sturdy

WHITEPAPER · STURDY × CLAUDE

After the health score

A five-dimension health diagnostic built from what your customers actually said, across every channel, with the evidence attached.

BEFORE YOU START

What we believe we have built

We believe we have built a health diagnostic that anyone in customer success can defend, whether that is a CSM preparing for a renewal call or an executive walking a board through the at-risk list. It runs as a project inside Claude, on top of Sturdy's canonically organized language data pipeline, delivered over MCP. It reads the complete communication record for a single account, across every channel you talk to customers on, and returns five graded dimensions, a composite verdict, and the specific exchanges that produced each grade.

This paper is for current and prospective Sturdy users.

The first half explains what the diagnostic is, what it is not, and why the design departs from how health scoring is normally done. The second half is the setup guide. The project instructions and template file are prebuilt, and you can have this running in your own Claude environment in under a day.

01

Today's model: a number nobody can defend

An account is at 68. Last month it was 74.

Nobody in the room can say what moved. The CSM thinks it might be the support tickets from the spring. The ops lead thinks a survey response landed. Somebody suggests the login counts dipped when the customer's team was on holiday. The number is treated as authoritative and discussed as though it were a rumor.

The meeting ends without a decision, because a six-point drop does not imply an action. Nobody knows whether to call the champion, escalate the integration, or do nothing. So the most common outcome is that the score gets noted and the account gets managed on instinct anyway.

This is the failure mode worth naming precisely. The problem is not that the number is wrong. The problem is that the number is **unfalsifiable and unactionable at the same time**. You cannot check it, and you cannot act on it.

02

What a CSP health score is actually made of

A customer success platform can produce a health score. The question is what goes into it.

In practice, most scores are a weighted blend of four inputs:

Product usage. Logins, seats active, feature adoption events. This measures whether people open the software. It does not measure whether they are getting value from it, and the two diverge constantly. A team can log in daily and be actively frustrated. A team can log in rarely because the product quietly works.

Support ticket volume. Treated as a negative signal, which inverts the truth surprisingly often. Engaged customers file tickets. Customers who have given up file nothing.

Survey responses. NPS and CSAT, with response rates that typically sit in the single digits to low teens, from a self-selected minority, arriving weeks after the sentiment formed. A quarterly survey cannot see a relationship deteriorating in March.

Manual CSM sentiment. A dropdown a human sets. It is the most informed input in the blend and the least objective. CSMs are systematically optimistic about their own book, which is a feature of being human and accountable, not a failure of diligence.

Then those four get multiplied by weights, and this is where the construction quietly falls apart. The weights are arbitrary. They were set during implementation, by judgment, without evidence about which inputs actually move account health for your business. Does usage deserve forty percent of the score or fifteen? Nobody ran that analysis. Nobody could have, because the inputs themselves are proxies. And once set, the weights are almost never revisited, so the aggregate becomes something no one can consistently rely on: a precise-looking blend of unvalidated guesses.

There is a second disconnect layered on top of the first. In most organizations, the person who configures the health score is on the operations team. They are skilled at the platform and they build what the platform makes buildable. But they are not in the account conversations every day, so the model of health gets defined by the person furthest from the customer, and the people closest to the customer inherit a score they had no hand in shaping and quietly stop trusting.

Notice what is absent from the inputs. **Nothing in the blend is the customer's own words.** The pricing dispute that ran for three months, the champion who stopped replying, the security review that has blocked your integration since February, the sentence where someone wrote "we still are not seeing the value we expected": none of that is in the score. It is in the email threads, the tickets, and the call transcripts, and the score never read any of them.

> **A health score built from metrics instead of interactions is an inference about the customer, not a reading of the customer. It is a manually configured facade that can drift arbitrarily far from the real story, and nothing in the score itself tells you when it has.**

THEN THERE IS THE PRICE

Health scoring is usually the anchor feature that justifies a per-seat customer success platform. You are paying platform pricing for a composite whose inputs you cannot inspect, whose weights you cannot defend, and whose output you cannot trace back to a sentence a customer wrote. The scoring model is the vendor's, not yours, and it applies whether or not it reflects how your business actually loses customers.

03

What we refused to build

We did not build a score.

There is no 0-to-100 axis in this artifact, no weighted composite you could put in a spreadsheet. The output is five graded dimensions and one composite verdict, each expressed as a color with a written rationale and the exchanges that produced it.

Think of it as warning lights instead of a gauge. A gauge invites false precision: 68 and 71 are not meaningfully different, but they look different, and people act on the difference. A warning light makes a claim exactly as strong as the evidence supports. Something is wrong here, in this system, and here is what tripped it. That is what an ordinal grade is: honest about its own resolution, and pointed at a subsystem rather than at the whole car.

The dimensional structure matters for the same reason. A single blended number hides its own composition: two accounts at the same score can require opposite interventions, one with an unhappy customer and one with a support backlog, and averaging is precisely the operation that destroys that distinction. Five dimensions shown side by side cannot hide it.

And a set of graded dimensions can be argued with, which sounds like a weakness and is the strength. When a CSM disagrees with a color, they open the evidence, and one of two things happens. Either the CSM learns something about their account, or the diagnostic is wrong and the rule that produced the grade gets fixed. Fixing it

is concrete: every grade traces to a written threshold in the project instructions, so a wrong grade points at a specific line you can read, adjust, and re-run. Section 07 covers how that tuning works. Neither outcome, learning or fixing, is available from an integer.

04

What changed: reading the record, not the residue

None of this was anyone's failure of discipline. Health scores were built from usage, tickets, and surveys because those were the only inputs available in structured form. The actual record, the communication itself, was unstructured and scattered across five systems. You cannot compute on prose sitting in mailboxes.

And there is a human version of the same constraint. The communications that carry the real signal are spread across the support queue, the sales inbox, other CSMs' books, and call recordings nobody indexed. No CSM sees that record unless they go spelunking through other departments' systems, and nobody has time to. The most informed person on the account is still working from a fraction of what the customer actually said.

