Recapp

Recapp, explained

The field taught from scratch, for a reader who is technically strong and does not work in it.

2026-09-02

Prepared by Copundit

AI for Advanced Ideation™

The field taught from scratch, for a reader who is technically strong and does not work in it. Every term is defined where it first appears. The pitch is elsewhere; this file only explains.

Recapp is a proposed phone application that records an in-person meeting through the phone's own microphones and hands back three written things within a few minutes of the meeting ending: a transcript of what was said, a summary of it, and a list of action items. The customer it aims at is the person who currently buys a dedicated pocket recorder to get exactly that: a card-sized object that sticks magnetically to the back of the same phone and sells for about 159 US dollars, with a subscription behind it.

Before anything else, two pieces of vocabulary, because everything below turns on them. **Automatic speech recognition** (ASR) is software that turns recorded speech into written words. **Word error rate**, written WER, is how its accuracy is reported: the percentage of words a transcript gets wrong, counting substitutions, deletions and insertions against a human reference. It is the field's universal number, and it is comparable between two systems only on the same test set under the same audio condition. That condition matters more than anything else in this document, and this field has three of them, not two. **Read speech** is somebody reading a prepared passage into a close microphone, and it is where the famous low-single-digit accuracy numbers come from. **Near-field** speech is a real conversation captured at somebody's mouth, by a headset, a lapel clip or a phone held to the face, and it is harder. **Far-field** speech is several people talking across a table in a room with echo, air conditioning and a corridor outside; it is harder again, and the field treats it as a separate discipline with its own name, **distant automatic speech recognition**, and its own competition series running since the early 2000s. One recent system measured on one corpus puts a size on the last step: the same meetings score 16.40 percent when captured by the participants' own headsets and 21.26 percent when captured by a single microphone in the middle of the table, on a metric that also credits words to the right speaker (defined below). The room costs about five points, and the near-field starting point is already 16 percent, not 2. A product that records a meeting from a table is doing the far-field job while borrowing the reputation of the read-speech one, and almost every confusing claim in this market comes from that swap.

Contents

Why this is not the note-taking bot that joins your video call

The product most readers already have a mental model for is the meeting bot: you schedule a video call, a software participant appears in the attendee list, and afterwards a transcript and a summary arrive by email. Fireflies is the clearest example, and Zoom, Microsoft Teams and Google Meet now ship their own versions inside the call. That is a different product from this one in four ways, each of them a design decision rather than a preference.

The input signal is physically different. A conferencing platform hands the bot a clean stream per participant, each captured near-field by that person's own headset or laptop microphone and already separated by speaker. A phone lying on a meeting-room table hands an application one mixed channel containing everyone at once, plus the room. The bot is solving the near-field problem, which is not itself a solved problem, as the 16.40 percent above shows; the phone is solving the far-field one. Error rates from one are not evidence about the other, in either direction.

There is no call to invite a bot to. The meeting this product is about is one you physically walk out of: a room, a corridor, a coffee shop. Nothing exists for a software participant to join, so the whole distribution mechanism the bot category grew on, arriving through the calendar invite and being seen by everyone in the meeting, is unavailable.

The consent surface is different, and it is the design decision with the largest legal consequence. The bot's presence in the participant list is itself the notice to the room. A phone face down on a table gives no such signal. That is not a detail: one of the four pending United States class actions in this category, *Chamberlain v. Granola* (Northern District of California, July 2026), pleads wiretapping over silent device-level capture with no visible bot, against a product that processes audio on the user's own device. Where the computation runs is not a defence against that theory.

What is being displaced is an object, not a plugin. The competitor here is a 159 US dollar piece of hardware with its own microphones, its own battery and its own subscription, sold in retail. That changes the economics on both sides, as the money section below sets out.

One more neighbour worth taking away, because the vocabulary invites it: this is not dictation. A voice-memo or dictation application solves the near-field, one-speaker problem, and solves it well. Nothing about that performance carries over to six people in a room.

The raw material: meetings that somebody recorded and then transcribed by hand

Everything measurable in this field rests on a small number of corpora, meaning recorded conversations paired with a human-written reference of what was actually said, who said it, and sometimes what it amounted to. They are the raw material in the literal sense: no error rate, no comparison between two devices and no claim about summary quality exists except as a number computed against one of them.

What one costs to produce. A usable record needs three things at once. It needs real conversation, not read speech, because people interrupt, trail off and raise their voices in noise in ways a script does not reproduce. It needs the far-field audio the product will actually see. And it needs a ground-truth reference, which in practice means a headset, lapel or in-ear microphone worn by every participant, so a transcriber can hear each person cleanly. That last requirement is why these corpora are rare: it costs the consent of every person in every meeting, recorded and kept. Annotation is then done by hand and reviewed; the one corpus that documents its procedure used three trained annotators plus a senior reviewer over every session. The creators of one of the field's newest corpora deliberately refused to let annotators start from machine output, on the ground that people accept a plausible but wrong machine guess in exactly the noisy segments that matter.

What exists, and how long each session runs. Session length is not decoration here: it is the axis on which every number in this document is a floor rather than an estimate, for the reason given at the end of this paragraph. The AMI Meeting Corpus is roughly 100 hours recorded on both headsets and an

eight-microphone table array, with 137 sessions carrying a four-part human summary; the sessions run 15 to 45 minutes, and they are **scenario** meetings, which is the field's word for role-play, in this case four people acting out roles in a fictitious electronics company. That matters more than it sounds: the action items in AMI are partly baked into the role-play's own rules, so they are not quite commitments that people made. The ICSI corpus is 75 academic meetings of about an hour each, about 72 hours, 61 of them summarised. AliMeeting is about 120 hours of Mandarin meetings of 15 to 30 minutes, recorded near-field and far-field simultaneously. NOTSOFAR-1 is real corporate office meetings across 30 rooms with a withheld evaluation split, and its sessions are truncated to about six minutes; note that its own two papers disagree about how many meetings it contains, giving 280 and 315, while the public repository holds 237. CHiME-6 is 20 dinner parties of at least two hours each. LOTUSDIS is 90 Thai sessions of 15 to 20 minutes, three speakers each, recorded simultaneously by nine separate single-microphone devices placed at measured distances from 0.12 to 10 metres, which makes it the closest published design to the question this project actually asks. Now the consequence: the meeting this category is bought for runs forty minutes to two hours, and nothing in a six-minute clip exercises what an hour does. Speaker groupings drift as voices warm and people move, a recogniser conditioned on its own previous window can carry one hallucination across a hundred windows, and the memory, storage and thermal budgets that decide whether a phone finishes the job never come under load at all.

What is missing, and the ceiling the sources put on their own answer. Nothing in the family was recorded on a smartphone. A recent survey's 36-row table of notable meeting corpora contains no smartphone-captured entry, and the field's own review of the last two challenge rounds lists 32 notable robust-recognition datasets and challenges since 2000 without naming a phone as a capture device. Two independent research sweeps assert the same null, and neither cites a catalogue behind it, so this is an asserted absence rather than a proven one. The nearest neighbours each fail for a stated reason: one academic group did build an ad-hoc array out of iPhones and never released the audio; one commercial vendor holds a 1,000-hour conversational smartphone corpus available only by purchase; one challenge corpus was recorded on smart glasses.

The gap is public, and the corresponding private holdings are not. This is where a reader should resist concluding that nobody has the data. The incumbent application companies do. Otter states its models are trained on millions of hours of audio from its own users; Gong holds a proprietary graph over business-to-business sales calls and internal meetings; Microsoft, Google and Amazon hold the audio inside Teams, Meet and Chime. All of it is real meetings of natural length captured on consumer microphones, and no outside researcher or competing developer has obtained any of it, on any terms that anyone has published. The public absence is therefore symmetric only for entrants; for the incumbents it is a holding.

Why the gap cannot simply be filled. Three routes are closed. Simulation is closed: the field accepts synthetic far-field audio for training a model and refuses it as evidence for a hardware claim, because mixing recorded voices together in software does not reproduce turn-taking, interruption, or the way people raise their voices in noise. It is closed from the other end too, because the leading simulated meeting set was built from 15,000 room measurements taken on the target hardware itself, so simulating a new device needs the recording sessions the simulation was meant to avoid. Repurposing existing archives is closed: podcasts are near-field, video-call recordings have been irreversibly processed by the platform's own noise suppression, and parliamentary proceedings use push-to-talk gooseneck microphones. And the one corpus that recorded exactly the right thing, real unscripted project meetings averaging over an hour, withheld its audio for privacy and released censored transcripts only. That is the same force that would apply to anything collected fresh.

