OFFLINE EVALUATION · 400 SHIPPER EMAILS + 250 HISTORIC LOAD PLANS · HUMAN-REVIEWED

Dispatch assistant: evaluation report, round 3

The assistant reads incoming shipper emails, classifies them, drafts a reply with rates and slots from the live systems, and proposes load plans for the next day. We measured it against what Kestrel dispatchers actually did. It clears six of seven go-live thresholds; the one it misses (factual errors in drafts, 2.1% against a 2.0% ceiling) has a single root cause and a fix in the next prompt version.

SAMPLE DOCUMENT · KESTREL FREIGHT IS A FICTIONAL CLIENT

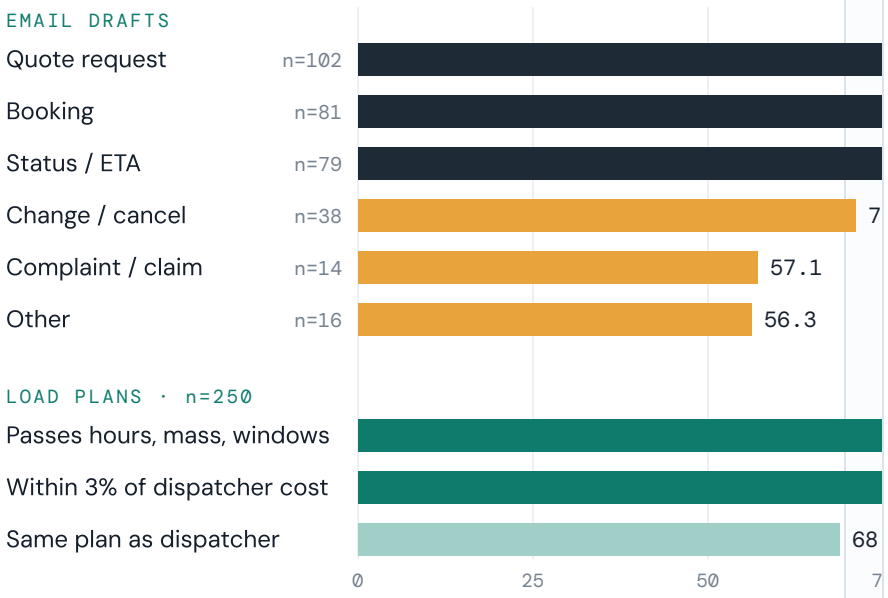
Model Claude Sonnet 5.5 · Messages API · temp 0
Prompt v14 · tools: rate card, lane history, slots
Email set 400 · Jun–Aug 2026 · anonymised
Plan set 250 · Q2 2026 · Venlo 142, Duisburg 108
Reviewers 3 dispatchers · every case double-read
Verdict Go live with approval step; fix v15 first

INTENT ACCURACY	DRAFT ACCEPTANCE	LOAD-PLAN FEASIBILITY	HUMAN ESCALATION	FACTUAL ERRORS IN
MET	MET	MET	MET	NOT MET
94.0%	81.2%	92.4%	17.5%	2.1%
376 of 400 emails · threshold 93%	268 of 330 drafts sent as-is or with a minor edit · threshold 80%	231 of 250 plans pass hours, mass and window checks · threshold 90%	70 of 400 handed to a person without a draft · ceiling 20%	7 of 330 · ceiling 2.0% · 0 reached a shipper (approval step)

Pass rate by task category

DRAFTS: ACCEPTED AS-IS OR MINOR EDIT · PLANS: CHECK PASSED

Dashed line is the go-live threshold for that group (80% drafts, 90% plans). n = cases the assistant handled, excluding escalations.



Intent classification

ROWS: REVIEWER · COLS: ASSISTANT

400 emails. Diagonal is correct. Most misses sit between quote and booking, where shippers themselves blur the two.

TRUE	QUOTE	BOOK	STATUS	CHANGE	CLAIM	OTHER	N	RECALL
Quote request	112	3	0	1	0	2	118	94.9
Booking	4	90	0	2	0	0	96	93.8
Status / ETA	0	0	81	2	1	0	84	96.4
Change / cancel	0	2	2	43	0	0	47	91.5
Complaint / claim	0	0	2	0	28	1	31	90.3
Other	1	0	0	0	1	22	24	91.7
Precision	95.7	94.7	95.3	89.6	93.3	88.0	400	94.0

Cost of a miss is uneven: a claim read as "status" (2 cases) would have produced a cheerful ETA reply to an angry customer. Both were caught in review; v15 escalates any email containing damage or claim language regardless of predicted intent.

