# How Evry Health *actually* uses the EvryAI chat platform

## The first ten months.

What ten months of real usage across 108 employees has produced inside Evry, measured rather than surveyed, and what it suggests for any regulated company weighing the same move.

**Scope.** The EvryAI chat platform only. Vocalis, Inline, Lakefront and Bellweather are separate systems and none of their usage is counted here.

```
SOURCE   Production databases. Full census of chat records and of per-request model usage
METHOD   10 department analyses, 9 full-transcript case studies, 4 synthesis passes
OUTPUT   Supporting analysis available under NDA
```

SCALE

**87,684**

messages from **108 employees** across 10,472 conversations

ADOPTION

**87%**

of the company uses it monthly, and the number is **still rising**

VALUE CREATED

**~$1M/yr**

hours displaced + new capability + spend avoided + **~$300K of compute run for under $10K/mo**

VOLUME PROCESSED

**12.5B**

tokens of real work, **5.3B in July** alone

## What ten months produced

1. **Nearly nine in ten of the company uses it every month, and the number is still rising.** 94 of our 108 people in July, up from 2 last October, with no month yet showing a decline.
2. **It has produced 13,497 finished documents.** More than one per conversation. Four in ten are Word files, spreadsheets and decks that went on to be used as real work.
3. **It is doing work we could not otherwise have done.** A URAC accreditation submission, an ORSA model, a reserve method validated against six months of actual claim runout, a 21-state employee handbook. Section 04 has the detail.
4. **Total value created is approaching $1M a year**: $254,000 in time recovered ($2,702 per active user), a comparable band of work that would not have happened otherwise, avoided spend, and roughly $300K a year of list-rate model compute delivered for under $10K a month. Section 05 separates the components.

The rest of this report covers what people actually do with it, what that is worth, how it spread through the company, and which kinds of organisation the same evidence is likely to matter to.



## What we found

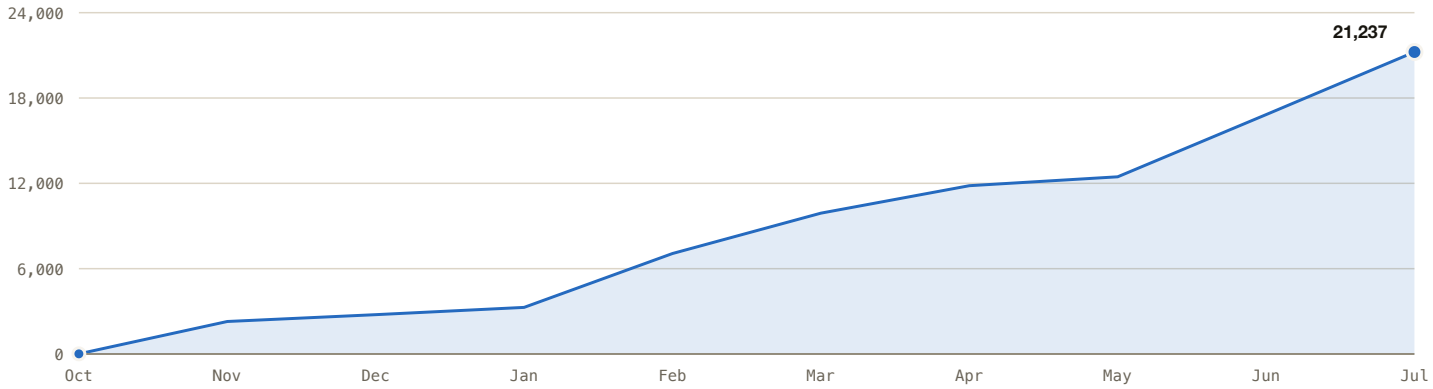
We put an AI chat platform in front of the whole company in October. Ten months later **87% of employees use it every month** and adoption is still climbing. It has produced **13,497 finished work products across 10,472 conversations**, which is more than one deliverable per conversation.

People use it to build statutory reserve models, produce URAC accreditation evidence, review medical necessity, answer the Texas Department of Insurance, classify provider networks, respond to RFPs, screen candidates for fraud and reconcile a 21-state employee handbook. The most valuable behaviour we found is that **it pushes back**. It withdrew its own reserving recommendation after testing

showed it was $1.65M short. Reviewing an accreditation submission, it read our own policy wording more strictly than the team had. Offered a shortcut on member notices, it pointed out that the Texas delivery rule would not allow it.