The third output is the least supplied of all. Action items exist as a labelled target in three places. In AMI, 101 meetings carry 381 action items, and those labels are derived by linking dialogue acts to the action-related part of a human summary rather than annotated directly, so the count is one team's floor rather than a census. The largest purpose-built set is Chinese: 424 meetings and 306,846 utterances carrying 1,506 action items, with annotator agreement at Cohen's kappa 0.47, which is moderate. A turn-level layer over 22 ICSI meetings finds 318 actionable turns out of 21,035, that is 1.5 percent. The earliest English action-item annotation, covering 18 ICSI meetings, is no longer publicly available at all. Licensing narrows this further: of the corpora above, only AMI and LOTUSDIS carry a licence permitting

commercial use that has actually been read from the holder's own terms; one widely used set is Non-Commercial; and for NOTSOFAR-1 three sources give three different answers, with the dataset paper itself stating no licence anywhere.

Why "the phone has four microphones" is not the answer

The first thing a technical reader proposes, correctly, is that a modern phone contains three or four good micro-electro-mechanical system (MEMS) microphones, which are the same class of component a pocket recorder uses, and that an application could therefore run the same multi-microphone processing a purpose-built device runs. This is the load-bearing part of the document. Read it with one fact held in front of everything else: **nobody has ever measured a phone against a dedicated recorder on the same meeting, in either direction.** No published word error rate exists anywhere for a smartphone recording a multi-speaker meeting, no corpus in the field was recorded on a phone, and no independent head-to-head recording exists; a search for one turned up vendor marketing and one enthusiast log comparing two models of the same recorder to each other. What follows is a mechanism argument about which condition each device competes in, and mechanism arguments do not settle quantities. Nothing below says the phone hears a meeting worse than the recorder does. That comparison is unmeasured, and the protocol that would settle it is described at the end of this section.

How much the audio condition matters is easiest to see inside a single engine. Apple's on-device recogniser scores 2.12 percent word error rate on clear read-aloud speech, and 14.0 percent on about twelve hours of real corporate earnings-call conversation. Same software, roughly a sevenfold difference. This is why a vendor claim such as "95 percent or better accuracy" or "98.86 percent across 58 languages", naming no test set, no microphone and no audio condition, is not evidence of anything in either direction. No vendor in this category publishes a quality figure that names its condition.

Meetings need a second number, because a meeting transcript has to say who spoke as well as what was said. The field's ranking metric is the **time-constrained minimum-permutation word error rate**, written tcpWER: it concatenates each speaker's words, tries every possible matching of system speakers to real people, reports the best one, and additionally requires the words to land in roughly the right place in time. A system that hears every word perfectly and credits it to the wrong person is still penalised. Where the same idea is scored without the timing constraint the field writes cpWER, the concatenated minimum-permutation word error rate, which is the metric behind the 16.40 and 21.26 percent figures in the lead.

Now the finding, stated exactly as the evidence supports it. No documented route exists on either mobile platform for a third-party application to obtain the raw per-microphone signals, and no handset has been shown to expose one. On Apple's platform this is a platform-level negative: input routing belongs to the audio session interface and there is no request for a synchronised four-channel stream from the built-in microphones, while the two-channel path that does exist is reported by developers to be forced down to 16 kHz. On Android the position is weaker and genuinely open. A constant exists for unprocessed audio, but the compatibility specification guarantees it only for the built-in system camera application, which leaves third-party access to each manufacturer's own hardware layer, and nobody has tested the fleet. Where a third-party application has asked for multi-channel unprocessed audio the reported result is one mono channel mirrored across two tracks, but that is a report, not a survey. If some manufacturer does expose true unprocessed multichannel audio, this finding flips from a blocker into an advantage on that handset, and which of the two it is remains open. A second open question sits beside it: the platform's own automatic gain control, noise suppression and echo cancellation are inserted when an application declares a voice-communication mode, and they destroy the phase relationship between microphones that spatial processing reads. Whether choosing a plain recording session instead avoids that chain has not been established either way.

So the relevant quantity is not how many microphones a device contains. It is how many **phase-coherent channels** reach the software: audio streams whose timing relationship to each other is intact. On the published evidence a third-party phone application should expect one processed channel. What the recorder gets is a per-model question, and the answers are not uniform: the 159 dollar card is reported

as two MEMS microphones plus a vibration sensor used for call audio, the more expensive Pro model as four microphones with a stated 5 metre pickup range and beamforming, and the pendant model as a single mono microphone. Those counts come from retailer listings, review sites and comparison blogs; nobody has published a physical examination of any of these devices, so treat the counts as reported rather than confirmed. They change the comparison rather than decorate it: against the mono pendant it is single-channel software against single-channel software, and the phone may be competitive; against the four-microphone Pro the same source expects the phone to lose badly. The card that sticks to the back of the phone occupies the same point in the room as the phone and may still get a different number of channels. A specification sheet listing a microphone count is not a statement about this.

What the extra channels are worth is measured, on a comparison that is not this one. On the same 170 real office meetings in the NOTSOFAR-1 evaluation set, each about six minutes long, the winning system scored 22.2 percent tcpWER reading one distant microphone stream and 10.8 percent reading a seven-microphone tabletop array. The published baseline for the single-channel condition was 41.4 percent. Read that gap carefully in both directions: it is a single-device-versus-conference-array gap, and a card-sized pocket recorder does not have a seven-microphone tabletop array either, so it pays some version of the same penalty. It is not a measurement of this product against that one, and 22.2 percent is a ceiling on what a phone could reach, not a prediction of what it does reach.

The spatial half of that gap is not recoverable in software, and there is a mathematical reason. The front end that produces most of the multi-channel advantage, guided source separation, builds a spatial model from the phase differences between microphones and then beamforms. One channel has no phase differences to build it from, which leaves a single-channel system restricted to separating voices by their spectral shape alone, and that fails exactly when two overlapping speakers sound alike or when reverberation smears the envelope. The challenge organisers describe the result as a significant gap that single-channel systems are currently unable to bridge, and they note it is worst precisely where meetings are hardest, in debate-style overlapping speech. Overlap is not an edge case: about 19 percent of AMI's duration is two or more people at once, 29.6 percent of the NOTSOFAR-1 evaluation set, and 34 to 42 percent of AliMeeting.

Distance costs more than most people expect. The Thai corpus mentioned above is the only published study that puts several single-microphone devices at measured distances on the same conversation. With an off-the-shelf model, word error rate ran 36.4 percent on a table condenser at 12 to 15 centimetres, 44.2 percent on a tabletop unit at 2 metres, 96.3 percent on a speakerphone at 3 metres and 104.2 percent at 10 metres. A rate above 100 percent is possible because the recogniser inserts words that were never said. Fine-tuning on distance-diverse in-domain data pulled the far-field average from 81.6 to 49.5 percent, so the penalty narrows and does not close. The language is Thai, so the absolute levels do not transfer to English, but the shape of the curve is about acoustics rather than language. Two further results from that study transfer without qualification and are counter-intuitive: classical dereverberation and denoising front ends made recognition worse at every distance, and a model fine-tuned on one microphone type collapsed on every other type unless reverberation augmentation was added.

Where the phone is put costs something too, and the pitch names the two worst places. A phone lying flat on a table picks up each voice twice, once directly and once a fraction of a millisecond later off the table surface, and the two copies cancel each other at predictable frequencies. Face down, the top microphone is physically covered by the table, which acts as a low-pass filter and destroys the high-frequency fricatives, the s, f and th sounds, that carry most of the information distinguishing one name from another. In a shirt or trouser pocket the sound has to cross fabric first, and fabric attenuates by up to 24 dB at 500 Hz and up to 28 dB at 1 kHz; because the loss grows with frequency, the vowels get through and the consonants between 2,000 and 8,000 Hz do not. A device worn on the outside of clothing, which is what the pendant recorders are, skips that loss entirely. Hold this next to the summarisation finding several sections below, which says the summary survives a noisy transcript only when the errors miss names, numbers and decisions: these are the same fact, and the placements the product depends on attack exactly the frequencies that names live in. What no one has is the number. Nobody has measured word error rate for a real phone in a real meeting face-up on a table, face-down, in a shirt pocket, in a trouser pocket or in a bag. The mechanisms are documented for all five and the measurement exists for none.

