Recapp

Data Landscape

What labeled data exists for turning the audio of an in-person meeting into a transcript, speaker labels, a summary and action items, who holds each ...

2026-09-01

Prepared by Copundit

AI for Advanced Ideation™

What labeled data exists for turning the audio of an in-person meeting into a transcript, speaker labels, a summary and action items, who holds each corpus, and on what terms it can be obtained.

Contents

Contents

Abbreviations

Abbr	Stands for	What it actually is (plain English)
AIMU	Actionable Items for Meeting Understanding	A 2016 layer of assistant-task labels added on top of 22 ICSI meetings
AMC-A	AliMeeting-Action Corpus	A Chinese meeting corpus whose sentences are marked as containing an action item or not
AMI	Augmented Multi-party Interaction	A 2005 corpus of staged meetings, recorded on headsets and on a table array at the same time
ASR	Automatic Speech Recognition	The software that turns recorded speech into written words
ATF	Acoustic Transfer Function	A measurement of how a room changes a sound between the mouth and the microphone
CC BY 4.0	Creative Commons Attribution 4.0 International	A public licence allowing any use, including commercial, if the source is credited
CC BY-NC-ND 4.0	Creative Commons Attribution-NonCommercial-NoDerivatives 4.0	Credit required, no commercial use, and no altered version may be shared
CC BY-NC-SA 4.0	Creative Commons Attribution-NonCommercial-ShareAlike 4.0	Credit required, no commercial use, and any derivative carries the same terms
CC BY-SA 4.0	Creative Commons Attribution-ShareAlike 4.0 International	Credit required, commercial use allowed, and any derivative carries the same terms
CDLA-Permissive	Community Data License Agreement, Permissive	A data licence allowing commercial use and imposing no terms on derived work
CER	Character Error Rate	The word error rate's equivalent for languages written without spaces, such as Mandarin
CHiME	Computational Hearing in Multisource Environments	The long-running challenge series that sets the benchmarks for meeting transcription
DAMO	Discovery, Adventure, Momentum and Outlook	Alibaba's research academy; its speech lab released the AliMeeting and AMC-A corpora
DASR	Distant Automatic Speech Recognition	Speech recognition when the microphone is across the room instead of at the mouth
DER	Diarization Error Rate	Percentage of speaking time attributed to the wrong person, or missed, or invented
DiPCo	Dinner Party Corpus	A 2019 corpus of four-person dinner conversations recorded on five table arrays
DUA	Data Use Agreement	A signed contract setting what a recipient may and may not do with a released dataset
ELITR	European Live Translator	A research project whose corpus holds meeting transcripts and hand-written minutes
GDPR	General Data Protection Regulation	The European Union's data protection law; a recording of a person is personal data under it
GLM	Global Mapping file	A scoring-time list saying which alternative wordings count as the same words
IAA	Inter-Annotator Agreement	How often two people given the same job produce the same label
ICSI	International Computer Science Institute	The Berkeley institute whose 2003 corpus of real research-group meetings is still a benchmark

Abbr	Stands for	What it actually is (plain English)
LDC	Linguistic Data Consortium	The University of Pennsylvania body that licenses and distributes speech corpora
LLM	Large Language Model	A text-generating model; here, the part that writes the summary and pulls out action items
LOTUSDIS	Thai far-field meeting corpus, 2025	Nine separate recording devices at measured distances, all on the same Thai meetings
MAPSSWE	Matched Pairs Sentence-Segment Word Error test	The standard test for whether one recogniser really beats another rather than got lucky
MC	Multi-Channel	Audio recorded by several microphones at once, which preserves where each sound came from
MIT	Massachusetts Institute of Technology	Here, the permissive software licence named after it, which allows commercial reuse
MMCSG	Multimodal Conversations in Smart Glasses	A 2024 challenge on conversations recorded by a head-worn consumer device
NIST	National Institute of Standards and Technology	The United States agency whose scoring toolkit and evaluations set this field's measurement rules
NOTSOFAR	Natural Office Talkers in Settings of Far-field Audio Recordings	Microsoft's 2024 corpus of short office meetings on tabletop array devices
PII	Personally Identifiable Information	Anything in a recording that could be traced back to a named person
QMSum	Query-based Meeting Summarization	A benchmark of question-and-answer summaries built on top of older meeting corpora
RT60	Reverberation time to minus 60 decibels	How long a sound keeps bouncing in a room before it dies away, in seconds
SC	Single-Channel	Audio recorded by one microphone, which throws away all directional information
SDM	Single Distant Microphone	The standard test condition of one far-away microphone, the closest public analogue to a phone
SUMM-RE	French meeting corpus, 2024	About a hundred head-mounted-microphone French planning meetings, discourse-segmented
tcpWER	time-constrained minimum-permutation Word Error Rate	Word error rate that also punishes attributing words to the wrong speaker or the wrong time
UEM	Un-partitioned Evaluation Map	A file saying which parts of a recording count towards a score
WER	Word Error Rate	Percentage of words a transcript gets wrong; the standard accuracy number for speech

Executive Summary

The signal family is **meeting-condition conversational audio**: several people talking, often over each other, into a microphone that is not at anybody's mouth, plus whatever was annotated on top of it. The family is rich at one end and empty at the other, and the emptiness is specific rather than general. **Verbatim transcription with speaker attribution is well supplied**: **AMI** gives about 100 hours of headset-and-array meetings, **NOTSOFAR-1** gives real office meetings across 30 rooms with a blind evaluation split, **AliMeeting** gives about 120 hours of Mandarin meetings recorded near-field and far-field at once, and

CHiME-6, **DiPCo**, **LOTUSDIS** and **Mixer 6** fill in dinner parties, Thai meetings and one-to-one interviews. Two things narrow that supply before any of it can be used here. Only the single-channel condition of any of them is reachable by a phone application at all, for the reason given in the primer below, and **none of what is left is a clean held-out English test set**: AMI and ICSI have been public for two decades and CHiME-8 lists AMI among the material a system may train on, NOTSOFAR-1's blind split has its reference transcripts withheld so nobody outside the challenge can score it, and LOTUSDIS, the one recent uncontaminated set, is Thai. **Nothing in the family was recorded on a smartphone.** Two independent sweeps assert that null and neither cites a catalogue behind it, but the field's own review lists 32 notable robust-recognition datasets and challenges since 2000 and none of them names a phone as a capture device; the nearest neighbours each fail for a stated reason, and the closest is a Waseda ad-hoc array of iPhones whose audio was never released. **Summaries come from four corpora and action items from three**, the largest of them Chinese: AMI and ICSI carry a four-part human abstractive summary in which **Actions is an optional section an annotator could leave empty**, and the published counts are **101 AMI meetings carrying 381 action items** (derived from dialogue acts, not annotated directly), **1,506 action items over the 424 Chinese meetings of AMC-A**, and **318 actionable turns over 22 ICSI meetings** in the AIMU layer. **No open corpus is a real business meeting of the target length with its audio attached**: AMI is 15 to 45 minute role-play, ICSI is academic seminars, NOTSOFAR-1 is truncated to about six minutes, CHiME-6 and DiPCo are two-hour dinner parties, **MeetingBank** is city council proceedings under a NonCommercial licence, and **ELITR**, the one corpus of real hour-long project meetings, **withheld its acoustic data for privacy**. **No corpus of any kind records whether a recording survived**, and the reason is structural: a session that failed mid-collection was discarded before release, so no published file is paired with a cause of failure. The largest holdings of exactly the target signal are commercial and none has a published price or a named outside user: **Appen**, **Defined.ai**, **Otter.ai**, **Gong.io**, and Microsoft's withheld NOTSOFAR-1 reference transcripts.

Signals and Labels Primer

A usable record in this field is an audio recording of a real conversation plus a hand-made written reference for it, and the three things that decide whether it is worth anything are what recorded the sound, where that recorder was, and how the reference was produced.

Far-field meeting audio with verbatim transcripts and speaker attribution. The signal is sound from several talkers arriving at a microphone metres away, carrying room echo, noise and, for a fifth to nearly half of the time, two people at once. Corpora capture it with a known array (a fixed ring or line of microphones with published geometry), with a set of separate consumer-grade devices, or with a single processed channel. Only the last of those three is reachable by a product like this one, and the consequence runs through every hour count in this document: a third-party application on either mobile platform receives one processed mono stream and never the raw per-microphone channels, so an array condition supervises a signal a phone application cannot produce, and the single distant microphone (SDM) or single-channel (SC) subset is the part of any corpus here that a phone can be measured against. The label is a human-typed transcript of every word, cut into utterances and attributed to a named speaker; the governing granularity is the **session**, meaning one meeting as heard by one device, because that is what a system is scored on, and it is why a corpus can report both a meeting-hours figure and an all-microphones figure several times larger. Ground truth comes from a headset, lapel or in-ear microphone worn by each talker: a clean per-person recording that a transcriber listens to, and that costs a wearer's consent in every meeting.

