AI Call Center Supervisor in a Live Company: Technical Report

Companion report to the article "An AI Supervisor Ran a Call Center Team for Three Weeks. Then Someone Checked Its Weekly Report". Case study under NDA.

Evgenii Arsentev, MD, PhD · Founder & CEO, Dad4Kids · ORCID 0000-0002-9120-7298

Technical report, version 1.0, November 2026 · DOI 10.5281/zenodo.23926851 · CC BY 4.0 · https://arsentev.ai/research/ai-call-center-supervisor

Cite as: Arsentev, E. (2026). *AI Call Center Supervisor in a Live Company: A Three-Week Case Study of an Autonomous AI Agent Supervising a Phone Team* (Technical report). Zenodo. https://doi.org/10.5281/zenodo.23926851

---

1. Scope

Setting. A US-facing telemedicine company runs a small remote team of account managers (callers, 4–8 people in a given week) who phone partner physicians and other partners from a CRM with a built-in auto-dialer. From day 1 of the observation window, an autonomous Claude agent acts as the team's supervisor. Humans stay in the loop: a human team supervisor, a department head, and company leadership, which writes the agent's rules.

What the agent does (as logged). Scheduled runs (morning plan, three daytime pulses, a nightly review; from the end of week 2 also a floor check every 30 minutes during calling hours), plus a run for every chat message addressed to it. It reads call records and call classifications, posts to the team chat, tracks open issues, escalates to humans and writes a weekly report.

What the agent may not do (its written rules, from week 2). Hire or let go of people. It may recommend; the decision belongs to a human.

Windows.

- Before the agent: weeks −4 to −1 (four full ISO weeks).
- With the agent: week 1 (the agent starts on Thursday of this week), week 2, week 3 (Monday–Friday; data cut on Friday evening, local time).
- No calendar dates are disclosed.

Unit of analysis. A call record in the CRM created by a caller and linked to a partner. "Answered" = not marked missed. Duration = CRM call duration in seconds. Weeks are ISO weeks in the team's local time zone. The main series excludes the human supervisor's account (see A3).

Data sources (read-only, extracted on the last day of the window):

- S1 — CRM call records (one row per call).

- S2 — automated per-call classification of answered calls (doctor / other person / silence / voicemail / screening robot / phone menu; plus a yes/no "real conversation" flag). Exists only for week 3; long calls from Monday, short calls from Tuesday.
- S3 — call transcripts stored in the CRM.
- S4 — the agent's run logs.
- S5 — the agent's own work log and open-issues file; the team chat log available to the agent.
- S6 — CRM user accounts (creation date, active flag) and the CRM partner table (assigned owner, funnel transition log).

2. What the agent did: activity counts

metric	value	source
model runs, 16 days	459	S4
— triggered by chat messages	≈279	S4
— floor checks (from end of week 2)	72	S4
— structured interview of the human supervisor about the dialer (week 2)	44	S4
— daytime pulses	31	S4
— nightly reviews / morning plans	13 / 11	S4
— interview briefs	9	S4
runs that did no work (all on Monday of week 3)	80 of 84 that day	S4
chat messages available to the agent (team chats and direct)	584	S5
— addressed to the agent	272	S5

Per-mode counts sum to 459 ± 1 (one run's mode label is ambiguous in the log).

3. Team output before and with the agent

3.1 Weekly series (callers only, partner calls)

week	callers active*	dials	answered	answered ≥60 s	talk time, h	partners reached	auto-dialed share
−4	2	1,386	192	76	9.0	141	87%
−3	4	1,963	505	254	22.6	336	78%
−2	5	3,307	666	298	29.2	479	78%
−1	6	2,044	532	254	27.6	398	70%
1 (agent from Thu)	8	1,054	422	220	25.9	307	21%
2	4	426	241	135	19.7	162	5%
3 (Mon–Fri)	6	3,208	965	417	52.5	575	84%

* callers with at least one partner call that week. Source: S1. Formulas: talk time = Σ duration of answered calls / 3600. Auto-dialed share = auto-dialed records / all records.

Derived figures used in the article.

- Talk time, week 3 vs best week before the agent (week −2): 52.5 / 29.2 = 1.80 → **+80%**.
- Same comparison including the human supervisor's account (A3): 54.1 / 40.8 = 1.33 → **+33%**. The article reports the range 33–80%.
- Talk time, week 3 vs average of weeks −4…−1 (22.1 h): 52.5 / 22.1 = **2.4×**.
- Answered calls ≥60 s, week 3 vs week −2: 417 / 298 = 1.40 → **+40%**.
- Talk time per active caller: week −2 29.2 / 5 = 5.8 h; week 3 52.5 / 6 = 8.75 h → **+50%**.
- Weeks 1 and 2 with the agent were **below** the best week before it (25.9 h and 19.7 h vs 29.2 h).
- The two callers who worked full time in both weeks 2 and 3: talk time 1,144 → 2,024 min → **+77%** (S1, per-caller sums; this per-caller extraction was taken later on the same evening and sums to 52.9 h for week 3, vs 52.5 h in the weekly table).