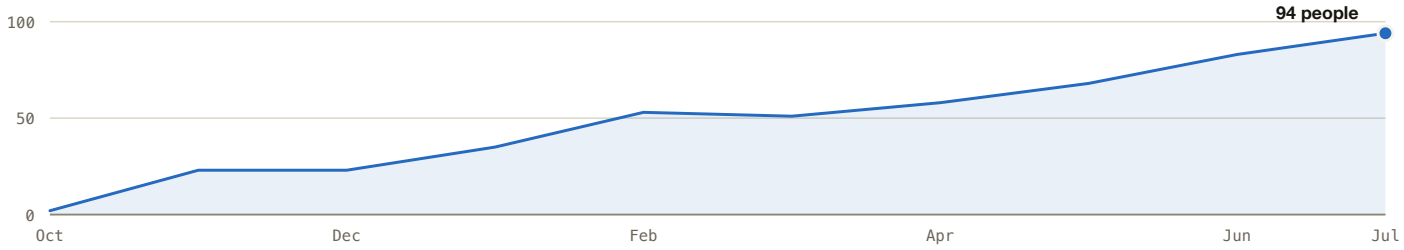
MESSAGES PER MONTH

Ten months of growth with no sign of levelling off. July was our largest month, at 21,237 messages.




PEOPLE USING IT EACH MONTH

From 2 people to 94, out of 108 employees. Shown separately from volume above rather than on a second axis, so the two are not confused.




03 FINDINGS

## Five things worth your attention

01

### People leave with finished work

13,497 documents against 10,472 conversations. **Four in ten recent files are Office or PDF documents**: 562 Word files, 570 spreadsheets, 285 decks, 241 PDFs. Staff are not asking questions and reading answers. They are walking away with the document they needed.

02

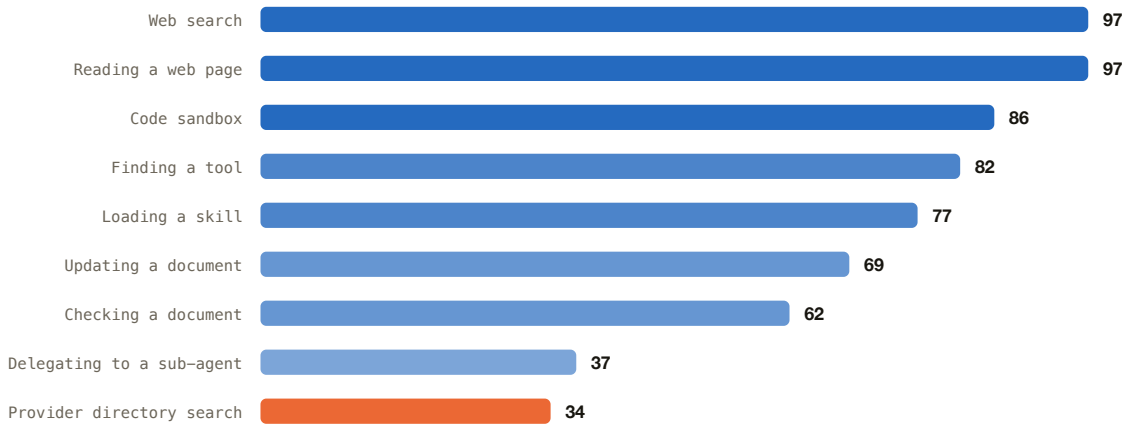
## Ordinary staff are doing analyst work

Our code sandbox is the most-used feature in the product, with **105,470 uses by 86 of our 108 employees**. That is nearly six times the size of the engineering team. **SQL is the most common thing it writes**, at 989 files. Claims specialists, care coordinators and underwriters are pulling answers out of the data warehouse without knowing how to query it.

Nine people outside engineering have built working internal tools. One of the strongest examples came from a new customer service hire with no coding background, who in her first three weeks built a call console, a Salesforce practice environment and interactive training material generated from pasted procedures.

HOW MANY OF OUR 108 EMPLOYEES USED EACH CAPABILITY

Breadth of use, not volume. The code sandbox reaching 86 people is the result worth noticing, because our engineering team is 15.




*Our own provider directory, queried in plain English by a third of the company*

03

## It is already doing regulated work

We found evidence of work touching **more than 60 named standards and rules**: 15 URAC standards, the Texas insurance code (28 TAC §21.3107, Chapter 1467 on QPA, prompt-pay rules, Market Conduct Audit), NAIC and ORSA and RBC, actuarial standards of practice, the No Surprises Act, mental health parity, ERISA and Form 5500, and HIPAA. Regulated member data passes through the platform every day, under the controls that work requires.

04

## It corrects itself when the evidence changes

We read all 43,763 responses the platform has sent and graded a sample of each behaviour by hand. **Roughly 108 conversations contain a genuine self-correction, and about three quarters of those were driven by evidence** rather than by the user pushing back: a query or a document fetch contradicted what it had said earlier, and it said so. The clearest case is the reserving work in section 04, where it withdrew its own recommendation after a back-test.