Meeting transcripts with human summaries, minutes and action items. The signal here is text, either a human transcript or one corrected by hand from machine output, and the audio may not be released at all. The label is prose written by an annotator after reading or attending: a free-form abstract, a set of structured sections, or formal minutes. The governing granularity is the **annotated meeting**, far smaller than the hour count suggests, because one meeting of an hour yields one summary. For action items the granularity drops again, to the **labeled sentence or turn**, and it drops hard: the target class runs from half a percent to four percent of all turns depending on the corpus, 0.5 percent in AMC-A, 1.5 percent overall in AIMU and 4.2 percent in AIMU's densest group of meetings. Ground truth is the

annotator's judgment, which is why this family has no accepted score and why every serious set reports an inter-annotator agreement figure beside its counts.

Simulated far-field meeting audio. The signal is built rather than recorded: clean single-speaker speech convolved with measured room responses, mixed with noise and with other talkers to a chosen overlap. Because the mixer knows exactly which words belong to whom and when, the label is exact and free at any volume, and the governing granularity is the **synthesised hour**. What it cannot supply is conversational behaviour, since turn-taking, interruption and the way people move around a room are not part of the recipe. The field's rule follows from that: simulated audio is accepted for TRAINING an acoustic model and rejected for EVALUATING a hardware claim.

Landscape at a Glance

Access verdicts, used in this table, in every profile header and in the coverage table: **OPEN**, downloadable now with no agreement and no fee; **APPLICATION**, behind an agreement or a committee, with a named outside group that has obtained it; **PARTNER**, held by named institutions with no external precedent found; **COMMERCIAL**, sold by a vendor; **NONE**, no located corpus pairs this signal with this label. The verdict is an availability column and not a permission column, and the two come apart here more often than not. A parenthesis after the verdict says what is established about the licence, and a reader should treat the qualifier as binding: a licence that was not read is not permission, and a licence two sources disagree about is not permission either. Only AMI and LOTUSDIS carry a licence permitting commercial use that anybody in this project has read.

Corpus	Signal and labels	N at label granularity	Verdict	Who else holds it
AMI Meeting Corpus	Headset and 8-channel array meeting audio; verbatim transcript, speakers, dialogue acts, extractive and abstractive summaries	137 scenario meetings with a four-part summary, about 65 hours; 101 meetings carrying 381 action items; about 100 hours of audio	OPEN (CC BY 4.0)	Everyone; public download
NOTSOFAR-1 recorded meetings	Tabletop and linear array office-meeting audio; word-aligned transcript and speaker attribution	280 meetings in the dataset paper (107 train, 36 dev, 137 eval), 315 in the challenge paper, 237 in the repository; about 28 meeting hours, about 260 across all microphones	OPEN (licence in dispute)	Everyone; public download
LOTUSDIS	Nine separate single-channel devices at 0.12 to 10 metres on the same Thai meetings; transcript, speakers, overlap mask	90 sessions of 15 to 20 minutes, 3 speakers each, 86 speakers; about 20 meeting hours, 114 across all microphones	OPEN (CC BY-SA 4.0)	Everyone; public download

Corpus	Signal and labels	N at label granularity	Verdict	Who else holds it
ICSI Meeting Corpus	Far-field academic meeting audio; verbatim transcript, dialogue acts, extractive and abstractive summaries	75 meetings, about 72 hours, 3 to 10 participants; 61 meetings with an abstractive summary	OPEN (licence not established)	Everyone via a preprocessed public release; also catalogued by the LDC
AliMeeting	8-channel array plus per-participant headset audio of the same Mandarin meetings; verbatim transcript and speakers	About 120 hours over roughly 220 to 240 sessions of 15 to 30 minutes, 2 to 4 participants, 481 speakers, 13 rooms	OPEN (licence not established)	Everyone via a public download
CHiME-6	Four-person dinner-party audio on 6 four-microphone Kinect arrays; verbatim transcript and speakers	20 parties, each at least 2 hours, split 16 train, 2 dev, 2 eval; 49:44 annotated hours	OPEN (licence not established)	Everyone via the challenge toolkit
DiPCo	Four-person dinner-party audio on 5 seven-microphone circular arrays; verbatim transcript and speakers	10 sessions, 32 speakers, 5:19 hours annotated across all splits	OPEN (licence not established)	Everyone via the challenge toolkit
Mixer 6 Speech	Two-person interviews captured by 10 heterogeneous far-field devices; verbatim transcript and speakers	450 interview portions annotated of 1,425 sessions; 20:54 fully annotated hours	APPLICATION (LDC)	LDC licensees; challenge participants during the challenge
AMC-A (AliMeeting-Action Corpus)	Mandarin meeting transcripts with every sentence marked for containing an action item	424 meetings, 306,846 utterances, 1,506 action items, 3.55 per meeting, agreement 0.47	OPEN (licence not established)	Everyone via a public repository

Corpus	Signal and labels	N at label granularity	Verdict	Who else holds it
ELITR Minuting Corpus	Meeting transcripts with hand-written minutes; audio not released	113 English and 53 Czech meetings, over 160 hours of content	OPEN (licence not established)	Everyone via a preprocessed public release; the audio, nobody
AIMU	Assistant-executable intents labeled on the turns of 22 ICSI meetings	21,035 turns, 318 with an actionable item (1.5 percent), 10 intent types	OPEN (licence not established)	Everyone, if the release address still resolves
MeetingBank	City-council meeting video and transcripts; official minutes, agendas and segment summaries	6,892 segment-level summarisation instances over 1,366 meetings, 3,579 hours	OPEN (NonCommercial only)	Everyone for research; nobody for a commercial product
NOTSOFAR-1 simulated training set	Synthesised far-field meeting mixtures matched to the array geometry; exact transcripts and speaker labels by construction	About 1,000 hours, built with 15,000 measured acoustic transfer functions	OPEN (licence in dispute)	Everyone; public download

Corpora by Signal Family

Far-field meeting audio with verbatim transcripts and speaker attribution

The best-supplied family in the field and the one every published transcription number comes from. Eight corpora carry real multi-talker audio with hand-made transcripts and speaker labels; two have a licence whose text was read, five have a working download and a licence claim nobody has verified against the holder's own page, and one sits behind a consortium. **AMI Meeting Corpus, NOTSOFAR-1 recorded meetings, LOTUSDIS, ICSI Meeting Corpus, AliMeeting, CHiME-6, DiPCo, Mixer 6 Speech.**

AMI Meeting Corpus | OPEN (CC BY 4.0)

- **What it holds:** about 100 hours of meeting recordings collected by the Augmented Multi-party Interaction project and held by the School of Informatics, University of Edinburgh. The corpus is counted three different ways in three sources read here: 137 scenario-driven meetings of 15 to 45 minutes, about 65 hours (Rennard 2023, arXiv 2208.04163; Prevot 2025, aclanthology 2025.sigdial-1.14); about 137 sessions in a corpus-comparison table (Tipaksorn 2025, arXiv 2509.18722, Table 1); and 171 meeting transcripts of which 145 are scenario-based and 26 naturally occurring (Liu 2023, arXiv 2303.16763). The count that governs a summary claim is the scenario count, 137; the count that governs an action-item claim is 101, for the reason under labels below.
- **Recording condition:** each talker wears a close-talking headset microphone and the room carries far-field microphones including an 8-channel circular array on the table, at 0.3 to 3 metres (AMI corpus page, groups.inf.ed.ac.uk/ami/corpus/; Tipaksorn 2025, Table 1). Four participants per scenario meeting, three instrumented rooms, 16 kHz. Speech overlaps naturally; the overlap ratio

most often quoted for AMI is about 19 percent of conversation time, which the sources read here attribute rather than measure.