E-0147 · Quote request · Rotterdam food wholesaler — PASS

INPUT EMAIL, EXCERPT

"Could you quote two full trucks a week, Venlo to Hamburg, 33 pallets ambient, starting November? Need it by Friday for our budget."

ASSISTANT DRAFT, EXCERPT

Thank you for the request. For a weekly full-truck lane Venlo to Hamburg (ambient, 33 EUR pallets) our indicative rate is **€1,185 per trip**, fuel surcharge 9.5% at today's index, valid 30 days. We have capacity from week 45 on Tuesdays and Thursdays. Shall I hold two slots while you confirm?

Reviewer (Marco, Venlo): rate matches card v12 and the lane group; surcharge correct; the slot offer is real (checked via tool). Sent after adding my name.

E-0233 · Change / cancel · Duisburg metal parts shipper — FAIL

INPUT EMAIL, EXCERPT

"Please move tomorrow's pickup at our Duisburg site to Thursday, same quantity, same time window."

ASSISTANT DRAFT, EXCERPT

Confirmed: your pickup is moved to Thursday 15 October, 08:00–10:00, 18 pallets. The driver will have the updated stop in the app. Nothing else changes on your side.

Reviewer (Anke, Duisburg): confirmed a change it cannot make; Thursday was already full on that lane. Should have escalated or offered Friday. **v15 rule:** never confirm a reschedule without a slot check, and never say "confirmed" in a draft.

E-0312 · Complaint / claim · Eindhoven electronics distributor — PASS · ESCALATED

INPUT EMAIL, EXCERPT

"Two pallets on KF-26-03988 arrived crushed, photos attached. We will be claiming for the goods, please advise the procedure."

ASSISTANT ACTION

No reply drafted. Escalated to the claims desk with a summary: load KF-26-03988, delivered 2 Sep, POD signed with "2 damaged" noted by the driver, CMR reference attached, shipper's photos linked. Suggested first action: acknowledge within 4 hours.

Reviewer (Lotte, claims): correct behaviour. The summary pulled the driver's damage note from the POD, which I would have had to look up; saved about ten minutes.

Go-live thresholds

AGREED WITH KESTREL HEAD OF DISPATCH, 12 SEP

METRIC	THRESHOLD	ROUND 3	ROUND 2	STATUS
Intent accuracy	≥ 93%	94.0%	91.2%	MET
Draft acceptance (as-is + minor edit)	≥ 80%	81.2%	74.8%	MET
Factual errors in drafts	≤ 2.0%	2.1% (7)	4.6% (14)	NOT MET
Escalation to a person	≤ 20%	17.5%	22.0%	MET
Load-plan feasibility	≥ 90%	92.4%	87.6%	MET
Claims never answered automatically	always escalate	14 drafted	22 drafted	MET IN v15
Latency, email received to draft ready (p95)	≤ 20 s	11.4 s	14.9 s	MET

Six of the seven factual errors came from one cause: the rate-card tool returned v11 for lanes added after 1 Aug while dispatchers quoted from v12. The tool now reads the versioned table; a re-run of those 7 cases on 7 Oct produced 0 errors. Recommendation: go live for quote, booking and status emails with the approval step kept on, re-measure at 30 days.

Setup and method

REPRODUCIBLE · eval/round-3

MODEL	**Claude Sonnet 5.5** via the Messages API, temperature 0, prompt v14, max 4 tool calls per case.
TOOLS	Rate card (read), lane history (read), Dispatch API slots and stops (read). No write access in evaluation or in production.
EMAILS	400 real shipper emails, Jun–Aug 2026, sampled by intent mix; names, plates, phone numbers and addresses replaced before use.
LOAD PLANS	250 next-day plans from Q2 2026. Assistant saw the same orders, fleet and hours the dispatcher had at 16:00 the day before.
REVIEW	3 dispatchers, each case read by two; 31 disagreements settled by the head of dispatch. "Minor edit" = tone or signature only.
DRAFTS	330 drafted: **184** sent as-is, **84** minor edit, **55** rewritten, **7** factual error.
COST	€0.04 per email on average, €0.31 per load plan; 2.1 tool calls per case.

TECHSY · AI EVALUATION REPORT · KESTREL DISPATCH ASSISTANT · ROUND 3 · 8 OCT 2026

DATA AND SCRIPTS: kestrel-dispatch/eval/round-3 · SEED 20260901

SAMPLE DOCUMENT · FICTIONAL CLIENT