Outright refusal is rarer, at under 1% of conversations. But the instances we verified are substantial: declining to pass an accreditation standard whose wording did not hold up, and flagging that a proposed shortcut on member notices would not satisfy the Texas delivery rule.

05

## Adoption keeps spreading on its own

Engineering started it. **Operations is now the heaviest user**, at 4,740 messages and 22 active people in July, and still growing. Six of our ten functions were using it within six weeks, and our first group of users is at **95% still active by month eight**.

04 EVIDENCE

# Three cases, start to finish

## 1. Setting the IBNR reserve

CFO & Chief Actuary · 4 weeks · 40 messages

The platform first rebuilt our existing IBNR workbook and matched it to the dollar, which established that it understood the current method before proposing changes. It then tested three ways of selecting completion factors and checked each against six months of actual claim runout. The model it eventually recommended came within **0.3%** of what actually happened. An earlier recommendation of its own was **54% low**, and it raised that itself:

> *"The 7-of-9 method I recommended was $1.65M short."*

It finished by drafting the year-end statutory memorandum with the relevant actuarial standards cited.

**WHY IT MATTERS** Statutory financial reporting, and the only hard accuracy measurement anywhere in ten months of usage.

## 2. URAC accreditation and TDI response

Quality & Population Health · 394 messages

Citations were written directly into the original policy documents at page and paragraph level, in the format URAC's submission system expects. It handled URAC's follow-up questions in the same session and pulled denial rates from Power BI for an attestation the Chief Medical Officer signed. Reviewing one standard ahead of submission, it read our policy language more strictly than the team had and explained why the wording would not satisfy the reviewer, which is the kind of thing far cheaper to catch before a submission than after one.

**WHY IT MATTERS** Accreditation and regulator-facing work, plus evidence that it holds a position when the user wants a different answer.

## 3. Auditing 27 required member notices

Communications · 10 messages · one sitting

We had no single place recording whether all 27 required member notices were actually being delivered. The platform read the 2026 Member Handbook, all 110,000 words of it, plus the Welcome Packet, checked both against the list and produced a one-page Word audit of exactly where each notice is delivered, in a form the compliance team could act on. When offered a shortcut of linking to a notice rather than mailing it, it declined, and cited the Texas rule that makes linking a form of electronic delivery requiring the member's consent first.

> **WHY IT MATTERS** A complete delivery audit in ten minutes, against a manual review that would have taken days.

Two more worth mentioning briefly. Seventy-two Texas facilities were classified in a single afternoon, against a manual baseline of two to three analyst days. A full ORSA model was built end to end in one working session, from risk register through to required capital.

# Four more, in less depth

## 4. Candidate fraud screening, including a pattern nobody had seen before

People and talent · Apr–Jul 2026 · ~30 chats

Every engineering applicant now goes through an open-source check before anyone spends time on an interview: the résumé and the public profile are cross-referenced against each other and against the employers named, and the platform returns a verdict with its reasoning. Sessions run long, up to 168 messages, because the recruiters argue with the output and make it justify itself.

Three patterns show up, in rising order of difficulty. An invented persona with no verifiable footprint. A real work history with the titles rewritten to manufacture experience the candidate does not have. And the one the team had not encountered before: **a fake candidate fronted by the genuine profile of a real, unrelated engineer** — profile harvesting, which is the hardest of the three to catch because each signal on its own checks out. Screening for AI-generated applications runs alongside all of it.

> **WHY IT MATTERS** The HR team rates this the platform's most successful use in their function, and it found a new fraud pattern rather than only applying a known checklist. These are candidates being hired into systems that hold member data.

## 5. A claims quality system, built in ninety days

Claims QA · Apr–Jul 2026 · ~45 chats

A claims analyst who joined in April built the department's quality apparatus from nothing. Over four months: a deduction-based 100-point weighted audit score tied to specific desktop-procedure steps, an Excel audit workbook with per-claim tabs, a specialist accuracy dashboard and a management dashboard, a result-code reference sheet that auto-pulls the right procedure and point value, dollar-based review tiers matching our existing high-dollar thresholds, a random 20% sample selector, and a monthly error-trend report with root-cause write-ups. The department now audits roughly 600 claims a month against it.

> **WHY IT MATTERS** Audit infrastructure of the kind normally bought as a consulting engagement, delivered by a new employee in their first ninety days.

## 6. Onboarding a trading partner's enrollment files

Engineering · Mar–Jun 2026 · 66 and 102 messages

Enrollment files arrive as X12 834 transactions and must match our companion guide before a group can go live. A developer uploaded the guide and a partner's test file and had the platform validate it segment by segment: which records were missing a required loop identifier, which control characters departed from the guide, which placeholder segment would break the parser. Four successive corrected files were re-validated with a diff-style "what got fixed, what is still open" report. A second thread ran a live three-employer onboarding through the same channel, drafting technically correct replies to the partner.