- **Labels and their ground truth:** orthographic transcription of every word, dialogue acts, topic segmentation and head movement, plus human extractive and abstractive summaries. The abstractive summary is written in four parts, Abstract, Decision, Problems and Actions, each up to 200 words, and annotators "were not obligated to provide summaries for all categories unless they felt it was motivated" (Rennard 2023). It supervises all three promised outputs: verbatim text, speaker attribution and an abstractive summary. For action items the count at label granularity is 101 meetings carrying 381 items, an average of 3.77 per meeting, and those labels are not a direct annotation: they are dialogue acts linked to the action-related part of the abstractive summary, treated as positive examples (Liu 2023, Table 1 and section 2.1). The average AMI summary runs 322 words (arXiv 2212.08206).
- **Holder and licence:** University of Edinburgh; Creative Commons Attribution 4.0 International (CC BY 4.0), whose text permits commercial use with attribution (AMI corpus page, read at kickoff). Two later sources restate CC BY 4.0 for AMI; one of them extends the same sentence to ICSI, which contradicts the Linguistic Data Consortium (LDC) account of ICSI below, so it adds no independent weight.
- **Access route and precedent:** direct download from the corpus page. CHiME-8 lists AMI among the external datasets participants may train on, so outside use is routine (Cornell 2024, arXiv 2407.16447).
- **Who else holds it:** everyone. A public download under a permissive licence is a commodity input, not a position.
- **Known limits:** the meetings are staged. The survey that describes the annotation warns that the designed scenarios and the fact that participants did not know each other produced "overly well-behaved" interactions, and that heavy use of slides and prototypes adds a visual channel the audio does not carry (Rennard 2023). The corpus dates from 2005, and two decades of use mean a modern pretrained recogniser may have seen it.

NOTSOFAR-1 recorded meetings | OPEN (licence in dispute)

- **What it holds:** real English office meetings averaging six minutes, recorded across 30 conference rooms with a disjoint evaluation split. Three meeting counts are on the record and no source read here reconciles them. The dataset paper's own Table 1 gives 107 training, 36 development and 137 evaluation meetings, 280 in total, 20 rooms for train and development plus 10 for evaluation, 22 speakers plus 10 (Vinnikov 2024, Interspeech, Table 1). The challenge summary paper presents a table it labels as reproducing that same table and gives 110, 35 and 170 meetings, 315 in total (Abramovski 2025, arXiv 2501.17304, Table 1). The public repository holds 237 (github.com/microsoft/NOTSOFAR1-Challenge, read at kickoff).
- **Recording condition:** four to eight participants seated in a real meeting room, each meeting captured simultaneously by about nine devices: roughly 5 single-channel devices each emitting one internally processed stream, and 4 multi-channel devices each emitting 7 raw streams from one central and six surrounding microphones. Close-talk microphones were used for quality control. One device recording one meeting is a "session", which is why the hour counts fan out: the dataset paper reports about 150 hours of single-channel and 110 hours of multi-channel audio across all sessions, against roughly 28 hours of distinct meeting time (Vinnikov 2024). Overlapped speech runs 31.7 percent of training duration, 16.7 percent of development and 29.6 percent of evaluation (Cornell 2025, arXiv 2507.18161, Table 3). Meetings are steered by a professional actor.
- **Labels and their ground truth:** utterance transcripts with word-level alignment and speaker attribution, plus per-meeting metadata tags for conditions such as a talker at a whiteboard. Transcription was done by human listeners with machine pre-transcription deliberately withheld, because annotators accept plausible machine guesses in noisy segments and so import the very model biases the corpus exists to test (Vinnikov 2024). It supervises verbatim text and speaker attribution only. It carries no summaries and no action items, so it cannot test two of the three promised outputs; the review that used it for summarisation had to generate reference summaries from the ground-truth transcript with a language model first (Cornell 2025).
- **Holder and licence:** Microsoft. The challenge repository states the data under CC BY 4.0 and the

code under the MIT licence (read at kickoff). A 2025 corpus-comparison table in a peer-reviewed corpus paper instead records NOTSOFAR-1 as CC BY-NC-ND 4.0, which would forbid both commercial use and derivative datasets (Tipaksorn 2025, Table 1). Neither statement is the holder's own licence file and the two cannot both be right.

- **Access route and precedent:** public download from the repository or through the challenge toolkit, which fetches CHiME-6, DiPCo and NOTSOFAR-1 in one command (Cornell 2024). Precedent is heavy: 32 submissions across the twin challenges, from the University of Science and Technology of China, NTT, Northwestern Polytechnical University, the Nara Institute of Science and Technology, Brno University of Technology and Johns Hopkins University among others (Cornell 2025).
- **Who else holds it:** everyone, on whichever licence turns out to govern.
- **Known limits:** a six-minute meeting is not the meeting this category is bought for, and it is too short to test long-context summarisation or topic segmentation over an hour-long agenda. The speaker pool is small, in the low tens. The single-channel devices supply only post-processed mono and the organisers note their internal processing can suppress a talker outright. The reference transcripts for the evaluation split are withheld.

LOTUSDIS | OPEN (CC BY-SA 4.0)

- **What it holds:** a Thai meeting corpus released in 2025 by the Speech and Text Understanding Research Team at NECTEC: 90 sessions of spontaneous unscripted dialogue, 15 to 20 minutes each, three participants per session, 86 distinct speakers, about 20 hours of unique meeting time and 114 hours across all microphones, split 88 hours training, 12.8 development and 13.3 evaluation (Tipaksorn 2025, arXiv 2509.18722).
- **Recording condition:** the reason this corpus matters here. Nine independent single-channel devices spanning six microphone types record the same conversation at fixed measured distances from 0.12 to 10 metres: three lavaliers and three table-mounted condensers at 12 to 15 centimetres for the near-field reference, a tabletop loudspeaker-microphone at 2 metres, and two Bluetooth speakerphones at 3 and 10 metres. Line of sight is maintained, synchronisation is by slate pulse verified by cross-correlation to sub-sample alignment, and there is no array processing at all. Overlap is frequent and natural.
- **Labels and their ground truth:** utterance-level transcripts with speaker labels and an explicit overlap mask, produced by three trained annotators to a unified guideline and reviewed session by session by a senior annotator. It supervises verbatim text and speaker attribution; no summaries. Its published baseline is the sharpest far-field number in this document: an off-the-shelf recogniser scores 81.6 percent WER on the distant microphones and 49.54 percent after fine-tuning on distance-diverse conversational data.
- **Holder and licence:** NECTEC, Thailand; released "under a permissive CC-BY-SA 4.0 license" in the corpus paper's own words. Commercial use is permitted and the ShareAlike clause binds derivative datasets to the same terms. This is the only licence in this document stated by the holder in a document read here.
- **Access route and precedent:** public download from github.com/kwanchiva/LOTUSDIS, with standard train, development and test splits and a reproducible baseline. No outside group's use is named yet.
- **Who else holds it:** everyone.
- **Known limits:** Thai, so it is acoustic and device evidence and not language evidence for an English product. Three speakers per session and a fixed room layout. The far-field devices are consumer speakerphones rather than phones, and no handset appears. Recent enough that no independent group has used it.

ICSI Meeting Corpus | OPEN (licence not established)

- **What it holds:** 75 naturally occurring meetings, about 72 hours, recorded at the International Computer Science Institute in Berkeley between 2000 and 2002: weekly research-group meetings of about an hour in which students and professors discuss technical work (Prevot 2025; Rennard 2023).
- **Recording condition:** participants wear close-talking microphones and the table carries four far-field microphones (Tipaksorn 2025, Table 1, records "close-talk worn + table top"; the corpus

sweep describes four tabletop microphones). Three to ten participants per meeting, six on average. Real interaction between people who already know each other, so overlap and interruption are natural rather than staged, and the corpus is named in the diarization literature as one of the two most speaker-congested sets, meaning the number of people talking in a short window frequently exceeds what a segmentation model assumes.

- **Labels and their ground truth:** gold transcripts either fully human-produced or human-corrected from machine output, plus topic segmentation, dialogue acts and extractive and abstractive summaries in the same four-part structure as AMI (Rennard 2023). Human-written abstractive summaries exist for 61 of the 75 meetings. It supervises verbatim text, speaker attribution and an abstractive summary. It does not supervise action items: the corpus was released without publicly available action-item annotations, and the one research annotation that added them covered 18 meetings and is reported as no longer publicly available (Liu 2023, section 2.2). The average ICSI summary runs 534 words (arXiv 2212.08206).
- **Holder and licence:** two incompatible accounts, neither read from a holder's own page. The corpus sweep places it with the Linguistic Data Consortium (LDC2004S02 and LDC2004T04) under the LDC User Agreement for Non-Members with a per-user signed licence and a fee, quoting 7,500 US dollars non-member and 3,750 reduced; the catalogue entry cited for that fee is Switchboard-2 Phase II, not ICSI's own, so the figure is not established. A corpus-comparison table records ICSI as CC BY 4.0 (Tipaksorn 2025, Table 1), consistent with the Edinburgh-hosted distribution at groups.inf.ed.ac.uk/ami/icsi/.
- **Access route and precedent:** a preprocessed public release of AMI and ICSI including the summary annotations is published at github.com/guokan-shang/ami-and-icsi-corpora (Rennard 2023); the LDC catalogue is the other route. Precedent for outside use is heavy: the survey leaderboard reports ICSI results from a dozen independent groups (arXiv 2212.08206).
- **Who else holds it:** everyone, through the preprocessed release, subject to terms nobody in this project has read.
- **Known limits:** academic research-group talk, dense with technical vocabulary and with shared background the transcript does not contain, which the survey names as a specific difficulty for summarisers. Formal corporate action items are rare in it by the nature of the meetings. It is a 2003 recording on that era's hardware.