3.2 Daily dials in week 3 (all partner calls incl. the human supervisor's account)

Mon	Tue	Wed	Thu	Fri (to evening)
269	453	798	1,055	657

Source: S1. Thursday / Monday = 3.9×.

3.3 What changed in week 3 at the same time (confounders)

1. The auto-dialer was back in use: auto-dialed share 5% (week 2) → 84% (week 3).
2. Headcount went from 4 to 6 active callers; three new callers made their first calls on Tuesday–Wednesday.
3. A call-screening setting on one caller's phone was turned off (Monday, §4.1).
4. The dialer's connection order was changed (Thursday afternoon, §4.2).
5. A rule against redialing the same number within 3 hours and a rule against rapid skipping were introduced.
6. The agent did no work for most of Monday (§2).

The design has no control group. **The week-3 increase cannot be attributed to the agent,** and the article does not do so.

4. Two defects, measured before and after

4.1 Call screening on a caller's phone (Monday of week 3)

One caller's phone had a call-screening feature turned on. Physicians who answered heard an automated assistant asking for their name and reason for calling, and many hung up. The CRM still recorded these as answered calls.

Detection rule (S3): the transcript contains one of the screening assistant's standard phrases (asking the caller to state their name or the reason for the call).

that caller, Monday	answered (all auto-dialed)	screening robot in transcript	answered ≥60 s	share ≥60 s
before the fix	54	47 (87%)	15	28%
after the fix (same afternoon)	18	0	9	50%

Other callers that day: 0 robot transcripts. The same caller on Tuesday–Friday: one or two a day.

- "Result doubled" holds for the **share** of answered calls lasting ≥60 s (28% → 50%, ×1.8).
- It does **not** hold for long calls per hour, which stayed about flat (≈2.8/h → ≈2.6/h; hours approximated from first and last call of each part of the day).
- The agent's report quoted "45 of 51"; a time-stamped log entry made when the problem was found reads "40 of 51". Both were taken before all transcripts were processed; the final count is 47 of 54.
- Who found it: a company executive reviewing transcripts with a separate AI session, not the supervisor agent, which was not running at that time (S4, S5).

4.2 Silence at the start of auto-dialed calls (Thursday of week 3)

In the original dialer mode the system rang the physician first and connected the caller after the physician picked up. Physicians heard 10–15 s of silence and many hung up. On Thursday afternoon the order was reversed (the caller is connected first).

Metric (S1 + S2): **real doctor conversations** = calls classified as "doctor" with the "real conversation" flag set. Share = real doctor conversations / all partner dials.

period	dials	classified	real doctor conversations	share of dials
Tuesday (old mode)	453	124	78	17.2%
Wednesday + Thursday before the change	1,603	350	134	**8.4%**
Thursday after the change + Friday	907	240	146	**16.1%**
Friday alone	657	164	97	14.8%

- "Silence" among classified calls: 16.8% Thursday morning → 1.3% Thursday afternoon; 6.1% on Friday.
- Absolute check, Thursday only: 47 real doctor conversations before the change vs 49 after, in calling windows of ≈6.3 h and ≈3.7 h → ≈7.5/h → ≈13/h (window lengths estimated).
- Caveats: (a) Tuesday, in the old mode, was already at 17.2%, so the change restored the earlier level rather than doubling a stable baseline; (b) after the change, dial volume per hour fell, which raises any per-dial share; (c) after the change, a call where the caller's own phone did not pick up is not logged, so the denominator shrinks; (d) other rules changed in the same 48 hours (§3.3); (e) the classifier was not validated against human labels, and its coverage of answered calls varies between 77% and 92% across the compared periods.

- Who found it: logged as a leadership change; the supervisor agent's logs do not contain the finding. In week 2 the agent ran a 44-session structured interview with the human supervisor about the dialer; it listed blind skipping, duplicate calls and dead numbers, not the silence.

