

# OSynth: A Longitudinal, Context-Sensitive Evaluation of Meeting Understanding

Ian Porada, Suraj Tripathi, Chaitanya Animesh, Leda Sari,  
Chaitanya Narisetty, Austin R. Matthews, Bohan Yu, Ramesh L. Thulasiram,  
Gregory Sell, Kean Chin, Dmitriy Genzel

Otter.ai

Mountain View, California, USA

{ian, stripathi, canimesh, kkchin, dmitriy}@otter.ai

## Abstract

Meeting notetakers are increasingly deployed in settings where longitudinal context is available including calendar metadata, prior meetings, emails, slides, and speaker enrollments; however, meeting corpora remain limited in terms of size, domain coverage, and available context. We present OSYNTH, a large-scale synthetic dataset of simulated organizations comprising 570 hours of meeting audio together with associated artifacts and a latent ground-truth state. In a pairwise comparison, human raters distinguish OSYNTH meetings from real meeting transcripts with only 56% accuracy based on excerpts of 150–300 words. Across state-of-the-art LLMs, additional context yields significant gains on several entity-sensitive dimensions, demonstrating limitations of contemporary approaches and the potential usefulness of OSYNTH in developing better-contextualized meeting-understanding systems.

## 1 Introduction

Meeting understanding has expanded from summarization of an isolated transcript to a longitudinal problem in which one can draw on surrounding organizational artifacts such as prior meetings, shared documents, slide decks, internal emails, and speaker enrollments (Asthana et al., 2025; Laskar et al., 2023; Arakawa et al., 2025). Existing meeting benchmarks were built largely before this shift: they consist of standalone sessions with limited surrounding contextual artifacts (Carletta et al., 2006; Janin et al., 2003; Zhong et al., 2021; Hu et al., 2023; Nedoluzhko et al., 2022), and the community consequently has no way to measure, at scale, whether longitudinal grounding actually improves summarization or downstream outputs.

We introduce OSYNTH,<sup>1</sup> a fully synthetic, longitudinally generated meeting corpus that closes

<sup>1</sup><https://github.com/otter-ai/osynth>

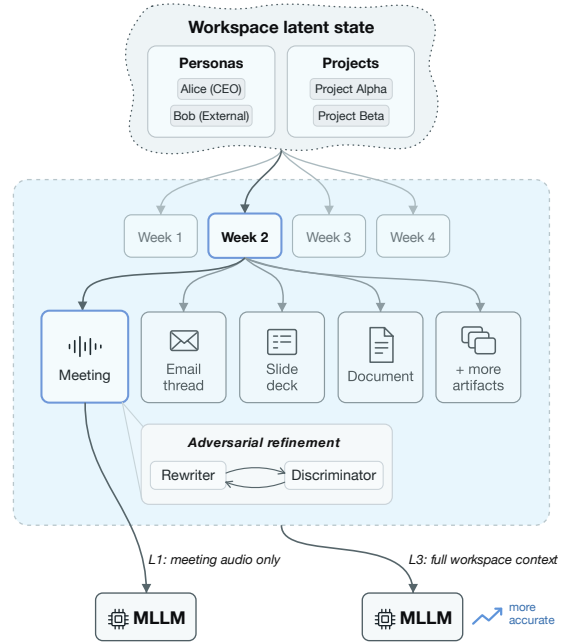


Figure 1: Each OSYNTH organization, termed a *workspace*, is generated from a single shared latent state from which all artifacts (audio, transcripts, etc.) are derived. Meeting transcripts are adversarially refined to improve realism, and in experiments multimodal large language models (MLLMs) improve significantly in certain dimensions when given access to broader workspace context (L3) as compared to a single meeting’s audio (L1).

this gap, and use it to show that as more context is made available, summary quality from state-of-the-art MLLMs improves significantly across entity-sensitive dimensions such as summary faithfulness, action-item assignee accuracy, and cross-meeting speaker-identity consistency.

Following existing work that uses language models to synthesize conversational data (Chen et al., 2023; Kim et al., 2023; Ding et al., 2023; Xu et al., 2023), OSYNTH is generated hierarchically by Claude Sonnet 4.6 (Anthropic, 2026b). To ensure that the generated transcripts resemble real meetings, we develop an adversarial refinement

process grounded in real AMI transcripts (Carletta et al., 2006), in which Claude Opus 4.7 (Anthropic, 2026a) refines drafts against Gemini 3.1 Pro (Google DeepMind, 2026) as a realism discriminator (§3.4). Each meeting’s context maps to a *workspace* which is a simulated organization with a latent state (personas, ongoing projects, prior decisions, outstanding tasks) from which every observable artifact is derived (Figure 1). Observable artifacts include: transcripts (turn-by-turn, with disfluencies), synthesized audio rendered with Qwen3-TTS (Qwen Team, 2026) and simulated acoustics, executive summaries, action items, outlines, internal emails, and slide decks.

Because the latent state is known by construction, evaluation metrics can be computed against an exact reference. OSYNTH also contains only synthetic data and is shareable in full. This is in contrast to published meeting corpora which must redact sensitive content—for instance, ICSI (Janin et al., 2003) anonymizes speaker identities and ELITR (Nedoluzhko et al., 2022) anonymizes named entities—or are relatively limited in size and scope. The final dataset spans 570 hours of audio across 160 workspaces, and the nature of the data creation pipeline allows for scaling to even larger sizes. The released corpus is English-only, although the pipeline is language agnostic, and we performed a preliminary data generation in five additional languages (Appendix R).

We use OSYNTH to evaluate two state-of-the-art systems (an OpenAI stack composed of GPT-4o and GPT-5.5 (OpenAI, 2026a,c), and a Gemini stack that pairs Gemini 3.1 Pro (Google DeepMind, 2026) for enrollment-based speaker identification with Gemini 3 Flash (Google DeepMind, 2025) for transcription and summarization) under three progressively richer levels of context: (L1) single meeting audio only; (L2) L1 plus the model’s own outputs on prior meetings; and (L3) full workspace artifacts (slides, emails, and speaker enrollments). Systems are scored both on isomorphic metrics derived directly from the latent state (atomic-fact coverage, named-entity spelling accuracy, and true-entity mention rate) and on standard summarization metrics such as LLM-as-a-judge (Zheng et al., 2023), which has been shown to underperform reference-grounded metrics in measuring factual accuracy (Tang et al., 2024).

We also validate the benchmark itself with a human-discrimination study (§3.5) in which annotators are shown a length-matched pair of one real

and one synthetic excerpt and asked to identify the synthetic side; they do so significantly less often for OSYNTH (56.3%) than for a control without our adversarial refinement (79.6%; p-value 0.001).

In summary, our contributions are: (i) the OSYNTH pipeline and associated dataset, a novel pipeline for generating synthetic meetings with aligned workspace artifacts and an exact latent-state reference; (ii) a validation of the dataset’s transcript-level realism through a pairwise human-discrimination study; and (iii) a context-sensitive evaluation of state-of-the-art, multimodal large language model (MLLM) families, showing that longitudinal context yields gains on entity-sensitive summarization and notetaking dimensions that existing benchmarks do not expose.

## 2 Related work

**Meeting and dialogue corpora.** Spoken meeting corpora such as AMI (Carletta et al., 2006), ICSI (Janin et al., 2003), MeetingBank (Hu et al., 2023), ELITR (Nedoluzhko et al., 2022), AliMeeting (Yu et al., 2022), and NOTSOFAR-1 (Vinnikov et al., 2024) provide important testbeds, but they are either narrow in domain, lack audio, lack longitudinal workspace artifacts, or focus on ASR/diarization rather than summarization and downstream meeting understanding (Table 1). The CHiME shared-task series (Barker et al., 2018; Cornell et al., 2023) has driven progress on far-field ASR, diarization, and SID for meeting-style audio, but is upstream of the tasks we study such as summarization. The AutoMin shared-task series (Ghosal et al., 2021, 2023; Shinde et al., 2025) instead targets the generation side directly, eliciting minutes from ELITR project meetings and European Parliament sessions and, in its third edition, question answering over meeting transcripts; participating methods moved from segment-then-summarize sequence-to-sequence pipelines to prompted LLMs. In both tasks the system input is the transcript of a single session, so the evaluation cannot isolate the effect of surrounding workspace context—the comparison OSYNTH is built to support. Dialogue-summarization resources such as QMSum (Zhong et al., 2021), SAMSum (Gliwa et al., 2019), and MediaSum (Zhu et al., 2021) broaden task format and domain, but do not expose the deployed workspace setting. OSYNTH, by contrast, combines longitudinal workspace context with an exact latent reference at a larger scale.

| Corpus         | Hours | Domain            | Reference              | Contextual artifacts                                                               |
|----------------|-------|-------------------|------------------------|------------------------------------------------------------------------------------|
| ICSI           | 72    | Academic          | gold + recorded audio  | prior meetings (recurring groups)                                                  |
| AMI            | 100   | Product design    | gold + recorded audio  | slides (subset), whiteboard video, prior meetings (scenario subset)                |
| AliMeeting     | 119   | Multi-domain (zh) | gold + recorded audio  | —                                                                                  |
| ELITR-Minuting | 80    | Technical         | gold, no audio         | minutes                                                                            |
| MeetingBank    | 3,579 | City council      | ASR + recorded audio   | agenda, prior meetings                                                             |
| NOTSOFAR-1     | 32    | Office meetings   | gold + recorded audio  | —                                                                                  |
| OSYNTH (ours)  | 570   | Multi-domain      | gold + synthetic audio | calendar, prior meetings, emails, docs, slides, speaker-identity voice enrollments |

Table 1: OSYNTH compared to existing meeting corpora. *Reference* encodes the transcription process (*gold* vs. *ASR*) and audio source (*recorded* vs. *synthetic*); OSYNTH is the only synthetic corpus in this table and is validated against real meeting transcripts by the human-discrimination study in §3.5.

**Evaluating context-augmented dialogue summarization.** Prior work shows both the need for and the difficulty of evaluating context-aware summarization: dialogue summaries often hallucinate (Tang et al., 2024), reference-free LLM judges can underweight factual precision (Liu et al., 2023b), and long-context benchmarks (Bai et al., 2024; An et al., 2024) do not isolate longitudinal workspace grounding. Retrieval-augmented meeting systems (Asthana et al., 2025; Laskar et al., 2023; Arakawa et al., 2025) therefore require benchmarks that can score whether added context improves factual outputs. This motivates our use of *isomorphic* latent-state metrics alongside conventional LLM-as-judge (Zheng et al., 2023; Liu et al., 2023a) scores.

**Synthetic conversational data.** LLM-prompted dialogue synthesis (Chen et al., 2023), role-play and game-based corpora (Rameshkumar and Bailey, 2020; Tomlin et al., 2025), and agent simulations with persistent memory (Park et al., 2023; Li et al., 2023; Hong et al., 2024) show that synthetic multi-party interaction can scale. OSYNTH adapts this idea to meeting evaluation by generating not only conversations but also the surrounding workspace artifacts and latent references needed for scoring. Related to our transcript generation are computational models of multi-party turn-taking dynamics (O’Bryan et al., 2025).

### 3 The OSYNTH dataset

OSYNTH is a corpus of fully synthetic workspaces. Each workspace is a simulated organization observed over a multi-week window, and is produced by a hierarchical LLM pipeline (§3.1, Figure 2) that first samples a latent state and then derives every observable artifact from that state. The latent state comprises an org chart with personas and

speaking styles, ongoing projects, a narrative arc of decisions and outstanding action items, and a calendar of recurring and ad-hoc meetings. The derived artifacts (§3.2) include clean turn-by-turn transcripts, ASR-ready multi-speaker audio rendered via TTS with voice cloning, per-meeting summaries, action items, outlines, inter-meeting emails and shared documents, slide decks, and per-speaker voice enrollments. Because every artifact is generated from the same explicit latent state, each observable can be scored against an exact reference. We conduct a two-alternative forced choice (2AFC) human-discrimination study which confirms that OSYNTH transcripts are difficult to distinguish from real meeting transcripts (§3.5).

#### 3.1 Hierarchical generation pipeline

The pipeline expands a workspace top-down across four levels of abstraction, with each level consuming the latent state produced by the previous one. We use `claude-sonnet-4-6` as the instruction-following LLM at every stage and pass a structured state representation through prompts to keep entities, decisions, and outstanding actions consistent across meetings. Appendix C gives additional implementation details.

**1) Workspace.** The first level samples a self-contained organization: a company description (industry, size, stage), an org chart with named personas, per-persona attributes (role, seniority, communication style, demographic metadata used later for voice assignment), an inventory of active projects, and a baseline schedule of recurring meetings (standups, reviews, 1:1s). Names and organizations are sampled from a generated pool rather than reused from real entities to avoid real instances of PII and ensure data diversity.

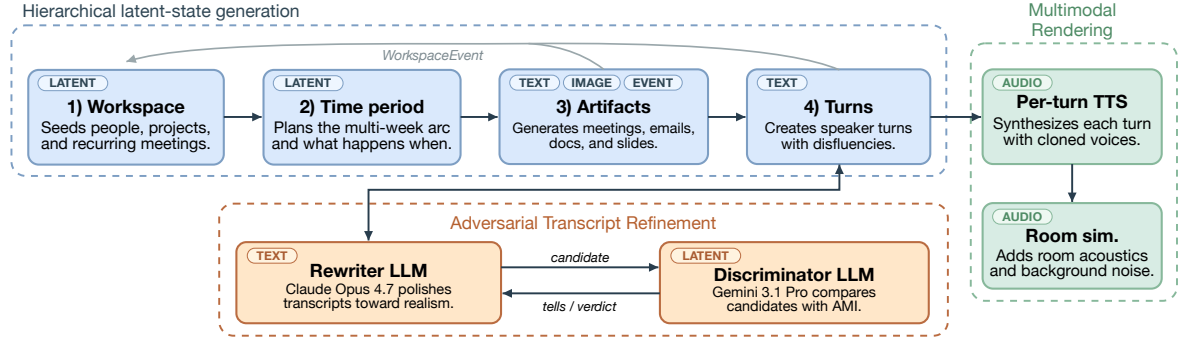


Figure 2: Overview of the OSYNTH synthetic data pipeline. A hierarchical LLM pipeline samples a workspace-level latent state, expands it into meetings and surrounding artifacts, renders the artifacts into audio and text, and maintains a rolling WORKSPACESTATE so that every observable can be scored against an exact reference. Artifacts emit WORKSPACEEVENTS that are considered when generating subsequent artifacts. Audio is synthesized turn-by-turn with voice cloning and the final audio channels are simulated (virtual, hybrid, or in-person acoustics) using pyroomacoustics (Scheibler et al., 2018).

**2) Time period and narrative arc.** The second level fixes a simulated calendar window and writes weekly plans: which projects make progress, which decisions are made, which action items are opened and closed, and which external events intrude.

**3) Artifacts.** The third level decomposes weekly plans into a concrete calendar of meetings together with the inter-meeting artifacts the same latent state drives: emails, shared documents, and slide decks tied to projects and meeting agendas. Each meeting spec is assigned an invite, agenda, attendees, and the slice of latent state it advances; each non-meeting artifact spec is anchored to its source project and (when applicable) to the meeting it prepares for or follows up on.

As each artifact spec is emitted, it also declares its cross-artifact references (e.g., a prep doc that points at an upcoming meeting, or an email thread that follows up on a decision); these references become edges in an explicit *dependency graph*, and we use Kahn’s algorithm with calendar-timestamp tiebreaking to obtain a topological generation order. Each artifact reads a scoped snapshot of the workspace built from its transitive dependencies’ events (termed WORKSPACEEVENTS) and emits new such events as it runs, so later artifacts (including later meetings) inherit a consistent history of decisions, opened and closed action items, and project status changes. A reducer function consolidates these events in timestamp order (with deterministic tiebreaking) to produce the final WORKSPACESTATE. Because the graph makes every conditional dependency explicit, artifacts that are mutually independent under it can be generated in parallel.