AliMeeting | OPEN (licence not established)

- **What it holds:** Mandarin office meetings collected by Alibaba for the ICASSP 2022 Multi-channel Multi-party Meeting Transcription challenge: 104.75 hours of training data, 4 hours for evaluation and 10 hours for test, about 120 hours in total (Shi 2023, arXiv 2211.00511). Session counts differ slightly by source: 212 training, 8 evaluation and 20 test sessions in one account, 220 in a corpus-comparison table, 240 in another (Tipaksorn 2025, Table 1). 481 distinct speakers.
- **Recording condition:** an 8-channel circular far-field microphone array on the table (Ali-far) recording the same meetings as a headset worn by each participant (Ali-near), at 0.3 to 5 metres, across 13 conference rooms of 8 to 55 square metres with reverberation times (RT60) from 0.3 to 0.6 seconds. Sessions run 15 to 30 minutes with 2 to 4 participants. Overlap is the highest of any real corpus located, 42.27 percent of training duration and 34.76 percent of evaluation. Participants were instructed to stay in the same seats throughout, which stabilises direction-of-arrival estimation and costs ecological validity.
- **Labels and their ground truth:** verbatim transcripts with speaker attribution, scored in speaker-dependent character error rate (CER) because Mandarin is written without spaces. It supervises verbatim text and speaker attribution and carries no summaries. Its transcripts are also the base layer under AMC-A, profiled below, which adds action-item labels to 224 of these meetings.
- **Holder and licence:** Alibaba, distributed through OpenSLR as resource 119. The corpus sweep states that commercial use is permitted and reaches that by reasoning that "datasets on OpenSLR generally utilize the Apache 2.0 or CC BY 4.0 framework", which is not a licence read. A corpus-comparison table records CC BY-SA 4.0 (Tipaksorn 2025, Table 1), whose ShareAlike clause would bind derivative datasets to the same terms.
- **Access route and precedent:** direct download from openslr.org/119. Precedent is the whole ICASSP 2022 challenge field plus later outside work, including the multi-channel speaker-

attributed recognition study read here (Shi 2023) and the Alibaba group that built AMC-A on top of it.
- **Who else holds it:** everyone.
- **Known limits:** Mandarin, so it transfers as an acoustic and array-geometry resource and not as language evidence for an English product. Seated immobility and an unusually high overlap ratio both push it away from a normal meeting. It carries no summaries.

CHiME-6 | OPEN (licence not established)

- **What it holds:** 20 dinner parties in real homes, each with four participants who are friends, each party lasting at least two hours and split into a kitchen, a dining and a living-room phase of at least 30 minutes each (Watanabe 2020, arXiv 2004.09249). Repartitioned for the later challenges into 16 training, 2 development and 2 evaluation sessions with no speakers shared: 40:05 hours of training audio with 79,967 utterances and 32 speakers, 4:27 development and 5:12 evaluation (Cornell 2025, Table 3).
- **Recording condition:** six Microsoft Kinect devices, each a linear array of four sample-synchronised microphones plus a camera, placed so at least two cover each room; each Kinect recorded to its own laptop. For transcription reference, each participant wears Soundman OKM II Classic Studio binaural in-ear microphones through a Soundman A3 adapter onto a body-worn Tascam DR-05 recorder (Watanabe 2020). Participants move between rooms. This is the noisiest and most overlapped scenario in the benchmark suite: 43.5 percent of development duration is two or more people at once (Cornell 2025).
- **Labels and their ground truth:** verbatim transcripts with speaker attribution and utterance segmentation, established from the in-ear recordings. It supervises verbatim text and speaker attribution and carries no summaries. Some personally identifying material was redacted after recording as part of the consent process, and background television and commercial music were disallowed to avoid capturing copyrighted content.
- **Holder and licence:** the University of Sheffield and Inria, distributed through the CHiME challenge series. The corpus sweep states that the Sheffield paid commercial licence no longer applies and the data is now under CC BY-SA 4.0, and cites an unrelated Hugging Face dataset for it; a corpus-comparison table independently records CC BY-SA 4.0 (Tipaksorn 2025, Table 1). No holder page was read, so the ShareAlike claim stands unverified, and if it is right the clause would bind any derivative dataset to the same terms.
- **Access route and precedent:** the challenge toolkit chime-utils downloads CHiME-6, DiPCo, Mixer 6 and NOTSOFAR-1 with a single command (Cornell 2024). Precedent is every CHiME-6, CHiME-7 and CHiME-8 participant.
- **Who else holds it:** everyone who can run the toolkit.
- **Known limits:** a dinner party is not a meeting; the lexicon is cooking, eating and socialising, and no business decision is ever taken in it. Inter-array synchronisation is imperfect, because CHiME-6 re-synchronised the CHiME-5 audio using video available only to the organisers and residual misalignment still runs to several thousand samples. An earlier scoring error, in which an unannotated first minute was scored anyway and inflated insertion errors, is corrected only by applying the evaluation map files the later challenges supply.

DiPCo | OPEN (licence not established)

- **What it holds:** the Dinner Party Corpus, released by Amazon in 2019: 10 sessions of four-person dinner conversation, 32 speakers, all recorded in one room. As repartitioned for CHiME-8, 1:12 hours of training with 1,379 utterances, 1:31 development and 2:36 evaluation with 3,405 utterances (Cornell 2025, Table 3).
- **Recording condition:** five far-field devices, each a 7-microphone circular array with a centre microphone, scattered around a single room, at 1 to 4 metres from the talkers, plus an on-speaker lapel microphone for each participant (Cornell 2025, section 3.1.2; Tipaksorn 2025, Table 1). All far-field signals are sample synchronised, unlike CHiME-6. Overlapped speech is 30.6 percent of development and 24.9 percent of evaluation.
- **Labels and their ground truth:** verbatim transcripts with speaker attribution, referenced to the lapel recordings. It supervises verbatim text and speaker attribution; no summaries.

- **Holder and licence:** Amazon Science. The corpus sweep states the code under Apache 2.0 and the audio and metadata under CC BY 4.0, and cites a Python package index page for it, which does not support the claim. A corpus-comparison table instead records CDLA-Permissive (Tipaksorn 2025, Table 1). Both would permit commercial use; neither was read from Amazon's own repository.
- **Access route and precedent:** direct download from github.com/amazon-science/dipco, or the same one-command challenge toolkit; precedent is the CHiME-7 and CHiME-8 participant field.
- **Who else holds it:** everyone.
- **Known limits:** small, and one room and one microphone placement shared by every session, which the challenge review calls a best case for adapting a system to a fixed environment and rare in real deployment. The lapel signals are badly misaligned with the far-field ones, by over a thousand samples and sometimes in the wrong direction, so the reference channel needs work before it can be used for alignment. Dinner-party content again.

Mixer 6 Speech | APPLICATION (LDC)

- **What it holds:** 1,425 recording sessions made in 2009 and 2010 with 594 native English speakers, each session containing entry questions, an interview of about 14 minutes, transcript reading and a phone call. The parent corpus totals 15,863 hours; the transcribed conversational part released for CHiME-8 is 80 hours.
- **Recording condition:** two people, an interviewer and a subject, in one of two rooms, captured by 10 heterogeneous far-field devices at once: arrays that emit a single processed signal, commercial single-microphone recorders and a camcorder, plus close-talk microphones, a lapel for the subject and a headset for the interviewer, at 13 different distances at 16 kHz. This is the only located corpus that puts different DEVICE TYPES on the same conversation in English, which makes it the field's device-comparison design. Overlap is the lowest of the suite, 19.5 percent of development and 13.9 percent of evaluation (Cornell 2025, Table 3).
- **Labels and their ground truth:** the original release transcribed almost nothing beyond a read-speech section. CHiME-7 annotated 450 of the 1,425 interview portions, the interviewer side semi-automatically with a large recogniser on the lapel channel followed by forced alignment and then a manual check, with CHiME-8 extending the manual check to development. The fully annotated result is 24 training, 35 development and 23 evaluation sessions totalling 20:54 hours, with a further 63:06 hours carrying subject-only partial annotation. It supervises verbatim text and speaker attribution; no summaries.
- **Holder and licence:** the Linguistic Data Consortium (LDC) at the University of Pennsylvania, as LDC2013S03 and, for the transcribed subset, LDC2025S07. The LDC User Agreement for Non-Members governs, each user signs, and the catalogue entry read by the corpus sweep quotes 3,000 US dollars non-member and 1,500 reduced-licence; commercial use is permitted once the agreement is signed and the fee paid.
- **Access route and precedent:** a catalogue purchase from the LDC. Named precedent: every CHiME-7 and CHiME-8 participant, who had free access for the duration of the challenge, and the annotation work itself, led from Johns Hopkins University (Cornell 2025).
- **Who else holds it:** LDC licensees, plus the challenge cohort for the window the challenge ran.
- **Known limits:** an interview is a two-person exchange, not a meeting, and only the interview portion counts as conversational speech. Most of the corpus is still unannotated. The data has been public for over a decade, which the CHiME-8 organisers name directly as the reason the older evaluation sets were "not really blind".