4.3 Issues the agent raised (S5)

week	what the agent raised	follow-up
1	a promised callback not made	posted to the caller
2	an account-provisioning error affecting a new hire	corrected the same day by a human
3	the auto-dialer ending sessions in the middle of live calls (also reported by the department head)	escalated to humans
3	8 callbacks promised by one caller and not made, found from call content at the department head's request	rule added the same day: the callback time from the call summary goes into the dialer queue
3	access of people who had left still to be revoked	escalated 3 times; not complete when the weekly report was checked

5. Who made the people decisions

while the agent was running (weeks 1–3)	count
callers who left	5
— let go by human decision (leadership, human supervisor, department head)	5
— let go by the agent	0
— recommendations to let someone go found in the agent's logs	0
new caller accounts created	10 (3 of them for one-day trials)

Five more callers whose last call was before the agent started are counted as pre-agent departures. Sources: S6 (creation dates; leave dates reconstructed from last call and S5 records, because the CRM keeps no deactivation history), S5 (who decided, as written in the logs; searched for dismissal-related phrases). The agent's nightly review includes a place for performance concerns; the logs contain one entry stating there were no dismissal candidates; every departure was decided by humans. Because the agent made no recommendation, its "recommend, never decide" rule was not tested.

6. Audit of the agent's weekly report (end of week 3)

The agent's report made nine checkable claims. Each was recomputed from S1–S6.

#	claim in the report	what the data shows	verdict
1	"Live conversation up 41%"	no calculation found in the agent's files; nearest: answered ≥60 s +40% vs week −2; talk time +80% vs week −2, +166% vs week 2	not reproduced
2	"Calls nearly quadrupled, 269 → 1,057"	daily, all accounts, lowest to highest day: Monday 269 → Thursday 1,055 (S1, 2 records fewer)	holds (a daily, not weekly, figure)
3	"4 people let go; all operator accounts cleaned up"	4 departures in week 3 — correct; access cleanup not complete at the time of the check	partly false
4	"A new caller is the team's best at conversation today"	on Friday, most long (≥60 s) answered calls, but not the most talk time (S1, per-caller daily query)	holds under one metric
5	"A second new caller shows good results"	no metric given	not testable
6	"Three candidates start trial days; all scored 90/100 in the AI interview"	the score is a fixed label stored for a verdict, not a measurement	misleading
7	"Call screening: 45 of 51 calls ended in nothing; after the fix the result doubled"	47 of 54 after all transcripts were processed; share ≥60 s ×1.8; long calls per hour flat	holds under one way of counting
8	"Dialer fix: real conversations 8% → 15%"	reproduced only with: denominator = all dials, "before" = Wed + Thu morning, "after" = Friday; Tuesday was 17%	holds under one way of counting
9	"N partners had no owner at the start of the week; far fewer now" (exact figures withheld: they size the partner base)	the starting figure was not found in any table and cannot be reconciled with the assignment log	not reproduced

Tally: holds as stated 1 (#2); holds only under one way of counting 3 (#4, #7, #8); partly false or misleading 2 (#3, #6); not reproduced or not testable 3 (#1, #5, #9).

Claim 9 detail. Even if every assignment logged during the week had come from the unassigned pool (and net of returns to it), the start-of-week figure would still be far below the one claimed.

7. The humans who check the agent (S5)

- In week 3 the department head addressed most of their team-chat messages to the agent.
- On Monday, one message gave the agent three standing tasks: alert when a caller's dialer is paused over 15 minutes, alert after 30 minutes without CRM activity, send a short evening report.
- On Wednesday the department head noted that the CRM's follow-up dates were useless for one caller (99% of that caller's partners showed as overdue) and asked the agent to find, from the calls, which promised callbacks had not been made. The agent found 8 (§4.3).
- On Friday the department head passed the agent a dialer bug with reproduction steps, reported independently by two callers.

8. Assumptions and limitations

- A1. No control group; before/after only. Week 1 is a transition week, week 2 had a headcount drop (4 callers), week 3 is Monday–Friday only.
- A2. Call classification (S2) exists only in week 3 and was not validated against human labels; coverage of answered calls is 77–92% across compared periods. No pre-agent comparison of conversation quality is possible, only of durations.
- A3. The human supervisor's account placed several hundred dials a week in weeks −4…−2 and almost none afterwards. Including it changes the pre-agent baseline (§2: +33% instead of +80%). The main series excludes it.
- A4. After the Thursday dialer change, unanswered caller legs are not logged; per-dial shares after the change are biased upward.
- A5. Departure dates are reconstructed; the CRM keeps no history of account deactivation.
- A6. Small numbers: the team is 4–8 people, so one person joining or leaving visibly moves weekly totals.
- A7. Business outcomes are not reported (NDA); the window is too short to judge them.

9. What is disclosed and what is not

Disclosed: counts and shares in this report, formulas, how call screening was detected, the agent's role and its written limits on people decisions, the audit of its weekly report.

Not disclosed: company name and market details; calendar dates; names or any identifying details of callers, candidates, physicians or partners; locations; business results; the agent's prompts, rule texts and technical setup beyond "Claude agents"; other vendors and internal operations; subscription prices. Raw data and scripts stay with the company.