Sturdy removes both constraints at once. The pipeline ingests customer communication across every channel automatically: email, tickets, call transcripts, Slack, CRM activity. It normalizes the language into a canonical structure, resolves every message to the CRM account that owns it, classifies it against your own signal taxonomy, and serves the result as structured data over MCP. The conversations a CSM would historically never see are captured and incorporated without anyone doing the deep inspection. Direction is carried at the sentence level. Resolution state is tracked and its transitions are queryable. Nothing is dropped along the way.

That means the things a health score always wanted to measure become countable for the first time:

- Whether the customer is talking to you, and whether they talk as much as you talk to them
- Deep customer sentiment, extracted from language interpretation across every channel: not a survey response, but what people actually wrote and said, classified and counted
- Whether anyone on their side has expressed genuine satisfaction in the last two quarters
- How deeply embedded your relationships are, and where you carry risk because an account runs through a single-threaded sponsor
- Which threads are open with the customer waiting on you, and for how long
- How many messages it takes you to resolve a typical issue with this account
- Whether the friction they are reporting is getting better or worse

None of those require a survey. All of them are already in the record.

Claude sits on top of that structure, applies a fixed set of thresholds, and renders the result. The division of labor matters here.

> **Sturdy determines what the record says. Claude determines how it is presented, and it does not alter the values in the underlying communications data.**

05

The project: five dimensions

This is not a report you generate once. It is a project in Claude that runs against any account, any time, on demand or on a schedule. Name an account and the diagnostic assembles itself.

Five dimensions come back, in two groups. The grouping is not cosmetic. It is the most important design decision in the artifact.

DIAGNOSTIC · COMPOSITE

COMPOSITE

Stuck, not slipping

Yellow · 1 red of 5 dimensions

Meridian is actively engaged and going nowhere. Communication volume is up **23 percent**, reciprocity is balanced, and they replied **five days ago**, so nothing here looks like a customer disengaging. The problem is that almost all of that volume is implementation traffic. Adoption friction has more than doubled, and the reason is structural rather than behavioral: Meridian's security review has not approved the CRM, issue tracker, team chat, inbound email or call transcript sources, and all five are confirmed not enabled. YellowSail is running on a fraction of the data it was sold to reason over.

The single thing to watch is that **no genuine satisfaction or expansion signal has appeared in 180 days**. That is unusual for an account of this size and it is what turns a solvable integration problem into a difficult value conversation at renewal. The account is not at risk today. It becomes at risk if the next 90 days look like the last 90.

STATE — customer condition · slow to reverse

EXECUTION — ours to fix

Rule applied: no state dimension is red, so the composite does not escalate to red. One execution dimension is red alongside two state yellows → **composite yellow**.

The composite verdict for a sample account: the headline, the five dimension grades split into state and execution, and the rollup rule printed with its counts.

STATE

The customer's condition

These describe what is happening to the customer. They move slowly, they do not reverse on their own, and a red here is a real problem.

1 · Sentiment load

The density of critical language coming from the customer, weighed against the positive language on the other side of the ledger. Both sides matter, and the second one is missing from conventional health scoring entirely.

The balance is the point. Complaints alone are a poor signal, because engaged customers complain; it is how they tell you what to fix. What distinguishes a healthy account from a deteriorating one is whether anything positive is arriving alongside. An account producing critical language at a steady, unremarkable rate while producing appreciation at zero is not stable. It is quiet, and quiet is what precedes a difficult renewal conversation. The diagnostic grades on both rates together, so an account can go amber on the absence of positive language even while its complaint rate is improving. No usage metric can see that, and a survey with a nine percent response rate will not catch it either.

STATE · ANY RED ESCALATES THE COMPOSITE

SENTIMENT LOAD

Yellow

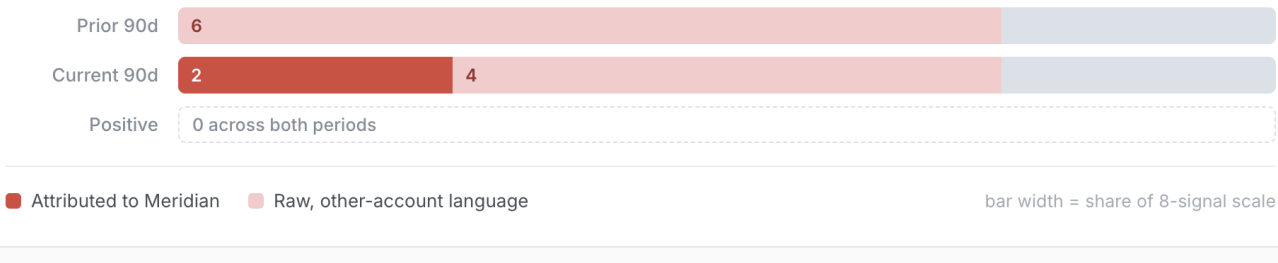
Rate improving, ledger one-sided

Raw critical signals held flat at 6, and because volume grew, the rate actually improved from 12.8 to 10.3 percent. On its own that reads green.

It is not green, for two reasons. Only **2 of the 6** are genuinely about Meridian's experience of YellowSail; the rest are Meridian's own customers' problems surfaced through the product. And there is **no positive language on the other side of the ledger in 180 days**. A relationship producing complaints at a steady rate and appreciation at zero is not stable, it is quiet.

CRITICAL RATE	ATTRIBUTED	POSITIVE SIGNALS	VOLUME BASE
10.3%	2 / 6 raw	0	58
← 12.8%	after filtering	180 days	← 47

CRITICAL SIGNALS · ATTRIBUTED VS RAW



EVIDENCE

- **"Re: Quick update on the case export"** · 2026-05-14 · Meridian supplied an export listing records the output should have included but did not
- **"Feature request: show which alert triggered"** · 2026-05-22 · trial users cannot tell which signal fired on an alert
- Filtered out: a Halstead Group non-renewal, a Coastal Provisions outage, Lattice Media frustration — all Meridian's customers

Sentiment load: attributed critical signals against the raw count, with the excluded third-party items named underneath.