Meeting transcripts with human summaries, minutes and action items

The thin family, and the one that decides whether the second and third promised outputs can be measured at all. Five entries carry a written label made by a person after the meeting; the largest forbids commercial use, the richest for action items is Chinese, the one with real hour-long project meetings has no audio, and the two English corpora that pair audio with summaries are profiled above with the audio they came from. **AMC-A (AliMeeting-Action Corpus), ELITR Minuting Corpus, AIMU, MeetingBank.**

AMC-A (AliMeeting-Action Corpus) | OPEN (licence not established)

- **What it holds:** 424 Mandarin meetings with manual action-item annotations on manual transcripts of the recordings, built by the Speech Lab of DAMO Academy, Alibaba Group, by extending 224 previously published AliMeeting meetings with 200 new ones (Liu 2023, arXiv 2303.16763). Each session is a 15 to 30 minute discussion by 2 to 4 participants on topics biased towards work meetings in various industries.
- **Recording condition:** for the 224 inherited meetings, the AliMeeting condition profiled above, an 8-channel table array plus per-participant headsets. The paper does not state the recording condition of the 200 added meetings and the annotation itself is on transcripts, not audio, so this corpus should be treated as a text label layer with audio available for part of it.
- **Labels and their ground truth:** every sentence is labeled for whether it contains action-item information, meaning a task description, a time frame or an owner. Counts at that granularity: 306,846 utterances, 1,506 action items, an average of 3.55 per meeting with a standard deviation of 3.97, split 295 train, 65 development and 64 test meetings. Ground truth is three independent annotators working from detailed guidelines on candidate sentences pre-highlighted for temporal expressions and action verbs, with an expert reviewing the majority vote where they disagreed; the average pairwise agreement is a Cohen's kappa of 0.47 (Liu 2023, Table 1). It supervises extracted action items and nothing else, and its authors call it the largest action-item detection corpus in any language.
- **Holder and licence:** Alibaba DAMO Academy, released through ModelScope as Alimeeting4MUG (modelscope.cn/datasets/modelscope/Alimeeting4MUG/summary), with the detection code at github.com/alibaba-damo-academy/SpokenNLP. No licence text was read.
- **Access route and precedent:** public download from ModelScope. No outside group's use is named in the sources read here.
- **Who else holds it:** everyone who can reach the repository.
- **Known limits:** Mandarin. The annotation is sentence-level binary classification, so it says a sentence contains an action item and does not say what the item is, who owns it or when it is due. Agreement at 0.47 is moderate, and the paper's own explanation is that action items are inherently subjective, citing a kappa of 0.36 on an earlier ICSI attempt.

ELITR Minuting Corpus | OPEN (licence not established)

- **What it holds:** transcripts of 113 technical project meetings in English and 53 in Czech, over 160 hours of meeting content, from the European Live Translator project at Charles University, distributed through LINDAT. It is one of only three English meeting-and-summary corpora the abstractive-summarisation survey could name, alongside AMI and ICSI, which together offer roughly 280 hours (Rennard 2023).
- **Recording condition:** not usable as an audio corpus. The original audio recordings are not released and sections of certain meetings are censored, both for privacy. The meetings themselves are natural, unscripted, work-based project meetings averaging over an hour, which is the condition every other corpus here lacks.
- **Labels and their ground truth:** hand-written minutes. Unlike AMI and ICSI annotators, ELITR annotators "were not provided with a structure for producing minutes", so the results vary widely from one annotator to the next, which the survey names as making the corpus a resource for studying that subjectivity (Rennard 2023). Gold transcripts are human-produced or human-corrected from machine output. It supervises an abstractive summary only, at a granularity of 166 minuted meetings; it cannot supervise verbatim text from audio, speaker attribution from audio, or action items as a separate field.
- **Holder and licence:** Charles University and the LINDAT repository. The corpus sweep states CC BY-NC-SA 4.0, which would forbid commercial use; that is the same licence string the kickoff pass read on MeetingBank's data card, the sweep states no licence for MeetingBank at all, and no LINDAT page was read here, so the attribution is not established.
- **Access route and precedent:** direct download from the LINDAT repository, and a preprocessed public release including the annotations at github.com/guokan-shang/elitr-minuting-corpus, published by a group outside the original project, which is itself the precedent (Rennard 2023).
- **Who else holds it:** everyone, for the text. The audio, the ELITR project alone.
- **Known limits:** no audio at all, unstructured minutes, and a third of the content in Czech. Its value

here is as evidence about what minute-writing looks like and about why real meeting audio does not get released, not as material anything acoustic can be trained or measured on.

AIMU | OPEN (licence not established)

- **What it holds:** an extended annotation layer over 22 named public ICSI meetings, published by Microsoft Research in 2016, marking the turns at which an automated meeting assistant could act (Chen and Hakkani-Tur 2016, aclanthology L16-1117).
- **Recording condition:** inherited from ICSI, profiled above: close-talking microphones plus four tabletop far-field microphones on academic research-group meetings of about an hour. The annotation itself is on turns of transcript, not on audio.
- **Labels and their ground truth:** 21,035 speaker turns, of which 318 carry an actionable item, 1.5 percent, spread very unevenly by meeting type (4.2 percent in one group of meetings, 1.3 and 1.5 percent in the other two). Ten intent types are defined across five domains, including create-reminder, create-calendar-entry, add-agenda-item, send-email and search, each with its own argument slots such as owner, date, time and reminder text; the paper positions conventional action items as a subgroup of these and does not report a count for that subgroup alone. Agreement was measured on two doubly annotated meetings: 0.644 on whether a turn contains an actionable item at all, 1.000 on which action it is once both annotators agree there is one, and 0.673 overall (Table 3 and Table 4). It supervises extracted action items in the broad sense, and nothing else.
- **Holder and licence:** Microsoft Research. The paper's data-availability statement gives research.microsoft.com/projects/meetingunderstanding/ as the release address. No licence text was read and no source read here confirms that the address still resolves.
- **Access route and precedent:** the address above. No outside group's use is named in the sources read here, and a sibling ICSI action-item annotation covering 18 meetings is reported as no longer publicly available (Liu 2023), which is a reason to check this one before relying on it.
- **Who else holds it:** everyone, if it is still there.
- **Known limits:** 318 positive examples is a very small target class, and the schema is an assistant's task list rather than a meeting's to-do list, so it labels a request to open a calendar the same way it labels a commitment. Academic meetings, and a twenty-year-old recording underneath.

MeetingBank | OPEN (NonCommercial only)

- **What it holds:** 1,366 city-council meetings and over 3,579 hours of video from the councils of Alameda, Boston, Denver, Long Beach, King County and Seattle, hosted through Archive.org (data card, huggingface.co/datasets/huuuyeah/meetingbank, read at kickoff).
- **Recording condition:** public civic proceedings in council chambers, with an agenda and a chair; microphone type, placement and speaker counts are not stated on the data card, and the audio arrives as broadcast video. A different acoustic and conversational regime from four people around a table.
- **Labels and their ground truth:** 6,892 segment-level summarisation instances, transcripts, the councils' own official minutes and agendas, and summaries from six systems alongside human annotations (data card). It supervises an abstractive summary and, through the transcripts, verbatim text; the segment count is the number that governs a summarisation claim, not the 3,579 hours. No action items as a separate target. The corpus sweep repeats the 6,892 figure with a citation that points back at our own question rather than at the data card, so the data card remains the only support for it.
- **Holder and licence:** released by the dataset authors on Hugging Face under CC BY-NC-SA 4.0. The NonCommercial clause forbids use in a product that is sold and the ShareAlike clause binds derivatives to the same terms (data card).
- **Access route and precedent:** immediate public download. No outside-group use is named in the sources read; one meeting-summarisation evaluation says explicitly that it considered and excluded MeetingBank to avoid imbalancing the meeting types in its study (arXiv 2404.11124).
- **Who else holds it:** everyone, for research and evaluation. Nobody, for a product that is sold.
- **Known limits:** the licence first. Then the regime: civic proceedings are formal, chaired, largely non-overlapping and already minuted by a clerk, which is the opposite of the acoustic and conversational problem an in-person business meeting poses.