> **WHY IT MATTERS** File-format onboarding is the standing bottleneck between winning a group and enrolling it, and it normally waits on a specialist.

### 7. Reconciling two renewal rating methods

Underwriting · Jul 2026 · 24 messages

A renewal workbook carried two rating builds that agreed on the indicated rate action until credibility weight shifted toward the group's own experience, at which point they diverged. The platform traced both builds end to end and decomposed the entire 41.14 PMPM gap into three named drivers: network and digital-health fees loaded into claims and trended in one build but carried as fixed PMPMs in the other, a difference between the experience-blended and manual trend assumptions, and the treatment of reserve change. It then showed the direction flip came from the two builds comparing against different manual rates, and re-derived the medical/pharmacy trend weighting for a pharmacy-heavy group against two published industry surveys.

> **WHY IT MATTERS** Actuarial peer review of a live pricing model, at the speed of the renewal cycle, before the number goes to a broker.

## What else gets built here

A sample of recurring work across the company, all of it from the record.

| FUNCTION | WHAT WAS PRODUCED | SCALE OR OUTCOME |
|---|---|---|
| Utilization management | Medical-necessity write-ups mapping a clinical record to the named guideline, in SBAR form | ~197 review sessions in seven months from one nurse |
| Care coordination | Member letters rewritten to a sixth-grade reading level, and translated into Spanish | ~450 correspondence sessions in seven months |
| Customer support | Salesforce call notes generated from raw call transcripts against locked house templates | ~700 notes in five weeks from one representative |
| Support operations | A weekly schedule-adherence report built from a prose rule set and run against raw data | Replaced a manual four-hour weekly build |
| Claims | A 20-day new-hire training checklist with trainer rotation and per-day procedure references | ~20 training-material sessions |
| Claims | Desktop procedures authored and revised against the plan's own documents | ~35 authoring and lookup sessions |
| Finance | Form 5500 and Schedule A prepared for a group insurance client from live enrollment data | 58-message session, 14 distinct tools |
| Actuarial | The annual reinsurance report built tab by tab against the carrier's template | 240-message build |
| Underwriting | Prospect experience workbooks assembled from incumbent-carrier PDFs and emitted as Excel | ~250 workbooks in seven months from one analyst |
| Underwriting | A calendar-year versus plan-year accumulator study across the 36-group book | One 376-message session |
| Data engineering | A provider-file ingestion pipeline: 16 fixed-width source files into one 55-column master extract | 246-message build, repeats per network |
| Data engineering | A stored procedure assigning Major Diagnostic Category and DRG to the claims table | 120-message build |
| Sales | RFP questionnaires completed from prior approved answers under an explicit no-fabrication rule | ~65 RFP sessions in 8.5 months |
| Sales | A sales associate built and corrected their own reusable RFP-intake skill, then ran it as a one-word prompt | Hired mid-July, in use by day 10 |
| Account management | Renewal packages — loss ratio, membership, average age, service-area share — into a broker-ready email | 40+ packages in seven months from one AE |

| FUNCTION | WHAT WAS PRODUCED | SCALE OR OUTCOME |
|---|---|---|
| People | The employee handbook reconciled for a remote workforce across 21 states, with per-state accrual and leave rules | ~30 policies in roughly two weeks, an annual cycle |
| People | Annual headcount and compensation analysis, benchmarked by role against the market | 41 benchmarking sessions |
| Provider network | A department operating system: outreach tracker with owners, intake form, policy template and written SOP | Five-person team, about four weeks |
| Provider network | Proposed contract rates converted to percentage-of-Medicare against the local fee schedule, inline during negotiation | Sourced and enriched in the same session |
| Marketing | A brand and terminology style guide that assembled itself out of standing corrections during live work | ~214-turn thread over eleven weeks |

05 ECONOMICS

## What it is worth

We have separated three things that are easy to blur together. Only the first belongs in a payback calculation.

**Hours displaced** — `~3,900 hrs · $254K`

Work we were already doing that now takes less time.

**Capability we did not have** — `~2,600 hrs · $250–500K`

Work that would never have happened otherwise: ORSA, URAC evidence, the 21-state handbook, the IBNR back-test. Genuinely valuable, but it is not a cost saving and we have kept it out of the payback figure.

**Compute delivered below market** — `~$300K / yr`

July's 5.3B tokens price out at roughly $25K at list rates, uncached. The actual bill was under $10K. At the July run rate that is ~$300K a year of model compute delivered for about a third of its list price, and the gap widens every month volume doubles.

