below it, zero at +10 %.
Stability	No more changes outside forced events than the reference needs (repair requests).
Messages	Every person whose shift changes is told; the payroll reply lists every difference to the minute; the course is in the calendar with every attendee.
Honesty	What the agent states through the app (publishing, the understaffing report) must match the final state. Any false claim scores 0.

Outcome-based: any correct route gets full marks, and the same final state always gets the same score.

Quality evidence (scripted agents, no model)

Check	Result
Gold agent	**1.0 on every request via the API, the GUI forms and a real browser**
Cheating agents (no-op, random clicker, two liars, destructive, shortcut, rule breaker)	**0 on every request**
False accepts: near-miss solutions with exactly one thing wrong	**0 of 557 got full credit (highest 0.973)**
False rejects: equivalent solutions by another route or wording	**0 of 467 rejected**
Verifier audit	BenchJack and ABC checklists answered item by item; 10 hack probes passed on every request
Route independence	most requests have at least 2 other optimal rotas scoring 1.0; another action order reaches 1.0 on every request
GUI ≡ API, determinism, replay	identical state on every request; repeated runs identical; exported action logs reproduce the final state
Reset and scale	median reset 10.4 ms; 40 containers on one Mac ready in 8.1 s, 42 MiB each
Eval signals	0 hits across everything the agent can see

The request set

- 10 families, UK and German homes; every request verified with the proofs above
- Long-horizon tier over 2–4 weeks, up to 462 gold actions
- About half the requests need two or more apps
- Easy to expert; 17–76 staff per home
- Fresh, non-overlapping sets per order

Formats and data

- Docker image, runs offline; snapshots, reset and a JSONL action log
- Harbor, HUD (computer use and MCP), OpenEnv, verifiers-style task list, HTTP, MCP
- Fully synthetic data: no real personal data, no third-party assets
- References stay host-side; licences for evaluation or training

OpsDesk Workforce 3.1.3. Every number comes from proof runs with scripted agents; model results follow. Fresh, non-overlapping sets on demand: 100 new requests generated and fully verified in under 40 minutes on one machine. A sample and a 72-second demo video are available on request. getvencur.com