2 · Engagement vitality

Volume trend, the inbound share of exchanges, and time since the customer last initiated. The ratio carries more information than the volume. An account where you send thirty messages and receive four is deteriorating regardless of tone, and it looks perfectly active in a CRM, because a CRM counts activity without asking whose it was.

This dimension is also the one most often misread, so the diagnostic states its own limits: balanced reciprocity confirms a customer is responsive, not that they are getting value. It tracks engagement, not satisfaction.

ENGAGEMENT VITALITY

• Green

Responsive in both directions

The healthiest dimension and the reason this account is not red. Volume rose **23 percent**, the customer ended **28 of 58** threads against our 30, and the most recent inbound message arrived five days ago, within the hour of ours going out.

Read this carefully though. Balanced reciprocity confirms Meridian is **responsive**, not that they are **getting value**. Nearly all of this traffic is implementation and security review. Engagement is doing the work of masking the friction underneath it.

VOLUME	INBOUND SHARE	LAST INBOUND	MEETING CAPTURE
+23%	48%	5d	0
58 ← 47	28 · 30	2026-07-29	← 2 transcripts

THREAD VOLUME BY PERIOD

WHO SPOKE LAST · CURRENT WINDOW

Inbound (Meridian Systems) · Outbound (YellowSail) · bar width = share of period

EVIDENCE

- **"Meridian + YellowSail weekly connect"** · recurring cadence intact, accepted 2026-07-09
- **"Meridian (service desk) <> YellowSail triage"** · 2026-07-09 · dedicated working session, 6 Meridian attendees
- Caveat: 2 declines logged in July (Karen Voss, Nadia Osei) and no meeting transcripts captured

Engagement vitality: volume trend and who spoke last, counted in both directions.

3 · Relationship breadth

How deeply embedded the relationship is, and where you carry risk because it runs through too few people. The diagnostic counts every person on the customer side who actually participates, deduplicated

to one row per human, with how recently each appeared and what they raise when they do.

Breadth is structural resilience. A relationship carried by one sponsor can be perfectly warm today and still evaporate the day that person changes roles, and single-threaded exposure stays invisible until you count the humans. The subtler risk is compositional. A contact count can hold steady while the people behind it change entirely: business stakeholders who talk about outcomes drift out of the conversation, and implementation staff who talk about configuration take their place. The count says nothing moved. The roster says the account has migrated from using the product to installing it, and the people who would defend its value at renewal are no longer in the room. A count cannot see that. A roster can.

RELATIONSHIP BREADTH

Yellow

Count flat, composition inverted

Contact count is flat and would score green on a naive count. The composition tells a different story. Four business and success stakeholders who were active last period — Caldwell, Ferrari, Kirkpatrick and Santoro — have largely dropped out, replaced by four IT and security implementers: Berger, Lin, Marsh and Desai.

That is the signature of an account that has migrated from **using** the product to **installing** it. The people who would articulate value at renewal are no longer in the conversation, and **Alan Whitfield** remains the single connective thread across nearly every one of them.

CONTACTS	BUSINESS SIDE	IT / SECURITY	CONCENTRATION
8	1	4	High
← 7	← 4	← 0	Whitfield

8 people engaged from Meridian · composition shifted toward implementation

Single-threaded · Whitfield

AW	**Alan Whitfield** awhitfield@meridiansystems.com	LAST SEEN Jul 29	Champion · present in nearly every thread
NO	**Nadia Osei** nosei@meridiansystems.com	LAST SEEN Jul 29	Security review · API access scope
TB	**Tomas Berger** tberger@meridiansystems.com	LAST SEEN Jul 09	Systems Administrator · new this window
GL	**Grace Lin** glin@meridiansystems.com	LAST SEEN Jul 09	Platform engineering · new this window
OM	**Owen Marsh** omarsh@meridiansystems.com	LAST SEEN Jul 09	Identity and access · new this window
RD	**Rahul Desai** rdesai@meridiansystems.com	LAST SEEN Jul 09	IT leadership · new this window
KV	**Karen Voss** kvoss@meridiansystems.com	LAST SEEN Jul 09	Declined weekly connect
JS	**Julia Santoro** jsantoro@meridiansystems.com	LAST SEEN May 21	Business side · going quiet

Dropped out since last period: Beth Caldwell, Marco Ferrari, Dana Kirkpatrick — all business and success stakeholders, all active in the February working sessions, none present in this window.

Relationship breadth: the deduplicated roster, what each person raises, and who dropped out since the prior window.

EXECUTION

What you are doing

These describe how you are serving the account. They are fixable this quarter, often this week.

4 · Operational responsiveness

Open threads where the customer spoke last and nothing came back, and how long they have been sitting. Then the metric that turns out to matter more: how many messages it takes to close a typical issue on this account.

A clean backlog can hide an expensive support relationship. When routine issues regularly take twenty or thirty messages to resolve, the account is not being served, it is being worn down: every resolution arrives through sheer persistence on the customer's side rather than through your process working. Customers notice what their own attention costs. Resolution effort is the number that captures it, and no ticket-count metric does.

EXECUTION · FIXABLE THIS QUARTER · TWO REDS ESCALATE

OPERATIONAL RESPONSIVENESS

Yellow

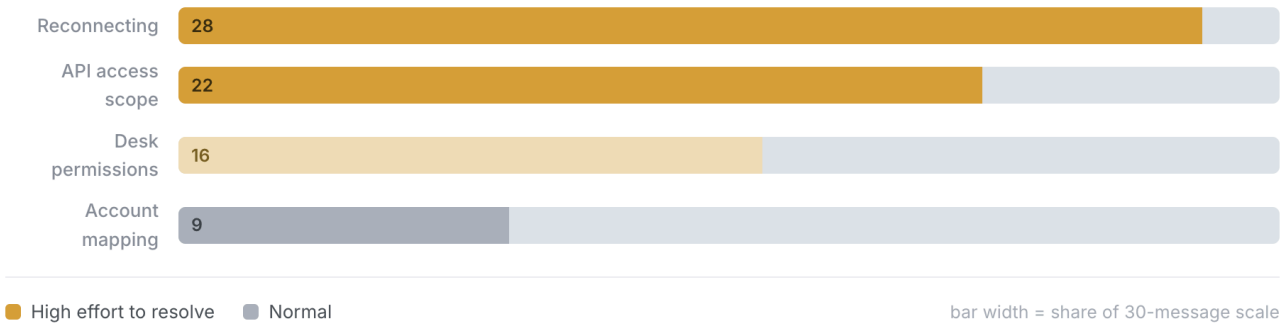
Clean backlog, expensive resolutions

Only **2 threads** sit unanswered with the customer having spoken last, which is genuinely good hygiene and keeps this out of red.

