Measuring Public Participation in Local Government with Audio-Based Speaker Classification

Menglin (Miley) Liu
Division of Computational Social Science, School of Humanities and Social Science
The Chinese University of Hong Kong, Shenzhen
Shenzhen, Guangdong, China
menglinliu@cuhk.edu.cn

J. Sophia Wang
Department of Political Science
Yale University
New Haven, Connecticut, USA
jinghong.wang@yale.edu

Ge Shi
Department of Computer Science
University of California, Davis
Davis, California, USA
geshi@ucdavis.edu

Abstract

Democratic legitimacy rests not only on periodic elections but on ordinary residents having an ongoing voice in the local decisions that affect them—and the public-comment period of city council meetings is where that voice is most directly exercised. Yet we still lack scalable, direct measures of who actually speaks in these meetings—and for how long—which has limited empirical research on grass-root democracy. We show that recent advances in audio processing and multimodal large language models make such measurement possible at scale. To our knowledge, this is the first study to measure public participation in local government directly from meeting audio. We develop and validate an audio-based pipeline that operates without transcripts or external identity records. The pipeline first applies speaker diarization to determine who speaks when. It then uses a multimodal large language model to classify each detected speaker as an official or a public participant and calculates measures including public floor-time share. We apply this approach to approximately 2,600 hours of audio collected from 1,565 meetings held in five U.S. cities between 2017 and 2023. Against human annotations, diarization achieves 89.4% accuracy, while speaker-role classification achieves 87.7% accuracy and a public-participant F1 score of 0.832 on 1,761 speakers. The resulting measures reveal that public participants constitute 29.9% of detected speakers but receive a median floor-time share of only 10.0%. Public floor-time share also decreased from 15.5% before 2020 to 4.2% in 2020–2021 and recovered only partially to 9.6% in 2022–2023. These findings extend previous research showing inequalities in who attends local meetings: even when public participants attend, their presence does not translate into proportional voice. Beyond this study, the approach can be scaled to additional cities and adapted to examine finer-grained roles, topics, interactions, and institutional features of public deliberation.

CCS Concepts

• **Applied computing** → **Law, social and behavioral sciences**; • **Computing methodologies** → *Machine learning*; • **Information systems** → *Data mining*.

Keywords

political participation, local government, city council meetings, speaker diarization, audio classification, measurement, multimodal LLMs

ACM Reference Format:
Menglin (Miley) Liu, J. Sophia Wang, and Ge Shi. 2027. Measuring Public Participation in Local Government with Audio-Based Speaker Classification. In *Proceedings of ACM KDD 2027 (KDD '27).* ACM, New York, NY, USA, 24 pages.

1 Introduction

Local government is where most Americans encounter the state directly. City council meetings—open forums where residents can address their elected representatives—are a fundamental channel of democratic participation [Einstein et al. 2020, 2019]. Yet we know remarkably little about who actually speaks in these meetings, for how long, and how participation patterns vary across cities, time periods, and institutional contexts. Theories of participatory and deliberative democracy hold that legitimate collective decisions require the direct involvement of those they affect, not merely periodic elections [Fung 2004; Mansbridge 1983; Pateman 1970], and public comment is one of the few institutionalized channels through which residents voice preferences to officials between elections.

The value of these forums is contested. A familiar critique, in both scholarship and popular culture, is that the open microphone is captured by a vocal and unrepresentative minority: those with the time, confidence, and resources to attend dominate the floor, amplifying narrow grievances while the broader public goes unheard [Einstein et al. 2020, 2019]. Whether public comment broadens democratic voice or narrows it is, at bottom, an empirical question about who holds the floor and for how long. Answering it at scale would let researchers ask whether these forums systematically privilege some groups over others and how institutional design shapes the distribution of voice—first-order questions for local democracy that remain out of reach.

The obstacle is a measurement barrier. Existing work on political participation has focused overwhelmingly on voting, campaign contributions, and contacting officials [e.g., Brady et al. 1995; Verba et al. 1995]. A growing body of work examines public comment in local government [Einstein et al. 2019; Sahn 2024], but it relies on labor-intensive manual coding of minutes or transcripts—a single meeting runs two to three hours with dozens of speakers—which

KDD '27, San Jose, CA, USA
2027.

confines it to small samples [Einstein et al. 2020]. The record itself is abundant—thousands of meetings are posted publicly as audio and video—but it has been prohibitively costly to convert into measures.

Audio-native multimodal models change what is measurable. This paper develops a measurement instrument that converts public-meeting audio into validated measures of participation. The instrument combines neural speaker diarization (identifying *who* speaks *when*) with a multimodal large language model (MLLM) that classifies each speaker as an official or a public participant directly from the audio signal, exploiting the meeting's own temporal structure to resolve ambiguous cases. We combine each predicted role with the speaker's diarized speaking duration and aggregate these durations by role to estimate public floor-time share. We validate every stage against human annotation and propagate the residual classification error into the reported measures. Applied to 1,565 meetings from five U.S. cities between 2017 and 2023, the instrument yields the first floor-time-share measures of public participation at this scale.

Our contributions are organized around the instrument. **Measurement:** we introduce an audio-native pipeline that measures public participation—floor-time share, not merely headcount—from meeting audio alone, without transcription or linkage to external voter or property records. **Methodology:** we show that the meeting's temporal structure can be exploited as a classification signal through temporal exemplar selection, iterative refinement, and a majority vote across complementary classifiers, and that these supporting components raise accuracy over a strong single-pass audio baseline. **Validation:** we validate diarization and classification against human annotations from five cities, report agreement that approaches the agreement among human annotators themselves, and carry the resulting measurement error through to the substantive estimates. **Domain finding:** using the instrument, we document that public participants receive a far smaller floor-time share than their headcount would suggest, that speaking time is highly concentrated among officials, and that public participation declined over the COVID-19 period. We release the pipeline as open-source software so that the same measurement strategy can be applied to other deliberative settings.

2 Related Work

Political participation and public voice. Research on political participation has traditionally examined voting, campaign contributions, contacting officials, and the resource and mobilization inequalities that structure these activities [Brady et al. 1995; Verba et al. 1995]. More recent work studies public comment in local government, finding that meeting participants are often older, whiter, and more likely to own homes than the populations they represent [Einstein et al. 2020, 2019; Sahn 2024]. This literature establishes who participates, but its reliance on sign-in sheets and manually coded minutes limits its scale and does not directly measure each speaker's floor-time share.

Computational measurement of political speech. Text-as-Data research generally treats transcripts as the primary unit of analysis [Grimmer and Stewart 2013]. When applied to spoken proceedings, this approach requires transcription, which adds computational cost and may reproduce documented disparities in automatic speech recognition [Koenecke et al. 2020]. Recent local-politics pipelines combine speech recognition and diarization to produce transcripts for downstream analysis [Barari and Simko 2023; Martin and Venugopal 2026; Xu et al. 2025]. Although these pipelines enable large-scale content analysis, transcription can omit acoustic, vocal, and turn-taking information that may help distinguish institutional roles.