The obvious next move, measuring it, is blocked in a specific way. Since simulation is not accepted as evidence for a hardware claim, settling the phone-versus-recorder question requires a new parallel recording: both devices adjacent at the centre of the table, a close-talking headset on every participant as the reference channel, alignment by cross-correlation afterwards because two independent devices share no hardware clock and drift over half an hour, and the identical recogniser decoding both streams, because a different recogniser on either device confounds the hardware test completely. The published shape of such a study is at least 20 sessions of at least 15 minutes, at least 20 speakers in groups of three to five, across at least four acoustic environments. It is worth knowing in advance what it can return. If the platform's own gain control and noise suppression cannot be avoided, then the thing being compared is the phone as a shipped ecosystem against the recorder rather than one microphone capsule against another, and the best result a practitioner would accept as a win is statistical equivalence, not superiority.

One last distinction that is invisible in every number as normally quoted. "Single-channel" carries two incompatible meanings in this literature. In the NOTSOFAR-1 family it means a commercial conference device's output stream *after* that device's own echo cancellation, dereverberation, beamforming and noise suppression; the dataset authors say so explicitly. In the Thai corpus it means one literal microphone element with no processing. A phone application sits closer to the second, and the second is where the higher error rates are.

There is also a difference between what is measured looking backwards and what would be measured looking forwards, and it separates the two halves of this product. Every accuracy number above was computed retrospectively, on archived audio, from meetings that were recorded successfully and cleanly truncated before release. The quantity that decides whether the product works at all, whether a phone actually holds the microphone for a whole meeting on a real handset, can only be measured prospectively, by instrumenting a deployed application. Nobody has published that measurement, and the corpora structurally cannot contain it, because a session that failed mid-collection was discarded before publication.

What the product actually computes, stage by stage

One decision is open, and the reader needs it before the stages make sense. Two questions have not been answered by this project's founders: whether the audio leaves the phone at all, and which mobile platform ships first. They are not details. The document's own latency figures, its cost per recorded hour, its battery draw, whether a token limit applies to the summary at all, and its exposure in the pending lawsuits all fork on the first, and the capture adversary, the store commission and which free pre-installed product the thing stands next to all fork on the second. Rather than pick quietly, here is what each branch costs and buys.

Audio leaves the phone, or does not. Sending audio to a hosted recogniser costs 0.21 to 0.62 US dollars per recorded hour and makes breakeven on a 15 dollar subscription arrive at roughly 32.5 recorded hours a month; running recognition on the phone costs about 0.05 dollars an hour in the usual hybrid arrangement and pushes breakeven past 250. The cloud path is measured at 10 to 20 minutes to a finished summary in its common batch form, against 1.5 to 5.5 minutes derived for a fully on-device path, because queueing and transport rather than compute are what the user waits for. The cloud path costs no battery worth naming; the on-device path is put at 25 to 40 percent of a charge an hour for a continuous pipeline. The cloud path has no token limit; the on-device path inherits its supplied model's, which on Apple's is 4,096 tokens per session. And the on-device path is a real privacy position only if the transcript stays too: the arrangement that produces the 0.05 dollar figure keeps recognition local and sends the full plaintext transcript to a cloud language model, which one source calls legally indistinguishable from sending the audio for healthcare, legal and finance users. Privacy in this category is not a free consequence of being an application. It is bought by accepting the token limit.

iOS first, or Android first. On Apple's platform the adversary is the operating system itself, in one well-documented place, and the neighbouring free product is Apple Notes. On Android the operating system is comparatively permissive and four handset manufacturers are the adversary, and the neighbouring

free product is Pixel Recorder, which is a considerably harder competitor: it records, transcribes offline in real time, labels speakers and summarises on the device, free, with no length limit beyond storage. The commission schedules also differ, in a way the money section takes up.

With that named, the pipeline itself is six stages, in the order that makes each one necessary. Each has an input, an output, and a reason the stage before it was not enough.

Capture: hold the microphone for an hour. Input, the room. Output, an audio file. This is first because it is the only failure that nothing downstream can repair: no model recovers audio that was never written.

On Apple's platform an application declares a background audio capability and holds an audio session, and what happens next is documented and deterministic rather than probabilistic. When a call comes in, the telephony system takes the microphone away at the moment the phone starts ringing, not at the moment the user answers, so declining a call does not protect the recording. When the interruption ends, the system sends a notification suggesting the application resume, and then the privacy subsystem refuses to let a backgrounded application reactivate the microphone; the attempt returns a specific error code. Unless the user unlocks the phone and brings the application to the foreground by hand, the rest of the meeting is silently lost. Calls are not the only trigger. Users of Apple's own Voice Memos report a recording stopped by a muted video autoplaying in a browser tab the user had switched to, with no warning, and a ninety-minute recording lost outright when the battery ran down because the file had never been finalised to disk. Everything else continues: screen lock, switching applications, low power mode and ordinary thermal throttling all leave the recording running.

On Android the operating system is not the main adversary; four handset manufacturers are. Samsung's sleeping-application tiers and memory eviction, Xiaomi's and Huawei's termination shortly after screen lock even with the required persistent notification visible, and Oppo's freezing of background services on screen lock each end a correctly declared recording service, and each requires the user to walk a different multi-step settings path that major updates have been observed to revert.

There are two different failures here and the difference decides what engineering can do about them. The first is a stop: the operating system suspends or kills the recording, and the file simply ends early. That one has a standard mitigation, which is to encode and flush audio to storage every few seconds rather than hold one buffer until the user taps stop, so a hard kill costs the last unwritten seconds instead of the whole meeting. The second is worse and the mitigation does not touch it. Developers report the microphone feed continuing to fire after a background anomaly while delivering buffers full of zeroes. The application is not stopped, receives no error, and writes a normal-looking file of the right length containing digital silence. A stop can be detected and salvaged; a file of silence looks exactly like a successful recording until somebody opens it.

Recognition: audio to words. Input, the captured audio. Output, a transcript with no speakers attached. Capture was not enough because an hour of audio is not a written record. This is where word error rate applies, and where the far-field numbers above land. The architectural choice here is streaming against batch: a streaming pipeline recognises speech while the meeting is still running, so at the moment the user taps stop only the summary remains to be produced, whereas a batch pipeline starts the whole job at tap-stop. Streaming is one of the two architectures that reach the walk-back deadline, the other being the fully on-device path set out under delivery below, and it is not free. It trades the capture problem for a connectivity problem, because the connection has to survive the whole meeting in exactly the conference rooms and basements where it is least likely to. It forecloses the on-device position, because the audio leaves the phone continuously rather than once and deliberately, and the meter runs for the full duration whether or not anyone reads the result. A recogniser denied future context is worse than the same model run offline, and how much worse on far-field meeting audio has been measured nowhere; no streaming system has been scored on NOTSOFAR-1, AMI, CHiME or any other named meeting corpus. And nothing in the field's separation or diarization stack runs online at challenge quality, so a streaming product is choosing the plain single-stream recogniser and giving up the multi-channel front end entirely.

Speaker attribution: who said which words. Input, the audio and the transcript. Output, the same words with names or at least stable labels attached. Recognition was not enough because a decision or a commitment is meaningless without knowing who made it. The field calls this **diarization**, and it is separately hard for reasons that are not obvious. It is scored by its own metric, **diarization error**

rate, the percentage of speaking time given to the wrong speaker plus missed speech plus speech invented where there was none, and the published values are wide: across the four evaluation splits of the CHiME-8 challenge one baseline ran from 10.3 to 60.0 percent, and a current system reports 25.9 percent on AMI, 22.9 on AliMeeting and 24.3 on MSDWild. Merely detecting that two people are talking at once tops out at 82.76 percent on the F1 score, the combined precision-and-recall measure, on the AMI test set. The dominant residual error in the leading systems is not mishearing but mis-counting: getting the number of people in the room wrong, and many front ends assume at most three speakers in any local window. That matters here because nothing in a two-tap interaction tells the application how many people are present.

The field's standard answer to mis-counting is speaker enrolment: store a voice embedding for each known person so that attribution becomes verification against a known vector instead of unsupervised guessing, reported at a 12.60 percent diarization error improvement on one conversational corpus and 14.01 on another, neither of them a far-field meeting. The price is exact and it is legal rather than technical, and it is taken up below: the technique that fixes attribution is the technique that creates a biometric record.