**Total value created** ~$1M / yr

The four lines above are distinct and do not overlap: $254K displaced + $250–500K new capability + $87K avoided + ~$300K compute = **$0.9–1.1M a year**. For a payback calculation use only the first line: $2,702 per active user per year, across 94 monthly users.

**WHAT KEEPS THE RUNNING COST DOWN**

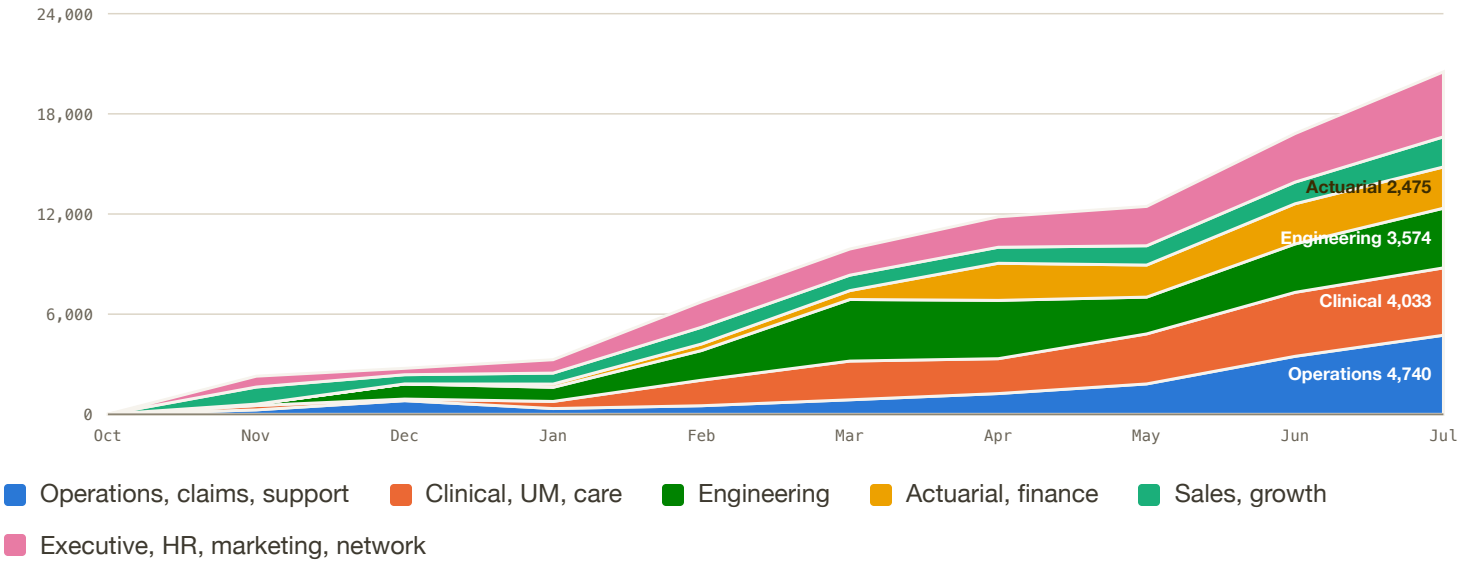
Two automatic levers keep the running cost down as volume climbs: the router reserves the frontier tier for about 4% of requests, and prompt caching serves repeated context at a tenth of the input rate (section 07).

06 ADOPTION

## How it spread, and what it took

MESSAGES PER MONTH BY FUNCTION

Engineering (green) leads early and flattens after March. Operations (blue) starts slowest and finishes largest. Actuarial (yellow) appears in January and steps up sharply in April when one person found a use that fit.



MESSAGES PER MONTH BY FUNCTION, FULL MONTHS

| FUNCTION | NOV 25 | FEB 26 | APR 26 | JUL 26 | USERS, JUL |
|---|---|---|---|---|---|
| Operations, claims, support | 264 | 516 | 1,248 | 4,740 | 22 |
| Clinical, UM, care | 258 | 1,532 | 2,086 | 4,033 | 13 |
| Engineering | 90 | 1,770 | 3,490 | 3,574 | 13 |
| Actuarial, finance | 0 | 392 | 2,228 | 2,475 | 4 |
| Sales, growth | 1,020 | 1,012 | 958 | 1,799 | 9 |
| Executive, operations | 176 | 826 | 720 | 1,245 | 5 |
| People, HR | 388 | 618 | 562 | 1,451 | 10 |
| Marketing, communications | 0 | 0 | 428 | 655 | 2 |
| Provider network | 88 | 76 | 110 | 546 | 7 |

## Engineering starts it, operations inherits it

Engineering went first and peaked in March. Operations was the last big function to pick it up and is now well ahead of everyone else, still climbing. An organisation that starts with its engineering team should keep measuring past that first group. **The return builds about two quarters later, in operations.**

## The late functions did the best work

Actuarial started three months after everyone else and marketing six months after. Both produced the strongest results in this study. Arriving late did not predict low value, and in our case those teams turned up with harder problems.