Simulated far-field meeting audio

One entry, and it has its own family because its labels are exact and its conversations are not real. It is how the current front ends were trained when nobody could record enough meetings. **NOTSOFAR-1 simulated training set.**

NOTSOFAR-1 simulated training set | OPEN (licence in dispute)

- **What it holds:** about 1,000 hours of synthesised far-field meeting audio, built to the same array geometry as the recorded set, using 15,000 real acoustic transfer functions (ATFs) that the team physically measured in various positions and rooms with the target hardware, and a mean-opinion-score-filtered clean speech source of about 500 hours (Vinnikov 2024; Abramovski 2025).
- **Recording condition:** simulated. Clean speech is convolved with the measured transfer functions and summed at varying offsets to produce overlap, with noise recorded on the same hardware injected. There is no real room, no real turn-taking and no real speaker movement.
- **Labels and their ground truth:** exact by construction, since the mixer knows which words belong to which source at which time. It supervises verbatim text and speaker attribution for training purposes; it settles no accuracy question and carries no summaries.
- **Holder and licence:** Microsoft, distributed with the recorded set, so it inherits the same unresolved licence question described in the NOTSOFAR-1 profile above.
- **Access route and precedent:** public download alongside the recorded set. Every multi-channel system in the NOTSOFAR-1 challenge trained its separation model on this data (Abramovski 2025).
- **Who else holds it:** everyone.
- **Known limits:** the challenge review's own finding is the limit: every submitted separation model was trained on simulated data only, and the organisers name fine-tuning on real meeting audio as the unexplored path. A later controlled study makes the size of the gap explicit: training a multi-talker model on synthetic mixtures alone reached 16.0 percent tcpWER on the AMI single distant microphone (SDM) condition and 20.1 percent on NOTSOFAR-1 single-channel, and adding a small fraction of real in-domain data moved those to 15.2 and 16.3 percent (Mind the Gap, arXiv 2605.15442). The transfer functions themselves were measured on the target hardware, so the simulator is device-specific and does not transfer to a different microphone geometry.

Holders of Closed Cohorts

The data in this field that never left the building, grouped by what it holds. The academic side publishes its corpora, so this section is mostly the industrial side, and the industrial side is where the volume is.

Smartphone-captured conversational speech. Ochi and colleagues at Waseda University published in 2016 on multi-talker recognition over "an ad hoc microphone array, which consists of smartphones... realized using iPhone and Dropbox". The method for synchronising several consumer handsets is therefore on the record; the audio behind it was never released as a public benchmark, and the paper was not obtainable in this pass, so the quotation above is all that is established. Appen holds a proprietary corpus titled "English (United States) Conversational Smartphone Speech" containing 1,000 hours of fully transcribed data, available only by commercial procurement (quoted from appen.com/speech-and-audio-training-data; the page itself was not read here). Nothing states its speaker count per recording, and telephony collections of this kind are historically two-party.

Real business meetings with the customer's own feedback attached. Otter.ai states that its models are trained on millions of hours of audio recordings gathered from its user base, and Gong.io captures and transcribes thousands of hours of business-to-business sales calls and internal meetings through what it calls its Revenue Graph. Both are real meetings of natural length captured on consumer microphones, both are locked behind business privacy and compliance commitments, and no outside researcher or competing developer has obtained either (quoted from the vendors' own pages and a third-party profile; no data-availability statement of any kind exists for either). Defined.ai builds and sells bespoke conversational datasets to order, including call-centre audio and a 225-hour annotated human-demonstration set, on commercial purchase agreements with no published price.

The withheld halves of published corpora. Microsoft holds the reference transcripts for the NOTSOFAR-1 evaluation split, deliberately, so that a challenge score cannot be overfitted, and it holds the 15,000 measured acoustic transfer functions behind the simulated set; the synthesised audio is released and no source read here says whether the measurements are. The ELITR project holds the audio behind its 113 English and 53 Czech meetings, withheld for privacy (Rennard 2023), and that is the sharpest single fact in this document: the right meetings were recorded, and privacy stopped their release. The CHiME-6 organisers hold the video recorded alongside the audio, used to re-synchronise the arrays and described as available only to the organisers (Cornell 2025). The Linguistic Data Consortium holds the unannotated remainder of Mixer 6, 975 of 1,425 sessions with only partial subject-only annotation, and the corpus's conversational portions were transcribed by the challenge organisers rather than by the holder.

Meeting audio with summaries. There is no located academic closed cohort here, and there is an explicit statement of why the public ones are so few: "the cost of producing such corpora, together with concerns about the privacy of meeting content, mean that there are very few such data sets available", after which the survey names three for English and no more (Rennard 2023). A second survey gives the same reason from the other side, that most meetings performed in industry are proprietary (arXiv 2212.08206).

Making Data That Does Not Exist

Four signal-label pairs the located corpora do not cover, and what producing each one consists of in this field, as facts about how the field has done it before.

Smartphone-captured meeting audio with verbatim transcripts and speaker attribution. The field's protocol for this is a parallel-capture study, and its parts are all documented. The topology is the devices under comparison placed adjacently at the same point on the table, with a close-talking lapel or headset microphone on every participant to produce the reference. The reference transcript is made by humans listening to the per-speaker channels, and machine pre-transcription is refused rather than merely discouraged, because annotators accept plausible machine guesses in noisy segments and thereby import the model biases the corpus exists to measure (Vinnikov 2024). Scoring conventions matter as much as the audio: the NIST scoring toolkit's Global Mapping (GLM) files decide which alternative wordings count as the same words, and a mapping that expands contractions in both reference and hypothesis double-counts errors. Because independent devices share no word clock, their sample rates drift over a 30-minute meeting, so the CHiME-6 baseline aligned array signals to the reference headsets by cross-correlation with sox, and LOTUSDIS synchronised nine devices by slate pulse verified to sub-sample alignment. The metric is a speaker-attributed word error rate, cpWER or tcpWER, and the significance test is the Matched Pairs Sentence-Segment Word Error (MAPSSWE) test in the NIST toolkit, with McNemar or Wilcoxon signed-rank as non-parametric cross-checks. The shape of a minimum credible study, as the field's own challenge splits imply it: at least 20 sessions of at least 15 minutes, at least 20 speakers in groups of 3 to 5, rotating through at least 4 rooms of different reverberation time, 10 to 15 hours of test audio, one recogniser decoding both device streams so the microphone is the only variable. The pass criterion is symmetric and stated in advance: the phone passes if its cpWER is significantly lower than the comparison device's, or if MAPSSWE returns no significant difference; it fails if the comparison device is significantly lower at 95 percent confidence. Room diversity matters more than duration: NOTSOFAR-1 used 30 rooms and short sessions, AliMeeting 13 rooms and long ones.

The simulation shortcut is closed in both directions and this is not a matter of preference. Simulated far-field audio is accepted by the field for TRAINING an acoustic model and rejected universally for EVALUATING a hardware claim, because linear summation of convolved sources does not reproduce the Lombard effect, non-linear pre-amplifier compression or clipping when two loud voices hit the converter at once. And building a simulator for a new device is not cheaper than recording: NOTSOFAR-1's simulator rests on 15,000 acoustic transfer functions physically measured on the target hardware in real rooms, which is the recording session the simulation was supposed to avoid.

The other route in this field is not a study at all, and it is the one every commercial holder named above

took. A product ships, a training-data opt-in sits in its terms of service, and the users' own meetings are retained under it together with the users' corrections to the transcript, the summary and the action items, which are the only labels for the second and third outputs that ever reach volume. It is how Otter.ai and Gong.io came to hold what they hold. Two properties make it a different instrument from the parallel-capture study rather than a cheaper version of it. It produces nothing before a product ships, so it cannot settle a capability question in advance. And its cost is legal rather than logistical: a voice recording is personal data, a speaker label makes it biometric, the consent a terms-of-service checkbox obtains binds the user and not the other people in the room, and the ethics guidance read here holds that presenting commercial development as academic research invalidates the lawful basis of the consent.