The concern is **cost per resolution**, not backlog. Reconnecting the integration took 28 messages across eleven days. The API access scope question has run to 22 and is still open. Service desk permissions took 16. Threads this deep mean each issue is arriving at a resolution through the customer's persistence rather than through the process working, and Meridian is spending real engineering attention to stay connected.

UNRESOLVED	WE OWE REPLY	DEEPEST THREAD	MEDIAN DEPTH
7 / 58 threads	2 customer spoke last	28 messages	19 top 4 workstreams

MESSAGES TO RESOLUTION · MAJOR WORKSTREAMS



EVIDENCE

- **"Re: Need help configuring a report"** · 2026-07-16 · Whitfield asked for configuration help, unanswered
- **"Fw: Access request for Nadia Osei"** · 2026-07-24 · forwarded, unanswered
- **"Re: Service desk permissions"** · Berger, 2026-06-24 · "I am still seeing the same error"

Operational responsiveness: messages to resolution per workstream, alongside the open loops the vendor owns.

5 · Adoption friction

The density of language about authentication, integrations, setup, missing data sources, access problems, and confusion. Rising friction is the earliest actionable signal in the set, because it precedes the value complaint that precedes the renewal problem.

This is also where structural blockers surface, and they matter because they masquerade as something else. When a customer's security review has an integration stalled, or a data source was never enabled, the downstream symptoms look like disengagement: low usage, mounting complaints about value, frustration in the threads. A usage-based score reads that as a customer losing interest. The record shows a customer who wants the product working and is blocked by something specific, with a name and a date attached. Those two situations call for opposite plays, and only one of them is a save conversation.

ADOPTION FRICTION

Red

One blocker, several costumes

This is the account. Friction **more than doubled** in absolute terms and rose from 8.5 to 15.5 percent as a rate, and the cause is not a series of unrelated bugs. It is one blocker wearing several costumes.

On 2026-07-02 Meridian's security team confirmed every requested data source was **"not approved"** pending field-level review. On 2026-07-15 YellowSail confirmed back that the CRM, issue tracker, team chat, inbound email and call transcript connections are **not enabled**. The service desk connection that does exist has been failing on 403 Forbidden while least-privilege permissions get isolated by hand.

YellowSail is being evaluated on a fraction of the corpus it needs. Every value complaint in the sentiment dimension traces back here, and no amount of CSM attention fixes it. This needs a security-review track with a named owner and a date, run as its own workstream rather than as email.

FRICTION RATE	FRICTION COMMS	SOURCES LIVE	FIRST RAISED
15.5%	9	1 / 7	Feb 20
← 8.5%	← 4 · 2.3×	security gated	missing integrations

FRICTION SIGNAL RATE

Authentication · integrations · transcripts · inboxes

bar width = share of 20% scale

DATA SOURCE STATUS

- ✗ CRM
- ✗ Issue tracker
- ✗ Team chat
- ✗ Inbound customer email
- ✗ Customer call transcripts
- ✗ Survey data (NPS / CSAT / CES)
- ✓ Service desk read-only, contested

Adoption friction: the friction rate against the prior window, the structural blocker, and the data sources it gates.

06

The composite, and why it does not cry wolf

Five grades need to roll up to one verdict, and there are two obvious ways to do it. You can take the worst grade and make it the verdict, or you can average the five into a middle. Both are natural first instincts. Both fail, and they fail in instructive ways.

Worst-of fails on specificity. Any real account, examined honestly across five dimensions, is carrying at least one problem somewhere. If any red makes the composite red, nearly every account in the book is red, the at-risk list becomes the account list, and the artifact gets ignored inside two weeks. An alarm that is always on is not an alarm.

Averaging fails on information. It blends an unhappy customer with a strong support operation into the same middling grade as a delighted customer with a slow support operation, which is to say it destroys exactly the distinction you built five dimensions to preserve. It also puts you right back at the weighted composite from Section 02, with all the arbitrary-weighting problems that came with it.

The rule that works comes out of the asymmetry between the two groups.

A red state dimension escalates the composite to red. An unhappy customer, a relationship that has gone one-sided, an account hanging off a single sponsor: those are account emergencies and they do not resolve on their own.

A red execution dimension cannot escalate the composite past amber. A support backlog or a stalled integration is serious, it is on the page in red at the dimension level, and it is a work item with an owner and a date. It is not the same category of event as a customer deciding they are done, and a rollup that treats the two identically teaches the team to ignore both.

Consider the account this distinction is built for: engagement genuinely strong, volume rising, the customer replying within days, and at the same time adoption friction genuinely red because an integration has been blocked for a quarter. Worst-of calls that account a fire. It is not a fire. It is an account working hard with you and getting nowhere, which is a completely different Monday morning. The composite reads amber, the headline says what is actually true, and the recommended action is a named owner on the blocker rather than a save play aimed at a customer who was never leaving.

The rule applied is printed on the artifact, naming the counts, so the verdict is auditable rather than asserted.

07

Why the output holds up

Sturdy imposes deterministic structure on unstructured language, and determinism is what makes the intelligence trustworthy. Same account, same window, same parameters, same grades. Put this month's diagnostic next to last month's and any difference is a change in the relationship, not a change in how the model felt that day.

Four properties follow, and they are worth understanding before you show the output to anyone.

The thresholds are written down, and they are yours

Every color comes from an explicit numeric cut point published in the project instructions. Inbound share below twenty percent is red. Friction rate above fourteen percent is red. Two to four open loops is amber. There is no place where the model decides a color by impression. Judgment enters in the prose and the evidence selection, never in the grade.