The size of the attribution problem relative to the recognition problem is visible in one pair of numbers: the winning single-channel system scored 22.2 percent with speaker credit required and 17.7 percent with speaker identity ignored, so roughly a fifth of the errors a user would see are attribution rather than misheard words. Improving one metric can make the other worse, and this has been reported repeatedly; in one system a diarization stage that improved diarization error rate from 23.43 to 21.07 percent made transcription worse, from 12.87 to 14.13 percent. Neither mobile platform exposes speaker attribution to third-party developers, while Google's own built-in Pixel Recorder has labelled speakers since its version 4.2 in 2022. An application that wants speaker labels must therefore ship and manage its own model. Attribution is also where the legal exposure concentrates: clustering voices anonymously inside one recording is described as the safe position, and tying a voice to an identity, whether by enrolment, a calendar invite or an email address, is the step into biometric territory under Illinois's statute, where a voiceprint is a named biometric identifier with statutory damages per violation, though a 2026 appellate decision on retroactivity leaves the current per-instance multiplier unsettled.

Summarisation: transcript to prose. Input, the attributed transcript. Output, a paragraph or a structured set of sections. A **large language model** (LLM) is a text-generating system; here it reads the transcript and writes the summary. This stage is simultaneously the reason the product works at all and the most dangerous thing in it, and the reason is one measurement. Across 1,760 system-session pairs from two challenge series, transcription error and the best automatic summary score correlate at only -0.51, and systems above 50 percent tcpWER produced summaries scoring roughly on par with systems near 11 percent. The mechanism is that a language model acts as an error corrector, emitting fluent prose that masks the recognition failure underneath. Read one way, that is the finding that makes this whole category commercially possible: the artifact the customer reads is far more robust than the raw transcript beneath it. Read the other way, it is the hazard: a product can look excellent to its own user while its transcript is poor, and the user has no way to tell.

Three cautions travel with it, and the second is the one this document has to be blunt about. First, the summary side of that correlation was scored by another language model acting as judge, with documented self-preference and position biases, and no human read those summaries. Second, the tolerance is conditional and the study that measured it was not run on meeting audio: the threshold is about 0.2 word error rate, with per-model values of 0.070, 0.195, 0.249 and 0.297, and the degraded audio in that study was text-to-speech output reverberated and noised in software rather than anything recorded in a room, with the resulting summaries again ranked by a model rather than by people. That is the same class of evidence this document told the reader the field refuses for a hardware claim, and it is being leaned on here because it is the only measurement of the threshold that exists. The conditional half survives independently and is the part that matters: tolerance holds only when the errors fall on filler words, false starts and conjunctions rather than on names, numbers and decisions. Far-field acoustics smear consonants, so proper nouns fail first, which means the errors concentrate in exactly the places that break the tolerance. Nobody has measured a **named-entity word error rate**, that is a word error rate counted only over names, numbers and acronyms, on far-field meeting audio. There is a shipping

lever aimed straight at that problem, and it is worth knowing what it costs: contextual biasing, meaning feeding the recogniser a list of the names it is likely to hear so they beat acoustically similar common words, is exposed by every commercial recognition interface in this category as custom vocabulary or keyterm boosting, and is reported at a 44 percent improvement in named-entity word error rate overall and 57 percent on rare personal names, on an unstated audio condition. Building that list means reading the contacts and the calendar, which is the data a privacy-positioned product would rather not touch. Third, even with a perfect transcript the summaries are incomplete: under expert human annotation of machine summaries of 35 meetings by five models, on clean human transcripts carrying zero recognition error, missing information appeared in 89 to 97 percent of them.

Extraction: commitments to a checklist. Input, the transcript and the summary. Output, action items with owners. Summarisation was not enough because a summary is prose and a commitment is a task. This is the stage with the least measurement in the entire field. Its characteristic failure is not invention from nothing but misattribution: binding a task to whoever produced most of the words describing it. No hallucination rate has ever been published for this task, and the numbers that circulate as proxies come from clinical notes and citation fabrication, which are different problems. There is a deeper reason the measurement is thin. Human annotators agree at Cohen's kappa 0.85 to 0.89 on where a discourse unit starts and ends in the same kind of speech, and at kappa 0.47 to 0.64 on whether a turn contains an action item. Structure is decidable; importance is not, and that gap belongs to the task rather than to the annotators.

Delivery: the clock. Input, all of the above. Output, three artifacts in the user's hand. The promise here is that everything is finished during the walk back from the meeting room, so latency is a product requirement rather than an implementation detail. The measured answer inverts most people's intuition. Data-centre silicon is roughly 40 to 60 times faster at raw inference than a phone, running the same recogniser at about 597 times real time against 10 to 16 times on a recent handset, and yet measured end-to-end delivery runs the other way: cloud batch products are timed at 10 to 20 minutes to a finished summary while on-device paths are derived at 1.5 to 5.5 minutes, because multi-tenant queueing and network transport, not compute, are what the user waits for.

That on-device derivation assumes a one-hour meeting and no thermal penalty, and the same source that supplies it also supplies a 30 to 50 percent throttling penalty which it never applies. Carried through its own measured 10 to 16 times real-time band, a 90-minute recording lands at 11 to 18 minutes on the phone, which misses the deadline outright. Both derivations come from the same place and only one of them is usually quoted. A second fact from the same evidence points the same way: shipping applications already defer batch transcription until the device is charging or above 50 percent battery, which forfeits the walk-back promise by design.

Published turnaround across the category varies accordingly, and the two products that beat everyone are the free pre-installed ones. Pixel Recorder produces its on-device summary in 5 to 15 seconds, and Apple's Voice Memos has the transcript within seconds because transcription runs while recording, though it produces neither a summary nor action items. Among the paid applications, Otter's help centre states 1.5 to 3 minutes to a finalised transcript and 2.5 to 6 minutes to a summary for a 30-minute meeting while an independent 2026 review timed the same product at 10 to 15 minutes at an unstated recording length, outside the range its own help centre publishes; MeetGeek was independently timed at about 17 minutes on a 33-minute recording; Fireflies publishes 15 to 20 minutes; and Fathom was independently timed at under one to two minutes to summary and action items, though on a video call rather than a room. No vendor in the category publishes a guaranteed maximum.

There is one hard constraint on the on-device route, and it belongs to a supplied model rather than to the phone. A **context window** is the number of tokens, roughly word pieces, a model can hold at once including its instructions, the transcript and its own output. Apple's developer documentation states a hard 4,096-token per-session limit for the on-device model it supplies, covering all three together; the figure circulating for Google's equivalent comes from a developer blog rather than vendor documentation. One hour of speech is about 9,000 words, which two independent sources put at roughly 11,000 to 13,000 tokens, about three times the window. An hour-long meeting therefore cannot be summarised in one pass by Apple's supplied model; it must be chunked and rolled up. That workaround has no published quality evaluation of any kind, and its characteristic failure is specific: a decision negotiated across a chunk

boundary is exactly what a chunk-local summary loses, and no metric in this field detects a dropped cross-chunk reference. Note also that the limit is not obviously a property of the device. Google's own Pixel Recorder summarises a 41-minute recording on device against a stated 4,096-token window, so either the first-party application chunks internally or the published figure is wrong, and nobody outside those teams knows which.

How companies in this field die

This field has a graveyard, and the deaths are instructive because most of them are not commercial failures in the usual sense. In two of the three cases below the company was lost while the customers were still there, or partly still there.

The named case: Vowel AI. Founded 2018, it raised 17.8 million US dollars including a 13.5 million Series A in September 2021, and built a video conferencing platform of its own with real-time transcription, collaborative agendas and automatic summaries. Two things killed it, and the order matters. Zoom and Microsoft Teams shipped free native meeting summaries, which removed the reason to pay for a separate product; a signed venture term sheet was then withdrawn against a high burn rate, which removed the runway to respond. It announced its shutdown in July 2023, and the remaining team was acqui-hired by Zapier in March 2024. The asset a thin layer over somebody else's platform holds is the layer, and when the host ships the layer there is nothing left to sell but the team. The same mechanism ended Regie.ai when customer-relationship platforms shipped their own built-in writing assistants, ended Viable when its own model supplier gave enterprises the tools directly, and retired Dragon NaturallySpeaking for Mac after Microsoft acquired Nuance.