Audio-as-Data methods. Audio-as-data research analyzes speech directly—pitch, emotional arousal, and prosody [Dietrich et al. 2019; Knox and Lucas 2021; Mestre and Ryan 2025]—exploiting paralinguistic and interactional signals, such as timing, turn-taking, and overlapping talk, that transcription may obscure and that conversation-analytic work treats as politically meaningful [van Burgsteden and te Molder 2022]. However, this literature primarily studies elite speakers whose identities are known in advance, and so does not address classifying unlabeled participants in public meetings. Neural speaker diarization can now identify who speaks when in raw audio [Bredin 2023; Park et al. 2022], while audio-capable multimodal large language models can analyze speech without an intermediate transcript [Gemini Team, Google 2023]. To our knowledge these capabilities have not been combined to measure mass participation in local government at scale; we do so here, using audio for role classification and leaving systematic prosodic analysis to future work.

Measurement gap. Political-science research provides the substantive question of how public voice is distributed, but existing data primarily measure who appears. We address this gap with a validated audio-based measurement pipeline that combines speaker diarization with multimodal speaker-role classification, requiring neither transcripts nor external identity records. The resulting measures extend the study of participation from who attends local meetings to how much speaking opportunity members of the public receive.

3 Data

We study five U.S. cities between 2017 and 2023: Asbury Park, Brockton, Inglewood, Jersey City, and New Brunswick. The sample is purposive rather than representative: cities were selected for publicly available YouTube recordings of council meetings, geographic spread across three states (California, Massachusetts, and New Jersey), variation in council size and meeting format, and sufficient meeting volume for longitudinal analysis. The cities were originally drawn from a sampling frame of municipalities with rent control policies, though the substantive analysis of housing policy is not the focus of this paper. The final analytic corpus contains 1,565 meetings from these five cities; per-city counts are summarized in Appendix C (Table 7).

Sample representativeness. These cities are not a representative cross-section of the roughly 30,000 U.S. general-purpose local governments, and we make no claim that our patterns generalize to all of them. All five analytic cities are dense, urban or suburban, and in Democratic-leaning states; the rent-control sampling frame and the requirement that a city post meetings to YouTube compound this skew, favoring larger jurisdictions with professionalized clerk offices and active civic constituencies, while rural and Republican-dominated jurisdictions are absent. We treat the study as a proof of concept for the measurement instrument rather than

a census of American local democracy (Section 7). Coverage is also uneven over time—the bulk of recordings fall in 2019–2022, with a pandemic-era surge in 2020–2021 and few Brockton recordings before 2020—which we account for in the longitudinal analysis of Section 6.

Meeting recordings were obtained from official city YouTube channels and government websites via automated scraping (Appendix C); all are public records of open meetings, and we assign de-identified speaker labels and classify only by institutional role (Section 8).

After speaker diarization (described in Section 4), the five-city corpus yields 199,027 speech segments from 29,244 meeting-level speaker clusters across 1,565 meetings, totaling approximately 2,600 hours of audio. The application analysis includes 29,238 speakers. The median meeting contains 16 identified speakers (mean = 18.7) and runs 76 minutes, though meetings range from brief special sessions to multi-hour hearings. At the speaker level, each speaker produces a median of 7 speech segments (mean = 6.8). City-level variation is considerable: Jersey City meetings average 30.6 speakers per meeting, reflecting that city's large council and active public comment periods, while Brockton averages 13.2. Per-city descriptive statistics are reported in Appendix C.5 (Table 8).

A full data statement for the corpus—consolidating purpose, provenance, scale, preprocessing, partitions, leakage controls, known biases, and access terms—is provided in Appendix C.5 (Table 9). The supplementary materials provide source manifests, retrieval and preprocessing code, releasable annotations and split files, and the prompts and exemplars used.

Human annotation and reference labels. Three independent research assistants listened to up to three short audio clips per speaker and labeled each speaker as an official or a public participant; disagreements were resolved by majority vote. The retained five-city evaluation set contains $N = 1{,}761$ speakers from 165 meetings, comprising 1,069 officials and 692 public participants (39.3% public-participant rate). Mean inter-rater agreement across the retained annotation batches is Fleiss' $\kappa = 0.723$ (range 0.525–0.868). Appendix F provides the complete protocol, reliability statistics, and per-city counts.

4 Method

We first state the classification problem formally, then describe the three-stage pipeline that solves it.

Problem statement. Let $\mathbf{x} \in \mathbb{R}^T$ denote a raw audio recording of duration T. Speaker diarization partitions $\mathbf{x}$ into segments $\{(s_i, e_i, k_i)\}_{i=1}^N$, where s_i and e_i are start and end times and $k_i \in \mathcal{K} = \{1, \ldots, K\}$ is an anonymous, meeting-specific speaker-cluster label. No speaker roster, identity labels, or value of K is supplied: K is the number of distinct speaker clusters inferred by diarization. The classification task is to learn a mapping $f : \mathcal{K} \to \{0, 1\}$, where 0 = *official* and 1 = *public participant* (a member of the public; we adopt this operational label rather than "citizen" because legal citizenship is not observed). Given M classifiers, we combine their per-speaker predictions by majority vote (our default), or by AND/OR rules that trade precision against recall; Appendix J gives the formal definitions of these combination rules.

Our pipeline consists of three stages: (1) speaker diarization, (2) per-speaker audio preparation, and (3) speaker classification, with a bootstrapping loop feeding predicted labels back as exemplars. Figure 1 summarizes the pipeline; Appendix A provides the algorithms.

4.1 Stage 1: Speaker Diarization

Speaker diarization is the task of partitioning an audio recording into segments labeled by speaker identity—answering "who spoke when." We use `pyannote.audio` 3.1 [Bredin 2023], an end-to-end neural diarization framework that jointly performs voice activity detection, speaker embedding extraction, and agglomerative clustering within a single pipeline. We choose `pyannote` over cascaded alternatives [Park et al. 2022] because it reaches competitive diarization error rates on standard benchmarks, handles overlapping speech—common when officials interrupt or the public speaks over the chair—without a separate overlap-detection module, and is free, open-source, and runs on a single consumer-grade GPU. Appendix C.4 details the hardware.

Before diarization, each audio file is resampled to 16 kHz mono via `librosa` to match the model's expected input format. After diarization, we discard segments shorter than 0.5 s, which typically correspond to microphone noise, coughs, or brief crosstalk rather than substantive speech. For each meeting, diarization produces a set of segments $\{(s_i, e_i, k_i)\}_{i=1}^N$, where s_i and e_i are the start and end times of segment i in seconds, and $k_i \in \{1, \ldots, K\}$ is the assigned speaker label.

As a transparent baseline for the MLLM, we also extract a 36-dimensional prosodic/spectral feature vector per segment with `librosa` (pitch, energy, zero-crossing rate, spectral shape, and MFCCs, each summarized by mean and standard deviation; Section 5). We validate diarization against human judgment (Section 5), keeping two quantities distinct: *inter-rater reliability*—agreement among annotators, which is high (mean pairwise Cohen's $\kappa = 0.84$, "almost perfect" [Landis and Koch 1977]) and bounds achievable accuracy—and *accuracy*—agreement between the system and the human consensus, 89.4% across the five validation cities. Residual errors are dominated by overlapping speech merged into one segment. Because the human annotations evaluate the same diarized speaker units classified by the MLLM, these errors are reflected in the end-to-end validation; Appendix G examines how prediction error affects downstream aggregate estimates.

4.2 Stage 2: Per-Speaker Audio Preparation

Speaking time varies substantially across speakers. If audio were sampled uniformly over the meeting timeline, speakers who talk longer would be more likely to appear in the classification inputs, while brief speakers could receive little or no representation. Formally, let $T_j = \sum_{i:k_i=j}(e_i - s_i)$ denote the total speaking time of speaker j. Under uniform temporal sampling, the probability of selecting speech from speaker j is proportional to T_j. We avoid this imbalance by constructing one audio input separately for every detected speaker.