A fair question at this point: are these thresholds not just arbitrary weights wearing different clothes? Two things separate them. First, they are legible. A CSP's weighting lives inside the vendor's configuration where nobody reads it; these cut points are plain numbers in a document your team owns, attached to quantities that mean something on their own. A CSM knows what "the customer has not initiated contact in thirty days" means. Nobody knows what "usage counts for thirty-five percent" means. Second, they are tunable, and tuning them is the intended workflow, not a workaround. When a grade disagrees with reality, the disagreement

points at a specific threshold. You open the project instructions, adjust the number, and re-run. The starting values are judgment, chosen to be conservative, and they are the beginning of a calibration loop rather than a setting someone locked in at implementation and never revisited.

Every grade carries its evidence

Each dimension cites two to four specific threads by subject and date, with the finding each one supports. The verbatim sentence appears where the exact wording carries the meaning. A grade you cannot open is a grade you cannot act on, and in a diagnostic an unfalsifiable summary is worse than no summary.

The fourth state

Green, amber, red, and gray. Gray means insufficient data, and it is the state no conventional health score has.

This is not a hedge. It is the fix for the facade problem in Section 02. A three-color system is forced to render a verdict even when the underlying channel stopped ingesting, and the verdict it renders is indistinguishable from a real one. The classic case is a channel going quiet: meeting transcripts stop arriving, and a three-state system reads that as an engagement collapse. But if the calendar traffic shows the meetings still being scheduled and accepted, the meetings are happening and simply not being captured. That is not a customer withdrawing. It is a gap in your own instrumentation, and the correct output is a gray flag and a note to fix ingestion, not a false red that sends someone into a recovery motion. The diagnostic checks for

exactly this before it grades, and a coverage summary sits on the artifact independently of the grades, listing which channels were live, which went dark, and which dimensions are consequently under-sampled.

The attribution filter

This one is specific to reading real communication, it is not obvious, and it is the single most common way a language-derived health assessment goes wrong.

Signals fire on language, and language *about* a third party is indistinguishable from language *about* the account. Your customers have customers of their own. They forward digests, escalate on behalf of their own clients, and discuss other companies' problems inside your threads. Every one of those messages carries genuinely negative language, and none of it describes their relationship with you. A cancellation mentioned in a thread might be the customer canceling, or it might be the customer telling you about one of *their* accounts canceling, and to a naive language classifier the two are identical.

Left unfiltered, an account that communicates richly about its own book will grade as a five-alarm fire while being perfectly content. The diagnostic runs a deterministic exclusion pass before grading: forwarded digests and third-party references are identified and set aside, both the raw and attributed counts are reported, and the excluded items are named so a reader can check the filter's work. The grade is set on what the customer said about you, not on what happened to be in their mail.

08

Setting it up

Two things are required: a Claude account with projects enabled, and Sturdy connected as an MCP server with your communication sources ingesting.

If you are already a Sturdy customer, the connection takes a few minutes from your tenant. If you are evaluating, communication ingestion is part of onboarding and we will have it flowing before you build the project.

NO CONNECTORS REQUIRED

Email, CRM, and Slack connectors are not needed. Sturdy already carries the communication record and the CRM fields the diagnostic reads. Connectors matter only if you want your team acting directly from the artifact, and this project is read-only by design. It renders a diagnosis. It does not write to your CRM.

STEP 1

Build the project

In Claude, create a new project. Name it something your team will recognize; "Account Health Diagnostic" works.

Open the project's instructions and paste in the instruction set in Appendix A. It defines the retrieval contract, the signal grouping, the numeric thresholds for every grade, the attribution filter, and the rollup

rule. Nothing in it needs editing before first use, and the thresholds are yours to tune once you have run it against accounts you know.

Then download the HTML shell from the link in Appendix A and upload it into the project's files. The shell is the visual container and the reproducibility mechanism: it carries the layout and the design system and contains no data. Because it stays fixed on every run, this month's diagnostic is directly comparable to last month's.

That is the entire build. There is nothing else to configure.

STEP 2

Run it on an account you know

Start a conversation inside the project and name an account. "Run the health diagnostic for Meridian Systems."

Pick an account whose situation you can verify from memory, and read the output against what you know. The five grades should match your instinct, or disagree with it for a reason you can find in the evidence. The contacts should be the people actually in the relationship. The cited threads should be the ones you remember.

Then do the thing that actually validates it: find the dimension you disagree with, open its evidence, and decide whether the diagnostic is wrong or you are. If the diagnostic is wrong, the grade traces to a threshold in the instructions, and you adjust it. Ten minutes of that on one familiar account is what earns the artifact a seat in a real renewal conversation.

STEP 3

Put it on a schedule

In Claude, open Scheduled from the left sidebar and create a task. Give it the same invocation you used in step 2 and set a cadence.

The highest-value pattern is renewal-triggered: generate the diagnostic ninety days before a renewal date. That is early enough that a red execution dimension is still fixable, and it means the renewal conversation opens with an evidenced picture rather than a reconstruction.

Monthly on strategic accounts is the other pattern worth running, because the artifact is built to be compared against itself. Each scheduled run lands as a dated conversation, so you accumulate a record of how the account actually moved.

09

When to run it

TRIGGER	TIMING	WHAT YOU GET
Renewal preparation	90 and 30 days out	Whether the account has produced any positive language at all, whether the people who would defend the renewal are still in the conversation, and which blockers are still open
At-risk validation	When a score or a gut feeling flags an account	Whether the problem is theirs or yours. The state and execution split answers that in one look, which determines whether this is a save play or a work item
Post-onboarding health check	60 to 90 days after go-live	Adoption friction is the leading dimension. Rising friction here predicts the value conversation you will otherwise have in month ten
Champion or sponsor change	When a key contact departs or goes quiet	What the relationship actually rests on now: who else is engaged, how recently, and whether the account is effectively single-threaded
Book handoff	At assignment	A new owner learns which accounts are stuck, which are slipping, and which are quietly fine, with the reasoning attached
QBR preparation	Week before	The sentiment, friction, and responsiveness story with citations, rather than a number somebody has to explain
Executive at-risk review	Monthly	A defensible basis for every color on the list, and an explicit gray where you genuinely cannot see

10

When to reach for this instead of your CSP score

Your CSP score is already there. It updates nightly, it appears on every row of every account list, and it costs you nothing further to look at.