Action items as an independently counted target. Two production routes exist and they cost differently. The AMI and ICSI route makes action items a section of a summary: one annotator writes up to 200 words under Actions and may leave it empty, which yields the 101 meetings and 381 items the literature has been able to derive, by treating dialogue acts linked to that section as positive examples. The AMC-A route makes them a label in their own right: sentence-level binary classification, three independent annotators on candidate sentences pre-highlighted for temporal expressions and action verbs, an expert adjudicating the majority vote, which yields 1,506 items over 424 meetings at a pairwise Cohen's kappa of 0.47. The pre-highlighting step is the documented cost lever and the agreement figure is the documented ceiling: the same paper reports a kappa of 0.36 for the earlier ICSI attempt and calls action items inherently subjective. On the summary side, the most rigorous published human protocol is Atomic Content Units, in which a reference summary is decomposed into binary atomic claims and each is checked against the system output; the benchmark that established it required over 150 hours of in-house annotation to produce 22,000 summary-level annotations, on news and chat data rather than meetings (RoSE, arXiv 2212.07981). One further external figure circulates, that multi-party annotation regularly exceeds 50 hours of labour per hour of audio, and no source read here carries a citation for it.

Long-session capture reliability. No corpus in any family records whether a recording completed, and the reason is structural rather than accidental: acoustic corpora ship cleanly truncated audio files, and a session that failed mid-collection was discarded before publication to keep the benchmark intact, so no release pairs a truncated file with a cause such as an application crash, a thermal shutdown or a microphone seized by an incoming call. This one cannot be produced from archives at all; the field's protocol for it is instrumentation, not annotation. The documented schema logs session start and end, buffer overrun and underrun counts, operating-system audio-session interruptions and watchdog-timer expiries, as structured records that never carry the audio itself, which keeps the telemetry out of biometric and personal-data scope. The shape of a minimum credible study, again from the field's own practice: several hundred real sessions from several dozen users on their own handsets across both platforms, with a defect defined as any session losing audio frames to a buffer underrun, an unhandled interruption or a watchdog timeout, and a defect-free completion rate above 99.0 percent as the pass criterion.

Buyer-stated reasons. What a recorder buyer is paying for is answered by interview or survey data about people, not by a corpus of audio, and none was located in any form. What exists is technology-reviewer opinion, which is a different instrument.

What consent and ethics look like when this data is made now. AMI and ICSI were collected before the General Data Protection Regulation (GDPR) was enforced; AMI went through the European Commission's ethics review procedure with a technical annex checklist rather than a United States institutional review board. Under current practice a voice recording is personal data and, where it is used to identify a person, special-category biometric data requiring explicit consent. Guidance read here separates the participant's informed consent from the processing privacy notice, treats true anonymisation of voice as impossible so that data is pseudonymised instead, and requires that participants be told their voice itself is an identifier; broad consent covers unknown future algorithmic uses and dynamic consent covers new phases. Two points bear directly on any collection done to settle a commercial claim: the ethics submission and the privacy notice must state the commercial intent, because presenting commercial development as purely academic research invalidates the lawful basis of the consent; and participants retain a right to withdraw up to publication or aggregation. Release practice has moved with it, from an open server to a Data Use Agreement (DUA) plus gated hosting that requires registration and contrac-

tually forbids re-identification, which is how NOTSOFAR-1 is distributed.

Dead Ends: Corpora That Cannot Be Used

A licence that forbids commercial use, with no located waiver. MeetingBank is released under CC BY-NC-SA 4.0; the NonCommercial clause bars use in a product that is sold and the ShareAlike clause would bind any derivative to the same terms. It remains usable for evaluation and for a paper. No route to a commercial licence was found and the data card names no rights holder to ask. The same licence string is attributed to ELITR by the corpus sweep, unverified, and if it holds the same bar applies there.

A ShareAlike clause that would follow the work out. If the CC BY-SA 4.0 attributions read here are right, CHiME-6, AliMeeting, AISHELL-4 and LOTUSDIS all permit commercial use and all require that a derivative dataset be released on the same terms. That is not a bar to using them and it is a bar to keeping quiet about what was built from them, which is a different decision and has to be made deliberately. None of the four was read from its holder's own licence file.

The audio was never released. The ELITR Minuting Corpus distributes transcripts and minutes only, with sections censored for privacy, so nothing acoustic can be trained or measured on it however good the minutes are. Route back: none found; the decision is described as a privacy one, and it is the same force that would apply to anything collected fresh.

The annotation went offline. The action-item annotations added to 18 ICSI meetings by Purver and colleagues, reported at a Cohen's kappa of 0.36, "are no longer publicly available" (Liu 2023, section 2.2). That is the earliest English action-item label set in this field and it cannot be obtained. AIMU's release address should be checked against the same risk before anything is built on it.

The access window closed. Free access to Mixer 6 was provided by the Linguistic Data Consortium to challenge participants for the duration of the challenge. Route back: the consortium's standard catalogue licensing, at the fee quoted in the profile above.

Public long enough that a held-out claim cannot be made. The CHiME-8 organisers write that in the previous edition the evaluation data was "not really blind", because only Mixer 6 was partially blind and that data had been public for more than a decade. This does not stop anyone using AMI, ICSI, CHiME-6, DiPCo or Mixer 6; it stops a claim that a number computed on them is a clean held-out result. NOTSOFAR-1's blind evaluation split is the located exception.

The wrong regime entirely. Read-speech and telephone corpora appear throughout this field's history and are allowed as external training material in the challenges, but a number computed on them is not meeting evidence. Corpora built by replaying clean speech through loudspeakers into a room, such as LibriCSS, sit in the same place: there is no natural overlap, the mixing is scripted, and the recordings carry no Lombard effect, which makes them useful for separation research and not a measurement of real conversation. The archive types considered and rejected for a device comparison are on the record with their reasons: podcasts (near-field dynamic microphones), video-call recordings (vendor noise suppression and automatic gain control irreversibly mask the raw microphone), body-camera audio (moving target, restricted), and parliamentary or council proceedings (push-to-talk gooseneck microphones, which misrepresent a single device on a table).

Coverage of the Pitch

One row per claim whose test needs data. Descriptive only: it records what exists, not which claim to keep.

Claim	Signal and label needed	Verdict	Nearest corpora
C1, an app on a phone can capture a meeting well enough for a usable transcript, summary and action items	Smartphone-captured in-person meeting audio with verbatim transcripts, speaker attribution and a summary	NONE	AMI supplies the meeting condition and all three label types, but on a headset and a table array; NOTSOFAR-1 and AliMeeting supply the acoustics and no summaries; no located corpus of any kind was recorded on a phone
C2, the phone's own microphones are better than the dedicated recorder's for this job	The same meetings captured simultaneously by a phone and by the comparison device, scored on one reference transcript	NONE	LOTUSDIS is the closest design, nine separate devices at 0.12 to 10 metres on one conversation, and contains no phone and no English; Mixer 6 puts 10 device types on a two-person interview; AliMeeting pairs near-field and far-field of the same meetings
C3, the three outputs are finished within the few minutes after the meeting ends	Wall-clock production time from end of recording to finished outputs, with recording length	NONE	No corpus records a production time; NOTSOFAR-1's blind evaluation audio is material a latency harness could be run on, and supplies no timing itself
C4, the phone keeps recording to the end of a real meeting	Long capture sessions with completion or loss recorded as the outcome	NONE	Nothing located records capture outcomes, and the reason is structural: failed sessions are discarded before release; CHiME-6's two-hour-plus parties are the longest real sessions located, on dedicated arrays
C5, the person buying a 159 dollar recorder would take the app instead	Stated purchase reasons from buyers of dedicated recorders	NONE	No corpus of any kind; the located material is reviewer opinion, not buyer research

What the Corpora Are Labeled For

The inventory, in no particular order. Each line is a signal-label pair the located corpora cover at a count that supports work, whether or not the pitch asked for it.