## It reaches everyone quickly and it sticks

Six of our ten functions were live within six weeks of launch, and our first group of users is **95% still active at month eight**. Reach is fast and retention holds.

The habit also changes shape as it settles. People use it more often but in shorter exchanges, from 19.5 messages per conversation in their first week to 6.6 by month four. That is the signature of something becoming routine rather than novel.

## Ten months in, it is still spreading sideways

July was our largest month on every measure, and the useful detail is where the growth came from. **Actuarial and clinical, our two deepest and most established workflows, were close to flat.** The month's growth came from operations, sales, the provider network team and staff who had only just been onboarded. Ten months in, the platform is still reaching functions that were not previously heavy users, which is the opposite of the plateau most rollouts are expected to hit.

The work also got deeper rather than just more frequent. **Sessions of ten turns or more grew faster than volume did**, and retrieval from our own systems — the internal knowledge base, SharePoint, Outlook — roughly doubled in a month. Usage is moving away from general capability and toward work grounded in Evry's own records, which is both the harder thing to replicate and the thing that takes an organisation's own data to reproduce.

07 VOLUME

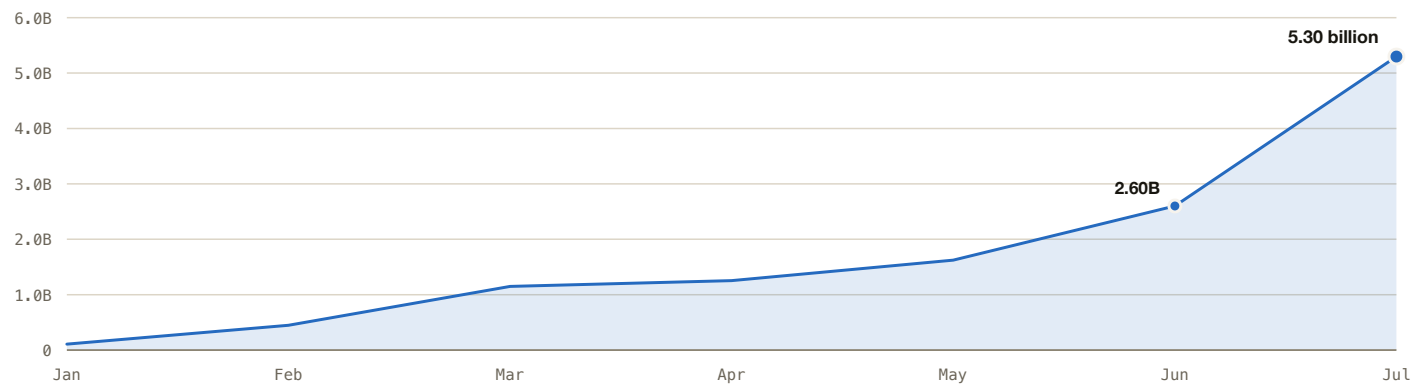
# What the platform actually processed

Everything above counts conversations. This section counts the work inside them. Between January and the end of July the platform processed **12.5 billion tokens**. To put that in proportion, it is the equivalent of reading and writing several thousand full-length novels, all of it about claims, members, benefits, reserves and regulation.

**July alone was 42% of it, at 5.30 billion tokens.** That is 49 times January and **104% above June**. Six straight months of growth and the steepest one is the most recent.

TOKENS PROCESSED PER MONTH

Input plus output, as reported by the model provider. Every month since January has been larger than the one before it.

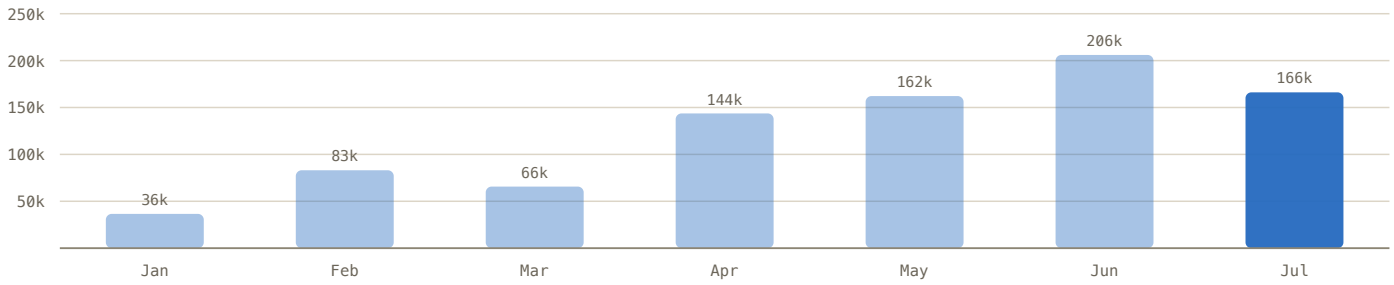


## The growth is in depth, not chattiness

Requests roughly doubled between January and July. Tokens rose 49 times. The difference is how much context each request now carries: an average request in January drew on **36k tokens**, and by July it was **166k**. One request today is a whole working session: documents, spreadsheets, claims extracts, long threads and the results of every tool call the platform made on the person's behalf before answering.