For each speaker, we arrange all diarized segments chronologically and concatenate them into a speaker-level audio sequence. If $T_j \leq 30$ s, we retain the complete sequence. If $T_j > 30$ s, we extract

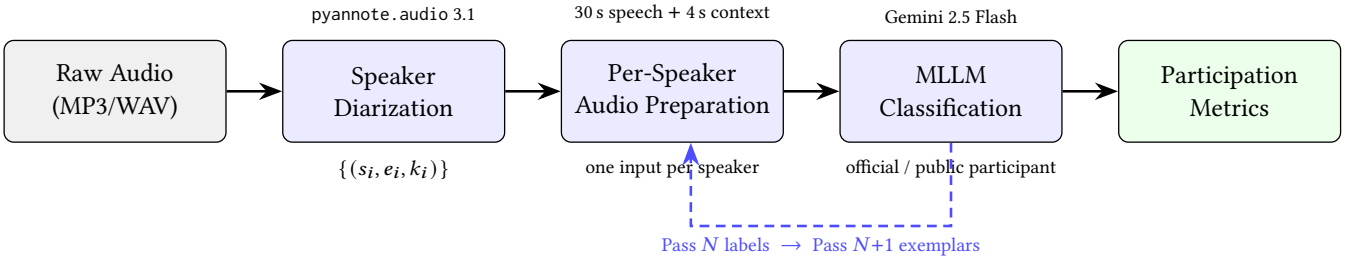

Figure 1: Audio-based measurement pipeline. Raw meeting audio is diarized into speaker segments, assembled into one bounded audio input per speaker, classified by the MLLM, and aggregated into participation measures. The dashed arrow shows the bootstrapping update described in Section 4.3.4.

10 s from its beginning, middle, and end and concatenate these excerpts chronologically. This limits the speaker's own audio to 30 s while representing different parts of their participation.

We also add up to 2 s of original meeting audio before the speaker's first segment and after the last segment, subject to the recording boundaries, to provide brief context for classification.

4.3 Stage 3: Speaker Classification

We classify each speaker as an *official* (elected representative or government staff) or a *public participant*(member of the public; ambiguous roles are coded by institutional role, Appendix F) using Google Gemini 2.5 Flash [Gemini Team, Google 2023], an MLLM that processes audio directly via its Files API.[1] All API calls use temperature = 0 to reduce sampling variability. The exact model identifier, API settings, retry and unparsable-response handling, run dates, token counts, and cost are in Appendix C.4. We use zero-shot classification as a diagnostic baseline and evaluate three production strategies: fixed 5-shot exemplars, bootstrapping with temporal exemplars, and sliding-window classification. We then combine their predictions using the ensemble rules defined in Appendix J.

4.3.1 Baseline: Zero-Shot and Few-Shot Classification. In the **zero-shot** setting the model receives a single per-speaker audio input and a prompt that defines the two roles and returns a one-word label, relying only on its pretraining. The **few-shot** setting prepends k labeled audio exemplars before the query; varying k, class balance, and wording on development data (Appendix D), the best configuration is $k = 5$ (4 officials + 1 public participant, matching the official-heavy base rate) with a minimal prompt. The response is parsed for the class token,[2] and unparsable outputs are dropped (per-strategy rates in Appendix C.4).

4.3.2 One Label per Speaker. Each method returns a single label per speaker: a speaker's representative clips (up to three) are sent *together* in one classification call, so the model integrates them internally rather than voting over separately-classified clips. Presenting multiple clips jointly reduces sensitivity to any single noisy clip and to diarization merge errors (where two speakers are grouped under one label).

4.3.3 Temporal Exemplar Selection. Rather than a fixed, meeting-external pool, we select exemplars by *temporal adjacency* within the same meeting. Because council meetings have role-clustered turn-taking (procedural business runs official-only; public comment alternates between officials and public participants or contains runs of consecutive public participants), a query's nearest-in-time neighbors are disproportionately the same role and highly informative about it—even when the speaker is ambiguous in isolation. When public participation is very sparse this can bias predictions toward "official" (observed in one official-heavy meeting; Appendix H), so we use temporal context only within the ensemble (Section 4.3.6) and test its effect in Section B.3. Algorithm 1 ranks the other speakers by first-appearance time—not per-turn adjacency, since officials recur throughout—and takes the $k = 4$ nearest; each contributes its audio and previous-pass label (Appendix A).

4.3.4 Bootstrapping: Iterative Refinement Without Oracle Labels. Temporal exemplar selection faces a chicken-and-egg problem: to classify a speaker from the labels of their temporal neighbors, we need those neighbors' labels—but those are precisely what we are trying to predict. We resolve this circular dependency through bootstrapping (Algorithm 2). In Pass 0, we label every speaker with the meeting-external baseline classifier, which ignores neighbors. In each subsequent pass, we re-classify every speaker using the *previous* pass's predicted labels for that speaker's temporal neighbors as exemplars. We run three update passes and stop earlier if no labels change; Section 5 reports the accuracy and number of label changes at each pass.

Appendix B (Figure 4) illustrates the iterative loop; related work uses repeated prompting with prior outputs as context [Sun et al. 2024].

4.3.5 Sliding-Window Context Classification. A third supporting component classifies each speaker from a small window of surrounding context. We order the meeting's *unique speakers* by first appearance—$[s_1, \ldots, s_N]$, the same per-speaker inputs produced in Stage 2, not raw diarization segments or speaking turns—and slide a length-three window across this speaker sequence, classifying the center speaker s_i from the audio of s_{i-1}, s_i, s_{i+1} (formalized in Appendix J). Figure 5 illustrates the procedure.

[1] At the time of analysis Gemini 2.5 Flash was among the few production MLLMs accepting raw audio. The classifier is an interchangeable component; we report absolute accuracies for a fixed model and release code so the benchmark can be re-run as models improve.

[2] The deployed prompt and parser use the word *citizen* as the class token. Throughout the paper we report this class as *public participant* (member of the public), since legal citizenship is not observed and the label denotes an institutional role, not legal status.

We evaluate two variants:

- **Variant 2a (audio only):** For each query speaker s_i, the model receives the audio of $[s_{i-1}, s_i, s_{i+1}]$ in chronological order and classifies the middle speaker; no neighbor labels are provided.
- **Variant 2b (audio + labels):** Same as 2a, but the model also receives the Pass 3 bootstrapped labels for the two neighbors (e.g., "Audio 1 was classified as: official"), combining audio context with prior predictions.

Adjacent windows overlap by two of their three speakers, so turn-taking patterns (e.g., the official → public → official transition) are detectable at every position. The first and last speakers use one-sided context, and because the sequence is over unique speakers each is classified in exactly one window—one label per speaker.

4.3.6 Ensemble Methods. The three production strategies—5-shot baseline, Pass 3 bootstrapping, and window 2b—make partially complementary errors, which we combine with majority-vote, AND, and OR rules. Window 2b conditions on Pass 3's neighbor labels, so the two are correlated and the vote gives Pass-3-aligned signal some extra weight; we retain the ensemble because window 2b is not a copy of Pass 3 (it also reads the neighbors' audio and disagrees with it on a substantial share of speakers) and because the label-free 5-shot classifier is an independent anchor. A leave-one-out check confirms both dependent methods contribute—the three-way majority reaches 87.7%, versus at most 86.7% when window 2b is dropped and 85.6% when Pass 3 is dropped (Appendix J). Majority vote is our default; the AND and OR rules provide alternative precision–recall tradeoffs for applications that weight false positives and false negatives unequally. Table 1 reports all pairwise ensembles; the pairs we cite—5-shot ∩ window 2b for peak precision and Pass 3 ∪ window 2b for peak recall—were selected *after* inspecting validation performance and characterize the frontier, whereas majority vote is the only rule fixed *a priori*. Performance across all methods and ensembles—evaluated with public-participant precision, recall, and F1—is reported in Section 5.