**4) Turns.** The fourth level generates turn-by-turn dialogue conditioned on personas, agenda, and the latent state. For each meeting, the initial transcript first passes through the adversarial refinement loop (§3.4); only the post-refinement turns are used to extract that meeting’s WORKSPACEEVENTS, so any structural rewrites the refiner accepts propagate into the final WORKSPACESTATE. We render three transcript variants from the same turn sequence: an initial *canonical* written form used for downstream references, a *semantic-verbatim* form with disfluencies and repairs, and a *TTS* form with non-verbal tokens for speech synthesis.

A final post-processing pass derives reference summary artifacts from the canonical transcript plus latent state, yielding gold references for the summarization metrics in §4.3.

### 3.2 Multimodal rendering

The TTS form of each transcript is rendered into multi-speaker audio with the Qwen3-TTS 1.7B base model and 12Hz tokenizer (Qwen Team, 2026). For voice cloning, we draw persona-matched spontaneous-speech reference clips from CommonVoice (Ardila et al., 2020) (Appendix L details the persona-to-voice matching) and use each clip once to synthesize a short per-speaker enrollment utterance; this enrollment is then frozen and reused as the voice-cloning prompt for every subsequent turn from that speaker. Notably, we perform significant filtering and preprocessing of voice prompts to reduce degenerate TTS outputs and use time and noise-based filtering to perform rejection sampling on final audio generations.

Turns are synthesized independently, mixed with realistic gaps and limited overlap, and post-processed with room simulation. Additional rendering, alignment, acoustic, and multilingual details appear in Appendices K–R.

### 3.3 Corpus statistics and release

The released corpus is English-only and comprises 160 workspaces, 2,126 meetings, and 570 hours of synthesized audio, spanning domains that include enterprise software development, sales pipelines, customer support, internal operations, medical, academic and non-business scenarios. Appendix A gives aggregate corpus statistics and per-workspace artifact counts, and Appendix B walks through one full workspace end-to-end with a representative instance of every released artifact type; Table 1 situates OSYNTH alongside existing meeting corpora along the axes of scale, available context, and the naturalness of the underlying recordings.

Meeting length averages 16.1 minutes of active speech ( $\sim 2,900$  spoken words per meeting at  $\sim 180$  wpm). Unlike popular meeting corpora (AMI, ICSI, MeetingBank), whose transcripts may have entered the pretraining data of modern LLMs, OSYNTH was newly generated for this work. Also, the pipeline can be re-invoked at low marginal cost to produce additional workspaces, so fresh held-out splits can be created as evaluator models continue to evolve.

### 3.4 Adversarial Transcript Refinement

LLM-generated transcripts possess a characteristic set of structural features that distinguishes them from real transcripts such as clean narrative arcs, perfect turn taking, and few dropped threads. To make the canonical transcripts realistic, we apply an adversarial structural realism pass (§3.2). For this, we use AMI (Carletta et al., 2006) excerpts as style references, and a multi-LLM-family loop pairing a gemini-3.1-pro-preview discriminator with a claude-opus-4-7 refiner that revises the transcript for up to five iterations. The discriminator is presented with a pairwise A/B comparison between the candidate and a length-matched AMI excerpt drawn from a held-out AMI split, with speaker labels swapped with generic identifiers (*Speaker 1*, *Speaker 2*, ...), so that it cannot anchor on transcript length or speaker names; this adapts the self-refinement structure of Madaan et al. (2023) by replacing self-feedback with a cross-family adversarial judge. We split the loop across

| Synthetic source        | Trials | Accuracy |
|-------------------------|--------|----------|
| Haiku (w/o obfuscation) | 54     | 79.6%    |
| OSYNTH                  | 378    | 56.3%    |
| Chance baseline         | —      | 50%      |

Table 2: Human discrimination accuracy in the 2AFC realism study. The OSYNTH condition is relatively closer to chance, supporting OSYNTH as a more credible proxy for real workspace meetings.

model families because in preliminary runs a same-family judge was fooled after a single iteration even when obvious structural tells remained; and we cap iterations in the interest of runtime rather than convergence, since test runs with  $N = 10$  continued to land accepted edits past iteration five. Appendix F gives the full procedure.

### 3.5 Realism: human discrimination study

To verify that OSYNTH is a credible proxy for real workspace meetings, we run a two-alternative forced-choice (2AFC) human-discrimination study. This approach was previously used to assess the realism of synthetic speech (Mai et al., 2023). Raters are shown one real excerpt (from AMI (Carletta et al., 2006) or NOTSOFAR-1 (Vinnikov et al., 2024)) and one synthetic excerpt (from either OSYNTH or a baseline Haiku w/o obfuscation control) and asked to identify the synthetic side. We retain only sessions that pass two attention checks; Appendix S describes the annotation interface, and Appendix G the protocol details together with a complementary full-transcript LLM-judge probe.

Among the final set of annotators (24 annotators, 432 non-attention trials), Haiku-w/o-obfuscation fakes were identified reliably: 43/54 (79.6%) were classified as LLM-generated. OSYNTH fakes were substantially harder: 213/378 (56.3%) were classified correctly. The gap is large (+23.3 pp; Fisher two-sided  $p = 0.001$ ), supporting the interpretation that annotators can catch obvious synthetic artifacts while realistic OSYNTH transcripts lack comparably obvious tells. We note that 56.3% is above chance: a naive binomial test against 50% gives  $p = 0.016$ , so OSYNTH excerpts are still detectable with above chance accuracy rather than indistinguishable. Table 2 reports annotator discrimination accuracy.

## 4 Context-sensitive evaluation

We use OSYNTH to evaluate two state-of-the-art, multimodal large language model (MLLM) fami-

| System | Ctx. | WER ↓             | Num. rec. ↑       | Term rec. ↑       | AI desc. ↑      | Assignee ↑        | Spk. cons. ↑      |
|--------|------|-------------------|-------------------|-------------------|-----------------|-------------------|-------------------|
| OpenAI | L1   | 0.162±.014        | 0.680±.056        | 0.684±.040        | 3.49±.06        | 0.485±.040        | 0.220±.045        |
|        | L2   | 0.155±.015        | 0.695±.056        | 0.708±.040        | 3.60±.06        | 0.504±.040        | 0.219±.048        |
|        | L3   | <b>0.144±.015</b> | <b>0.883±.038</b> | <b>0.778±.036</b> | <b>3.96±.05</b> | <b>0.690±.035</b> | <b>0.476±.067</b> |
| Gemini | L1   | 0.099±.015        | 0.960±.022        | 0.809±.031        | 3.55±.05        | 0.069±.024        | 0.265±.050        |
|        | L2   | 0.099±.015        | 0.963±.021        | 0.796±.032        | 3.60±.05        | 0.071±.023        | 0.254±.049        |
|        | L3   | <b>0.095±.015</b> | <b>0.975±.017</b> | <b>0.841±.028</b> | <b>3.80±.06</b> | <b>0.630±.040</b> | <b>0.349±.068</b> |

Table 3: Main results across context levels L1 to L3 on the evaluation set. Values are means with 95% bootstrap percentile interval (1,000 resamples). Bold marks the best context level for each system. Both systems improve mostly monotonically on the entity-sensitive dimensions (number recall, technical-term recall, assignee accuracy, speaker consistency). The OpenAI stack uses gpt-4o-transcribe-diarize, gpt-audio, and gpt-5.5; the Gemini stack uses gemini-3-flash and gemini-3.1-pro-preview.

| Metric                 | Type | L1         | L3         | Δ     |
|------------------------|------|------------|------------|-------|
| Number recall          | iso. | 0.680±.056 | 0.883±.038 | +0.20 |
| Technical-term rec.    | iso. | 0.684±.040 | 0.778±.036 | +0.09 |
| AI assignee accuracy   | iso. | 0.485±.040 | 0.690±.035 | +0.20 |
| Spk.-consistency rate  | iso. | 0.220±.045 | 0.476±.067 | +0.26 |
| LLM-judge faithfulness | std. | 3.10±.09   | 3.40±.10   | +0.30 |
| LLM-judge coherence    | std. | 4.60±.06   | 4.91±.04   | +0.31 |
| AI description quality | std. | 3.49±.06   | 3.96±.05   | +0.47 |

Table 4: L1 to L3 gain on isomorphic latent-state metrics versus standard LLM-as-judge summarization metrics for the OpenAI system on the evaluation set including 95% bootstrap percentile intervals (1,000 resamples). The **Type** column marks each row as either *iso.* (isomorphic, scored by exact comparison against OSYNTH’s latent state) or *std.* (standard meeting-summarization metric, scored by LLM-as-judge over the produced artifact). Gains concentrate on the entity-sensitive isomorphic dimensions (assignee accuracy, number recall, speaker consistency), which standard summary-side LLM-as-judge scores systematically compress.

lies as meeting understanding systems (§4.1) across three progressively richer levels of workspace context (§4.2). Metrics cover transcription, diarization and speaker identification, and downstream summarization artifacts (§4.3), split into *isomorphic* metrics scored against OSYNTH’s latent state and *standard* meeting-summarization metrics. We therefore treat meeting summarization as the concrete output surface through which broader meeting understanding can be evaluated. All numbers reported in §5 are computed on a held-out evaluation set of 238 meetings drawn from a sample of 15 workspaces; each meeting is scored in each of the three context settings for each system. Point estimates are reported with 95% bootstrapped confidence intervals (1,000 resamples).

#### 4.1 Systems evaluated

These systems are chosen to reflect the current state of the art in transcription, diarization, speaker iden-

tification, and summarization. For reproducibility, Appendix V additionally reports a fully open-source SpeechBrain-based cascade on a subset of the same evaluation set.

**OpenAI.** A single-vendor cascade built from OpenAI’s current API surface: gpt-4o-transcribe-diarize (OpenAI, 2026a) produces speaker-attributed transcripts, gpt-audio (OpenAI, 2026d) performs enrollment-based speaker identification, and gpt-5.5 (OpenAI, 2026c) performs jargon post-correction plus prompted summary, action-item, and outline generation.

**Gemini.** A stage-for-stage mirror of the OpenAI pipeline via Vertex AI. Diarized transcription, jargon post-correction, and prompted summary / action-item / outline generation are performed by gemini-3-flash (Google DeepMind, 2025), and enrollment-based speaker identification is performed by gemini-3.1-pro-preview (Google DeepMind, 2026), which we use for the SID stage because of its stronger audio-grounded reasoning on short enrollment clips.

#### 4.2 Context levels

We define three context levels, constructed from OSYNTH’s latent state, that progressively expand the workspace signal available to a system. Each level is a strict superset of the previous one, so improvement curves across L1–L3 are interpretable as marginal gains from additional context rather than from a change of input modality.

**L1 — No context.** The system sees only the meeting as raw audio. This is the standard meeting-summarization setting reproduced from prior benchmarks (Carletta et al., 2006; Janin et al., 2003; Hu et al., 2023).

**L2 — Prior-meeting outputs.** L1 plus the system’s own outputs (transcript, summary, action items, outline) for prior meetings in the same workspace, presented as in-context history. L2 isolates the contribution of approximated, longitudinal meeting context without introducing any external workspace signal.

**L3 — Workspace artifacts.** L2 plus the workspace artifacts that surround a meeting: per-speaker voice enrollments and name lists, slide decks tied to the meeting agenda, and inter-meeting emails referencing relevant projects and people. L3 is the deployed-system setting that motivates this work, and the level at which a context-aware system should outperform a context-free one.

Context is encoded as additional prompt fields or model inputs depending on the system architecture; speaker-identity voice enrollments are routed to the speaker-identification stage when a system supports them.

### 4.3 Metrics

We compute metrics in four families: transcription quality, contextualized transcription, diarization and speaker identification, and downstream summarization artifacts. LLM-as-judge metrics use a fixed judge model (claude-sonnet-4-6 (Anthropic, 2026b)). Because this is the same model family used to generate the corpus, we repeat the judged metrics with an open-weights judge in Appendix W; the L1→L3 improvement is preserved, while the absolute range and calibration of the scores is judge-dependent. We split metrics into *isomorphic* metrics, which are deterministic functions of OSYNTH’s latent state, and *standard* reference-free LLM-as-judge metrics. OSYNTH’s latent state makes it possible to score entity-sensitive ASR and summarization behavior that existing meeting benchmarks cannot expose: attendee names and nicknames, project and product names, numbers and dates, technical terms, named-speaker attribution, action-item assignees, and longitudinal entity tracking.

We define each metric reported in Tables 3 and 4 below; Appendix H gives the corresponding scoring details. **Word error rate** (WER) is computed against the semantic-verbatim transcript (§3.2), which matches the synthesized audio rather than the cleaner canonical written form; both reference and hypothesis are passed through Whisper’s EnglishTextNormalizer (Radford et al., 2023)

before scoring. **Number recall** and **technical-term recall** (iso.) score the fraction of oracle-transcript numbers and dates, and of workspace-jargon terms (acronyms, product names, domain-specific terminology), recovered in the hypothesis; reference term sets are extracted directly from the latent state and intersected with the oracle transcript. **Cross-meeting speaker-identity consistency** (iso.) is the rate at which a workspace persona is mapped to the same predicted name across meetings within the workspace, exposing labeling-stability failures that a per-meeting attribution score misses. **Action-item assignee accuracy** (iso.) scores whether each predicted action item is assigned to the correct persona, under partial-credit substring matching against the canonical assignee name for OpenAI (Table 4) and exact match for Gemini (Appendix U). **Action-item description quality** (std.; “AI desc.” in Table 3) is an LLM-judge 1–5 rating of how completely and concretely each predicted action-item description captures the underlying commitment, scored against the gold action-item list derived from the latent state. **LLM-judge faithfulness** and **coherence** (std.) score per-meeting summaries on the corresponding 1–5 dimensions following Liu et al. (2023a) and Zheng et al. (2023) against the gold summary; see Appendix I for details.

## 5 Results

We report per-system results across the three context levels L1–L3 on the metric families defined in §4.3, computed on the 238-meeting evaluation set. Table 3 summarizes the main results; Table 4 compares the context gains visible through isomorphic metrics and standard LLM-as-judge metrics. The open-source SpeechBrain cascade is reported separately in Appendix V. Two robustness checks accompany these results: the judged metrics recomputed with an open-weights judge (Appendix W), and an anchoring study on real AMI meetings (§6).

### 5.1 Trends across context levels

For the OpenAI system, summary-side artifact quality and contextualized-transcription recall improve monotonically from L1 to L3 (Table 3). Gains concentrate on the contextualized-transcription and entity-sensitive summarization dimensions: number recall (+0.20 absolute, L1→L3), technical-term recall (+0.09), action-item description quality (+0.47 points on a 1–5 scale), action-item as-

signee accuracy (+0.20, scored under partial-credit substring matching), and cross-meeting speaker-identity consistency (+0.26). Concretely, the +0.20 number-recall jump is driven by recoveries such as L1’s “PAX *eight*” for the ISO 20022 identifier “pacs.008” (anchored at L3 by the workspace project document); a side-by-side excerpt is given in Appendix J. The Gemini stack shows the same broad pattern (Table 3), with the largest gains again on entity-sensitive dimensions: assignee accuracy, number recall, and technical-term recall all rise from L1 to L3. The assignee jump is expected because L3 adds per-speaker voice enrollments, turning assignee recovery from an open-ended name-generation problem (where names must be inferred from contextual mentions) into a closed-set lookup against workspace identities. On the OpenAI side, the +0.20 assignee gain is concretely illustrated by the L1→L3 trajectory *Bevan* → *Devin* → *Devon Ashby* (an invented name at L1, a different invented name at L2, the canonical workspace identity only at L3); the full block is in Appendix J. Overall WER improves only modestly with context (OpenAI L1→L3: 0.162 → 0.144; Gemini: 0.099 → 0.095), and in both cases the bootstrap 95% confidence intervals for L1 and L3 overlap. The entity-sensitive dimensions move several times as far over the same increment, consistent with residual errors routing primarily through entity-level surface forms rather than general word recognition, which is precisely where L3 yields the largest gains.