- Far-field multi-talker meeting audio with verbatim word-level transcripts and speaker attribution: NOTSOFAR-1 (280 to 315 meetings depending on the source, about 260 hours across all microphones), AMI (about 100 hours), AliMeeting (about 120 hours), CHiME-6 (49:44 annotated hours), LOTUSDIS (114 hours across nine devices), ICSI (75 meetings), DiPCo (5:19 hours), Mixer 6 (450 annotated interview portions).
- The same conversation heard simultaneously at the head and across the room: AMI (headset and 8-channel array), AliMeeting (headset and 8-channel array), NOTSOFAR-1 (close-talk on train and development), CHiME-6 (binaural in-ear), DiPCo (lapel), LOTUSDIS (lavalier), Mixer 6 (lapel and headset).
- The same conversation heard simultaneously by different device types at known distances: LOTUSDIS (9 devices, 6 microphone types, 0.12 to 10 metres, distances measured and reported), Mixer 6 (10 heterogeneous far-field devices at 13 distances), NOTSOFAR-1 (4 array devices and about 5 single-channel devices per meeting).
- Overlapped-speech ratios per split as a labeled property of the audio: the four CHiME-8 scenarios, from 13.9 to 43.5 percent of duration; AliMeeting at 34.76 to 42.27 percent; LOTUSDIS ships an explicit per-utterance overlap mask.
- Human extractive and abstractive summaries of meeting transcripts, structured into abstract, decisions, problems and actions: AMI (137 meetings, 322 words average) and ICSI (61 of 75 meetings, 534 words average).
- Machine-generated meeting summaries labeled by human annotators for error type, including hallucination and attributing a statement to the wrong participant: the Kirstein 2024 annotation set (175 summaries, 35 QMSum meetings by five systems, Krippendorff's alpha 0.76 to 0.83 for error detection).
- Sentence-level action-item labels as a target in their own right: AMC-A (424 meetings, 306,846 utterances, 1,506 action items, agreement 0.47).
- Turn-level assistant-executable intents with argument slots: AIMU (22 ICSI meetings, 21,035 turns, 318 actionable, 10 intent types).
- Action items derivable from a summary section: AMI (101 meetings, 381 items).
- Segment-level abstractive summaries of civic meetings, with official minutes as an independent reference: MeetingBank (6,892 instances over 1,366 meetings), research use only.
- Hand-written meeting minutes with no imposed structure: ELITR (113 English and 53 Czech meetings).
- Query-focused summaries of meetings: QMSum (1,808 query-summary pairs over 232 meetings drawn from AMI, ICSI and parliamentary committees).
- Dialogue acts and topic segmentation over meeting transcripts: AMI and ICSI.
- Elementary discourse units over spontaneous meeting speech: SUMM-RE (73 meetings, about 24 hours manually corrected and annotated, French).
- Synthesised far-field mixtures with exact source and speaker labels for training separation front ends: NOTSOFAR-1 simulated set (about 1,000 hours, 15,000 measured room responses).

Also Found, Not Profiled

Every corpus the sweep located that did not earn a profile, alphabetical, carrying what was already established.

Corpus	Holder	What it holds	What we know
AISHELL-4	AiShell	Mandarin office meetings on an 8-channel circular table array, with video for audio-visual diarization	120 hours, 211 sessions, 4 to 8 participants, 61 speakers, 0.6 to 6 metres, headset plus circular array, CC BY-SA 4.0 per a corpus-comparison table (Tipaksorn 2025, Table 1); listed as real-world long-form far-field in the CHiME review's dataset table
CHiME-5	CHiME challenge series	The same dinner-party recordings as CHiME-6	Shares the exact same data as CHiME-6, before the inter-array resynchronisation (Cornell 2025)
CID (Corpus of Interactional Data)	Aix-Marseille University	Eight one-hour French dialogues between friends, discourse-segmented	Eight hours total, two-party, not meetings; cited as the prior French resource SUMM-RE supersedes (Prevot 2025)
Ego4D	Ego4D consortium	Egocentric video and audio from head-worn devices	Named as a 2022 real-world, long-form, multi-speaker, far-field, multi-domain set in the CHiME review's dataset table; no card read
kiransarv action-item dataset	GitHub, individual	2,750 statements labeled for whether they contain an action item	Used as the action-item classifier's training data in an AMI summarisation study (arXiv 2312.17581); no holder, licence or provenance established

Corpus	Holder	What it holds	What we know
Kirstein meeting-summary error annotations	University of Goettingen	175 machine-generated summaries of 35 QMSum meetings, each labeled by a human annotator for eight error types including hallucination, wrong references and missing information	35 general-summary samples from the QMSum test set summarised by five systems; Krippendorff's alpha 0.76 to 0.83 for error detection; the paper states the annotations, the annotator guidelines and the code are released at github.com/FKIRSTE/emnlp2024-Meeting-Sum-Metrics (arXiv 2404.11124, read in full and indexed in LITERATURE.md); no licence established and the repository was not read
LibriCSS	Microsoft	Clean read speech replayed through loudspeakers into a conference room to create controlled overlap	10 hours, 10 sessions, 8 speakers per session, 40 speakers, 0.3 to 4 metres, 7-channel circular array, CC BY 4.0 per Tipaksorn 2025 Table 1; no natural overlap and no Lombard effect, so evaluation material for separation only
LibriSpeech	OpenSLR	Read audiobook speech	Permitted external training data in CHiME-7 and 8; read speech, so not meeting evidence
MISP	MISP challenge organisers	Far-field conversation with several sensing modalities	Named as a 2022 real-world multi-domain challenge set in the CHiME review's dataset table; no card read

Corpus	Holder	What it holds	What we know
QMSum	Yale LILY	Query-focused summaries built on top of AMI, ICSI and parliamentary committee meetings	1,808 query-summary pairs over 232 meetings: 137 AMI, 59 ICSI, 36 committee; average meeting 9,069.8 tokens, average summary 69.6 tokens (arXiv 2212.08206, arXiv 2404.11124). The corpus sweep states an MIT licence and cites an ACL events index for it, which does not support the claim; no licence established
Santa Barbara corpus	University of California, Santa Barbara	Recorded American English conversation	The earliest entry in the CHiME review's dataset table, 2000, marked real-world, long-form, multi-speaker, far-field and multi-domain
SUMM-RE	LINAGORA Labs	About 100 sessions of three 20-minute French event-planning meetings, 2 to 4 participants, one task per session including delegating the practical work	Most sessions recorded face to face on head-mounted microphones, a few over Zoom; the whole corpus automatically transcribed, with 73 meetings and about 24 hours manually corrected and discourse-annotated; at huggingface.co/datasets/linagora/SUMM-RE (Prevot 2025). Head-mounted capture means it is near-field, not a far-field resource
VoxCeleb 1 and 2	University of Oxford	Interview audio used to train speaker-discriminative models	Permitted external training data in CHiME-7 and 8, used for speaker-identity embeddings

Corpus	Holder	What it holds	What we know
WeCanTalk	Linguistic Data Consortium	Cantonese, Mandarin and English telephone conversations plus self-recorded video from 202 bilingual speakers in Hong Kong	At least 10 telephone calls of 8 to 10 minutes and at least 3 videos per speaker, collected June to October 2020 for the NIST 2021 Speaker Recognition Evaluation, to be published in the LDC catalogue; the corpus paper is itself under CC BY-NC 4.0. Two-party telephony and selfie video, not multi-party in-person meetings, and the smartphone reference in it is about handset penetration in Hong Kong, not about the capture device

Watchlist

Dated 2026-09-01. A fired trigger means this document is owed a refresh.

What to watch	Corpus	What it would change	Where it shows up
A challenge edition or corpus release whose capture devices include a phone on the table alongside an array	NOTSOFAR-1 or a successor in the CHiME series	It would create the first public phone-versus-device comparison and turn the microphone question from unanswerable into measured	The CHiME challenge task and data pages
Which of CC BY 4.0 and CC BY-NC-ND 4.0 actually governs	NOTSOFAR-1 recorded and simulated sets	The difference is between the field's current benchmark being usable in a commercial product and being unusable in one	The dataset's own licence file on the repository and the Hugging Face card
The licence and access terms of the smart-glasses conversation set	CHiME-8 MMCSG	It is the nearest thing to a consumer wearable corpus; its terms decide whether a wearable-capture baseline can be built at all	The CHiME-8 Task 3 data page

What to watch	Corpus	What it would change	Where it shows up
Whether the release address still resolves	AIMU	It is the only located English turn-level action-item annotation, and its sibling ICSI annotation has already gone offline	research.microsoft.com/projects/mee and any successor Microsoft Research page
An English counterpart, or an English annotation layer built the same way	AMC-A	It would move action items from a derived label with 381 examples to a directly annotated target with four figures of examples	ModelScope, and the meeting-understanding challenge tracks that use it
A second corpus adopting the measured-distance multi-device design	LOTUSDIS	It is the only located design that scores the same conversation across separate consumer-grade devices at stated distances, which is the shape a phone comparison needs	The corpus repository and the far-field recognition literature that cites it
Access terms outside a challenge window	Mixer 6 Speech	It is the only located English corpus with ten different capture devices on one conversation, so its terms decide whether device comparison is possible on real data	The Linguistic Data Consortium catalogue entry
A commercial-use variant or a licence revision	MeetingBank	It would move the largest located summarisation corpus from evaluation-only to something a product can be built on	Its Hugging Face data card
A release of the measured acoustic transfer functions behind the simulated set	NOTSOFAR-1 simulated training set	It would let anyone synthesise matched training audio for a different device geometry, including a phone's	The NOTSOFAR-1 repository
A reconciliation of 237, 280 and 315	NOTSOFAR-1 recorded meetings	Every per-meeting number this field quotes on that benchmark, including the overlap ratios and the hours, rests on which count is right	The repository contents against the dataset paper's own table