Humane. About 230 million US dollars raised against a peak valuation of 850 million, for a wearable pin with its own camera, microphone, projector and cellular connection, launched April 2024 at 699 US dollars and later 499, with a mandatory 24 dollar monthly plan. Returns exceeded new sales by the summer of 2024 and fewer than 10,000 units shipped. HP (Hewlett-Packard) bought the assets in February 2025 for a reported 116 million US dollars, which is 0.14 times the last private valuation. The ending is the instructive part: every device stopped working at noon Pacific time on 28 February 2025, and owners whose device shipped before 15 November 2024 received no compensation.

Limitless. A 99 US dollar wearable pendant that recorded continuously. Acquired by Meta in December 2025; the device was pulled from sale, the predecessor desktop product was shut down, and service ended entirely in several countries. Buyers report the pendants going inert. It is tempting to read this as a working product switched off by an acquirer, and the reading does not survive its own users: before the acquisition, owners reported severe transcription degradation, conversations the product invented outright and poor battery life, and left for software or for open-source hardware, with at least one abandoning it over having to explain what it was to everyone he met. Note the direction of travel all the same: this company started as software on a device the user already owned, added hardware, and the hardware is what the acquirer wanted. Amazon bought Bee, a 49.99 dollar pendant sold in front of a 24 dollar monthly subscription, in July 2025 and continued it. Three of three exits in this category went to platform owners, and only one price was ever disclosed.

Practitioners in this field describe the recurring ways companies here lose, and they are worth stating as patterns rather than as anecdotes.

Platform absorption. The operating system or host platform periodically ships the category's core feature natively and free. Both mobile platforms have now shipped recording, transcription and summarisation at no charge, and Apple additionally exposes recognition and summarisation, though not speaker attribution, to third parties at no per-hour cost. That exception is not a quibble: attribution is one of the six stages above, it accounts for roughly a fifth of the errors a user sees, and it is the one stage the platform keeps for its own application. An application whose entire value is a thin layer over models the platform gives away is still the most absorbable shape in software. The historical survivors of absorption in adjacent categories are instructive: weather, note-taking, document scanning, password management and call screening all produced survivors, and not one of them survived by being a better single-player utility on the same phone. Each left the phone, left the single user, or left the feature.

The second object. Humane and, in customer terms, Limitless. Both asked a buyer to acquire, charge and carry a device for a job the phone in the same pocket could do. This is the pattern a software-only product exists to exploit, and it comes with a second-order lesson that applies to software too: a product whose value lives on somebody else's servers can be switched off.

Distribution owned upstream of the store. ByteDance's Lark records an in-person meeting from its mobile application and routes the transcript, summary and action items straight into the chat and customer records the company already uses, free to individuals at 300 transcription minutes a month and unlimited on paid enterprise tiers at 6 to 39 US dollars per user per month; its parent has also shipped a ten-gram wearable button that feeds audio into the same suite, which is hardware treated as an accessory to software. SK Telecom's assistant sits inside the native dialer for its own subscribers, passed ten million of them by mid-2025, and never touches an app store. A bundled competitor acquires the market at an acquisition cost of zero and never appears on a comparison page. Two things temper the pattern. Whether the three largest Western bundlers, Microsoft, Google and Zoom, do the same for in-person mobile capture has never been established either way, which is a hole in the answer rather than evidence of absence. And where the Western products do ship the feature it is gated rather than free: Microsoft's intelligent recap needs a Teams Premium add-on or a Copilot licence, Google Meet's note-taking is gated on an eligible Workspace edition or AI plan held by the organiser, and Zoom's included tier caps summaries at three hosted meetings a month.

Undifferentiated crowding. Both app stores hold many near-identical recorder-and-summary applications built on the same rented models, at 9 to 19 US dollars a month, and several give the core interaction away. When products cannot be told apart, ranking and advertising decide, and acquisition cost climbs past what a subscriber is worth. The interaction itself is not the differentiator a newcomer expects, either. For a returning user, one of the incumbents already reaches one to two taps through a persistent live notification, a watch complication or an Android quick-settings tile, and the pre-installed recorder reaches one to two through a widget. What nobody has made short is the first run, which for any third-party application is roughly seven steps: launch, create an account, pass a biometric prompt, grant the microphone permission, grant the notification permission, dismiss the subscription upsell, record. A buyer in the 64-statement sample of hardware owners discussed below counts the application path as unlock, open, tap, against one pinch on the device.

Unit-economics inversion. Cloud transcription is metered by the recorded hour while subscriptions are flat, so the heaviest and most attached users are the least profitable. The free tier is the same arithmetic pointed at people who never pay at all: at the category's assembled 0.21 to 0.62 dollars a recorded hour, a 300-minute monthly free cap costs the vendor up to roughly 3 dollars a month for every free user who consumes it, and the one worked cohort anybody has published has 100 free signups consuming up to 200 dollars of inference a month while the three to five who convert generate 45 to 75, loss-making before the store takes its share. The standard responses, usage caps and price rises, churn exactly the users who care most.

Trust death by a single lost recording. A meeting recorder is bought as insurance, and one silently lost or truncated recording of an important meeting ends the relationship permanently. This is the pattern the capture section above describes mechanically, and it is the one where a dedicated recorder is structurally advantaged: it has its own microphone and its own battery, and no incoming call can take either. The advantage is not a clean record, though, and the same source that describes it lists the recorder's own failures: forgetting to charge it, forgetting to bring it, an obstructed microphone at the bottom of a bag, file-system corruption on hard power loss, and failed Bluetooth or USB transfer leaving hours of audio stranded on the device. Neither side of this comparison has a published failure rate.

The covert-capture marketing trap. Four United States class actions are pending against note-taker vendors and none has resolved: the consolidated Otter matter arising from *Brewer v. Otter.ai* (August 2025), *Cruz v. Fireflies.AI* (December 2025), *Basich v. Microsoft* (February 2026) and *Chamberlain v. Granola* (July 2026). One signal exists and it points one way: the privacy theory reportedly survived a motion-to-dismiss hearing in April 2026. Every vendor pushes the consent burden onto the user in its terms, which protects it in a contract dispute and does nothing against non-users who signed nothing, which is what these suits are. The legal ground underneath is not uniform: the federal floor allows one party to a conversation to record it, and across the 50 states plus the District of Columbia the split

is roughly 36 one-party, 10 all-party and 5 hybrid, but re-derived for an in-person meeting recorder specifically the binding set is 12 jurisdictions, possibly 13 on a cautious reading of Missouri. Beware the shorthand: whether a state is "all-party" is a property of the state *and* the medium, and five states split by medium, two of them in the opposite direction from the other three. Nevada, routinely listed as all-party, requires all-party consent for telephone calls and one party for an in-person conversation; Oregon and Hawaii require all-party for exactly the in-person case. On top of that the judicial consensus is that the strictest applicable standard governs, so one participant dialling in from another state pulls the whole meeting under the stricter law, which makes announcing to the room the operating default for anything shipped nationally.

Outside the United States it is sharper, and one duty is already in force rather than pending. The European Union's AI Act imposes transparency duties on any AI notetaker used with participants in the Union, enforceable since 2 August 2026, alongside its prohibition on emotion recognition in the workplace. Germany's criminal code makes recording the non-publicly spoken word an offence punishable by up to three years, applying to participants and not only to eavesdroppers, and a workplace deployment there additionally needs a Works Council agreement; France requires all-party consent with a proportionality test. Under European data protection law the household exemption does not survive a work purpose such as minutes, consent must be obtained from each participant independently and cannot be given by the host, which is a direct constraint on a two-tap interaction, and one named person must be able to have their voice and their contributions isolated and deleted on request.

Two sizings are worth carrying, one at each end. On the downside, the Illinois biometric statute sets statutory damages at 1,000 dollars per negligent and 5,000 per intentional violation with no proof of harm required; against that, there is no judgment, no settlement and no regulator fine anywhere in this field, and the one dollar figure attached to any of the four cases is a litigation-analytics vendor's projection. On the upside, compliance is the only demonstrated pricing power in this category: a German vendor sells at 24 to 47 euros a month, two to three times the category, on audited information-security certifications (ISO 27001 and SOC 2 Type II), European-only servers, audio deleted the moment the transcript exists and a contractual guarantee that customer data never trains a model. The certification is the product.