## 5.2 Isomorphic versus standard metrics

Isomorphic latent-state metrics and standard summarization metrics tell partially different stories (Table 4; Gemini breakdown, on the earlier 50-meeting subset, in Appendix U). Both metric families agree that L3 improves over L1 on the main entity-sensitive dimensions, but LLM-as-judge summary-side scores compress the gains visible in the isomorphic metrics. The Gemini stack has one non-monotonic standard cell: LLM-judge faithfulness dips slightly at L3 despite uniform improvement on every isomorphic dimension, reflecting the known tendency of reference-free protocols to underweight factual precision and over-penalize informative workspace-canonical expansions (Tang et al., 2024).

## 5.3 Which workspace artifacts matter?

An OpenAI-stack ablation separates the L3 increment into per-speaker voice *enrollments* and *workspace text* artifacts (member lists, emails, slides, and agendas). Workspace text explains almost all of the number-recall gain, while assignee accuracy benefits from both streams: enrollments anchor diarization clusters to canonical identities, and workspace text supplies the names and entities the summarizer writes into action items. Appendix D gives the full table and discussion.

Appendix E further analyzes OpenAI L1 errors using a six-way ASR taxonomy. The dominant categories are named entities, numbers/dates, and technical terms; L3 mainly helps the entity-sensitive cases, while filler, function-word, and lexical reanalysis errors remain largely orthogonal to workspace context.

## 6 Anchoring on real meetings

OSYNTH’s isomorphic metrics require a ground-truth latent state, which real corpora do not provide, so they cannot be transferred to recorded meetings. The judged summary metrics can be, which lets us check the direction of the context effect on real data. We randomly selected two meeting series from the ES subset of AMI (Carletta et al., 2006), ES2002 and ES2003, of four meetings each, and constructed a corresponding workspace state by aggregating the AMI metadata (an anonymous identity map held across each series), the gold transcripts, and the human-written abstractive summaries. We ran the Gemini system in the L1 and L3 settings, using for L3 the system’s own outputs on prior meetings in the series together with the meeting title, and scored the resulting summaries with the judged summary metrics (Table 5). Faithfulness improves by +0.13; coherence is at the ceiling in both settings. The direction agrees with the main result, which we take as convergent validity for the context effect rather than as a second measurement of its size.

A major caveat of this study is the limited size, but it further supports the validity of our results by providing convergent evidence on real, recorded audio.

## 7 Discussion

Both state-of-the-art systems show monotonic improvement on the entity-sensitive isomorphic dimensions as workspace context is added, with

| Metric                 | L1   | L3   | $\Delta$ |
|------------------------|------|------|----------|
| LLM-judge faithfulness | 3.50 | 3.63 | +0.13    |
| LLM-judge coherence    | 5.00 | 5.00 | 0.00     |

Table 5: Anchoring study on eight real AMI meetings (series ES2002 and ES2003), Gemini system, judged summary metrics only.

the largest gains on action-item assignee accuracy, number recall, technical-term recall, and cross-meeting speaker-identity consistency. Notably, these are all dimensions that prior meeting benchmarks cannot easily expose due to limited context, anonymization of named entities, and lack of the extensive oracle latents we use for evaluation. These gains do not mostly come from better acoustic recognition. They come from workspace-specific evidence: voice enrollments map diarization clusters to people, while workspace names and inter-meeting emails help resolve the entities that should appear in summaries. As a result, benchmarks that evaluate only isolated meetings, for instance, with no workspace context, understate the quality that a deployed workspace-aware system can achieve. The ablation in §5.3 further shows that the two L3 streams (enrollments and workspace text) are complementary rather than redundant.

This pattern bears directly on a long-running critique that WER “does not reflect text understanding and only weakly correlates with human judgement of a transcript’s quality” (as summarized by Kuhn et al., 2023), originally argued by Wang et al. (2003), Mishra et al. (2011), and Favre et al. (2013). By weighting every token equally, WER absorbs context-driven gains into the long tail of low-value words, even when, as in our setting, those gains concentrate on the entities that determine the usefulness of a transcript or downstream summary. The isomorphic, entity-grounded metrics enabled by OSYNTH’s latent state expose precisely the dimensions WER does not expose. Thus, the L1 to L3 measurement scales from a modest, aggregate WER improvement into a substantially larger, quantifiable improvement on the other measurements that we have considered.

## 8 Conclusion

We present OSYNTH, a longitudinal synthetic meeting corpus with aligned workspace artifacts and an exact latent-state reference, and use it to evaluate state-of-the-art meeting notetaking systems across

progressively richer context levels. Human discrimination results support OSYNTH as a relatively credible proxy for real meeting transcripts as compared to naive LLM generation, and the evaluation shows that workspace context improves summary artifacts and contextualized transcription on entity-sensitive dimensions that prior benchmarks cannot equivalently expose.

## Limitations

OSYNTH is a fully synthetic corpus and therefore cannot fully reproduce the long-tail acoustic, sociolinguistic, and interactional phenomena present in recorded workspace meetings; we make no claim that results reported here transfer numerically to natural recordings. We make this choice deliberately: a fully synthetic pipeline is what unlocks the longitudinal, multi-artifact workspace setting at releasable scale and with an exact latent-state reference. Scaling our AMI grounding study to more convergent results on natural spoken meeting data is a worthwhile point for future investigation.

Additionally, the released TTS audio is rendered with a single multi-speaker model and a single voice-cloning prompt per persona; even when the cloning prompt is drawn from spontaneous conversational speech, the resulting per-turn renderings can still exhibit a perceptibly read-aloud quality (i.e., steady prosody) relative to fully spontaneous speech, and downstream ASR and diarization numbers should therefore be read as relative comparisons within OSYNTH, not as absolute estimates of performance on natural recordings.

The reference clips themselves are drawn from the Mozilla Common Voice *Spontaneous Speech* corpus rather than the scripted releases, so the read-aloud quality is introduced downstream: measured against the cloning prompts, OSYNTH turns are paced more uniformly (Appendix N). A second limitation of the artifacts is that they are clean by construction: workspace emails, slides, and documents contain no typos or transcription noise beyond the ASR output itself, and injecting such noise is left to future work.

Finally, the evaluation in §4 fixes a single prompt template per task and a single decoding configuration per system; absolute scores will shift under heavier prompt engineering. We furthermore ensure diversity only via the workspace random initialization which would become a limitation as dataset size is scaled up.

## Ethical considerations

OSYNTH is generated end-to-end by an LLM-based pipeline. All personas, organizations, project names, and email addresses are sampled from generated pools rather than drawn from real entities, and TTS voices are cloned from a curated subset of CommonVoice (Ardila et al., 2020) whose speakers have consented to broad reuse. Because these names are generated without reference to real people or companies, any resemblance a generated name bears to an actual person or organization is coincidental and unintended; references to public standards and to the bodies that maintain them appear only as background domain context.

That being said, the CommonVoice terms of use prohibit any attempt to re-identify the speakers behind the released voice clips, and we make no such attempt at any stage of the OSYNTH pipeline: cloning prompts are used solely as anonymous acoustic conditioning for the TTS model, are stripped of all CommonVoice metadata before use, and are paired with synthetic personas that bear no relation to the underlying contributor. Downstream users of the released audio inherit the same condition and should not attempt to link voices to real individuals.

Because the corpus is synthetic, it may encode biases of the generating LLM in persona attributes, dialogue style, and decision dynamics. The corpus is intended for evaluation of meeting-summarization and notetaking systems; it should not be used as training data for systems whose claims of generalization rest on naturalistic conversational behavior.

Finally, an LLM-based coding assistant was used to copy-edit and copy-check the final draft of this manuscript.

## Acknowledgments

We thank Sam Liang and Yun Fu for their support of this work, the annotators who contributed to the human-discrimination study, and the anonymous reviewers for their constructive feedback.

## References

Chenxin An, Shansan Gong, Ming Zhong, Xingjian Zhao, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. 2024. [L-Eval: Instituting standardized evaluation for long context language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long*

*Papers)*, pages 14388–14411, Bangkok, Thailand. Association for Computational Linguistics.

Anthropic. 2026a. System card: Claude Opus 4.7. <https://anthropic.com/claude-opus-4-7-system-card>.

Anthropic. 2026b. System card: Claude Sonnet 4.6. <https://anthropic.com/claude-sonnet-4-6-system-card>.

Riku Arakawa, Hiromu Yakura, Kei Akuzawa, and Shizuma Kubo. 2025. [AI for meeting minutes: Promises and challenges in designing human-AI collaboration on a production SaaS platform](#). In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*, pages 1–6, New York, NY, USA. Association for Computing Machinery.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2020. Common Voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC)*, pages 4218–4222, Marseille, France. European Language Resources Association.

Sumit Asthana, Sagih Hilleli, Pengcheng He, and Aaron Halfaker. 2025. [Summaries, highlights, and action items: Design, implementation and evaluation of an LLM-powered meeting recap system](#). *Proceedings of the ACM on Human-Computer Interaction (CSCW)*, 9(CSCW1).

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024. [LongBench: A bilingual, multi-task benchmark for long context understanding](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3119–3137, Bangkok, Thailand. Association for Computational Linguistics.

Jeong-Uk Bang, Seung Yun, Seung-Hi Kim, Mu-Yeol Choi, Min-Kyu Lee, Yeo-Jeong Kim, Dong-Hyun Kim, Jun Park, Young-Jik Lee, and Sang-Hun Kim. 2020. [KsponSpeech: Korean spontaneous speech corpus for automatic speech recognition](#). *Applied Sciences*, 10(19):6936.

Jon Barker, Shinji Watanabe, Emmanuel Vincent, and Jan Trmal. 2018. [The fifth 'CHiME' speech separation and recognition challenge: Dataset, task and baselines](#). In *Proceedings of Interspeech 2018*, pages 1561–1565.

Arnaldo Cândido Junior, Edresson Casanova, Anderson Soares, Frederico Santos de Oliveira, Lucas Oliveira, Ricardo Corso Fernandes Junior, Daniel Peixoto Pinto da Silva, Fernando Gorgulho Fayet, Bruno Baldissera Carlotto, Lucas Rafael Stefanel Gris, and Sandra Maria Aluísio. 2023. [CORAA ASR: a large corpus of spontaneous and prepared speech](#)

- manually validated for speech recognition in Brazilian Portuguese. *Language Resources and Evaluation*, 57(3):1139–1171.
- Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner. 2006. *The AMI meeting corpus: A pre-announcement*. In *Machine Learning for Multimodal Interaction: Second International Workshop, MLMI 2005*, volume 3869 of *Lecture Notes in Computer Science*, pages 28–39. Springer.
- Maximillian Chen, Alexandros Papangelis, Chenyang Tao, Seokhwan Kim, Andy Rosenbaum, Yang Liu, Zhou Yu, and Dilek Hakkani-Tur. 2023. *PLACES: Prompting language models for social conversation synthesis*. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 844–868, Dubrovnik, Croatia. Association for Computational Linguistics.
- Samuele Cornell, Matthew Wiesner, Shinji Watanabe, Desh Raj, Xuankai Chang, Paola Garcia, Yoshiki Masuyama, Zhong-Qiu Wang, Stefano Squartini, and Sanjeev Khudanpur. 2023. *The CHiME-7 DASR challenge: Distant meeting transcription with multiple devices in diverse scenarios*. In *Proceedings of the 7th International Workshop on Speech Processing in Everyday Environments (CHiME)*, pages 1–6.
- Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. *Enhancing chat language models by scaling high-quality instructional conversations*. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3029–3051, Singapore. Association for Computational Linguistics.
- Benoit Favre, Kyla Cheung, Siavash Kazemian, Adam Lee, Yang Liu, Cosmin Munteanu, Ani Nenkova, Dennis Ochei, Gerald Penn, Stephen Tratz, Clare Voss, and Frauke Zeller. 2013. Automatic human utility evaluation of ASR systems: Does WER really predict performance? In *Proceedings of INTERSPEECH 2013*, pages 3463–3467.
- Tirthankar Ghosal, Ondřej Bojar, Marie Hledíková, Tom Kocmi, and Anna Nedoluzhko. 2023. *Overview of the second shared task on automatic minuting (AutoMin) at INLG 2023*. In *Proceedings of the 16th International Natural Language Generation Conference: Generation Challenges*, pages 138–167, Prague, Czechia. Association for Computational Linguistics.
- Tirthankar Ghosal, Ondřej Bojar, Muskaan Singh, and Anja Nedoluzhko. 2021. *Overview of the first shared task on automatic minuting (AutoMin) at Interspeech 2021*. In *First Shared Task on Automatic Minuting at Interspeech 2021*, pages 1–25.
- Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. *SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization*. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.
- Google DeepMind. 2025. Gemini 3 Flash model card. <https://deepmind.google/models/model-cards/gemini-3-flash/>.
- Google DeepMind. 2026. Gemini 3.1 Pro model card. <https://deepmind.google/models/model-cards/gemini-3-1-pro/>.
- Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiaowu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. 2024. MetaGPT: Meta programming for a multi-agent collaborative framework. In *The Twelfth International Conference on Learning Representations (ICLR)*.
- Yebowen Hu, Timothy Ganter, Hanieh Deilamsalehy, Franck Dernoncourt, Hassan Foroosh, and Fei Liu. 2023. *MeetingBank: A benchmark dataset for meeting summarization*. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16409–16423, Toronto, Canada. Association for Computational Linguistics.
- Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, and Chuck Wooters. 2003. *The ICSI meeting corpus*. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages I-364–I-367, Hong Kong.
- Hyunwoo Kim, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Le Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, and Yejin Choi. 2023. *SODA: Million-scale dialogue distillation with social commonsense contextualization*. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 12930–12949, Singapore. Association for Computational Linguistics.
- Korbinian Kuhn, Verena Kersken, Benedikt Reuter, Niklas Egger, and Gottfried Zimmermann. 2023. *Measuring the accuracy of automatic speech recognition solutions*. *ACM Transactions on Accessible Computing*, 16(4):Article 25.
- Md Tahmid Rahman Laskar, Xue-Yong Fu, Cheng Chen, and Shashi Bhushan TN. 2023. *Building real-world meeting summarization systems using large language models: A practical perspective*. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 343–352, Singapore. Association for Computational Linguistics.

- Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. CAMEL: Communicative agents for “mind” exploration of large language model society. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023a. [G-Eval: NLG evaluation using GPT-4 with better human alignment](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2511–2522, Singapore. Association for Computational Linguistics.
- Yixin Liu, Alex Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir Radev. 2023b. Revisiting the gold standard: Grounding summarization evaluation with robust human evaluation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4140–4170.
- Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, pages 46534–46594.
- Kimberly T. Mai, Sergi Bray, Toby Davies, and Lewis D. Griffin. 2023. [Warning: Humans cannot reliably detect speech deepfakes](#). *PLOS ONE*, 18(8):e0285333.
- Taniya Mishra, Andrej Ljolje, and Mazin Gilbert. 2011. [Predicting human perceived accuracy of ASR systems](#). In *Proceedings of INTERSPEECH 2011*, pages 1945–1948.
- Anna Nedoluzhko, Muskaan Singh, Marie Hledíková, Tirthankar Ghosal, and Ondřej Bojar. 2022. ELITR minuting corpus: A novel dataset for automatic minuting from multi-party meetings in English and Czech. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC)*, pages 3174–3182, Marseille, France. European Language Resources Association.
- Lisa R. O’Bryan, Madeline Navarro, Juan Segundo Hevia, and Santiago Segarra. 2025. ML-SPEAK: A theory-guided machine learning method for studying and predicting conversational turn-taking patterns. *Journal of Personality and Social Psychology*, 129(6):1257.
- OpenAI. 2026a. GPT-4o Transcribe Diarize model. <https://developers.openai.com/api/docs/models/gpt-4o-transcribe-diarize>. OpenAI API documentation, accessed August 2026.
- OpenAI. 2026b. GPT-5.4 Thinking system card. <https://openai.com/index/gpt-5-4-thinking-system-card/>.
- OpenAI. 2026c. GPT-5.5 system card. <https://openai.com/index/gpt-5-5-system-card/>.
- OpenAI. 2026d. GPT-Audio model. <https://developers.openai.com/api/docs/models/gpt-audio>. OpenAI API documentation, accessed August 2026.
- Joon Sung Park, Joseph C. O’Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. [Generative agents: Interactive simulators of human behavior](#). In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST ’23)*, New York, NY, USA. Association for Computing Machinery.
- Qwen Team. 2026. Qwen3-TTS: Multilingual, multi-speaker text-to-speech. Model release, Alibaba Cloud. <https://github.com/QwenLM/Qwen3-TTS>.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, pages 28492–28518.
- Revanth Rameshkumar and Peter Bailey. 2020. [Storytelling with dialogue: A Critical Role Dungeons and Dragons dataset](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5121–5134, Online. Association for Computational Linguistics.
- Elena Rastorgueva, Vitaly Lavrukhin, and Boris Ginsburg. 2023. NeMo forced aligner and its application to word alignment for subtitle generation. In *Proceedings of Interspeech 2023*, pages 5257–5258.
- Robin Scheibler, Eric Bezzam, and Ivan Dokmanić. 2018. Pyroomacoustics: A Python package for audio room simulation and array processing algorithms. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 351–355. IEEE.
- Kartik Shinde, Laurent Besacier, Ondřej Bojar, Thibaut Thonet, and Tirthankar Ghosal. 2025. [Findings of the third automatic minuting \(AutoMin\) challenge](#). Preprint, arXiv:2509.13814.
- Liyan Tang, Igor Shalyminov, Amy Wing-mei Wong, Jon Burnsky, Jake W. Vincent, Yu’an Yang, Siffi Singh, Song Feng, Hwanjun Song, Hang Su, Lijia Sun, Yi Zhang, Saab Mansour, and Kathleen Mckeown. 2024. [TofuEval: Evaluating hallucinations of LLMs on topic-focused dialogue summarization](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4455–4480, Mexico City, Mexico. Association for Computational Linguistics.

Nicholas Tomlin, Naitian Zhou, Eve Fleisig, Liangyuan Chen, Téa Wright, Lauren Vinh, Laura X. Ma, Seun Eisape, Ellie French, Tingting Du, Tianjiao Zhang, Alexander Koller, and Alane Suhr. 2025. [Characterizing language use in a collaborative situated game](#). arXiv preprint arXiv:2512.03381. *Preprint*, arXiv:2512.03381.

Alon Vinnikov, Amir Ivry, Aviv Hurvitz, Igor Abramovski, Sharon Koubi, Ilya Gurvich, Shai Peer, Xiong Xiao, Benjamin Martinez Elizalde, Naoyuki Kanda, Xiaofei Wang, Shalev Shaer, Stav Yagev, Yossi Asher, Sunit Sivasankaran, Yifan Gong, Min Tang, Huaming Wang, and Eyal Krupka. 2024. [NOTSOFAR-1 challenge: New datasets, baseline, and tasks for distant meeting transcription](#). In *Proceedings of Interspeech 2024*, pages 5003–5007.

Ye-Yi Wang, Alex Acero, and Ciprian Chelba. 2003. [Is word error rate a good indicator for spoken language understanding accuracy](#). In *2003 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 577–582.

Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. 2023. [Baize: An open-source chat model with parameter-efficient tuning on self-chat data](#). *Preprint*, arXiv:2304.01196.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.

Fan Yu, Shiliang Zhang, Yihui Fu, Lei Xie, Siqi Zheng, Zhihao Du, Weilong Huang, Pengcheng Guo, Zhijie Yan, Bin Ma, Xin Xu, and Hui Bu. 2022. [M2MeT: The ICASSP 2022 multi-channel multi-party meeting transcription challenge](#). In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6167–6171. IEEE.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhonghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023) Datasets and Benchmarks Track*, pages 46595–46623.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. 2021. [QMSum: A new benchmark for query-based multi-domain meeting summarization](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 5905–5921, Online. Association for Computational Linguistics.

Chenguang Zhu, Yang Liu, Jie Mei, and Michael Zeng. 2021. [MediaSum: A large-scale media interview dataset for dialogue summarization](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 5927–5934, Online. Association for Computational Linguistics.

## A Corpus statistics

|                                     |       |
|-------------------------------------|-------|
| Workspaces                          | 160   |
| Meetings                            | 2,126 |
| Hours of audio                      | 570   |
| Meetings / workspace (mean)         | 13.3  |
| Personas / workspace (mean)         | 12.9  |
| Email threads / workspace (mean)    | 4.3   |
| Shared documents / workspace (mean) | 1.5   |

Table 6: OSYNTH corpus statistics. Each workspace bundles meeting audio, aligned transcripts, downstream notetaking artifacts, inter-meeting workspace artifacts, and per-speaker enrollments derived from a shared latent state. Workspaces are generated with varying workspace size and time horizons, thus subsamples of the data follow distinct distributions in terms of length and content.

## B Workspace walkthrough example

To give a concrete picture of how a single OSYNTH workspace appears across personas, projects, meetings, and artifacts, we walk through one representative workspace: a synthetic Series-C fintech mid-market organization. The organization is observed over a two-week window.

**Org chart and personas.** The active org chart contains ten speaking personas plus two external contacts (Charlotte Bewick of Brindlewood; Tomás Herrera of Nexbridge Capital). Reporting lines are: Renata Sokolowski (CEO, p1) over Marcus Delacroix (VP Eng, p2; reports: Simone Okafor, Rafael Quintero, Devon Ashby), Priya Nambiar (Dir. Product, p3; report: Alejandro Vásquez), Dillon Hargrove (VP Sales, p4; reports: Adriana Castellanos, Everett Cho), and Yuki Tanaka (Dir. Ops/Compliance, p5; report: Devansh Pillai). Two personas (*Marcus*, *Dillon*) carry an explicit nickname policy (*Marc*, *Dill*; Appendix T), which the transcript LLM uses to vary how other speakers refer to them.

**Active projects.** Three projects are simultaneously live in this window: (i) *ISO 20022 Payment Message Migration* (owner: Yuki Tanaka; blocked

| Date        | Title                               | Dur | Att |
|-------------|-------------------------------------|-----|-----|
| 05-18 09:00 | Executive Leadership Sync           | 60  | 5   |
| 05-18 09:15 | Engineering Daily Standup           | 15  | 4   |
| 05-18 11:00 | Project Brindle Kickoff             | 45  | 4   |
| 05-19 09:15 | Engineering Daily Standup           | 15  | 4   |
| 05-19 10:00 | <b>Marcus–Simone 1:1</b> (★)        | 30  | 2   |
| 05-19 10:00 | ISO 20022 Migration Sync            | 60  | 5   |
| 05-19 10:30 | Marcus–Rafael 1:1                   | 30  | 2   |
| 05-19 11:00 | Marcus–Devon 1:1                    | 30  | 2   |
| 05-20 09:00 | SOC 2 Package Prep Session          | 60  | 2   |
| 05-20 14:00 | Orchestrate Failover Review         | 60  | 5   |
| 05-20 15:00 | CAS Locking Strategy Review         | 90  | 3   |
| 05-21 10:00 | BDW Sandbox Scope (Devon & Adriana) | 30  | 2   |

Table 7: Three-day slice of the example workspace meeting calendar around the focus meeting (marked with ★). *Dur* is scheduled duration in minutes; *Att* is attendee count. The full workspace contains 28 meetings across the two-week observation window.

on Payments Canada draft fields; target 2026-07-31), (ii) *Clearflow Orchestrate Failover Hardening* (owner: Simone Okafor; blocked on locking strategy; target 2026-06-19), and (iii) *Enterprise Prospect Acceleration — Brindlewood Commerce Group* (owner: Dillon Hargrove; SOC 2 and BDW-sandbox prerequisites; target 2026-05-31). Each project carries its own document set, calendar entries, and WORKSPACEEVENT stream (§3.1).

**Meeting calendar.** The workspace contains 28 meetings across the two-week window, spanning weekly executive syncs, daily standups, recurring 1:1s, and ad-hoc design and project reviews. Table 7 shows a representative slice of the first three days centered on the focus meeting used in the artifact excerpts below.

**Representative artifacts.** The focus meeting (*Marcus ↔ Simone 1:1*, 2026-05-19 10:00) is a 30-minute recurring 1:1 between the VP of Engineering and a senior engineer.

(1) *Transcript excerpt (canonical, 6 turns).*

**Marcus:** Hey. Failover.  
**Simone:** Race condition’s isolated. It’s the locking now — registry, concurrent writes.  
**Marcus:** Right.  
**Simone:** Global lock’s out, P99 was like 120.  
**Simone:** Devon’s on CAS. The harness doesn’t replay reconnects, so we don’t actually know what happens to in-flight transitions when a node drops.  
**Marcus:** When’s he have something reviewable? Not merged.  
**Simone:** End of week realistically. Brindlewood’s pulling on him.

(2) *Gold summary (1 paragraph).*

Marcus and Simone covered the status of the Orchestrate Failover Hardening project, which remains blocked pending Devon’s lock-free CAS prototype review; Simone committed to syncing with Devon before end of week and flagging Marcus by Thursday if the approach doesn’t look viable. They briefly touched on the ISO 20022 SettleMap work, which is in a holding pattern on three camt.054 sub-fields awaiting Payments Canada guidance, and discussed Devon’s split availability between the CAS prototype and the Brindlewood sandbox environment.

(3) *Gold action items (selected).*

**[Marcus Delacroix]** Handle Devon’s competing priorities with Brindlewood to keep him focused on CAS work (due before EOW).

**[Simone Okafor]** Look into whether SettleMap mappings that don’t touch camt.054 can be sequenced and progressed independently despite XSD cross-references.

**[Simone Okafor]** Check the Thursday meeting invite and confirm whether Karol moved it to 2 or 2:30, then tell Hwang whether attending remote or in-room.

(4) *Calendar invite (subject and body).*

**Subject:** Marcus ↔ Simone 1:1

**When:** Tuesday, 2026-05-19, 10:00–10:30 ET (30 min)

**Where:** Conference Room 1 (hybrid: Marcus in-room, Simone remote)

**Body:** Recurring 1:1. Touch base on Orchestrate failover hardening status, ISO 20022 SettleMap progress, and Devon’s availability on Brindlewood / CAS work.

(5) *Email thread (1-message excerpt).*

**From:** Yuki Tanaka (yuki.tanaka@clearflowpaymentinfrastructure.com)

**To:** Charlotte Bewick (Brindlewood) **Cc:** Dillon Hargrove, Devansh Pillai

**Subject:** SOC 2 Type II package delivered to Brindlewood — FYI + Q&A call offer

“Hi Charlotte, I wanted to confirm that Clearflow’s SOC 2 Type II attestation package has been shared with you via our secure document portal. No exceptions were noted that would materially affect availability, security, or processing integrity. Happy to set up a 30-minute Q&A call with myself and Devansh whenever it’s useful. Best, Yuki.”

(6) *Shared document excerpt (project doc).*

Clearflow Payments — SOC 2 Type II Compliance Posture Summary. Prepared for: Brindlewood Commerce Group (Procurement & Vendor Risk). Audit Period: April 1, 2024 to March 31, 2025. Independent Auditor: Hartwell & Crane LLP. Trust Service Criteria: Security, Availability, Confidentiality, Processing Integrity. Audit covered Clearflow’s core payment infrastructure platform in its entirety (Clearflow Orchestrate, Clearflow Settle, Clearflow Shield). One qualified opinion (CF-CM-04, UK change-management workflow); fully remediated as of 2026-02-03.

(7) *Enrollment metadata*. Per-persona enrollment clips are short (7–13 s), drawn from CommonVoice (Ardila et al., 2020) after the curation pipeline of Appendix K.

## C Generation-pipeline details

The schedule is frontier-parallel: at each step, every artifact whose predecessors have completed is dispatched concurrently against a *scoped* workspace state built by replaying only its transitive-dependency events. Artifacts emit typed WorkspaceEvent payloads (decisions, action items opened or closed, project status changes), which a final reconciliation pass globally sorts by timestamp and folds through a pure reducer  $\text{reduce} : \text{State} \times \text{Event} \rightarrow \text{State}$ .

Transcript prompts receive a calendar-anchored word budget (*floor, target, ceiling*) derived from scheduled duration and a target words-per-minute range. Transcripts below the floor are regenerated once with a deficit-aware addendum, and the downstream realism rewrite (§3.4) is gated against the same floor so that structural realism cannot collapse a transcript below its scheduled-duration contract.

## D Workspace-artifact ablation details

To isolate which components of L3 drive improvement over L2, we group the L3 increment into two streams—per-speaker voice *enrollments* (routed to the speaker-identification stage) and *workspace text* artifacts (the workspace member list, inter-meeting emails, slide decks, and meeting agendas)—and run the OpenAI stack with each stream individually enabled on top of L2 (Table 8).

The three target dimensions show qualitatively different complementarity patterns. **Number recall** is dominated almost entirely by workspace text: text-only already captures the full L3 gain ( $0.637 \rightarrow 0.975$ ), while enrollments-only barely moves the number above L2 ( $0.637 \rightarrow 0.605$ ). **Technical-term recall** shows a similar but weaker text-dominated pattern (text-only  $0.723 \rightarrow 0.739$  vs. full  $0.723 \rightarrow 0.760$ ). **Action-item assignee accuracy** is the cleanest case of two-stream complementarity: workspace text alone improves from  $0.524 \rightarrow 0.752$ , enrollments alone from  $0.524 \rightarrow 0.698$ , and full L3 exceeds either single stream ( $0.524 \rightarrow 0.792$ ). Enrollments anchor SID cluster labels to canonical workspace identities, while workspace text supplies the canonical names,

| Configuration            | Num. rec.    | Term rec.    | Assignee     |
|--------------------------|--------------|--------------|--------------|
| L2 (no L3 streams)       | 0.637        | 0.723        | 0.524        |
| + enrollments only       | 0.605        | 0.657        | 0.698        |
| + workspace text only    | 0.975        | 0.739        | 0.752        |
| <b>L3 (both streams)</b> | <b>0.973</b> | <b>0.760</b> | <b>0.792</b> |

Table 8: OpenAI-stack ablation on the L2→L3 increment, computed on the earlier 50-meeting evaluation subset. Per-speaker voice *enrollments* (routed to SID) and *workspace text* artifacts (member list, emails, slides, agendas) contribute differently across dimensions: workspace text accounts for essentially the entire number-recall gain, while enrollments and workspace text contribute complementarily on assignee accuracy. Number- and term-recall figures use exact matching; assignee accuracy uses partial-credit substring matching.

projects, and entities that the summarization stage writes into action items and contextualized terms.

## E Error analysis

To understand what drives the headline numbers in §5, we cluster reference–hypothesis disagreements in the OpenAI L1 output and codify a six-way taxonomy of ASR error classes: **NE** (named-entity confusion), **NUM** (numeric/date verbalization), **TERM** (acronym or compound-boundary errors), **FILL** (disfluency insertion or deletion), **FUNC** (function-word reshaping), and **LEX** (common-word bias or local reanalysis).

The three most prevalent categories (**NE**, **NUM**, and **TERM**) together account for roughly two thirds of extracted span pairs and dominate the word-error trends in §5.1. L3 context mainly helps the entity-sensitive classes: in one representative passage, BLE and ARCBUS are recovered only once the workspace jargon list is available, and enrollment-based speaker identification corrects an action-item assignee from a hallucinated name to *Darren Okafor*. NUM errors are less affected because the workspace artifacts contain little numeric metadata, while FILL, FUNC, and LEX are mostly orthogonal to context. Appendix J gives the full qualitative examples.