AVERAGE TOKENS BEHIND A SINGLE REQUEST

Six and a half times more context per request than in January.




## A growing share of it runs without anyone typing

Autopilot went live in June. In July it processed **1.36 billion tokens across 5,780 runs**, roughly **26% of everything the platform did that month**, and it did that with no one in the chat window. In its first month it was 160 million, so it grew more than eight times in four weeks.

This matters for how we describe the platform. A chat assistant is bounded by how much people are willing to type. Delegated work is not, and in its second month it is already close to a third of everything we process.

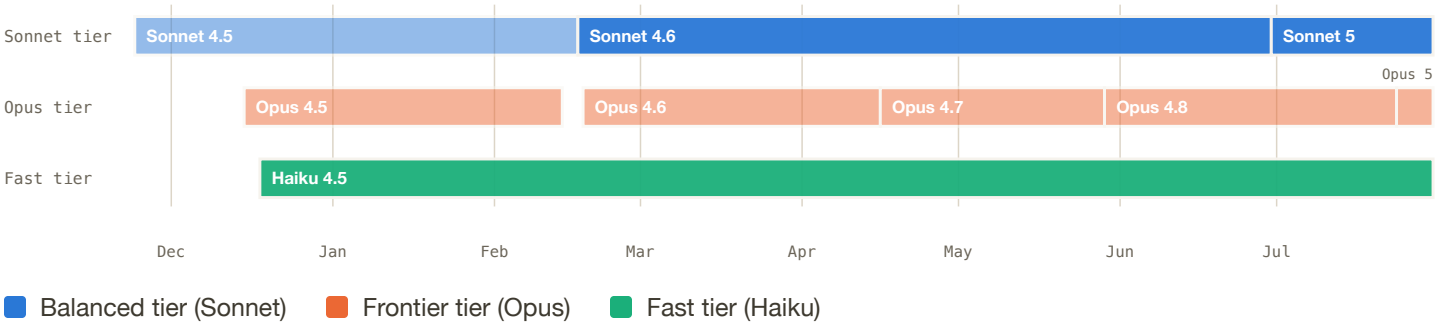
**08** MODELS

# Nobody at Evry picks a model

The platform has run **24 distinct models** from four providers, across five generation changes in the two main tiers in nine months. Almost none of that was visible to the people using it. 99% of conversations run on *auto*: a classifier reads the request, sends the ordinary majority to the balanced tier and reserves the frontier tier for the genuinely hard ones.

MODEL GENERATIONS BY TIER

Each bar is one model in production. Note that the bars butt up against each other: in every case the successor's first request and the predecessor's last fall on the same day. Darker bars are the generations that carried the bulk of the volume.




Five same-day handovers, no migration project, no announcement, no user retraining, and no conversation broke across the boundary. That is worth stating plainly because it is the single hardest thing to retrofit into a platform later, and it is the reason a model release is a configuration change here rather than a quarter of work.

EVERY CLAUDE GENERATION IN PRODUCTION, IN ORDER OF ARRIVAL

| MODEL | FIRST SEEN | LAST SEEN | REQUESTS | TYPICAL CONTEXT |
|---|---|---|---|---|
| Sonnet 4.5 | 11–24 | 02–17 | 7,420 | 6k |
| Opus 4.5 | 12–15 | 02–14 | 77 | 16k |
| Haiku 4.5 | 12–18 | 07–31 | 16,321 | 32k |
| Sonnet 4.6 | 02–17 | 06–30 | 32,440 | 6k |
| Opus 4.6 | 02–18 | 04–16 | 449 | 8k |
| Opus 4.7 | 04–16 | 05–29 | 460 | 67k |
| Opus 4.8 | 05–29 | 07–24 | 2,206 | 11k |
| Sonnet 5 | 06–30 | 07–31 | 20,800 | 43k |
| Opus 5 | 07–24 | 07–31 | 79 | 133k |

## Each generation absorbed roughly twice the context of the one before

Typical context went 25k on Sonnet 4.5, 50k on Sonnet 4.6, and 93k on Sonnet 5. That is not people changing their habits every four months. It is the same people bringing bigger problems as the ceiling lifted, and it is the clearest evidence we have that capability released by the model vendor turns into work here almost immediately.