What actually changed recently, and what did not

This is the section where a document is most tempted to flatter its own premise, so it is worth being blunt: the project's original implied reason for building this now did not survive contact with the evidence, and a different one did.

The founder's stated premise was that the phone underneath the recorder already has better microphones, a battery and a data connection. The claim about microphones is the one that died, and it died on mechanism rather than on a measurement. Neither platform documents a route for a third-party application to reach raw per-microphone channels, and no handset has been shown to offer one, so on Apple's platform and on every untested Android handset the application should expect to compete in the single-channel condition, while the recorder is reported to have low-level access to its own two or four. The narrower question the claim could have meant, one phone microphone against one recorder microphone, has no measurement in either direction, because no corpus contains a phone. So the honest statement is not "the phone is worse"; it is that the phone competes in the harder condition on everything anyone has documented, that the microphones being good does not change which condition it is in, and that the Android half of this has never been tested across the fleet, so a handset that does expose unprocessed multichannel audio would be a moat rather than a blocker.

The capture layer has been moving in the same unhelpful direction, with one exception. Every documented background-audio change on either platform but one has restricted capture further. Android 12 blocked background microphone-service starts; Android 13 forced active foreground services into a unified task manager; Android 14 made a microphone foreground-service declaration mandatory with a hard exception thrown on omission; Android 15 tightened background starts further. Apple's iOS 15 and 16 cemented the background-restart refusal and the always-visible microphone indicator, and iOS 17 added only finer interruption reasons for debugging. The one loosening is useful and narrow: Android

15 exempted microphone foreground services from its new six-hour cumulative timeout, so the length of a meeting is not the constraint. Nothing else opened.

What did change, and it is dated. The recognition layer opened. Apple's iOS 26 exposes a new speech framework to any third-party application: long-form transcription, short-utterance dictation and voice-activity detection, running entirely on the device, with an on-device foundation model available for summarisation through a separate framework. Google's equivalent is narrower and later: its on-device generative speech recognition is in alpha, with a basic mode on Android 12 and above across 15 locales and a generative mode hardware-gated to Pixel 10 and Pixel 11. That is a real, dated, citable change, and it does three things rather than the two usually claimed. It removes the per-recorded-hour cost of transcription, it shortens the clock, and it costs accuracy: on twelve hours of real earnings-call conversation the on-device Apple recogniser scores 14.0 percent word error rate and an on-device Whisper variant 12.8, against a cloud comparator around 6.0 in the same table, and that cloud figure is uncited in its source, so the on-device numbers are sourced and the thing they are worse than is not.

And here is the sting, which is the same fact wearing the other sign. The identical shipment put the identical capability into the platform vendors' own free applications. Apple Notes records, transcribes live and summarises; Pixel Recorder records, transcribes in real time entirely offline, labels speakers and summarises on device, free, with no subscription and no length limit beyond storage. Google reported a 24 percent engagement lift on its recorder application after that summarisation feature shipped. The reason this became newly possible and the largest single threat to it are not two findings; they are one fact wearing two signs. Two things bound the threat rather than removing it. Pixel Recorder is Pixel-only, so it reaches the Pixel installed base and not the Android one; and on Apple's side the pre-installed voice recorder has no speaker labels, no summary and no action items, so the pre-installed threat there is Notes rather than Voice Memos. One gap remains in both, and it is undocumented rather than demonstrated: neither is documented as producing action items as a distinct checklist rather than as lines inside a summary.

Where the money is, in both directions

There is no payer layer in this field, and that absence is the first thing to understand about it. No insurer, no employer benefit, no reimbursement code, no regulator classification. Every dollar travels a commercial rail and every entrant pays its own acquisition cost with no category subsidy in front of it. General startup rails do exist and are open to anyone, cloud-provider credits, the United States Small Business Innovation Research programme, the European Innovation Council Accelerator and the Israel Innovation Authority among them, but none is specific to this field and none pays for a recorded hour. The one public subsidy anyone found in the category points the other way: Japan's tax code lets a business expense a recorder priced under 100,000 yen within the fiscal year, which is a demand subsidy for the hardware. A reader arriving from a regulated field looking for the reimbursement section will not find one.

What a person pays. The individual paid tier has converged at roughly 8 to 19 US dollars a month, and that spread is mostly one billing cadence against another rather than two different prices: Otter, for instance, lists 16.99 month-to-month and 8.33 per month billed annually for the same 1,200 transcription minutes. Free tiers have not converged at all, and each one meters a different thing: 300 minutes a month with a 30-minute per-conversation cap at Otter, 100 minutes a week with a 30-day history cap at Voicenotes, 30 minutes a month at Wave, 120 minutes a month with a 3-minute per-file cap at Notta, unlimited recording with a 30-day history cap at Granola, and unlimited recording and transcription at Fathom with the summaries capped at the first five calls a month. What the seven have in common is that each meters whatever its own architecture makes expensive, which is a more useful fact than a shared number would have been.

What a company pays. Team and enterprise seats run from about 20 US dollars per user per month, invoiced directly and off the app stores; a 98 dollar top of that band circulates with no vendor page behind it anywhere and should not be used. This is where every scaled incumbent's growth actually comes from. Otter drives growth through its Business and Enterprise tiers; one competitor's valuation

is anchored to penetration of Fortune 500 companies rather than to the 100,000 consumer accounts it adds weekly. The individual consumer subscription is the rail every vendor lists first and none of the scaled ones grows on.

What the store takes. Apple's standard commission is 30 percent, falling to 15 under its Small Business Program for developers under 1 million US dollars of annual earnings, and for auto-renewing subscriptions in their second year. Two things are in motion, and a reader only needs their direction. In the United States, a 2025 antitrust ruling currently lets a developer send buyers to a web checkout with no Apple commission, which nets more than the in-app small-business rate; the position is provisional, because an appeals court has held that Apple is entitled to charge something and the fee has not been set. In the European Union a unified Digital Markets Act schedule takes effect on 1 October 2026, with separate rates for in-app purchase, alternative payment and an external link-out, each roughly ten points lower for small-business participants, plus a 5 percent core technology commission. Google Play is the one rate that cannot be stated here, and the reason is a disagreement between two accounts rather than an absence. One records a schedule cited to Google's own developer support page, under which an auto-renewing subscription of this shape meets an all-in 15 percent either way, as 10 percent plus a 5 percent billing fee on the first million dollars of annual earnings in the markets that schedule has reached, or a flat 15 percent in the ones it has not. The other records the schedule as not established at all. They contradict each other and the question is open.

What the hardware side takes in. The dedicated recorder at 159 US dollars is not a substitute for the subscription; it is in addition to it. The device line runs wider than the headline price, at about 174.90 for the pendant models and 179 to 181.90 for the Pro, and the software ladder behind all of them is free at 300 transcription minutes a month, 17.99 US dollars month-to-month or 8.33 billed annually for 1,200 minutes, 29.99 or 19.99 annualised for unlimited, and 35 per user per month for a team. That 1,200-minute tier is within a dollar of the leading application's price for the same allowance at the same cadence. Setting the device's annual plan against an application's month-to-month plan is what makes the hardware look half price, and that is a billing-cadence artifact rather than a price difference. Shipments are large and the figures do not form a series: more than 1.5 million devices reported in January 2026 against more than two million by mid-2026, with domestic Chinese sales estimated at under 100,000 units in 2024, so this is Western demand. There is price headroom underneath it: white-label clones of the same card form factor sell in China at about 300 Chinese yuan, roughly 42 US dollars, so roughly three quarters of the 159 dollars is not bill of materials. Revenue is reported two ways by the same tracker that do not reconcile: 100 million US dollars of software-only annual recurring revenue as of June 2026, and a 250 million total annualised run-rate calculated in September 2025. Treat the whole set as a magnitude. Three things that would make sense of it are disclosed nowhere: the attach rate, meaning the share of device buyers who pay for software at all, which decides whether the object sells the subscription or the subscription is a thin tail on a hardware business; the return rate and repeat-purchase rate, against a 30-day no-questions return policy that implies a vendor modelling remorse; and which device each buyer bought, which matters because the largest single block of stated purchase reasons is a job this project does not target. Notably, the company shipping those units has disclosed only about 5 to 6 million US dollars of venture funding, most recently a 4.75 million convertible note led by Carbide Ventures in April 2025; reports of a large strategic investment describe discussions at a valuation between 1 and 2 billion US dollars, not a completed round. One reading, which is reasoning rather than measured fact, is that the object is the acquisition channel: it is cash-positive at the point of sale, so it pays to find the customer without dilution, and a software-only entrant has nothing in its place. Read that alongside one product fact: since January 2026 the same vendor also ships a software-only desktop notetaker, so the incumbent has already crossed into software without giving up the object.