## F Adversarial transcript refinement details

LLM-generated transcripts possess a characteristic set of structural features that distinguishes them from real transcripts: a clean narrative arc, efficient speaker handoffs, explicit action-item recaps, and near-total absence of dropped threads. The adver-

sarial structural realism pass targets these discourse-level tells only; surface disfluencies (filled pauses, repetitions, truncations) are injected separately by the spoken-form formatter (§3.2).

A `claude-sonnet-4-6` rewriter first restructures the transcript using a real AMI dataset (Carletta et al., 2006) excerpt as a style reference, breaking the narrative arc (raising and dropping at least one thread, derailing the main topic for several turns, omitting any end-of-meeting recap), adding structurally costly interruptions, replacing expository dialogue with oblique shared-knowledge shorthand, blurring per-speaker roles, and letting the meeting trail off without resolution. The rewriter is allowed to split, merge, or locally reorder turns but is prevented from adding or removing speakers or discarding key decisions.

The rewritten transcript then enters a refinement loop using a cross-family discriminator (`gemini-3.1-pro-preview`). The discriminator performs a pairwise A/B comparison against an AMI meeting (Carletta et al., 2006) excerpt drawn from a holdout split reserved for the OSYNTH realism loop, with speaker labels replaced by generic identifiers and the candidate trimmed to match the AMI excerpt’s line budget inside the judge call; this neutralizes both naming-distribution and length-based shortcut features that would otherwise dominate the verdict regardless of structural realism. The synthetic side is randomized per (meeting, iteration) by a meeting-deterministic RNG so the verdict is reproducible. If the discriminator is fooled, the loop exits early. Otherwise it returns the reasoning for the remaining tells, and a `claude-opus-4-7` refiner revises the transcript to address them directly. The loop repeats for up to  $N = 5$  iterations. If the discriminator cites narrative efficiency, the refiner is permitted to permanently delete a topic thread, prioritizing structural realism over content completeness; however, when the candidate is already at or near its calendar-anchored word floor (§3.1), this permission is withdrawn for that meeting so that the refinement loop cannot collapse a transcript below its scheduled-duration contract, and the refiner is instructed to restructure rather than shrink.

**Why a cross-family loop.** In early development we tried a same-family configuration in which a single Claude model played all three roles (rewriter, refiner, and discriminator). On those runs the discriminator typically declared the candidate indis-

tinguishable on the very first iteration, causing the loop to exit immediately even when the candidate still exhibited the characteristic tells the pass is meant to remove (clean recaps, perfect handoffs, no dropped threads). Switching the discriminator to a different model family from the rewriter/refiner removed this trivial fixed point: the Gemini judge consistently flagged structural tells that the same-family judge accepted, and accepted candidates only after multiple rounds of edits.

**Iteration cap.** We chose  $N = 5$  as a runtime cap rather than as an empirical convergence point. In test runs we extended the loop to  $N = 10$  and observed that a non-trivial fraction of meetings continued to receive accepted refiner edits past iteration five, so the cap reflects the per-meeting cost budget for the full corpus generation rather than a saturation in the discriminator’s verdicts.

**Length consideration.** We noticed that our original implementation resulted in adversarial refinement reducing the length of the transcript at each iteration. We updated the prompt and added a minimum length (shorter lengths being rejected), resulting in two possible implementations of adversarial refinement: a naive implementation and a length-preserving implementation.

## G Additional realism-study details

In the 2AFC human-discrimination study (§3.5), annotators were recruited from a pool of professionals working in the meeting-transcription domain. Annotators were informed before beginning that their responses would be used to evaluate synthetic meeting transcripts and reported in aggregate in a research publication. Each session interleaves realistic OSYNTH trials, Haiku-w/o-obfuscation control trials, and two attention-check trials whose correct answer is unambiguous; only sessions in which both checks were answered correctly are retained for the headline analysis.

As complementary supporting evidence at the meeting level, we run a single-stimulus LLM-judge probe in which `gpt-5.4` (OpenAI, 2026b) (reasoning\_effort=low) is shown one full meeting transcript at a time and asked to classify it as real or synthetic. On a balanced sample of 100 transcripts (50 real AMI, 50 OSYNTH), the judge achieves 59% overall accuracy with 100% recall on real meetings and only 18% recall on OSYNTH meetings, calling 91/100 transcripts “real” overall

(41 of those being false positives). This probe has a strong “real” prior on isolated full transcripts, so we read it as consistent with the 2AFC finding that realistic OSYNTH excerpts lack obvious tells rather than as independent proof of realism. A qualitative analysis of the free-text justifications raters provided when correctly identifying synthetic excerpts surfaces a small set of recurring cues: over-articulated agendas, near-absence of overlapping speech, and unusually clean topic transitions.

## H System and metric details

Both single-vendor systems use vendor-recommended defaults for decoding and a fixed prompt template per task. For cascaded systems, context is encoded as additional fields in the prompted summarization, action-item, and outline stages; for one-shot MLLMs, context is passed as additional inputs to the single call. Speaker-identity voice enrollments are passed to the speaker-identification stage of cascaded systems and as additional audio inputs to MLLMs that support multi-clip audio inputs; for MLLMs that do not, we pass an ASR transcript of each enrollment together with the speaker’s name.

**OpenAI.** The `gpt-4o-transcribe-diarize` endpoint produces a speaker-attributed transcript with cluster labels in a single call. The `diarize` endpoint does not accept Whisper-style prompt biasing, so workspace vocabulary is instead applied as a downstream post-correction at L3. `gpt-audio` performs cluster-level speaker identification by comparing a per-cluster representative audio clip against the workspace’s speaker enrollments. `gpt-5.5` performs jargon post-correction together with prompted summarization, action-item, and outline generation, consuming the resulting transcript and the level-conditioned workspace context.

**Gemini.** The Gemini baseline mirrors the OpenAI pipeline via Vertex AI but is split across two Gemini models. Stage 1 (`gemini-3-flash` (Google DeepMind, 2025)) ingests the meeting audio directly and emits a JSON array of speaker-tagged segments with cluster labels. Stage 2, enabled at L3, uses `gemini-3.1-pro-preview` (Google DeepMind, 2026) for enrollment-based speaker identification, comparing a per-cluster representative clip against the workspace’s enrollment clips and returning the matching display name;

we use the larger Pro model for this stage because it is the SID stage that most relies on fine-grained acoustic-comparison reasoning. Stage 3 (`gemini-3-flash`) consumes the resulting transcript together with the workspace context and produces jargon post-corrections, summaries, action items, and outlines.

**Transcription quality.** We report standard word error rate (WER), scored against the semantic-verbatim transcript (§3.2), with reference and hypothesis passed through Whisper’s `EnglishTextNormalizer` (Radford et al., 2023) before scoring.

**Contextualized transcription.** OSYNTH’s latent state makes it possible to score ASR on word categories that disproportionately affect downstream user experience. We report two such categories: numbers and dates (NUMBER RECALL), and technical and domain-specific terms (TECHNICAL-TERM RECALL). Reference term sets for both are extracted directly from the workspace’s latent state and intersected with the oracle transcript, so a term counts against a system only if it was actually spoken in the meeting being scored.

**Diarization and speaker identification.** We report cross-meeting speaker-identity consistency: the rate at which a workspace persona is mapped to the same predicted name across the meetings of its workspace. Because the released alignments tie every word to a canonical persona (Appendix O), this scores whether words are routed to the correct *named* speaker rather than merely to a self-consistent per-meeting label, and it exposes labeling-stability failures that a per-meeting score cannot.

**Summarization artifacts.** Per-meeting summary quality is scored by an LLM judge on FAITHFULNESS and COHERENCE following Liu et al. (2023a) and Zheng et al. (2023), with reference content drawn from the gold summary derived from the latent state. Action-item extraction is scored against the latent action-item list by assignee accuracy, together with an LLM judgment of how completely and concretely each predicted description captures the underlying commitment.

## I Prompt templates

For reproducibility, we reproduce a representative prompt (or prompt-template sketch) for each major

prompting stage of the evaluation pipeline. The summarization, action-item, and outline prompts are shared verbatim across the OpenAI and Gemini stacks (modulo the model handle); the enrollment-based SID and vocabulary post-correction prompts are stack-specific. The adversarial-realism loop (§3.4) uses two additional model-family-specific prompts, reproduced below. Verbatim prompt text is shown in monospace blocks; templating placeholders are written {name}.

**Summarization (L1 / L2 / L3).** The OpenAI (gpt-5.5) and Gemini (gemini-3-flash) stacks share a single summarization prompt template (Figure 3). The user message contains an optional CONTEXT block (empty at L1, populated at L2/L3) followed by a TRANSCRIPT block. At L2 the CONTEXT block contains the system’s own prior-meeting outputs (summaries and action items, each truncated); at L3 it additionally contains the workspace name list (canonical name + role per persona), a condensed digest of recent workspace emails, and the meeting’s agenda title and type. Each L3 add-on is rendered under a byte budget so the assembled context fits in a single inference call.

**Action-item extraction.** The OpenAI and Gemini stacks share an action-item-extraction prompt template (Figure 4); the outline prompt has the same shape, with the instruction body replaced by a request for a markdown-heading outline. The same CONTEXT block as in summarization is used.

**Enrollment-based speaker identification (L3).** Both stacks identify the speaker behind each diarization cluster by comparing a per-cluster representative audio clip against the workspace’s enrollment clips and returning the matching display name. The OpenAI stack uses gpt-audio and the Gemini stack uses gemini-3.1-pro-preview; the system instruction and user-content scaffolding (Figure 5) are essentially identical across the two stacks. The user content is multimodal: a short framing string, then one (Enrollment – {display\_name}: + audio) pair per enrolled persona, then a final (Target turn: + audio) pair and a closed-list constraint. The response is a single line that we match back to the enrollment list with a tolerant first-name fallback.

**Vocabulary post-correction (L3, OpenAI).** The OpenAI L3 stack runs a per-turn vocabulary-correction LLM pass over the upstream ASR transcript, prompted with the workspace’s jargon list

(acronyms, product names, person names); see Figure 6. The response must be a JSON array of corrected text strings, length-equal to the input turns list; length-mismatched responses are rejected and the original turns are kept. The open-source SpeechBrain L3 stack (Appendix V) reuses this exact prompt with claude-haiku-4-5 as the rewriter, which is what drives the silent-summarization failure mode documented there.

**Vocabulary post-correction (L3, Gemini).** The Gemini stack uses a stricter variant that adds explicit anti-insertion and anti-expansion guards on top of the OpenAI prompt (Figure 7). These constraints were added after we observed that the Gemini-flash rewriter is more aggressive than the OpenAI rewriter at applying vocabulary corrections to already-correct spans; the stricter prompt restores the intended “vocabulary patch, not paraphrase” behavior.

**Adversarial-realism discriminator.** The structural-realism judge (gemini-3.1-pro-preview, Appendix F) is shown a pairwise A/B comparison between the candidate and a length-matched AMI excerpt; its prompt (Figure 8) explicitly scopes the verdict to discourse-level signals and bans the surface tells that the spoken-form formatter owns.

**Adversarial-realism refiner.** The refiner (claude-opus-4-7) is then told what the judge cited and asked to address those tells specifically. The full prompt (Figure 9) is long; we reproduce its key structural elements verbatim and elide ancillary boilerplate. The split between a same-family rewriter / refiner (Claude) and a cross-family discriminator (Gemini) is the empirical choice motivated in Appendix F: same-family judges tend to declare the candidate indistinguishable on the first iteration even when obvious structural tells remain.

**Summary-quality judge.** The FAITHFULNESS and COHERENCE scores reported in Tables 3 and 4 are produced by the prompt in Figure 10, executed by claude-sonnet-4-6 with content-hash caching (§4.3). The prompt elicits five dimensions in a single pass; we report these two and release the remainder with the corpus. The prompt is a direct-scoring LLM-as-judge in the spirit of Zheng et al. (2023): the judge writes integer 1–5 scores directly into a JSON object that we parse, with hand-written per-dimension rubric anchors (we do

```

You are generating a meeting summary from a transcript.
{context_block}TRANSCRIPT:
{transcript}
Write a concise summary covering the key discussion points, decisions made, and outcomes. Use bullet
points for clarity. Keep it under 500 words.

```

Figure 3: Shared summarization prompt template used by both the OpenAI (gpt-5.5) and Gemini (gemini-3-flash) stacks across all three context levels. {context\_block} is empty at L1 and populated at L2/L3 as described in the surrounding prose.

```

You are extracting action items from a meeting transcript.
{context_block}TRANSCRIPT:
{transcript}
Extract all action items mentioned in the meeting. For each, provide:
- description: What needs to be done
- assignee: Who is responsible (use the speaker name if mentioned, otherwise "Unassigned")
Respond with ONLY a JSON array of {"description": "...", "assignee": "..."}

```

Figure 4: Shared action-item-extraction prompt template. The outline prompt has the same scaffolding, with the instruction body replaced by a request for a markdown-heading outline of the meeting topics.

```

[System instruction]
You identify which speaker uttered a target audio clip by comparing it to a small set of labeled
enrollment clips. Respond with ONLY the label of the best-matching enrollment, exactly as given. Do
not add punctuation, explanation, or commentary.

[User content, multimodal - text + audio interleaved]
Below are enrollment clips, each labeled with a speaker name, followed by a target turn clip.
Identify which enrolled speaker produced the target clip. Reply with only the name.
Enrollment - {display_name_1}: [audio Part]
Enrollment - {display_name_2}: [audio Part]
...
Target turn: [audio Part]
Which enrollment label matches the target turn? Choose exactly one of: {comma-separated display
names}.

```

Figure 5: Shared enrollment-based speaker-identification prompt (system instruction + user content scaffolding) used by both the OpenAI gpt-audio stage and the Gemini gemini-3.1-pro-preview stage at L3. The user content is multimodal: each [audio Part] placeholder is an inline audio attachment (the enrollment clip or the target turn).

```

You are correcting a meeting transcript for vocabulary errors only.
EXPECTED VOCABULARY (terms likely spoken in this meeting):
{vocab}
INSTRUCTIONS:
- Replace likely misspellings with the matching expected term (e.g., "B.L.E." → "BLE", "Arc bus"
→ "ArcBus", "Yuki Tanaka" stays as-is).
- Do NOT paraphrase, reorder, summarize, or fix grammar.
- Do NOT change speaker labels, turn boundaries, or punctuation beyond spelling fixes.
- Output ONLY a JSON array of strings, one entry per input turn, in the same order. Each entry is
the corrected "text" for that turn.
INPUT TURNS:
{turns_json}

```

Figure 6: OpenAI L3 vocabulary post-correction prompt (gpt-5.5). The SpeechBrain L3 pipeline reuses this prompt verbatim with claude-haiku-4-5 as the rewriter; on high-WER upstream input that pass tends to silently summarize around the transcript rather than patch it, which is the failure mode documented in Appendix V.

not use auto-generated chain-of-thought steps or token-probability-weighted scoring as in Liu et al., 2023a). The judge is reference-grounded — inputs are the gold transcript, the gold summary, the workspace roster, and the system summary — rather than scoring the system summary in isolation. Three design choices in the prompt are load-

bearing for how Tables 3 and 4 should be read. (i) The prompt explicitly decouples length from every dimension except conciseness, so longer or shorter summaries are not rewarded under recall, precision, faithfulness, or coherence; reference length is not a target. (ii) The prompt marks any workspace-canonical-name expansion that is consistent with

[OpenAI prompt body, Figure 6, then:]

ADDITIONAL CONSTRAINTS:

- The vocabulary list is REFERENCE ONLY. Most input tokens are already correct – leave them alone.
- DO NOT replace a token just because a similar vocabulary entry exists. If the input token is already a valid English word, leave it unchanged.
- DO NOT change capitalization.
- DO NOT expand abbreviations or acronyms.
- DO NOT insert words. Output word count for each turn MUST be equal to or less than the input word count.
- If you are unsure whether a token is a misspelling, leave the turn unchanged. Conservative is correct.