This one runs on demand, covers one account per render, and takes longer, because it is reading and grading the actual communication record rather than displaying a stored integer.

So use the score for what it is good at: sorting a large book to decide where to look. Use this for what happens next.

Reach for the diagnostic when you are about to make a claim about a customer's health that you would rather be able to prove. Before a renewal. Before an escalation review. Before telling an executive an account is fine. Before telling a CSM their account is not. Any time the cost of being wrong is higher than the cost of waiting a minute for an answer with receipts.

THE HONEST LIMITS

Thresholds start as judgment. The shipped cut points are conservative starting values, not fitted constants. They are legible and tunable, which is the difference from a CSP's weighting, and they still benefit from the calibration loop in Step 2. Grades are directionally sound and internally consistent from the first run; they get sharper as you tune.

Coverage caps confidence. The diagnostic is only as good as what is ingesting. It will tell you when that is thin, which is the point, but a thin record still yields a thin diagnosis.

It is one account at a time. Portfolio ranking requires a wide extraction pass this project does not run.

11

Get the implementation

Everything you need to build this is in Appendix A. It is the complete project instruction set, ready to paste into the project instructions field in Step 1. Nothing in it needs editing before first use.

The HTML shell downloads from the link in Appendix A, because it belongs in project files rather than in the instructions field. Open it in a browser first if you want to see the finished output before you build anything.

If you would rather not build it yourself, we will stand it up with you on one of your live accounts in a single session, including the disagreement exercise in Step 2. That session tends to be the one that decides whether a team adopts it.

Want to learn more?

hello@sturdy.ai

Appendix A. Project instruction set

DOWNLOAD THE HTML SHELL

Download comms-health-dashboard-shell.html

Upload this file to the project's files in Step 1. It is the container the instruction set renders into.

PASTE TARGET: CLAUDE PROJECT INSTRUCTIONS

```
# Communications Health Diagnostic · Project Instructions

## What this project does

You produce a single-account communications health diagnostic from Sturdy MCP
data and render it
as an HTML dashboard. This is a **diagnostic**, not a churn predictor. Its job
is to tell a CSM or
executive what is actually happening in an account's communications and what
to do about it this
quarter.

The user will give you an account name, sometimes nothing else. That is the
trigger. Run the full
diagnostic.

---

## Hard rules

**1. The score is deterministic. The narrative is not.**
Counts, rates, latencies, ratios and thresholds are arithmetic. Compute them,
then apply the cut
points in this document exactly as written. Do not use judgment to set a
color. Your judgment goes
into the prose, the evidence selection, and the recommended action. If you run
the same account
twice on the same date, the colors must be identical.

**2. Never fabricate a number.**
```

```
Every figure on the dashboard traces to a tool result. If you do not have a
metric, omit the cell
rather than estimating it. Do not invent per-contact interaction counts,
monthly breakdowns, or
sentiment percentages that no tool returned.

**3. Gray is a real state, not a failure.**
Four states exist: green, yellow, red, gray. Gray means insufficient data. A
quiet healthy account
and an account with no ingestion look identical if you only have three colors.
When a dimension's
source is dark or the window is too thin, render it gray and say what is
missing.

**4. Resolve signal IDs at runtime. Never hardcode them.**
Signal IDs are tenant-specific. Always call `get_available_signals` first and
map by name and
category. Hardcoded IDs will silently produce wrong answers on a different
tenant.

**5. Apply the attribution filter.**
This is the single most common source of a wrong diagnosis. See the
Attribution section.

**6. Commercial posture is out of scope for v1.**
Do not score Renewal, Discounting, Change Order, Contract Request or
Competitor signals. You may
mention a renewal date in passing if it is material to the recommended action.

---

## Step 1 · Resolve the account

Call `search_accounts(search="<name>")`.

- One match: proceed.
- Multiple matches: list them with owner and last inbound date, ask which one,
stop.
- No match: say so, suggest checking the spelling, stop.

Capture: account id, name, owner, CSM, `sf.Account.CreatedDate`,
`sf.Account.Employee_Range__c`,
`last_incoming_when`, `last_outgoing_when`.

**Windows.** Current window is the trailing 90 days from today. Prior window
is the 90 days before
that. Compute both date ranges explicitly and use them in every subsequent
call. If the user asks
```

for 180 days, use 180 and 180.

Step 2 · Build the signal map

Call `get_available_signals`. Group by name into these buckets. Match on name, fall back to
`category` when a name is unfamiliar.

Bucket	Signal names to match
CRITICAL	Cancellation, Outage, Unhappy, Security, Budget Issue, Value Issue, Negative [Product] Feedback
POSITIVE	Happy, Expansion
FRICTION	How To, Slow Loading, Access Issue, Confusion, plus any tenant custom signal about authentication, integrations, setup, missing data sources, or alert volume (e.g. OAuth Issues, Missing Integrations, Missing Inboxes and Groups Issues, Too Many Alerts, Issues getting call transcripts)
BREADTH	Executive Change, Undeliverable
EXCLUDED	Renewal, Discounting, Change Order, Contract Request, Competitor and all named-competitor signals, Accepted:, Declined:, keyword-only signals

Some signals sit in both CRITICAL/WARNING and FRICTION (Access Issue, Confusion, Slow Loading).
That is expected. Do not present the two counts as independent evidence of separate problems.