Crowdfunding is the one public quantitative demand series for the hardware, and it is a winners-only sample. It covers three campaigns, all of which succeeded, and no failed or unshipped campaign is recorded anywhere in it. The leading recorder raised 1,108,181 US dollars from 7,564 backers on Kickstarter between 27 June and 16 August 2023; a competing device raised 554,444 dollars from 2,431 backers in late 2023; an earlier wearable raised about 100,000 dollars from about 1,000 backers in 2017.

What the software side has raised. Granola has taken 192 million US dollars in total, most recently a 125

million Series C led by Index Ventures with Kleiner Perkins at a 1.5 billion post-money valuation in March 2026. Otter has raised about 70 million and reported about 100 million US dollars of annual recurring revenue in March 2025 with 56.3 percent year-over-year growth. Read AI has raised 81 million at a 450 million valuation. Fireflies has raised about 19 million and publishes four figures that do not reconcile with each other: a 1 billion dollar valuation announced on its own blog in June 2025, about 11 million dollars of annual recurring revenue on a third-party estimate, and over 16 million users. A bootstrapped path also exists: Fathom reached an estimated 16 million US dollars of annual recurring revenue by the end of 2023 with no venture backing, then raised a 17 million Series A that reserved ten percent for its own users and took over 3.2 million dollars from 2,148 retail investors at a 73 million valuation, followed by a 43 million Series B in May 2025. A bootstrapped application in the same shape, with no outside funding at all, is estimated by a third-party tracker at 360,000 US dollars of monthly recurring revenue on 25,000 monthly downloads.

What a recorded hour costs to serve, and which architecture each figure belongs to. Sending the audio to a hosted recogniser costs 0.21 to 0.62 US dollars an hour all in, dominated entirely by transcription: one vendor prices at 0.0043 dollars a minute, that is 0.258 an audio hour, another lists 0.37, and self-hosted recognition on provisioned accelerators falls to about 0.12. The summarisation call on a 12,000-token transcript is about 0.002 dollars, roughly 130 times smaller, so the language-model half of the bill is negligible against the speech half. At a 15 dollar subscription netting about 13 after commission and a 0.40 dollar median hour, breakeven arrives at roughly 32.5 hours a month, and a user recording two to three hours a working day costs the vendor about 20 dollars. The 0.05 dollars an hour that pushes breakeven past 250 hours is not the fully local architecture; it is the hybrid one, capture and recognition on the phone with the full plaintext transcript still going to a cloud language model, which is the arrangement its own source calls bot-free but not private. The battery figures belong to yet another architecture, the continuously streaming one, and they are the worst-measured quantity in the field: recording alone is close to free at about 0.66 percent of charge an hour, but that comes from a single self-report with no device model and no method; a fully offline continuous pipeline is put at 25 to 40 percent an hour and recognition alone at 10 to 15 percent, both from single blog posts. A widely circulated figure of 3 percent of battery per minute of audio is arithmetically impossible, since it consumes 180 percent of a battery over the one-hour meeting its own source describes.

Zero marginal cost is not a moat here. One Chinese listed manufacturer sells recorders at 140 to 299 US dollars with lifetime free on-device transcription and no subscription at all, funded off a business with 27.82 billion Chinese yuan of annual revenue. Every product in this field with a one-time price or a free unlimited tier is a local-processing product, and they are not only the operating systems' own: one third-party phone application ships a wholly free mode running local models with no minute cap, and an open-source desktop assistant offers unlimited local transcription free. Local processing is a pricing weapon, and the weapon points at the price of the whole category.

What the borrowed customer economics say, and why they are borrowed. The figures that circulate for this category are free-to-paid conversion of 2 to 5 percent, acquisition cost of 20 to 40 US dollars, monthly churn of 5 to 8 percent and a resulting lifetime value of 180 to 300. Carried through, they give a lifetime-value to acquisition-cost ratio of roughly 4:1 to 6:1, which is a healthy consumer software business, and the mechanism the same source credits for holding acquisition cost down is sharing a summary or an action-item list branded with the application's logo with the other people who were in the meeting. That is worth pausing on, because it is the same act the legal section above treats as the exposure: the growth loop and the liability are the same gesture. All four inputs are borrowed from consumer productivity software generally, and the source that supplies them states outright that category-specific financial data for private meeting applications is rarely audited publicly.

What is absent from the world rather than from this document. Direct usage figures are missing across the field: no retention, no monthly churn, no share of installs that record a second meeting is published by anyone, and every traction number above is a raise, a valuation, an install count or a revenue estimate. What is not absent, and what the field itself treats as checkable, is store evidence: ratings volume, review velocity and category rank on both stores. Some of it is already on the record. The application that matches this interaction most exactly shows 4.8 stars from about 6,700 iOS reviews; a neighbouring study-note application carries 17,000; one vendor claims over 8 million users and another claims hun-

dreds of thousands of daily active users, both self-reported. Nobody has pulled that series over time for the five applications this document names. Nor does anyone publish a quality figure. A rail in this field can be sized by what it pays; it cannot be sized by what it converts.

What is not yet true

These are the things that are unproven, and they are worth separating from the things above, which are documented.

Nobody knows how often a phone loses a meeting. The failure mechanisms are documented and deterministic; the frequency is measured nowhere. Vendors with the telemetry treat it as proprietary, an aggregate Android rate would be close to meaningless under handset fragmentation, and the academic corpora structurally cannot contain it because failed sessions were discarded before release. Two percentages circulate and both were invented as illustrations in a discussion thread. The smallest experiment that would settle it needs no new product: take the recorder applications already shipping, run them across handset models and manufacturer skins rather than across people, force an incoming call, a sleeping-application tier, a memory reclaim and a competing microphone request, and count the recordings that survive intact. The proposed bar is above 99.0 percent defect-free completion across at least 500 sessions from at least 50 users on both platforms; without a bar the experiment cannot come back negative. Two things are missing around it. Nobody has published that measurement, and if the incumbents pass it cleanly the strongest remaining reason for a new entrant in this category goes away. And no shipped recorder of any kind, hardware or software, has ever published a capture defect rate, so there is nothing to judge the bar against.

Nobody knows whether the phone or the recorder hears a meeting better. The parallel-capture protocol described earlier is the only route, no such study exists in either direction, and the best result it can return for the phone is equivalence rather than superiority. The narrower placement question is equally open: no word error rate has been measured for a phone face-up on a table, face-down, in a shirt pocket, in a trouser pocket or in a bag.

The number the summarisation argument rests on has never been measured. The tolerance of a summary to a noisy transcript holds only if the errors avoid names, numbers and decisions, and far-field acoustics break those first. A named-entity word error rate on far-field meeting audio appears in no paper and no benchmark. The nearest published work reports relative reductions on far-field voice-assistant commands, which is a different acoustic and conversational problem.

Two of the three promised outputs have no defensible accuracy claim available to anyone. There is no accepted metric for action-item quality, and for summaries the automatic metrics actively mislead: across nine metrics scored against expert human error annotations, none correlated strongly with any error type, about a third of the metric-and-error pairs either ignored or rewarded the error, perplexity rewarded wrong speaker references at +0.44 and one metric rewarded structural disorganisation at +0.45. The only pair behaving as a product team would want detects gross omission and nothing finer. Any quality claim in this category has to rest on a rubric and human raters; there is no number available to buy it with.

Nobody knows whether the person who buys the recorder would take an application instead. No published survey, panel study or analyst report exists on that question. The only primary evidence located anywhere is a convenience sample of 64 explicit purchase-rationale statements from verified retailer reviews and six named forums, distributed as capturing a cellular phone call 28.1 percent, interruption and distraction mitigation 20.3, discretion and bot avoidance 17.1, battery preservation 14.0, offline and ambient capture 12.5, and the tactility of a physical start and stop 7.8. Not one of the 64 statements is about transcription accuracy or audio quality. Three things must be said about that sample in the same breath. Its largest block is a job no third-party application can do at all, because neither platform permits any application to touch the microphone during a cellular call and Google banned the accessibility-service route to it in May 2022. It is drawn from hardware-owner forums, so by construction it records why buyers stayed with hardware rather than what a marginal buyer would choose against a better application. And it does not record which device each statement came from, so the share of these buyers

who bought the card that lies on a table for an in-person meeting, as against a wearable pendant or the phone-call job, is unknown. Nobody has asked buyers directly.