Figure 7: Gemini-stack L3 vocabulary post-correction prompt. The instructions are the OpenAI body (Figure 6) plus the explicit anti-insertion / anti-expansion constraints listed above.

You are a forensic judge. You will see two meeting transcripts, labeled Transcript A and Transcript B. One is real; one is synthetic. Decide which is which based on structural signals only, then explain the tells that drove your decision.

SCOPE – structural signals ONLY. EXPLICITLY IGNORE: word-level disfluencies, single-word stutters, mid-word truncations, ASR transcription errors, diarization quirks ("Unknown Speaker", generic labels). A downstream pass owns those.

CONSIDER (in-scope structural tells): (1) NARRATIVE ARC; (2) DROPPED THREADS; (3) ACTION-ITEM RECAP; (4) INTERRUPTION COST; (5) BANNED TROPES (mute, pets, deliveries, scripted forgetfulness, closing recap); (6) EXPOSITORY DIALOGUE; (7) SPEAKER ROLES; (8) ENDING.

Transcript A:  
{{TRANSCRIPT\_A}}

—

Transcript B:  
{{TRANSCRIPT\_B}}

Reason briefly (≤500 chars) about the structural tells. Then output exactly one JSON object on the final line:  
{"verdict": "A"|"B", "reasoning": "...", "structural\_issues": ["...", ...]}

'verdict' is which transcript you believe is synthetic. 'structural\_issues' is drawn from: clean\_arc | no\_dropped\_threads | explicit\_action\_items | costless\_interruptions | banned\_trope\_present | expository\_dialogue | clean\_speaker\_roles | clean\_ending.

Figure 8: Adversarial-realism discriminator prompt (gemini-3.1-pro-preview). One transcript is the candidate, the other is a length-matched AMI excerpt; the synthetic side is randomized per (meeting, iteration) by a meeting-deterministic RNG so verdicts are reproducible.

the roster (e.g. *Brindle* → *Brindlewood Commerce Group*, first name → full name, codename → full project description) as faithful rather than as a fabrication; the alternative would penalize precisely the L3 behavior the experiment is meant to measure, and is what drives the Gemini L3 faithfulness dip on stacks whose roster-expansion behavior is more aggressive (§5.2). (iii) Speaker attribution is excluded from faithfulness, since a separate metric (CROSS-MEETING SPEAKER-IDENTITY CONSISTENCY) scores it; speaker labels are treated as opaque tags by the judge. We reproduce the prompt body verbatim and abridge only the per-dimension Likert-anchor rubrics.

## J Qualitative error examples

To make the error categories in §E concrete, we walk through representative spans drawn from the OpenAI system on OSYNTH at L1, L2, and L3, all taken from a single pilot meeting.

**Transcript-level TERM recovery.** The same passage shows progressively cleaner handling of hardware-domain terms as workspace context is added (see Table 9 for the contrast):

**GT:** *Yuki:* Already doing it. PMIC and BLE, nothing there needs the seven-one-four-two. [...] The pipeline's tied to the timing and resolution profile—ArcBus gives us some abstraction, but the digital filter coefficients are dictated by the analog characteristics of the front-end.

**L1:** *Unknown speaker:* Already doing it, PMIC and VLE Nothing there needs the seven one four two. [...] ARC plus gives us um some abstraction...

**L2:** *Unknown speaker:* Already doing it. PMIC and PLE Nothing there Needs the SN7142. [...] Arcus gives us um some abstraction...

**L3:** *Darren Okafor:* Already doing it, PMIC and BLE Nothing there needs the seven one four two. [...] ArcBus gives us um some abstraction...

**Summary-level NE recovery.** The same NE / TERM failures propagate into the summary. Ex-

You are a transcript style editor. You previously rewrote a synthetic meeting transcript to look structurally real. A forensic judge correctly identified your rewrite as the synthetic one. Revise the transcript to directly address the tells the judge cited.

{{LENGTH\_BUDGET\_CLAUSE}}

SCOPE – structural realism only (arc, dropped threads, unresolved blockers, action-item recap, interruption cost, banned tropes, expository dialogue, blurred speaker roles, trailing endings, structural disfluencies).

OUT OF SCOPE – do NOT add filler sounds, single-word stutters, or mid-word truncations (a downstream pass handles them).

RULES: all key info preserved; do NOT add or remove speakers; you MAY split / merge / reorder turns within a small local window, or drop a short turn if "the moment didn't land".

JUDGE'S REASONING: {{JUDGE\_REASONING}} Structural issues flagged: {{STRUCTURAL\_ISSUES}}

CHECKLIST (verify each before outputting): derail the main topic for 4+ turns at least once; add at least one place where someone talks past another; at least one speaker abandons a clause midway; when someone drops off, make it cost something; remove every banned interruption trope; replace expository dialogue with shorthand; blur at least one speaker role; trail off the ending with logistics – NO end-of-meeting recap.

Real transcript style reference: {{REAL\_REFERENCE}}

Previous rewrite: {{PREVIOUS\_REWRITE}}

Self-check (mandatory): identify the single most synthetic-looking feature of your current draft and fix it before outputting. Return ONLY a JSON array of turns of shape {"speaker\_id": "...", "speaker\_name": "...", "text": "..."}.

Figure 9: Adversarial-realism refiner prompt (claude-opus-4-7). The discriminator's per-iteration verdict reasoning and structural-issue tags are spliced into the prompt so the refiner addresses the cited tells directly; the CHECKLIST encodes the structural realism targets used throughout the corpus.

| Term   | GT     | L1       | L2     | L3       |
|--------|--------|----------|--------|----------|
| BLE    | BLE    | VLE      | PLE    | BLE ✓    |
| ArcBus | ArcBus | ARC plus | Arctus | ArcBus ✓ |

Table 9: Progressive resolution of TERM errors (BLE, ARCBUS) across L1, L2, and L3 from a single passage. L2's self-biased prior outputs recover some tokens (PMIC, SN7142) but also introduce confident hallucinations (ARCTUS); the explicit L3 jargon list recovers the remainder.

cerpts from the same meeting's L1 vs. L3 summaries:

*L1 (excerpt): Sony Tech SN714 slip / EVT impact.* Two SN714 samples have slipped by at least three weeks due to fab allocation. Team is treating three weeks as a minimum delay...

*L3 (excerpt): Sonitec SN7142 supply slip / EVT impact.* Darren flagged that Sonitec SN7142 samples have slipped at least three weeks; team should treat that as a floor, not an optimistic estimate. Switching to a second source was judged likely worse than waiting: AFE redesign, new HAL driver work, and re-characterization. Outcome: continue work not dependent on the SN7142—Gen 2 bring-up, PMIC, BLE, firmware.

L3 correctly names *Sonitec SN7142* (vs. L1's *Sony Tech SN714*), attributes the flag to *Darren* by name, and adds the explicit outcome decision absent from L1.

**Action-item assignee correction.** The transcript-level NE failures also drive incorrect attribution in

extracted action items. From the same meeting:

**GT:** *Darren:* Deliver Arc enclosure stack-up by Friday.

**L1:** *[Ryan]* Deliver the odd enclosure power stack-up by Friday.

**L3:** *[Darren Okafor]* Deliver the Arc enclosure tolerance stack-up by Friday.

L1 hallucinates *Ryan* (a name not present in the meeting) and corrupts the artifact name (*odd* for *Arc*); L3's enrollment-based SID and jargon list jointly recover both the assignee and the artifact.

**NUM recovery (number recall, +0.20).** The +0.20 absolute jump in number recall (§5.1) is driven by L3's ability to anchor ASR-fragile identifiers against the workspace text artifacts. A representative span from the *Marcus* ↔ *Simone* 1:1 (2026-05-19, fintech-payments workspace) discusses ISO 20022 message-family identifiers (pacs.008 and camt.054), which are spelled out by the active project spec under the heading "ISO 20022 Payment Message Migration":

**GT:** *PACS.008 is decent. Three of the CAMT.054 sub-fields are still draft, Payments Canada hasn't published.*

**L1:** *PAX eight is decent, three of the CAMTWOH five four sub-fields are uh still draft Payments Canada hasn't hasn't published.*

**L3:** *pacs.008 is decent. Three of the camt.054 subfields are, uh, still draft Payments Canada hasn't hasn't published.*

You are evaluating the quality of a meeting summary generated by a speech-to-text system.

You are given: (1) the WORKSPACE CONTEXT (canonical roster of company, people, and projects – the source of truth for full canonical names, affiliations, and project descriptions); (2) the REFERENCE TRANSCRIPT; (3) the REFERENCE SUMMARY; (4) the SYSTEM SUMMARY.

WORKSPACE CONTEXT: {workspace\_context} REFERENCE TRANSCRIPT: {transcript} REFERENCE SUMMARY: {ref\_summary} SYSTEM SUMMARY: {sys\_summary}

Evaluate the SYSTEM SUMMARY on five dimensions, scoring each from 1 (worst) to 5 (best). Use the full 1–5 range – pick 2 or 4 when the summary sits between the anchors below.

IMPORTANT – length is not quality. The system summary may be substantially shorter or longer than the reference; reference length is not a target. Judge each dimension on what the content does, not how much text it occupies:

- Recall: count distinct important facts covered, not bullets, paragraphs, or word count. Extra prose that does not add new facts does not raise recall.
- Precision: a short summary that includes only important points scores 5; do NOT penalize it for being brief (omissions are recall’s job, not precision’s).
- Faithfulness: score only on whether claims are supported by the transcript and/or workspace context. Do not reward length, polish, or apparent “thoroughness.”
- Coherence: score only on structure and flow. Do not reward longer summaries for appearing more comprehensive.
- Conciseness is the ONLY dimension where length directly matters.

**Faithfulness (1–5):** Evaluate factual content only: events, decisions, action items, deadlines, numbers, technical claims, products / projects / features, dependencies, and outcomes. IGNORE speaker attribution. Treat speaker labels (real names, “Speaker 1”, etc.) as opaque tags – do NOT check whether a statement is assigned to the right person. Speaker-attribution quality is measured by a separate metric and is out of scope here.

Workspace-consistent entity resolution is faithful, NOT a fabrication. When the summary expands a short form to its canonical workspace name (e.g., Brindle → Brindlewood Commerce Group, first name → full name, abbreviation → full project / product name, project codename → full project description) and that expansion matches the WORKSPACE CONTEXT, treat it as accurate. Only flag entity names or facts that are absent from BOTH the transcript and the workspace context, or that conflict with the workspace context.

[Per-dimension 1–5 anchor rubrics for Recall, Precision, Faithfulness, Coherence, Conciseness follow the standard G-Eval pattern (Liu et al., 2023a); elided here for space.]

Respond with ONLY a JSON object: {“recall”: <int>, “precision”: <int>, “faithfulness”: <int>, “coherence”: <int>, “conciseness”: <int>, “recall\_explanation”: <brief>, “precision\_explanation”: <brief>, “faithfulness\_explanation”: <brief>, “coherence\_explanation”: <brief>, “conciseness\_explanation”: <brief>}}

Figure 10: Summary-quality LLM-judge prompt (claude-sonnet-4-6). Reference-grounded direct-scoring LLM-as-judge in the spirit of Zheng et al. (2023), with hand-written per-dimension 1–5 rubric anchors; we do not use the auto-generated chain-of-thought steps or token-probability-weighted scoring of Liu et al. (2023a). The judge is explicitly instructed to (i) treat roster-consistent canonical-name expansions as faithful rather than as fabrications, (ii) decouple length from every dimension except conciseness, and (iii) exclude speaker attribution from faithfulness (scored separately). Per-dimension Likert anchors are abridged for space; full prompt is released with the evaluation code.

L1 parses pacs.008 as “PAX eight” (a phonetic guess that drops both the prefix and the dotted subtype) and camt.054 as “CAMTWOH five four” (a syllable-level mangling of the spelled-out form). L3 recovers both via the workspace jargon list and project-spec excerpt, which contain the canonical surface forms. The per-meeting number-recall score moves from 0.00 at L1 to 1.00 at L3 on this meeting; the L2 hypothesis already recovers both identifiers via the vocab post-correction LLM pass even without the workspace name list, consistent with the ablation finding (§5.3, Appendix D) that workspace text—not enrollments—drives the number-recall axis.

**Assignee correction, second meeting.** To rule out a single-meeting cherry-pick of the assignee gain, we give a second OpenAI assignee example from a different meeting in the same workspace—the Marcus ↔ Devon 1:1 (2026-05-19)—where the canonical assignee is Devon Ashby (workspace person p8). The L1, L2, and L3 hypotheses trace out an *invented name* → *invented name* → *canonical* trajectory rather than a flat accuracy lift:

**GT:** [Devon Ashby] Document idempotency-key behavior during failover/retry overlap, including whether the key survives a gateway switch, in the CAS ADR.

**L1:** [Bevan] Trace the idempotency key behavior during failover, especially when a CAS retry overlaps a gateway switch, to confirm whether duplicate charging is possible.

|          | Meeting $n$        | Meeting $n+1$      |
|----------|--------------------|--------------------|
| Title    | Marcus–Devon 1:1   | BDW Sandbox Scope  |
| Date     | 2026-05-19         | 2026-05-21         |
| L1 label | <i>Bevan</i>       | <i>Devin</i>       |
| L3 label | <i>Devon Ashby</i> | <i>Devon Ashby</i> |

Table 10: Cross-meeting labels for canonical person p8 (*Devon Ashby*) in two near-consecutive meetings in the same workspace, under OpenAI L1 versus L3. L1 hallucinates a different phonetically plausible English given name in each meeting because per-meeting diarization clusters do not persist and the L1 SID stage has no workspace name list to anchor on; L3 fixes both meetings to the same canonical identity via the workspace name list and enrollment-based SID against the persona’s prompt clip.

**L2:** [*Devin*] Trace the idempotency-key behavior when a retry overlaps a gateway failover transition, including what the downstream sees after the gateway switch.

**L3:** [*Devon Ashby*] Trace idempotency-key behavior during failover/retry overlap, including whether the key survives a gateway switch and what the downstream sees.

The L1 SID stage has no workspace name list, so it autocompletes the diarization cluster to a phonetically plausible but invented English name (*Bevan*); the L2 vocab-correction pass nudges the guess toward a different invented neighborhood (*Devin*) but still cannot resolve to any canonical workspace identity. Only at L3 do the workspace name list and the enrollment-based SID jointly anchor the cluster to person p8 (*Devon Ashby*). The per-meeting assignee accuracy is 0.00 at L1 and 1.00 at L3, and the same pattern holds across six near-identical Devon-assigned action items in the meeting, so the correction is not a single-item cherry-pick.

### Cross-meeting speaker consistency (+0.26).

The cross-meeting speaker-identity consistency metric (§4.3) flips most visibly when the same person appears in two near-consecutive meetings whose per-meeting diarization clusters carry incompatible labels. Picking up the same *Devon Ashby* (p8) as in the assignee example above, Table 10 shows the labels assigned by the OpenAI L1 versus L3 systems in two meetings two days apart in the fintech-payments workspace:

The two L1 labels disagree both with the canonical name and with each other—a single-meeting accuracy metric would mark them merely as “wrong”, but the cross-meeting consistency metric exposes the structurally worse failure that no downstream

consumer of the L1 output can stitch the two meetings into a single *Devon* entity. L3 recovers consistency on both axes because the per-meeting cluster-to-canonical-identity step is anchored to the same workspace name list and the same enrollment clip across meetings (Appendix B).