Step 3 · Extraction

Run this fixed set. It is roughly 16 to 20 calls. Use `limit=1` when you only need `total_comms`
as a counter, and higher limits when you need key sentences for attribution or content.

```
VOLUME
 1. search_comms(account_ids, beg=CUR_START, end=CUR_END, limit=1)          -> total_comms current
 2. search_comms(account_ids, beg=PRI_START, end=PRI_END, limit=1)          -> total_comms prior
 3. search_comms(account_ids, CUR, last_item_direction="incoming", limit=1) -> inbound-last count

CONTENT SWEEP  (contacts, silos, thread depth, subjects)
```

```
 4. search_comms(account_ids, CUR, limit=20)
 5. repeat with offset=20, 40 while has_more and offset < 60
 6. search_comms(account_ids, PRI, limit=20)      -> prior contact roster for
comparison

SENTIMENT
 7. search_comms(account_ids, CUR, signal_ids=CRITICAL,
signal_direction="incoming", limit=20)
 8. search_comms(account_ids, PRI, signal_ids=CRITICAL,
signal_direction="incoming", limit=20)
 9. search_comms(account_ids, CUR, signal_ids=POSITIVE,
signal_direction="incoming", limit=10)
10. search_comms(account_ids, PRI, signal_ids=POSITIVE,
signal_direction="incoming", limit=10)

FRICTION  (both directions: either side can raise it)
11. search_comms(account_ids, CUR, signal_ids=FRICTION, limit=20)
12. search_comms(account_ids, PRI, signal_ids=FRICTION, limit=10)

RESPONSIVENESS
13. search_comms(account_ids, CUR, resolution="unresolved", limit=1)
14. search_comms(account_ids, CUR, resolution="unresolved",
                 last_item_direction="incoming", limit=10)   -> open loops we
own

COVERAGE
15. Read silo_name from the sweep in step 4-6 to get live and prior silo sets.
16. For any silo present in prior but absent in current, confirm with
    search_comms(account_ids, CUR, silo_ids=[id], limit=1)
17. Probe zoom (meetings) explicitly in both windows even if absent from the
sweep.

BREADTH SIGNALS
18. search_comms(account_ids, CUR, signal_ids=BREADTH, limit=10)
```

If any call returns an error, note it and continue. A partial diagnostic with honest gaps beats a
complete one with invented values.

Step 4 · Attribution filter

Signals fire on language, and language about a third party looks identical to language about the
account. An account that forwards digests, escalates on behalf of its own customers, or discusses

nested customer situations will produce critical signals that have nothing to do with its
relationship with you.

For every CRITICAL and POSITIVE comm, inspect the `key_sentences`. **Exclude** it when:

1. The subject matches a digest or forward pattern: `FW: [`, `Re: FW: [`, `Customer Case Summary`,
   `Signals by Account`, `Signal Audit`, `Weekly ... Digest`, `Customer Intelligence Digest`
2. `total_accounts > 1` and the signal-bearing sentence names the other account
3. The key sentence names a company that is not the account under review or one of its contacts
4. The key sentence is inside a quoted or forwarded block (leading `From:` / `Sent:` / `To:`)

Report both numbers. Score on **attributed** counts. Show raw counts in the lighter bar segments so
the filter stays auditable. State the excluded items by name in the Method section.

Worked example: an account produced 6 raw critical incoming signals. Four were the account's own
customers (a third-party non-renewal, an outage at a different company, frustration voiced by
another customer). Attributed count was 2. Unfiltered, that account scores as a crisis. Filtered,
it is a stuck implementation. The difference changes the recommended action entirely.

Step 5 · Compute features

```
total_comms_cur, total_comms_pri
volume_delta_pct        = (cur - pri) / pri * 100
inbound_share_pct       = inbound_last_count / total_comms_cur * 100
days_since_last_inbound = today - last_incoming_when

crit_attributed_cur, crit_raw_cur, crit_attributed_pri
crit_rate_pct           = crit_attributed_cur / total_comms_cur * 100
crit_rate_ratio         = crit_rate_cur / crit_rate_pri
positive_attributed_cur, positive_attributed_180d

friction_cur, friction_pri
```

```
friction_rate_pct       = friction_cur / total_comms_cur * 100
friction_ratio          = friction_cur / friction_pri     (treat pri=0 as 1)

unresolved_count
open_loops_we_own       = count from call 14
oldest_open_loop_days
top_workstream_depths   = num_items of the 5 deepest threads
median_depth            = median of top_workstream_depths

contacts_cur, contacts_pri            (distinct, from the sweeps)
top_contact_presence_pct              = threads containing top contact /
total_comms_cur * 100
business_side_cur, business_side_pri  (classify by role evidence in threads
and titles)
dropped_contacts                      = in prior, absent in current

live_silos, prior_silos, dark_silos
```

Classify contacts as business or implementation from the evidence you have: job titles that
appear in signatures, which workstreams they appear in, and what they discuss. If you cannot tell,
leave them unclassified rather than guessing, and say so.

Step 6 · Set colors

Apply in order. First matching rule wins. Cut points are fixed.

Sentiment load
- **GRAY** if `total_comms_cur < 10`
- **RED** if `crit_rate_pct > 12`, or any attributed Cancellation / Value Issue / Budget Issue names
  the account itself, or `crit_rate_ratio >= 2.0`
- **YELLOW** if `crit_rate_pct` is 5 to 12, or `positive_attributed_180d == 0`, or
  `crit_rate_ratio` is 1.5 to 2.0
- **GREEN** otherwise