5 Validation

5.1 Data Partitions and Leakage Controls

We distinguish three data roles. A development set from Albany and Chino (4 meetings, approximately 51 speakers) was used for prompt development and diarization checks. The five-city human-labeled evaluation set contains 1,761 speakers and is used to report classification performance; it was also used to select the three-pass bootstrapping cap and majority-vote rule. The unlabeled application corpus contains 29,238 speakers and is held out from all pipeline-selection decisions. The classifier fits no task-specific model weights, but the prompt, fixed exemplar pool, number of bootstrapping passes, and ensemble rule are data-dependent operating choices. Appendix B (Table 4) records each stage and its overlap with the evaluation set. The principal exposure is that pass and ensemble selection use the same labeled set as final evaluation, so the reported operating-point accuracies may be mildly optimistic. We limit hindsight selection by using a fixed three-pass cap and majority vote rather than selecting the highest reported accuracy; temporal exemplars use predicted, not human, neighbor labels. Independent verification against rosters, agendas, minutes, or speaker cards remains an important next step.

Why Pass 3, when Pass 2 scores higher? Table 5 shows Pass 2 edging Pass 3 on the labeled set (86.9% vs. 86.7% accuracy; F1 0.826 vs. 0.824)—a gap of roughly three speakers, within noise. We use a fixed cap of three update passes rather than selecting the highest-accuracy pass on the same set used for reporting. Label changes decrease across passes (171 → 93 → 83), and the net change from Pass 2 to Pass 3 is near zero. Pass 3 is then included as one of the three inputs to the majority-vote ensemble. This rule avoids choosing Pass 2 solely by hindsight, although selecting the three-pass cap on the evaluation set remains a limitation.

5.2 Diarization Accuracy

Human validation on 483 speaker-pair judgments across five cities (the Chino development meetings and four production cities) shows the diarization stage tells distinct speakers apart with 89.4% overall accuracy (85.4% same-speaker, 94.1% different-speaker), ranging from 85.0% in Asbury Park to 94.2% in Jersey City; the per-city breakdown (Table 11), full protocol, and inter-rater reliability are in Appendix E.

Standard diarization metrics. The diarization literature also reports *Diarization Error Rate* (DER) and *Jaccard Error Rate* (JER), which require dense, timestamp-level reference annotations; three research assistants densely annotated seven complete meetings (five double-annotated) and we scored pyannote with `pyannote.metrics`. Pyannote achieves a speech-weighted DER of 7.6% and JER of 21.9% (Appendix E.6, Table 13), consistent with the 89.4% pairwise agreement and competitive with reported benchmark accuracy. Errors concentrate almost entirely in Brockton's noisier subcommittee audio (DER 17–22%; other cities ≤6.8%), yet Brockton still attains among the highest role-classification accuracy, because the clip-level majority vote absorbs the diarization splits and merges that inflate JER rather than propagating them to speaker-role labels.

5.3 MLLM Classification Accuracy

We evaluate the three production methods—5-shot (fixed exemplars), Pass 3 bootstrapping, and window 2b (audio + labels)—and their ensembles on the $N = 1{,}761$ speakers with human-consensus reference labels described in Section 3; all runs use Gemini 2.5 Flash at temperature 0. Table 1 reports the full results. Among individual methods, Pass 3 is best (86.7% accuracy, public F1 = 0.824), consistent with temporal context adding signal, with window 2b (85.2%) and the 5-shot baseline (84.7%) close behind. Because the three make complementary errors—the 5-shot baseline favors precision (0.878/0.710) while Pass 3 balances precision and recall (0.859/0.792)—their **majority vote** is the best overall method: **87.7% accuracy,** $\kappa = 0.736$**, public F1** = 0.832. The AND and OR rules provide additional precision–recall tradeoffs: AND (5-shot ∩ window 2b) gives the highest public precision (0.930), OR (Pass 3 ∪ window 2b) the highest recall (0.854), each at a cost on the other axis (definitions in Appendix J).

Per-pass stability and error structure. Accuracy rises from 84.8% (Pass 0) to 86.9% by Pass 2 and is 86.7% at Pass 3, while per-pass label

Table 1: Classification Performance on the Human-Labeled Evaluation Set (Gemini 2.5 Flash, N = 1,761 speakers, 5 cities)

Method	N	Accuracy	κ	Pub. P	Pub. R	Pub. F1
Individual methods						
5-shot (fixed exemplars)	1,761	0.847	0.669	0.878	0.710	0.785
Pass 3 (bootstrapped)	1,761	0.867	0.718	0.859	0.792	0.824
Window 2b (audio + labels)	1,761	0.852	0.685	0.844	0.766	0.803
AND ensembles						
AND: 5-shot ∩ Pass 3	1,761	0.848	0.665	0.924	0.668	0.775
AND: 5-shot ∩ Window 2b	1,761	0.844	0.654	**0.930**	0.652	0.766
AND: Pass 3 ∩ Window 2b	1,761	0.859	0.692	0.919	0.704	0.797
OR ensembles						
OR: 5-shot ∪ Pass 3	1,761	0.867	0.721	0.828	0.834	0.831
OR: 5-shot ∪ Window 2b	1,761	0.856	0.698	0.812	0.824	0.818
OR: Pass 3 ∪ Window 2b	1,761	0.860	0.710	0.803	**0.854**	0.828
Majority vote						
Majority (2/3)	1,761	**0.877**	**0.736**	0.895	0.777	**0.832**

changes fall 171 → 93 → 83 (Appendix B, Table 5). Across the three updates, the corrections are net-positive and disproportionately rescue true public participants misclassified as officials (127 of 190 fixes), unwinding the zero-shot majority-class bias. Window 2b's errors are partially orthogonal to those of the other two methods, which is what makes it a useful third ensemble member despite its lower solo accuracy.

Performance varies across cities: the majority-vote ensemble achieves 90.8% accuracy in Brockton (N = 283) and 89.0% in Inglewood (N = 318), with public F1 ranging from 0.711 in Brockton (where the low public-participant rate of 17.3% in the human-labeled sample makes classification harder) to 0.873 in Jersey City (N = 534). These per-city differences underscore the value of evaluating on geographically diverse human annotations rather than a single development site.

5.4 Comparison with Supervised Audio Baselines

A natural question is whether the MLLM is doing real work, or whether much simpler audio methods would classify speakers just as well. To answer it we ask: how well can a conventional, purpose-trained classifier do on the same task, using only measurable properties of the audio signal? This matters because the simpler methods are cheaper and fully transparent; if they matched the MLLM, the added complexity and opacity of a large model would be hard to justify. Appendix B (Table 6) reports the full comparison, and Appendix H reports the leading feature-importance rankings.