## K TTS rendering strategy

Recent multi-speaker TTS systems can render an entire multi-party conversation in a single forward pass, taking as input a single script with interleaved speaker tags and emitting one mixed waveform together with implicit turn timings. We deliberately do *not* use this single-call rendering mode in the released pipeline. The convenience of a one-shot call is offset by three concrete losses that matter for downstream evaluation: (i) the absence of an explicit per-turn audio segment makes word- and turn-level alignment back to the canonical transcript ambiguous, and any subsequent speaker-attributed reference (RTTM, word-level alignments, per-speaker enrollments) has to be re-derived by a forced-alignment pass with its own error budget; (ii) per-turn acoustic parameters that vary by persona (rate, pitch range, vocal effort, accent) are difficult to control independently when all turns are produced in a single decoding pass, which compresses the cross-speaker acoustic diversity that diarization and speaker-identification systems are meant to be tested on; and (iii) when a single decoding pass mishandles one turn (a degenerate repetition, a swallowed word, an off-target voice), the whole meeting must be re-rendered, whereas a per-turn pipeline localizes the failure to a single turn that can be regenerated independently. We therefore synthesize each turn independently (§3.2) with the persona’s reference clip, the persona’s target acoustic parameters, and the turn’s TTS-form text, and assemble the mixed waveform from the resulting per-turn segments. Per-turn segments are retained alongside the mixed waveform so that downstream users can re-mix, re-spatialize, or compose individual turns into new scenarios.

The per-turn pipeline does introduce its own failure mode: voice-cloning prompts drawn from a third-party corpus occasionally trigger degenerate behavior in Qwen3-TTS (Qwen Team, 2026), including stuck loops, silence, or off-target speaker identity. We mitigate this by pre-processing the CommonVoice (Ardila et al., 2020) reference pool offline before any meeting is rendered: each can-

didate clip is passed through an ASR round-trip check against its own transcript, a duration and SNR filter, and a short stability probe in which the clip is used to synthesize a fixed neutral utterance which is then re-transcribed; clips for which the round-trip WER exceeds a threshold, or for which the stability probe degenerates, are dropped from the pool. The surviving pool, which we refer to internally as the voice cache, is the only set of clips that personas are sampled from. This preprocessing turns a per-meeting failure mode into a one-time pool-curation cost and, in practice, eliminates the long-tail TTS failures that we observed when sampling reference clips on the fly.

For inference at corpus scale, we serve the 1.7B-parameter variant of Qwen3-TTS (Qwen Team, 2026) on a  $32 \times A100$  80 GB cluster using batched per-turn synthesis through a vllm-omni backend, ultimately sustaining a real-time factor of approximately 0.5 (one second of synthesized speech per half-second of wall-clock time across the batched job). This throughput is what makes rendering the 570 hours of audio in the released corpus tractable in a single pass and what makes the regeneration cost of individual failed turns (rather than whole meetings) negligible in practice.

## L Persona-to-voice assignment

Voices are assigned from the curated pool of Appendix K by matching a single attribute: gender. Each persona carries a gender assigned by the LLM during workspace sampling (§3.1), and the curated Common Voice pool carries a gender label per voice ID (*male*, *female*, or *unknown*). Personas are partitioned by their assigned gender and drawn from the corresponding subset of voice IDs; a persona whose gender is non-binary or unset draws from the union of all subsets. Within the matching subset the draw is uniform and without replacement inside a workspace, so no two personas in a workspace share a voice. No other persona attribute (age, seniority, role, accent, or communication style) currently conditions the assignment, which is a deliberate simplification rather than a modelling claim: richer demographic alignment would require voice-pool metadata that Common Voice does not consistently provide.

## M Conversational prosody

Two prosodic phenomena that are central to natural conversational audio (turn-taking gaps and

backchannels) are produced by the pipeline rather than left to the TTS system. We give them an explicit treatment because they are easy to omit and because both of them strongly affect how realistic the mixed waveform sounds to a human listener and how diarization systems behave on it.

**Floor-transfer offsets.** For each consecutive speaker change we sample a floor-transfer offset (FTO), the signed gap between the end of speaker *A*’s turn and the start of speaker *B*’s turn, from a mixture distribution fit to publicly available conversational corpora; complementary lines of work model turn-taking sequences themselves as a function of team-member trait composition (O’Bryan et al., 2025), whereas we leave who-speaks-next to the LLM-driven turn-generation pass (§3.1) and target only the timing of the resulting transitions here. The mixture concentrates short positive gaps in the range typically reported for spontaneous multi-party meetings, with a tail of longer gaps to model topic-boundary pauses and a negative-gap component that produces brief overlaps at turn boundaries. Same-speaker continuations (a single speaker holding the floor across multiple turns) draw from a separate distribution dominated by short intra-turn pauses. The exact distributional parameters used for the released corpus are recorded alongside the audio and can be re-fit if a user wants to target a different conversational register.

**Substantive turn-level overlap.** Beyond the small negative-gap overlaps produced at turn boundaries by the FTO distribution, an LLM annotator that runs after the spoken-form pass scans the formatted turn list and emits, for each non-first turn, an optional *overlap anchor* pointing at a specific word of the immediately preceding turn together with a mode label: INTERRUPT, where the anchor turn is truncated at the anchor word and the over-lapper takes the floor; or CONCURRENT, where the anchor turn continues to completion while the over-lapper speaks over it. Audio placement is two-pass: non-overlap turns are placed left-to-right by sampled FTO from the previous non-overlap turn’s offset, then overlap turns are placed at the anchor word’s forced-alignment onset and the cumulative timeline pointer is not advanced. A final per-speaker monotonic sweep iterates to a fixed point, dropping any annotation that would cause a speaker to overlap with themselves, so the released audio satisfies a per-speaker no-self-overlap invariant by construction (further checked in the

validation stage, Appendix Q). Annotator failures (malformed payloads, exceptions) fall back to FTO placement; the audio pipeline never blocks on overlap.

**Backchannels.** Single-token backchannels (*mm-hm*, *yeah*, *right*) are generated as ordinary turns by the transcript LLM and then identified and placed by two dedicated annotators that run after the spoken-form pass. A sparse classifier scans the formatted turn list and emits, for each turn it judges to be a backchannel, an explicit `bc_kind` label drawn from {AGREEMENT, ACKNOWLEDGMENT, CONTINUER, REACTION}; substantive turns are non-BC by absence. A second pass anchors each identified backchannel to the most recent prior non-backchannel, non-overlap turn and forces `mode=concurrent` so the backchannel overlays the floor-holder’s audio without advancing the meeting clock. Per-speaker backchannel priors (each persona’s vocabulary of customary backchannel phrases) condition the classifier so that backchannel forms vary across speakers. Backchannel rows are preserved in the transcript and RTTM under the original speaker’s identity, with an `is_backchannel` flag that allows downstream metrics to filter them in or out (e.g., opt-in WER scoring). The classifier and anchor pass each have heuristic fallbacks on truncation or parse failures and never block the pipeline.

## N Speech pacing relative to the cloning prompts

The reference clips used for voice cloning are spontaneous speech (Appendix X), yet the rendered turns can still sound read. To quantify how much of this comes from the TTS stage rather than the prompts, we ran the NeMo forced aligner (Rashtorgueva et al., 2023) over both the cloning prompts and the generated turns and macro-averaged per speaker (Table 11). OSYNTH turns are spoken faster and, more importantly, are paced far more uniformly: the inter-word pause ratio falls by 60% and pauses longer than 200 ms become roughly three times rarer per minute. We attribute the residual read-aloud character primarily to this pacing uniformity, and secondarily to a domain mismatch, since even spontaneous Common Voice speech is recorded in isolation in response to prompts rather than in live multi-party conversation. We expect this effect to diminish as TTS models improve.

| Measure              | Prompts | OSYNTH | $\Delta$ |
|----------------------|---------|--------|----------|
| Active-speech WPM    | 134.5   | 180.9  | +34.5%   |
| Pause ratio          | 47.9%   | 18.9%  | −60.5%   |
| Pauses >200 ms / min | 44.5    | 15.6   | −65.0%   |

Table 11: Pacing of the voice-cloning prompts versus the rendered OSYNTH turns, from NeMo forced alignments macro-averaged per speaker. The generated turns are faster and markedly more uniform in pausing.

## O Alignment procedure and validation

Every released meeting ships with word-level and turn-level alignments back to the canonical transcript (§3.2); these alignments are what make the RTTM file, the per-speaker enrollments, and the contextualized-transcription metrics of §4.3 well-defined. Because each turn is rendered independently from a known TTS-form string, we obtain initial per-word timings directly from the TTS aligner that ships with Qwen3-TTS (Qwen Team, 2026), anchored to the turn’s offset in the mixed waveform. We then refine these per-turn timings with a forced-alignment pass against a small open-source acoustic model on the unmixed per-turn audio, and propagate the refined timings through the mixing offsets and the pyroomacoustics convolution (Appendix P); the convolution introduces a known per-speaker delay that is subtracted analytically rather than re-estimated.

To validate the alignment quality, we evaluate on a held-out sample of meetings against per-utterance reference timings derived from the per-turn rendering offsets (which are exact by construction). We report onset mean absolute error, onset RMSE, duration mean absolute error, and word-coverage rate. We additionally measure the rate at which a word’s predicted span lies entirely within the correct speaker’s RTTM interval, which is the property the speaker-attributed evaluation cares about. Across the validation sample, onset MAE is well below 0.25 s and word-coverage exceeds 99%; we report the exact numbers and the validation procedure in the released code repository.

## P Audio post-processing

After per-speaker TTS synthesis and mixing (§3.2), each meeting is passed through a pyroomacoustics (Scheibler et al., 2018) post-processing stage that places the speakers in a simulated environment and renders the recording as if it had been captured by one or more microphones.

For every meeting we sample a shoebox room geometry (length, width, and height drawn from ranges typical of small-to-medium meeting rooms), a wall absorption coefficient drawn from a range consistent with carpeted, lightly furnished office spaces, and a microphone position near the center of the table. Each persona is assigned a fixed seat around the table for the duration of the meeting; we then compute the image-source-model impulse response from each seat to the microphone and convolve each speaker’s TTS turns with the corresponding response before mixing the convolved turns onto the final track. The resulting waveform exhibits reverberation, inter-speaker level differences, and spectral coloration consistent with a shared microphone recording, while preserving the exact word-level alignment back to the canonical transcript. The geometry, seat layout, and absorption parameters used for each meeting are released alongside the audio.

**Hybrid (in-room plus remote) calls.** A substantial fraction of workspace meetings are not fully co-located: one or more attendees join over a video-conferencing client from a separate room, a home office, or a phone. To cover this case, a sampled fraction of OSYNTH meetings are rendered in a *hybrid* configuration in which a subset of attendees is placed inside the shared conference room as above, while the remainder are rendered as remote participants. Each remote participant is placed in their own (smaller, more reverberant or more dampened) room with its own microphone, convolved with that room’s impulse response, and then passed through a lightweight codec / bandwidth simulation that approximates the spectral compression and occasional level mismatch of a typical VoIP downlink before being mixed into the shared track at a per-participant gain. In-room and remote participants therefore arrive at the final mix with markedly different reverberation profiles and bandwidths, which is the property that distinguishes hybrid recordings from fully co-located ones and that downstream diarization and speaker-identification systems must contend with in deployment. The hybrid-vs-co-located flag, the per-participant room assignment, and the codec configuration are released alongside each meeting so that users can subset the corpus by acoustic condition.

## Q Pipeline validation

Every persisted artifact in the pipeline is validated against a typed schema both at generation time and again on read, which catches malformed LLM outputs (null action items, missing fields, nicknames emitted in identifier slots) at the boundary they originate from rather than corrupting downstream metrics silently. A final pipeline-validation stage additionally checks a small set of cross-artifact invariants on every released meeting: every active attendee has a join mode consistent with the meeting’s mode (*virtual* meetings reject in-room participants, *in-person* meetings reject remote participants, and *hybrid* meetings with two or more active attendees require at least one of each); every meeting’s audio length is within a fill-ratio band of its scheduled calendar slot (§3.1), with under-filled meetings flagged in an aggregate `duration_fill.json`; and no speaker overlaps with themselves in the released RTTM (Appendix M), with violations counted in an aggregate `overlap_stats` block. The validator emits warnings rather than aborting the run, but the aggregate counts and per-meeting flags are released alongside the corpus so that users can subset out any meeting that fails an invariant.

## R Multilingual generation

The OSYNTH pipeline (§3.1–3.2) is language-agnostic: the LLM prompts that produce the latent state and turn-by-turn dialogue accept a target language as an argument, and the TTS stage uses Qwen3-TTS (Qwen Team, 2026), which supports voice cloning across all of its training languages. To verify this, using language-matched voice pools and the same demographic-matching procedure as the English release, we generated workspaces in five additional languages: German, Spanish, French (reference voices from CommonVoice (Ardila et al., 2020)), Portuguese (CORAA (Cândido Junior et al., 2023)), and Korean (Kspon-Speech (Bang et al., 2020)). We do not include these workspaces in the main corpus or evaluation: the human-discrimination study (§3.5) is English-only, so we cannot make comparable realism claims for the non-English data. The main release is therefore English-only; we expect to release non-English workspaces in a follow-up once the corresponding validation has been run.

## S Annotation interface

Figure 11 shows the recorded-trial view of the web interface used by annotators in the human-discrimination study (§3.5). Annotators were recruited from a pool of professionals working in the meeting-transcription domain. Before the recorded trials began, each session presented a written instructions panel followed by three practice trials with immediate correctness feedback (an extreme attention-check-style fake, a realistic OSYNTH fake, and a Haiku-w/o-obfuscation fake, each paired against the same real excerpt); practice responses were not stored. The verbatim instructions text is reproduced below. Each recorded trial then presents two length-matched transcript excerpts side by side (one real, one synthetic), with speaker labels replaced by generic identifiers (*Speaker 1*, *Speaker 2*, ...) to prevent name-based shortcuts and a word-count badge above each excerpt to make the length-matching transparent. A two-button forced-choice control beneath the excerpts records the annotator’s selection of which excerpt is LLM-generated. Trial progress (e.g., “Item 6 of 20”) is shown at the top of the panel, and attention-check trials are interleaved with realistic and Haiku-w/o-obfuscation trials in a fixed per-session order.

**Instructions shown to annotators.** The instructions panel preceding the practice and recorded trials read, verbatim:

Each item shows two short transcript excerpts of similar length. One excerpt is actual meeting audio manually transcribed by a person. The other is an LLM-generated transcript or a deliberately modified attention-check transcript.

Your job is to select the LLM-generated transcript.

Do not decide from domain terms, project names, or acronyms alone. An actual meeting can involve a fictional or roleplayed task, and an LLM-generated transcript can include realistic technical vocabulary.

Some recorded items are attention checks. If one option clearly contains task-aware or deliberately edited text, select that option.

The examples below are practice only and are not saved. You will get immediate feedback before the recorded trials begin.

You should spend no more than 2 minutes on each recorded item. You can choose after 20 seconds.

## T Nicknames

Real workspace conversation rarely uses canonical full names for familiar colleagues, and ASR

systems whose post-correction is biased toward an org-chart roster of canonical names will systematically miss the informal forms that actually appear on the audio. To make this dimension measurable, every persona in OSYNTH carries an explicit nickname policy that is propagated through both the latent state and the rendered transcript.

Each persona record carries two nickname fields in addition to its canonical name: an optional nickname string and a `nickname_usage` policy drawn from {none, sometimes, mostly}. The distribution is controlled at workspace-sampling time by the people-sketch prompt: most personas receive none, a minority receive a plausible informal nickname with sometimes, and a smaller minority receive one with mostly. The transcript-generation prompt receives the policy as part of the attendee roster and is asked to vary how other speakers refer to a person accordingly: mostly personas are referred to by nickname in nearly all informal turns, sometimes personas appear under a natural mix of canonical and informal forms, and none personas always appear by canonical name.