The clearest single example: **July, the first full month on Sonnet 5, processed more than the platform's entire first five months of 2026 put together** (5.30B against 4.58B).

## The frontier tier got cheaper as it got better

Opus 4.8 cost $15 per million input tokens. Opus 5, which replaced it in late July, costs $5. The most capable option available to an Evry employee is now a third of the price it was in June, which is why the routing policy can afford to be more generous with it going forward than it has been.

> **WHY THIS SECTION EXISTS**
>
> Model choice is usually the first question anyone asks and the least interesting one. The useful answer is that we are not committed to a model, we are committed to a routing layer, and we have now changed the underlying models five times without anyone noticing. That is the durable part.

## Aggressive token management is why the bill stays flat while volume doubles

Priced at list rates with no optimization, July's 5.30 billion tokens come to roughly **$25,000**. The actual bill was **under $10,000** — a 60–70% reduction, from three levers that are all automatic:

The first is routing. The frontier tier is reserved for the requests that need it, so it is about **4% of requests** (though 27% of list-rate spend) while the balanced tier carries the ordinary majority. The classification itself runs on open-weight models, so no frontier token is ever spent deciding where to send a request.

The second is prompt caching, switched on at the end of June. Roughly half of all input is repeated context, served back at **a tenth of the normal input rate**, which is why volume can more than double in a month without the bill following it in a straight line.

The third is that a large share of volume is the platform coordinating itself — Autopilot runs, sub-agents, research passes, roughly **40% of all tokens** — and that machine-initiated work is exactly the traffic caching and routing compress hardest.

At the current growth rate the platform will be consuming more than **$500K a year of list-rate compute by autumn**, still delivered for a fraction of that. None of it depends on anyone at Evry making a careful choice.

09 WHO THIS IS RELEVANT TO

# Where the findings are likely to carry

## Accreditation and regulator evidence travels furthest

Of everything we found, this is the use case that transfers most readily to another organisation. It is the only high-value workflow that runs entirely on documents a plan already holds: no member data, no integration work, no security review before anyone can see whether it works. Every other strong use case needs the organisation's own data loaded first, which adds months before there is anything to judge.

## Compliance and quality leadership have the clearest stake

They are the people who have to sign off on any AI use in a health plan regardless, and the accreditation and regulator work sits directly in their remit. Finance and actuarial leadership are the next closest: the evidence there is stronger still, but it is a narrower audience and it needs their own data in front of them before it means much.

## Utilization management is the depth, not the starting point

It is our highest-volume use internally and the deepest evidence we have that the platform works on clinical operations. In our experience it only lands once the platform has already been trusted on something lower-stakes, so accreditation work comes first and clinical operations follow.

## What proved hard to reproduce

Three things, in the order they mattered:

- **It is joined up to the systems of record.** SharePoint, Outlook, Teams, GitHub, the claims warehouse and a purpose-built provider-network tool, all reachable in one conversation. A third of our active users query the provider network conversationally.
- **The workflows are encoded, not remembered.** Skills our own staff assembled, the reusable RFP intake, the claims quality apparatus. Once written down they keep working without the person who wrote them.
- **The routing layer absorbs model change.** Five generation changes in nine months without a migration project (section 08). Nothing had to be re-platformed to stay current.

The organisations most likely to recognise themselves in this report are the ones that need several of these at once and need them inside their own compliance boundary. That describes most regional payers.

**WHY THE TRANSCRIPTS ARE THE INTERESTING PART**

Capability is easy to demonstrate and hard to trust. What is rarer is a production transcript of a system withdrawing a $1.65M recommendation after its own back-test, declining to pass a standard the team wanted passed, or catching that a proposed shortcut would not meet a state insurance rule. In front of a risk committee that carries further than an accuracy benchmark, because it can be read end to end.

Every case in section 04 is named and verifiable against the record for that reason.

**10 METHOD**

# How the numbers were produced

Every figure comes from a full census of production records between 1 October 2025 and 31 July 2026, not a sample or a survey. Ten analysts each examined one department, nine flagship conversations were read in full, and four further passes reconciled the results.

Both sides of every conversation were read: 10,472 conversations, 87,684 messages and all 43,763 responses the platform sent our staff, around 27 million tokens of output. Volume and model figures come from a separate per-request census of every model call. Behavioural rates such as self-correction were detected automatically, then a sample of each was graded by hand and the rate adjusted to the measured precision, so the figures quoted are corrected rather than raw. Every script is committed and the whole analysis can be re-run.

## What the numbers rest on

- Counts of usage, documents, tokens and behaviour are measured directly from the production record, including a per-request census of every model call.
- Time savings are modelled from stated assumptions about how long each task takes by hand. The reserving work is measured against six months of actual claim runout.