Alternative input modalities and models. A prior question is whether the input *modality*—audio—is doing the work, or whether transcripts, video, or an open model would classify speakers just as well. Table 2 compares the audio-native pipeline against three single-pass alternatives, each stripped of our temporal refinements to isolate the effect of input modality and model. Three findings stand out. First, transcription is not lossless: an ASR-plus-text pipeline (Whisper-large-v3 → Gemini) reaches 81.8%, below the audio-native model on the same speakers (85.3%), and degrades further on noisy public-comment audio (Appendix I). Second, video alone is weaker still: 75.7% accuracy, and only for the 80% of speakers the camera actually frames (it holds a fixed dais shot for the rest; the on-screen rate ranges 45–100% by city). Third, the leading open audio-LLM (Qwen2-Audio-7B) collapses to a single class and provides no discriminative signal—despite transcribing the audio accurately—indicating that audio-native role classification is not yet reproducible with open models, even as our closed-model pipeline reaches 87.7%. Our temporal refinements add a further +3 points over the audio-native base (84.7%→87.7%).

The supervised baselines require hand-crafted acoustic features and pretrained speaker embeddings, which were extracted for the subset of the validation speakers drawn from annotation rounds R2 and R4–R7 (N = 1,081; 653 officials, 428 public participants). We therefore report this comparison on that common N = 1,081 subset, on which the MLLM majority-vote ensemble scores 89.3% (versus 87.7% on the full N = 1,761 set); the two rates differ only because the subset omits the later, more public-comment-heavy rounds, and the ranking of methods is unchanged. Appendix B (Table 6) reports the comparison on these N = 1,081 speakers. Two families of supervised methods are evaluated: prosodic feature classifiers (79 hand-crafted acoustic and temporal features aggregated from diarization output) and pretrained speaker embedding classifiers (ECAPA-TDNN and Wav2Vec2). The best supervised model is the prosodic Gradient Boosting classifier at 80.3% accuracy (κ = 0.580, public F1 = 0.736), followed by ECAPA-TDNN + Random Forest at 77.4% (κ = 0.510). All supervised baselines fall substantially below the zero-shot MLLM 5-shot method (85.3%) and the majority-vote ensemble (89.3%).

6 Application

We apply the full pipeline to 1,565 meetings across five cities (Asbury Park, Brockton, Inglewood, Jersey City, and New Brunswick),

Table 2: Input-modality and model comparison. Alternatives are single-pass to isolate modality and model effects; sample sizes differ with source availability.

Method	Input	Model	N	Acc.
Open audio-LLM	audio	Qwen2-Audio-7B (open)	599	—[a]
Video-only MLLM	video frames	Gemini 2.5 Flash	381	0.757[b]
ASR → text LLM	transcript	Whisper-lg-v3 + Gemini	593	0.818
Audio-native base (5-shot)	audio	Gemini 2.5 Flash	1,761	0.847
Audio-native + bootstrapping	audio	Gemini 2.5 Flash	1,761	0.867
Audio-native + majority ensemble	audio	Gemini 2.5 Flash	1,761	**0.877**

[a]The open model collapses to one class. [b]Accuracy among speakers shown on screen. On the shared 593-speaker ASR subset, audio-native classification scores 0.853 versus 0.818 for ASR followed by text classification. Detailed matched-subset and audio-quality analyses appear in Appendix I.

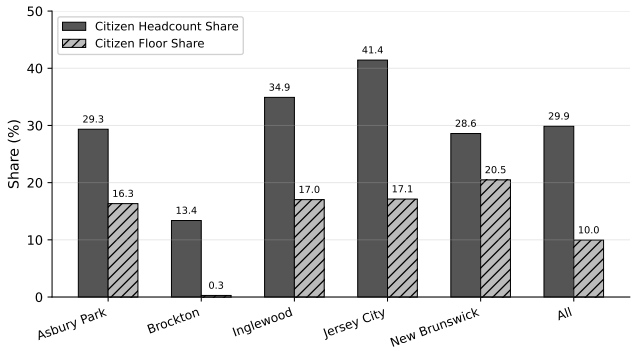

Figure 2: The Participation Gap: public headcount share vs. public floor-time share by city. In every city, the floor-time share allocated to public participants falls well below their share of speakers. The aggregate gap—29.9% of speakers accounting for 10.0% of speaking time—indicates that "showing up" does not translate into proportionate voice.

classifying 29,238 unique speakers.[3] Table 3 summarizes the cross-city results.

Overall, 70.1% of detected speakers are officials and 29.9% are public participants (observed plug-in proportions; Appendix G reports a pooled bias correction for public-participant prevalence). But the gap in floor-time share far exceeds the headcount gap: the median public floor-time share is just 10.0%, so nearly a third of speakers account for about a tenth of speaking time (Figure 2). Concentration is pronounced—the top speaker is an official in 94.2% of meetings and the top three in 76.5%—and public participants enter later (median 48% mark vs. 25% for officials). This extends Einstein et al.'s (2019) *who shows up* to *how much room they get*: substantially less than headcounts suggest.

Cross-city variation illustrates the pipeline's ability to detect institutional differences (reported descriptively, not causally). Public floor-time share ranges from 20.5% in New Brunswick to just 0.3% in Brockton—a nearly 70-fold difference. Brockton's near-zero figure appears substantively real rather than an artifact: its corpus is dominated by school-committee and subcommittee sessions (Section 3) with little open public comment, its classification accuracy is in line with other cities (Table 1), and its public headcount share (13.4%) is low but non-zero—so the few public participants who speak are detected but account for little speaking time. Conversely, Jersey City has the highest public headcount share (41.4%) yet a public floor-time share of only 17.1% (25.8% of its meetings have zero public speakers). These differences are consistent with institutional design—council size, public-comment placement and duration, and meeting format—which the pipeline can now detect at scale.

The COVID-19 period illustrates the pipeline's capacity for longitudinal measurement. Public speaker share dropped from 27.4% pre-pandemic (2017–2019) to 12.5% during 2020–2021, with only partial recovery to 18.2% afterward (2022–2023). The decline in public floor-time share was even steeper: from 15.5% to 4.2%, with recovery to just 9.6% (Figure 3). These patterns are consistent with Einstein et al.'s (2022) evidence that virtual meetings did not diversify participation. We describe them as longitudinal patterns rather than causal estimates: the recovery was incomplete as of 2023, but our design does not identify the pandemic as their cause. Martin and Venugopal [2026], exploiting the staggered withdrawal of remote-participation options across California cities, provide complementary quasi-experimental evidence that meeting access costs shape the volume of public participation. The figures report point estimates. Appendix G.1 states the assumption required to transport the classifier's error rates across periods and identifies city- and period-specific, meeting-clustered, and duration-weighted uncertainty analyses as necessary refinements.

7 Discussion and Conclusion

We have developed a measurement instrument that distinguishes officials from public participants using meeting audio alone—neural diarization plus an audio-native MLLM that exploits the meeting's temporal structure—reaching 87.7% accuracy against human-consensus reference labels from five cities. Applied to 1,565 meetings it yields participation measures (large gaps between public headcount and floor-time share, cross-city variation, and a COVID-era decline) that manual coding could not produce at scale.

These results extend Einstein et al.'s (2019) finding that local meetings are dominated by a narrow, unrepresentative public along a new dimension: where prior work measured *who shows up*, our measure captures *how much room they get*. The near-70-fold spread

[3]Classification uses the majority-vote ensemble of Pass 3, 5-shot, and window 2b labels for each speaker.