Identity remains anchored to the stable `person_id` throughout the pipeline: the transcript LLM only emits (`speaker_id`, `text`) pairs, and the human-facing `speaker_name` label is derived downstream (nickname when `usage = mostly`, canonical otherwise). Downstream artifacts (summaries, action items, emails, documents) continue to be addressed by the same `person_id`, so a nickname is never used as an alias key in the data model.

This separation is what keeps name scoring well-defined under either convention. The release carries both the canonical names from the org chart and persona profile and whichever forms (canonical or nickname) actually appear in the oracle transcript for a given meeting, so a scorer can be required to match either. The two diverge most for personas with `usage = mostly`, where a system that recovers only stored names misses the forms that were actually spoken.

## U Gemini: isomorphic vs. standard metrics

Table 12 reports the L1→L3 breakdown of isomorphic and standard metrics for the Gemini stack, complementing the OpenAI-side breakdown in Table 4. The qualitative pattern matches OpenAI: gains concentrate on the entity-sensitive isomor-

Which transcript excerpt is LLM-generated?
Item 6 of 20, transcript excerpts

Transcript A
249 words

Transcript B
249 words

Speaker 1: Actually what's the I in TFIDF? Really? Okay.

Speaker 2: I did do it once, I do have a Java class that does it for something, I don't know whether it'll work with this. But.

Speaker 1: Oh yeah?

Speaker 3: Alright. Because that would be really handy then, we can actually test it on the user, as opposed to just doing it and not using it. 'Cause if you just have the segmentation, that's great, but we can't we can't compare it to the the hand annotated, you know, the hand segmented tool.

Speaker 1: Alright. Yeah.

Speaker 2: Yeah Yeah Yeah. also key-words gives you a a whole new type of search. You do keyword search. But you could do key-word search could be topic search, can they can be the same thing. Instead it would just search for key-words when it when they you tell him that with topics, but actually get searching with key-words.

Speaker 1: Huh.

Speaker 2: For each do you see what I mean?

Speaker 3: Yeah, that would work. That'd be really nice.

Speaker 2: But I suppose even calculating the the the the what's-its-faces themselves would be too much too long. The easy bit is it's probably the easiest to calculate them based upon in their whole occurrences in in the corpus than it is to calculate them per topic. 'cause you don't have to integrate as much.

Speaker 1: Active authorization scoping, more precise timeline by end of week.

Speaker 2: Okay.

Speaker 1: Partners.

Speaker 2: Good news first First Tier one signed, ladder finalized, two Silver conversations going well. Pipeline feels right. ArcAlliance is solid, the the blockers are operational. Yuna the outside counsel piece, you know, the regulated-market language. I I need it for onboarding.

Speaker 3: One to two weeks. Counsel queue is full, I don't have, a lot of levers.

Speaker 2: So the package slips.

Speaker 3: Yes.

Speaker 2: Fine, I can work around it for the first couple onboards. The other thing is the portal I mean, deal reg needs the CRM backend, and, you know, DeShawn hasn't.

Speaker 4: Marc you cut Marc?

Speaker 5: He's gone.

Speaker 4: Connection?

Speaker 5: Probably. He was, glitching before too.

Speaker 1: Keep moving.

Speaker 4: Playbook library is moving. IR workflows mostly

Selection

A is LLM-generated
B is LLM-generated

You can choose the LLM-generated transcript excerpt.

Figure 11: Web interface used in the 2AFC human-discrimination study (§3.5). Each trial displays a pair of length-matched transcript excerpts (one real, one synthetic), with speaker labels anonymized to generic identifiers and a word-count badge above each excerpt; the annotator selects which excerpt is LLM-generated using the two-button control at the bottom. The transcript textboxes are scrollable.

phic dimensions (assignee accuracy, number recall, technical-term recall, speaker consistency) and are systematically compressed by the standard LLM-as-judge scores. The single non-monotonic cell is LLM-judge faithfulness ( $-0.14$ ), which we explain in §5.2 as a scoring asymmetry against Gemini’s workspace-canonical-name expansions rather than a quality regression.

## V Open-source SpeechBrain baseline

The main-paper evaluation in §4 focuses on the two state-of-the-art single-vendor systems (OpenAI and Gemini). For reproducibility, we additionally evaluate a fully open-source cascade on a subset of the same evaluation set; we report its system description and per-context-level results here so that the main-paper tables can focus on the state-of-the-art comparison.

**System.** The cascade is built around the speechbrain toolkit. CRDNN voice-activity detection (vad-crdnn-libriparty) produces speech regions; wav2vec2 acoustic posteriors

| Metric                 | Type | L1    | L3    | $\Delta$ |
|------------------------|------|-------|-------|----------|
| Number recall          | iso. | 0.913 | 0.983 | +0.07    |
| Technical-term recall  | iso. | 0.756 | 0.802 | +0.05    |
| AI assignee accuracy   | iso. | 0.101 | 0.666 | +0.57    |
| Spk.-consistency rate  | iso. | 0.326 | 0.401 | +0.08    |
| LLM-judge faithfulness | std. | 3.02  | 2.88  | -0.14    |
| LLM-judge coherence    | std. | 4.74  | 4.88  | +0.14    |
| AI description quality | std. | 3.53  | 3.79  | +0.26    |

Table 12: Gemini-stack analog of Table 4: L1→L3 gain on isomorphic latent-state metrics versus standard LLM-as-judge summarization metrics, on a smaller 50-meeting subset of the evaluation data. Same qualitative pattern as OpenAI (isomorphic gains compressed by standard summary-side scores), with a single non-monotonic cell: LLM-judge faithfulness dips slightly at L3 despite uniform improvement on every isomorphic dimension. We explain this as a scoring asymmetry of standard LLM-judge protocols against Gemini’s workspace-canonical-name expansions (§5.2). Assignee accuracy here is scored under exact match; the OpenAI counterpart in Table 4 uses partial-credit substring matching.

(asr-wav2vec2-commonvoice-14-en) are decoded with a KenLM-aware CTC beam search, optionally augmented at L3 with workspace-jargon hot-word biasing through a parallel pyctcdecode BeamSearchDecoderCTC sharing the encoder’s KenLM and beam parameters; ECAPA-TDNN embeddings (spkrec-ecapa-voxceleb) feed both spectral-clustering diarization and, at L3, cosine-similarity speaker identification against per-speaker voice enrollments. Prompted summarization, action-item extraction, and outline generation are then performed on the resulting speaker-labeled transcript by a text LLM (claude-haiku-4-5). The cascade uses the same level-conditioned CONTEXT block as the OpenAI and Gemini systems (§4.2); L3 additionally enables the hot-word biasing and enrollment-based SID stages described above.

**Results.** Table 13 reports the SpeechBrain cascade’s per-context-level numbers on the same metric slice as the state-of-the-art systems in Table 3. WER is roughly 4× higher than the OpenAI stack across all three context levels, and the cascade trails both state-of-the-art systems on every metric reported here. The L3 increment that monotonically helps OpenAI and Gemini does not transfer cleanly to this cascade: enrollment-driven metrics (cross-meeting speaker-identity consistency and action-item assignee accuracy) jump sharply at L3 once SID is enabled, and number recall and action-item description quality also improve, but WER degrades from 0.557 to 0.638 and technical-term recall regresses from 0.361 to 0.302. We do not draw any context-sensitivity claims from these numbers; the cascade is included only as an open reference point that future work can re-run end-to-end without commercial API access.

**Why L3 hurts WER and term recall in the cascade.** The L3 vocabulary-correction pass is implemented as a constrained LLM rewrite over the upstream ASR transcript, prompted to fix workspace-jargon spellings while explicitly preserving non-jargon content (“do not paraphrase, reorder, or summarize”). On the OpenAI and Gemini stacks this prompt sees nearly clean input (WER ~0.10) and the rewrite is essentially a vocabulary patch. On SpeechBrain it sees input at WER ~0.56, which the rewriter treats as noise and silently compresses around: across 10 evaluation meetings (409 turns), the mean L3/L1 word-count ratio per turn is 0.75 (i.e., the rewriter

deletes ~25% of words on average), 31% of turns lose more than 30% of their input words, and 13% lose more than 50%. Because technical-term recall is computed as a substring scan over the rewritten transcript, every deleted ground-truth term costs a hit; WER moves the same direction because deleted reference-aligned words score as deletions. The same pass on stronger upstream ASR is essentially a no-op, which is why the corresponding L3 cells in Table 3 do not show this regression. This points to a deployment-level lesson rather than a benchmark artifact: prompted post-correction passes that are safe at low WER can become silent-summarization stages at high WER, and cascades that pair a weak ASR front-end with an LLM post-corrector should either gate the rewrite on per-segment confidence or replace it with a token-level constrained edit. We leave that to future work and report the unmitigated cascade numbers here.

**Qualitative example.** A representative turn from the SpeechBrain L1 transcript and its L3 vocabulary-corrected counterpart on the same audio shows the failure mode concretely:

**L1 (raw SpeechBrain):** “RACE CONDITIONS ISOLATED IT S THE LOCKING NOW REGISTRY CAN CURRENT WRITES GLOBAL LOCKS OUT P NINETY NINE WAS LIKE ONE TWENTY DEVONSONCIS...”

**L3 (after vocab-correction LLM):** “RACE CONDITIONS ISOLATED LOCKING NOW REGISTRY GLOBAL LOCKS OUT P99 LIKE TWENTY DEVON...”

The rewriter’s intended behavior is visible (P NINETY NINE → P99, a workspace-jargon recovery), but it co-occurs with the silent compression mode the appendix is documenting: function-word spans (IT S THE, CAN CURRENT WRITES, WAS, ONE) and a partially garbled name (DEVONSONCIS → DEVON) are removed without flagging. Because the upstream WER is already ~0.56, the rewriter cannot tell garbled-but-recoverable spans from genuine acoustic noise and discards both, so every deleted reference-aligned word becomes a deletion error in WER and a missed hit in technical-term recall. The same prompt over OpenAI’s ~0.10-WER input has nothing comparable to discard.

## W Robustness to the choice of judge

The judged metrics in §4.3 use claude-sonnet-4-6, which is also the family used to generate the corpus (§3.1). To check

| Metric                   | L1    | L2    | L3    |
|--------------------------|-------|-------|-------|
| WER ↓                    | 0.559 | 0.557 | 0.638 |
| Number recall ↑          | 0.259 | 0.259 | 0.397 |
| Technical-term recall ↑  | 0.364 | 0.361 | 0.302 |
| AI description quality ↑ | 2.08  | 1.88  | 2.13  |
| Assignee accuracy ↑      | 0.012 | 0.017 | 0.465 |
| Speaker consistency ↑    | 0.244 | 0.215 | 0.478 |

Table 13: Open-source SpeechBrain cascade on a 27-meeting single-workspace subset of the evaluation set used in Table 3, and reported as a reproducible open-source reference point only; the subset is smaller than the 238-meeting main evaluation, so these numbers are not directly comparable cell-for-cell. The cascade trails both state-of-the-art systems on every metric reported here. Enrollment-driven metrics (assignee accuracy, speaker consistency) jump sharply at L3 once SID is enabled, but WER and technical-term recall regress at L3 because the L3 vocabulary-correction LLM pass silently summarizes around the cascade’s high-WER upstream output (see prose above).

| Metric                 | L1             | L3             | $\Delta$ |
|------------------------|----------------|----------------|----------|
| Faithfulness           | $4.00 \pm .10$ | $4.43 \pm .10$ | +0.43    |
| Coherence              | $4.99 \pm .02$ | $5.00 \pm .00$ | +0.01    |
| AI description quality | $3.87 \pm .05$ | $4.05 \pm .04$ | +0.18    |

Table 14: Judged metrics recomputed with Qwen3-235B-Instruct as the judge, on the same 238-meeting evaluation set and with the same prompts as Table 4 (95% bootstrap percentile interval, 1,000 resamples). The direction of the L1→L3 effect is preserved; coherence is saturated at the 5.00 ceiling under this prompt, which also truncates its interval.

that the reported context effect is not an artifact of that overlap, we re-ran the judged metrics over the OpenAI system’s outputs on the same 238-meeting evaluation set, with the same prompts and an open-weights judge, Qwen3-235B-Instruct (Yang et al., 2025) served locally through vLLM (Table 14). Faithfulness and action-item description quality improve from L1 to L3 as they do under the Claude judge. Coherence does not: under this prompt the Qwen judge awards 5.0 in the large majority of cases, leaving no headroom, so the +0.01 difference reflects saturation rather than an absence of improvement. We read this as evidence that the L1→L3 improvement does not depend on the judge, whereas the range and calibration of the assigned scores is model-family dependent.

## X Licenses of open-source models and datasets

Table 15 lists the licenses of the open-source models, software packages, and datasets used in the construction and evaluation of OSYNTH. All assets are used in a manner consistent with their stated licenses and intended use. The released OSYNTH corpus itself contains no derivatives of the audio content of these datasets: Common Voice clips are used solely as short voice-cloning prompts that condition Qwen3-TTS (Qwen Team, 2026) at synthesis time, and AMI (Carletta et al., 2006) and NOTSOFAR-1 (Vinnikov et al., 2024) excerpts appear only as the real-meeting side of the human-discrimination study (§3.5) and, in the case of AMI, as a style reference to the adversarial rewriter (§3.4).

For the Common Voice data, we use the English subset of the Mozilla Common Voice *Spontaneous Speech* corpus, version 3.0 (sps-corpus-3.0), chosen over the larger *Scripted Speech* releases because its conversational, unscripted prompts yield voice-cloning references whose prosody better matches the spontaneous register of workspace meetings. The corpus is distributed under CC0 1.0 (Ardila et al., 2020) and, as required by the Mozilla Data Collective terms, we do not attempt to identify speakers and propagate this restriction to downstream users via the OSYNTH release notes.

| Asset                                                               | Type     | License    | Use in OSYNTH                                                                                                     |
|---------------------------------------------------------------------|----------|------------|-------------------------------------------------------------------------------------------------------------------|
| Qwen3-TTS (Qwen Team, 2026)                                         | Model    | Apache 2.0 | Multi-speaker TTS with voice cloning (§3.2)                                                                       |
| SpeechBrain toolkit                                                 | Software | Apache 2.0 | Open-source cascade baseline (Appendix V)                                                                         |
| vad-crdnn-libriparty                                                | Model    | Apache 2.0 | Voice-activity detection in the cascade                                                                           |
| asr-wav2vec2-commonvoice-14-en                                      | Model    | Apache 2.0 | Acoustic posteriors in the cascade                                                                                |
| spkrec-ecapa-voxceleb                                               | Model    | Apache 2.0 | Diarization and SID in the cascade                                                                                |
| pyctcdecode                                                         | Software | Apache 2.0 | CTC beam-search decoding with hot-word biasing (L3)                                                               |
| pyroomacoustics (Scheibler et al., 2018)                            | Software | MIT        | Room-acoustics post-processing (Appendix P)                                                                       |
| meeteval                                                            | Software | MIT        | WER variants in the released evaluation harness                                                                   |
| Common Voice Spontaneous Speech 3.0 – English (Ardila et al., 2020) | Dataset  | CC0 1.0    | Voice-cloning reference clips per persona (§3.2)                                                                  |
| AMI Meeting Corpus (Carletta et al., 2006)                          | Dataset  | CC BY 4.0  | Style reference for the adversarial rewriter and real side of the 2AFC and LLM-judge realism studies (§3.4, §3.5) |
| NOTSOFAR-1 (Vinnikov et al., 2024)                                  | Dataset  | CC BY 4.0  | Real side of the 2AFC realism study (§3.5)                                                                        |

Table 15: Licenses of the open-source models, software packages, and datasets used in the construction and evaluation of OSYNTH. SpeechBrain pretrained checkpoints listed individually inherit the Apache 2.0 license of the SpeechBrain toolkit and are distributed via the corresponding Hugging Face repositories. For NOTSOFAR-1, we follow the CC BY 4.0 license stated in the official GitHub repository.