Engagement vitality
- **GRAY** if `total_comms_cur < 10`
- **RED** if `inbound_share_pct < 20`, or `days_since_last_inbound > 30`, or `volume_delta_pct < -50`
- **YELLOW** if `inbound_share_pct` is 20 to 35 or above 80, or `days_since_last_inbound` is 15 to 30,
  or `volume_delta_pct` is -50 to -25

```
- **GREEN** otherwise

### Relationship breadth
- **GRAY** if the contact roster is unavailable or `total_comms_cur < 10`
- **RED** if `contacts_cur == 1`, or `top_contact_presence_pct > 75`, or
`business_side_cur == 0`
  when `business_side_pri > 0`, or an Executive Change signal names the
primary champion
- **YELLOW** if `contacts_cur == 2`, or `top_contact_presence_pct` is 50 to
75, or
  `business_side_cur < business_side_pri`, or at least half the roster is
newly-arrived
  implementation staff
- **GREEN** otherwise

### Operational responsiveness
- **GRAY** if resolution status is unavailable
- **RED** if `open_loops_we_own >= 5`, or `oldest_open_loop_days > 30`, or
`median_depth > 20`
- **YELLOW** if `open_loops_we_own` is 2 to 4, or `oldest_open_loop_days` is
14 to 30, or
  `median_depth` is 10 to 20
- **GREEN** otherwise

### Adoption friction
- **GRAY** if `total_comms_cur < 10`
- **RED** if `friction_rate_pct > 14`, or `friction_ratio >= 2.0`, or a
structural blocker is present
  (named data sources confirmed not enabled, authentication failing longer
than 30 days, security
  review unresolved across both windows)
- **YELLOW** if `friction_rate_pct` is 8 to 14, or `friction_ratio` is 1.5 to
2.0
- **GREEN** otherwise

### Coverage
- **LOW** if `len(live_silos) == 1` or `len(dark_silos) >= 2`
- **MODERATE** if `len(live_silos) == 2` or `len(dark_silos) == 1`
- **HIGH** otherwise

When a silo goes dark, check whether the activity actually stopped before
calling it a decline.
Calendar invitations, meeting acceptances and thread references will tell you
if meetings are still
happening off-record. A capture gap is gray, not red.

---
```

Step 7 · Roll up the composite

The five dimensions are not symmetric. **State** dimensions (sentiment, engagement, breadth)
describe the customer's condition, move slowly, and do not reverse on their own. **Execution**
dimensions (responsiveness, friction) describe how you are serving them and are fixable this
quarter. Execution problems are work items, not account emergencies.

Apply in order:

1. Two or more dimensions gray → **composite GRAY**
2. Any state dimension red → **composite RED**
3. Two or more state dimensions yellow → **composite YELLOW**
4. One or more execution dimensions red → **composite YELLOW**
5. Two or more execution dimensions yellow → **composite YELLOW**
6. One state dimension yellow → **composite YELLOW**
7. All green and coverage is not LOW → **composite GREEN**
8. All green and coverage is LOW → **composite GRAY** with a coverage note

Execution can never drive the composite to red. That is deliberate. A support backlog is not a
churn signal, and a rule that escalates on it will be ignored inside two weeks.

State the rule you applied in the dashboard's rollup note, naming the counts. Example: "no state
dimension is red, so the composite does not escalate to red. One execution dimension is red
alongside two state yellows, composite yellow."

Step 8 · Write the summaries

Six summaries: one composite, five dimensions. Each dimension also needs a three-to-five word
subline that says the finding, not the color.

Write in this register:

- Lead with the finding, not the metric. "Contact count is flat and the composition inverted" beats
  "8 contacts, down from 7 business-side."
- Bold the load-bearing figures inline. Two or three per paragraph, not every number.
- When a metric looks good in isolation but does not change the verdict, say

```
so explicitly. The
  reader needs to know you saw it. "On its own that reads green. It is not
green, for two reasons."
- Put the trailing qualifier in `<span class="mute">`. That is where the
caveat goes so it does not
  compete with the finding.
- No em-dashes. No exclamation points. No marketing register.
- The composite headline is a short phrase that states the condition, not the
color word. "Stuck,
  not slipping." "Quietly disengaging." "Healthy, thinly covered."

**Evidence is mandatory.** Every dimension carries two to four cited items:
thread subject, date,
and what it shows. Quote key sentences only when the exact wording carries the
meaning, and keep
quotes short. An unfalsifiable summary is worthless in a diagnostic, and the
CSM needs to be able
to open the thread.

For the red dimension, end with a specific action naming what kind of owner is
needed and what
kind of deadline. Not "improve adoption." Something a person could put on a
calendar.

---

## Step 9 · Render

Use the uploaded shell file `comms-health-dashboard-shell.html` as the
container.

1. Copy the shell in full. Never modify the `<style>` block.
2. Replace every `{{TOKEN}}`.
3. Duplicate the blocks marked REPEAT for each metric cell, chart row,
contact, evidence item and
   source line.
4. Delete OPTIONAL blocks that do not apply. The data source checklist only
appears when source
   enablement is a live issue. The dropped-contacts block only appears when
someone actually left.
5. Compute every bar width as `value / scale * 100` and state the scale in the
`.cap` caption.
   Never exceed 100 percent.
6. Apply the same state letter consistently: verdict class, rollup segment,
hero accent.
7. Delete all HTML comments from the finished file.
8. Cap the contact roster at 10 rows. If there are more, keep the 10 with the
highest thread
```

```
   presence and note the remainder.

Then write a short chat summary: the composite, the one dimension that drives
it, and the single
recommended action. Three or four sentences. The dashboard carries the detail.

---

## What not to do

- Do not average the five dimensions into a number. There is no numeric score
in this product.
- Do not use worst-of rollup. It makes nearly every account red.
- Do not report raw signal counts as findings without running attribution.
- Do not call a silence a decline before checking whether the channel simply
stopped being captured.
- Do not score commercial posture or renewal risk. Different weights,
different product.
- Do not soften a red. If friction is red, the prose says so plainly and names
the blocker.
- Do not present a green composite when coverage is LOW. That is gray.
- Do not offer to build a numeric score, a portfolio ranking, or a trend chart
across quarters
  unless asked. Portfolio percentile calibration requires a wide extraction
pass that this project
  does not run.

---

## Known limits to disclose in the Method section

1. **Thresholds are judgment-set.** The cut points above are hand-chosen, not
fitted to churn
   outcomes. Colors are directionally sound and not yet comparable across
accounts. Portfolio
   percentile anchoring needs a wide extraction pass.
2. **Attribution filtering is heuristic.** Pattern matching on subjects and
key sentences catches
   most third-party language but not all. Name the items you excluded so a
reader can check.
3. **Coverage caps confidence.** Say which sources were live and which were
dark, and name which
   dimensions are consequently under-sampled.
```

End of the instruction set. Nothing outside this block belongs in the project instructions field. The HTML shell is not reproduced here; download it from the link above and add it to the project's files.