Table 3: Cross-City Summary of Speaker Classification Results

City	Meetings	Speakers	Officials (%)	Public participants (%)	Public Floor-Time Share (%)	Gini
Asbury Park	364	7,270	70.7	29.3	16.3	0.541
Brockton	499	6,600	86.6	13.4	0.3	0.531
Inglewood	141	2,346	65.1	34.9	17.0	0.456
Jersey City	299	9,143	58.6	41.4	17.1	0.548
New Brunswick	262	3,879	71.4	28.6	20.5	0.529
All	**1,565**	**29,238**	**70.1**	**29.9**	**10.0**	**0.533**

Note: Public Floor-Time Share is the median percentage of total meeting speaking time attributable to public participants. Gini is the median within-meeting Gini coefficient of speaking time across all speakers.

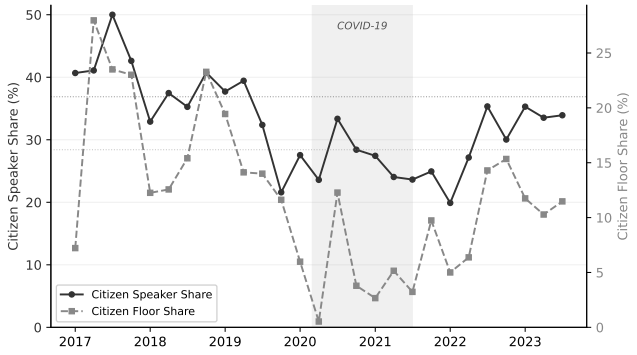

Figure 3: Public participation over the COVID-19 period: quarterly public speaker share and public floor-time share, 2017–2023. The shaded region marks the pandemic period (March 2020 – June 2021). Dotted lines indicate pre-COVID baselines. Public participation dropped sharply during the pandemic and had not fully recovered by 2023.

in public floor-time share across our five cities is too large to be demographic noise; it is consistent with design features that structure voice (comment format, time limits, council size, remote participation). Measuring these shares at scale lets researchers move from documenting that participation is unequal toward explaining *why*—the goal of ongoing work the instrument makes possible.

Methodologically, we bring the "Audio-as-Data" paradigm [Grimmer and Stewart 2013; Mestre and Ryan 2025] from elite speech [Dietrich et al. 2019; Knox and Lucas 2021; Tarr et al. 2023] to mass participation, showing that role classification needs no transcription and matches transcription-based pipelines [Martin and Venugopal 2026] without their cost or bias [Koenecke et al. 2020]. Three insights emerge: meeting structure is an informative signal (temporal context yields the largest gains); bootstrapping is net-corrective across the evaluated passes (rescuing public participants misclassified as officials); and the classifier generalizes across cities without retraining (leave-one-city-out accuracy 89.5% vs. 70.0% for prosodic baselines; Section B.3).

Natural extensions include scaling to the LocalView corpus [Barari and Simko 2023] for national coverage, coding institutional rules to estimate (with a causal design) which features expand public floor-time share, and finer-grained speaker taxonomies via end-to-end models such as SpeakerLM [Yin et al. 2025]. The measurement pipeline can in principle be applied to other structured-turn-taking settings—school boards, planning commissions, and utility hearings—subject to validation in each setting.

8 Limitations and Ethical Considerations

The substantive findings are descriptive and derive from five purposively selected urban and suburban cities; they are neither nationally representative nor evidence that the pandemic caused the observed decline. Classification accuracy may also change across institutional formats, accents, recording conditions, and hosted-model versions. Human-consensus labels validate agreement with trained judgment rather than objective role identity, demographic subgroup fairness cannot be assessed with the available annotations, and the audio-versus-text comparison uses a subset of speakers. Appendix K details these scope, validity, and durability limits.

The recordings are public records of open meetings, and the project received an institutional determination that it did not constitute human-subjects research. Public availability nevertheless does not eliminate privacy or misuse risks associated with identifiable voices. We use de-identified speaker codes, classify institutional role rather than identity or demographic attributes, do not redistribute raw audio, and restrict intended use to aggregate research and institutional comparison. Paid-tier Gemini processing and transient Files-API storage introduce third-party data handling, while unequal performance across accents or audio conditions could bias results. Identification, profiling, surveillance, or re-contact of individual speakers is outside the validated use of the pipeline; data-governance and fairness details appear in Appendix K.

9 Generative AI Usage

Generative AI is integral to the method: the instrument classifies speaker roles with Google `gemini-2.5-flash` (Sections 4–5, Appendix C.4). Beyond this, the authors used **Claude Code** to assist with software development and to proofread the manuscript. No generative AI generated, fabricated, or altered research data, results, or scientific claims: all measurements were produced by the authors' code and validated against human-consensus reference labels, and all interpretive claims are the authors' own. Generative AI also assisted in refining the manuscript's writing, while the authors remain fully responsible for all content.

References

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *Advances in Neural Information Processing Systems*, Vol. 33. 12449–12460.

Soubhik Barari and Tyler Simko. 2023. LocalView, A Database of Public Meetings for the Study of Local Politics and Policy-Making in the United States. *Scientific Data* 10, 1 (2023), 135.

Henry E. Brady, Sidney Verba, and Kay Lehman Schlozman. 1995. Beyond SES: A Resource Model of Political Participation. *American Political Science Review* 89, 2 (1995), 271–294.

Hervé Bredin. 2023. pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In *Proc. Interspeech*.

Marquis de Condorcet. 1785. *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. Imprimerie Royale, Paris.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. In *Interspeech*. 3830–3834.

Bryce J. Dietrich, Matthew Hayes, and Diana Z. O'Brien. 2019. Pitch Perfect: Vocal Pitch and the Emotional Intensity of Congressional Speech. *American Political Science Review* 113, 4 (2019), 941–962.

Naoki Egami, Musashi Hinck, Brandon M. Stewart, and Hanying Wei. 2023. Using Imperfect Surrogates for Downstream Inference: Design-based Supervised Learning for Social Science Applications of Large Language Models. In *Advances in Neural Information Processing Systems*, Vol. 36.

Katherine Levine Einstein, David M. Glick, and Maxwell Palmer. 2020. *Neighborhood Defenders: Participatory Politics and America's Housing Crisis*. Cambridge University Press, Cambridge.

Katherine Levine Einstein, David M. Glick, Maxwell Palmer, and Luisa Pei. 2022. Still Muted: The Limited Participatory Democracy of Zoom Public Meetings. *Urban Affairs Review* 58, 6 (2022), 1595–1622.

Katherine Levine Einstein, Maxwell Palmer, and David M. Glick. 2019. Who Participates in Local Government? Evidence from Meeting Minutes. *Perspectives on Politics* 17, 1 (2019), 28–46.

Archon Fung. 2004. *Empowered Participation: Reinventing Urban Democracy*. Princeton University Press, Princeton, NJ.

Gemini Team, Google. 2023. Gemini: A Family of Highly Capable Multimodal Models. *arXiv preprint arXiv:2312.11805* (2023).

Justin Grimmer and Brandon M. Stewart. 2013. Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis* 21, 3 (2013), 267–297.

Dean Knox and Christopher Lucas. 2021. A Dynamic Model of Speech for the Social Sciences. *American Political Science Review* 115, 2 (2021), 649–666.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial Disparities in Automated Speech Recognition. *Proceedings of the National Academy of Sciences* 117, 14 (2020), 7684–7689.

J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33, 1 (1977), 159–174.

Jane J. Mansbridge. 1983. *Beyond Adversary Democracy*. University of Chicago Press, Chicago.