Nobody knows whether the free products that already do this are actually used. Applications with this exact interaction and these exact three outputs already ship on both stores, several of them free, and the platform vendors ship the feature pre-installed. The one that matches the interaction most exactly is unfunded and bootstrapped, prices unlimited use at 9 US dollars per user per month, and has one traction figure it publishes itself, a cumulative one-million-downloads badge, which is an install count and not active use; the one independent signal on it is a store rating, 4.8 stars from about 6,700 reviews, which measures satisfaction among people who reviewed and not retention. No user, subscriber or revenue number for it exists anywhere. "The free product exists and nobody kept using it" is not ruled out, and it is the single number that would decide whether there is room here at all.

Nobody knows whether the user presses start. The recording law above is taught in this field as vendor liability, and it has a demand-side twin that is not measured: products in this category have lost usage not because capture failed but because the user stopped daring to press record in front of colleagues. The legal teaching in this document, all-party as the operating default and announce to the room, makes that the expected interaction rather than an edge case. Nobody has observed that behaviour in a real room and written it down.

Two further unknowns are procedural rather than empirical, and both bear on whether the product can ship in this shape at all. The platform review guideline governing background audio reportedly drives rejections of applications that declare the capability without a visible feature that needs it, and no one has established what minimum visible surface passes. And what the App Store review guidelines and the Google Play developer policies actually say about recording applications has never been established at all. A platform policy is the one rule that can remove a product from sale.

The idea, as it currently stands

Stated, not defended. This is the pitch as it currently stands:

> People pay 159 dollars for a card-sized recorder they stick to the back of their phone, just to leave meetings with a transcript and a summary written for them. The strange part: that phone, face down on the table or in a pocket, already has a battery and a data plan. Recapp does the recorder's job as an app. Tap start, tap stop, and before you reach your desk you have the transcript, the summary, and the action items. No new device to charge, carry or forget. Nothing works unless that phone keeps recording to the end of the meeting.

Five claims sit under it, with the status the evidence gave each one. A reader who has come this far can grade them without help, which is the reason for teaching first.

- **A phone running an application can capture an in-person meeting well enough to produce a usable transcript, summary and action items.** EVIDENCED, and only as an existence claim: five applications examined here deliver exactly these three outputs from the phone's own microphones today and both platform vendors ship the same feature pre-installed. What is not evidenced is "well enough" as a number, because no public benchmark measures any of this on a smartphone and no vendor publishes a quality figure.
- **The phone's own microphones are better than the dedicated recorder's for this job.** KILLED, on mechanism rather than on measurement, for the channel-release reason set out above: no documented route to the raw channels exists on either platform, against a recorder reported to reach its own two or four. The claim is dead as written rather than reversed, since no measurement exists in the other direction either, and the Android half of the mechanism has never been tested across the handset fleet.
- **The transcript, the summary and the action items can be finished in the few minutes between the end of the meeting and the user's desk.** EVIDENCED, on competitors' own published and independently timed turnaround rather than on any measurement of this product, and conditional

on architecture: recognising during the meeting meets the deadline and the category's cloud-batch default does not.
- **The phone can be relied on to keep recording to the end of a real meeting.** ASSERTED. This is now the load-bearing claim. The failure mechanisms are documented on both platforms and the failure rate is published by nobody.
- **The people currently buying a 159 dollar recorder would take the application instead.** KILLED as stated, on the 64-statement sample and its distribution of reasons, with the selection objection above recorded against it. It was the load-bearing claim before the evidence arrived, and no claim was written to replace either of the two that died, because the evidence supports none.

The project also wrote down, before any evidence came in, the three conditions under which it would abandon the idea: that the application version already exists at low or no cost while the hardware sells anyway; that the phone cannot reliably do it; and that the hardware buyer's stated reason is something an application cannot supply. The first and the third both fired on their own wording. What did not survive the evidence is the first condition's other half, "and is not winning", since the applications in this category are demonstrably winning.

Terms used here

Term	What it means
Action item	A commitment or task extracted from a meeting, with an owner. The least measured of the three outputs; no accepted accuracy metric exists
All-party consent	A legal regime requiring every participant to agree before a conversation is recorded, as opposed to a one-party regime where one participant may consent for the whole conversation
AMI Meeting Corpus	About 100 hours of scenario meetings of 15 to 45 minutes, recorded on headsets and an eight-microphone table array, with human summaries on 137 of them; the field's most-used English meeting reference
ASR, automatic speech recognition	Software that turns recorded speech into written words
Beamforming	Combining several microphone signals so the device listens more sensitively in one direction; it needs two or more phase-coherent channels
Capture survival	The fraction of recording sessions that run intact from start to finish. No published figure exists on either mobile platform, for a phone or for a dedicated recorder
Channel, phase-coherent	One audio stream whose timing relationship to the others is intact. The count of these an application receives, not the count of microphones the device contains, is what decides how well overlapping speakers can be separated
Context window	The number of tokens, roughly word pieces, a language model holds at once, counting instructions, input and output together. A property of the model, not of the device
Contextual biasing	Feeding the recogniser a list of names it is likely to hear, from contacts or a calendar, so they beat acoustically similar common words. Shipping on every commercial recognition interface here, and it costs access to personal data
cpWER	Concatenated minimum-permutation word error rate: words scored with speaker credit required, without the timing constraint tcpWER adds
DER, diarization error rate	The percentage of speaking time given to the wrong speaker, plus missed speech, plus speech invented where there was none
Diarization	See speaker attribution

Term	What it means
Distant automatic speech recognition	Recognition of speech reaching a microphone across a room rather than at a mouth; a separate research discipline from dictation
Far-field	The condition where the microphone is metres from the talkers, with echo, noise and overlap. The condition this product works in
General Data Protection Regulation	The European Union's data protection law; a voice recording of an identifiable person is personal data under it, its household exemption does not survive a work purpose, and consent must come from each participant rather than from the host
Guided source separation	The multi-microphone front end behind most winning meeting systems. It builds a spatial model from phase differences between microphones, so it does not run on one channel
ICSI Meeting Corpus	75 recorded academic meetings of about an hour each, about 72 hours, 61 of them carrying a human summary
LLM, large language model	A text-generating system; here, the component that writes the summary and extracts the action items
MEMS, micro-electro-mechanical system microphone	The chip-scale microphone used in both phones and pocket recorders
Named-entity word error rate	Word error rate counted only over names, numbers, dates and acronyms. Unmeasured on far-field meeting audio, and the number the summarisation argument actually rests on
Near-field	The condition where the microphone is at the talker's mouth: a headset, a lapel clip, a phone held to the face. Easier than far-field by about five points on the one within-system measurement available, and not itself a solved problem
NOTSOFAR-1	A corpus and challenge of real corporate office meetings across 30 rooms, sessions truncated to about six minutes, with a withheld evaluation split; the source of the 22.2 and 10.8 percent figures
Overlapped speech	Two or more people talking at once. Roughly a fifth to two fifths of meeting duration depending on the corpus, and where single-channel systems fail worst
Read speech	Someone reading a prepared passage into a close microphone. The easiest condition in the field and the source of most low-single-digit accuracy claims
Scenario meeting	A recorded meeting in which participants act out assigned roles to a brief rather than meeting for their own reasons. The AMI corpus is scenario-based, which is why its action items are partly a property of the brief
Speaker attribution	Deciding who spoke when, and attaching the right name or label to the right words. Scored separately from recognition, and neither mobile platform exposes it to third-party developers
Speaker enrolment	Storing a voice embedding per known person so attribution becomes verification rather than guessing. The technique that fixes attribution is the technique that creates a biometric record
Streaming recognition	Recognising speech while the meeting is still running, so only summarisation remains when the user taps stop; the opposite of batch, which starts the whole job at the end. It buys the clock and costs connectivity, the on-device position and an unmeasured amount of accuracy

Term	What it means
tcpWER	Time-constrained minimum-permutation word error rate: the ranking metric of the meeting challenges. It scores the words, who was credited with them, and whether they landed at roughly the right time
WER, word error rate	The percentage of words a transcript gets wrong. Comparable between systems only on the same test set under the same audio condition