Olivia Martin and Amar Venugopal. 2026. Participation and Representation in Local Government Speech. *arXiv preprint arXiv:2604.21202* (2026).

Rafael Mestre and Matt Ryan. 2025. Potential and Pitfalls of Audio as Data for Political Research. *Political Analysis* (2025).

Tae Jin Park, Naoyuki Kanda, Dimitrios Dimitriadis, Kyu J. Han, Shinji Watanabe, and Shrikanth Narayanan. 2022. A Review of Speaker Diarization: Recent Advances with Deep Learning. *Computer Speech & Language* 72 (2022), 101317.

Carole Pateman. 1970. *Participation and Democratic Theory*. Cambridge University Press, Cambridge.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust Speech Recognition via Large-Scale Weak Supervision. In *International Conference on Machine Learning (ICML)*. 28492–28518.

Alexander Sahn. 2024. Public Comment and Public Policy. *American Journal of Political Science* (2024). Forthcoming.

Jiashuo Sun, Yi Luo, Yeyun Gong, Chen Lin, Yelong Shen, Jian Guo, and Nan Duan. 2024. Enhancing Chain-of-Thoughts Prompting with Iterative Bootstrapping in Large Language Models. In *Findings of the Association for Computational Linguistics: NAACL 2024*. 4074–4101.

Alexander Tarr, JungHwan Hwang, and Kosuke Imai. 2023. Automated Coding of Political Campaign Advertisement Videos: An Empirical Validation Study. *Political Analysis* 31, 4 (2023), 554–574.

Lotte van Burgsteden and Hedwig te Molder. 2022. Shelving Issues: Patrolling the Boundaries of Democratic Discussion in Public Meetings. *Journal of Language and Social Psychology* 41, 6 (2022), 685–715. doi:10.1177/0261927X221079615

Sidney Verba, Kay Lehman Schlozman, and Henry E. Brady. 1995. *Voice and Equality: Civic Voluntarism in American Politics*. Harvard University Press, Cambridge, MA. https://www.jstor.org/stable/j.ctv1pnc1k7

Tianliang Xu, Eva Maxfield Brown, Dustin Dwyer, and Sabina Tomkins. 2025. PublicSpeak: Hearing the Public with a Probabilistic Framework. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence (AAAI-25)*. 28520–28529.

Han Yin, Yafeng Chen, Chong Deng, Luyao Cheng, Hui Wang, Chao-Hong Tan, Qian Chen, Wen Wang, and Xiangang Li. 2025. SpeakerLM: End-to-End Versatile Speaker Diarization and Recognition with Multimodal Large Language Models. *arXiv preprint arXiv:2508.06372* (2025).

A Algorithms

Temporal exemplars: what they supply, and why first-appearance time. Each temporal exemplar contributes *both* modalities—the neighbor's per-speaker audio clip and its previous-pass predicted label—presented as an (audio, label) pair (Section 4.3.3); the label is used only from Pass 1 onward, so Pass 0 is a pure audio baseline. Neighbors are ranked by the absolute difference in *first-appearance* time rather than by adjacency between individual speaking turns, and this is deliberate. Classification is at the speaker level—each speaker has one representative clip and one timestamp—and officials recur throughout a meeting; a turn-adjacency criterion would therefore make a recurring official a temporal neighbor of nearly every other speaker and wash out the signal. First-appearance time instead assigns each speaker a single, stable position that reflects the meeting's role structure: officials enter early (roll call, staff reports) and members of the public enter later (public comment). A query's nearest-in-first-appearance neighbors are thus disproportionately of the same role, which is precisely the prior the exemplars are meant to supply.

B Additional Validation Detail

B.1 Data Partitions and Leakage

Table 4 lists the data used at each pipeline stage and whether it overlaps the final evaluation set (referenced from Section 5.1).

B.2 Bootstrapping Across Passes

Table 5 reports per-pass performance of the bootstrapping procedure on the human-labeled evaluation set (N = 1,749–1,761 speakers across five cities); Pass 0 is a zero-shot classification and subsequent passes use the previous pass's predicted labels for temporal exemplar selection. Temporal context systematically shifts the error pattern: the zero-shot baseline has lower public-participant recall (0.722), which bootstrapping raises to 0.792 by Pass 3 while holding public-participant precision near-constant (0.869 to 0.859). Across the three passes the procedure is net-corrective—it fixes 190 previously-incorrect labels while newly breaking 157 (net +33), and the corrections disproportionately rescue true public participants (127 of 190). The regressions are roughly balanced across classes (80 true officials, 77 true public participants), and we detect a modest within-meeting propagation effect: speakers whose nearest labeled neighbors carried an incorrect prior-pass label regress about twice as often as those who do not (3.7% vs. 1.6%). Absolute label churn is substantially larger over the full five-city production corpus, reflecting the unlabeled population at scale rather than the labeled evaluation set.

Statistical comparison. We assess pairwise performance differences using McNemar's test with continuity correction. Pass 3 significantly outperforms 5-shot ($\chi^2 = 5.87$, $p = 0.015$), confirming the value of temporal bootstrapping over fixed exemplars. The majority-vote ensemble significantly outperforms both 5-shot ($\chi^2 = 25.01$, $p < 0.001$) and window 2b ($\chi^2 = 15.61$, $p < 0.001$), but does not significantly improve over Pass 3 alone ($\chi^2 = 2.75$, $p = 0.097$). The Pass 3 vs. window 2b comparison is likewise non-significant ($p = 0.082$), consistent with their similar overall accuracy.

Error complementarity. On the 1,761 speakers with human-consensus reference labels, Pass 3 makes 234 errors, 5-shot 269, and window 2b 260. Critically, a substantial share of each method's errors are unique: 24% of Pass 3 errors are not made by either other method, compared to 29% for 5-shot and 30% for window 2b. Pairwise inter-method agreement (Cohen's κ) ranges from 0.725 (5-shot vs. window 2b) to 0.751 (5-shot vs. Pass 3), indicating that the methods agree substantially but leave room for ensemble gains. The high proportion of unique errors—despite all three methods using the same underlying MLLM—reflects their distinct input representations (fixed exemplars, temporal neighbors, sliding window) and explains why majority voting improves over any individual method.

B.3 Robustness Across Speaker Characteristics

A potential concern is that classification accuracy may vary systematically with speaker characteristics, introducing bias into downstream measurements. We examine robustness along two dimensions: total speaking duration and temporal entry position.

Speaking duration. We bin human-labeled speakers by total speaking time from the diarization output (not the clips sent to the MLLM). The majority-vote ensemble performs consistently across duration bins: 87.5% for speakers with <30 s of total speaking time (N = 312), 89.2% for 30–120 s (N = 574), 84.9% for 120–300 s (N = 511), and 89.3% for >300 s (N = 364). The lack of a monotone relationship between duration and accuracy suggests that the pipeline does not systematically favor verbose speakers over brief ones—important because public participants tend to speak less than officials.

Temporal position. We compute each speaker's relative entry position (first appearance divided by meeting length) and bin speakers into tertiles. The majority-vote ensemble achieves 91.8% accuracy for early speakers (first third of the meeting), 86.0% for middle, and 85.2% for late. The higher accuracy for early speakers reflects the concentration of officials at the start of meetings during procedural business, where the classification signal is strongest. Late-meeting speakers—predominantly public participants during public comment—are classified less accurately but still at 85.2%, well above the 60.7% majority-class baseline. All three individual methods show the same monotone pattern, confirming that temporal position affects difficulty rather than introducing method-specific artifacts.

We compare against two families of supervised baselines. Because these baselines require extracting acoustic features and speaker embeddings from source audio, we evaluate them—and, for a like-for-like comparison, the MLLM ensemble—on the N = 1,081-speaker subset (annotation rounds R2, R4–R7) for which those features were computed. The first family uses *hand-crafted prosodic features*: numeric summaries of the sound itself—how high or low the voice is (pitch), how loud (energy), how the sound's frequency content is distributed (spectral measures and MFCCs, standard descriptors of voice timbre)—together with where and how much the speaker talks in the meeting (first appearance, entry position, total duration, segment count), 79 features in all. The second family uses *pretrained neural speaker embeddings*: numeric voice "fingerprints" produced by deep models built to recognize *individual* voices—ECAPA-TDNN [Desplanques et al. 2020], a 192-dimensional speaker-verification

Algorithm 1 Temporal Exemplar Selection

Require: Query speaker q with first-appearance time t_q in meeting m; set of speakers $\mathcal{S}_m$ in m (excluding q); number of exemplars k
Ensure: Exemplar set $\mathcal{E} \subset \mathcal{S}_m$, $|\mathcal{E}| = k$

for each speaker $j \in \mathcal{S}_m$ **do**
  $t_j \leftarrow$ first-appearance time of j in m
  $d_j \leftarrow |t_j - t_q|$
end for
Sort $\mathcal{S}_m$ by d_j in ascending order
$\mathcal{E} \leftarrow$ first k speakers from sorted list
return $\mathcal{E}$

Algorithm 2 Iterative Bootstrapping

Require: Segments $\mathcal{S} = \{s_1, \ldots, s_N\}$ from meeting m; baseline classifier C_0 (5-shot, random exemplars); number of temporal neighbors k; maximum passes T
Ensure: Final label assignment $\mathbf{L}_T$

$\mathbf{L}_0 \leftarrow C_0(\mathcal{S})$ ▷ Pass 0: baseline classification
for $t = 1$ to T **do**
  for each segment $s_i \in \mathcal{S}$ **do**
    $\mathcal{E}_i \leftarrow \textsc{TemporalExemplars}(s_i, k)$ ▷ Alg. 1
    for each exemplar $e \in \mathcal{E}_i$ **do**
      Assign label $\mathbf{L}_{t-1}(e)$ to exemplar e
    end for
    $\mathbf{L}_t(s_i) \leftarrow \textsc{Classify}(s_i, \mathcal{E}_i)$
  end for
  if $\mathbf{L}_t = \mathbf{L}_{t-1}$ **then break** ▷ Convergence
  end if
end for
return $\mathbf{L}_T$

Table 4: Data used at each pipeline stage and its overlap with the final evaluation set. The pipeline fits no model weights, but prompt, exemplar, pass, and ensemble choices were selected on labeled data; the production corpus is held out from every selection stage.

Stage	Data used	Overlaps eval set?	Role
Prompt development	Development sites (Albany, Chino), 4 meetings, ≈51 speakers	No	Tune prompt wording/format and select the 5-shot configuration.
Fixed few-shot exemplar pool	A fixed pool of validated exemplar clips (4 officials + 1 public participant per query)	No	Supply meeting-external reference clips to the 5-shot classifier.
Temporal exemplars (bootstrapping)	Same-meeting neighbors of each query, labeled by the *previous pass's predictions*	Speakers yes, labels no	In-context neighbors for Pass 1–3; uses predicted, not human, labels—so no human-label leakage.
Pass & ensemble selection	Human-labeled evaluation set (N = 1,761, 5 cities)	**Yes**	Choose the three-pass cap ($T = 3$) and the majority-vote rule.
Final evaluation	Same human-labeled evaluation set (N = 1,761)	—	Reported accuracies (Tables 1, 5).
Production application	29,238 speakers, 5 cities, unlabeled	Held out	Participation measures (Section 6).

Table 5: Bootstrapping Performance Across Passes on the Human-Labeled Evaluation Set ($k = 4$ temporal neighbors, Gemini 2.5 Flash, $N \approx 1{,}761$, 5 cities).

Pass	N	Acc.	κ	Pub. P	Pub. R	Pub. F1
0 (zero-shot)	1,759	0.848	0.672	0.869	0.722	0.789
1 (temporal + Pass 0 labels)	1,749	0.864	0.711	0.862	0.781	0.819
2 (temporal + Pass 1 labels)	1,759	**0.869**	**0.721**	0.864	0.790	**0.826**
3 (temporal + Pass 2 labels)	1,761	0.867	0.718	0.859	**0.792**	0.824

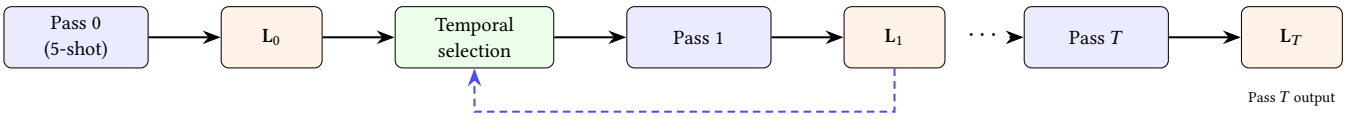

Figure 4: Bootstrapping loop. Pass 0 classifies segments with the baseline 5-shot classifier. Subsequent passes use the previous pass's predicted labels as exemplar labels for temporal selection through the configured pass cap.

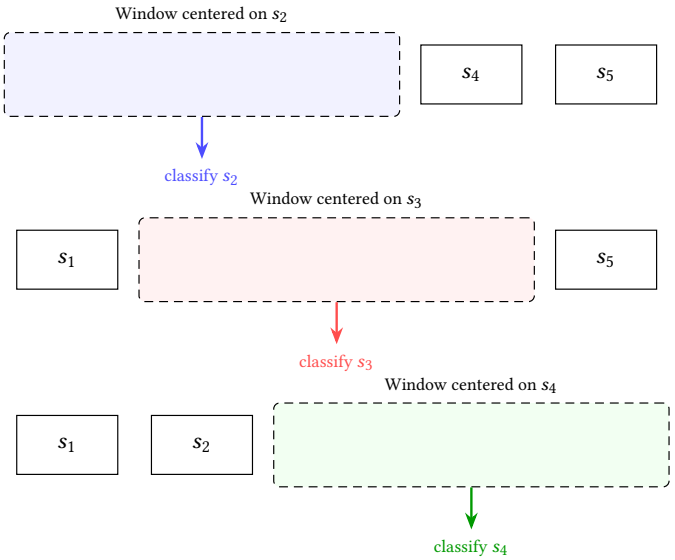

Figure 5: Sliding-window classification. Each window contains three consecutive speaker segments, and the MLLM classifies the middle speaker. Adjacent windows overlap by two segments, enabling detection of local turn-taking patterns.

Table 6: Supervised Baselines vs. MLLM Methods (grouped 5-fold CV on N = 1,081 human-labeled speakers, 5 cities)

Method	Acc.	κ	Cit. P	Cit. R	Cit. F1
Prosodic features (79 features from diarization output)					
Logistic Regression	0.714	0.410	0.629	0.678	0.652
Gradient Boosting	0.803	0.580	0.784	0.694	0.736
Speaker embeddings					
ECAPA-TDNN + Random Forest	0.774	0.510	0.772	0.610	0.681
Wav2Vec2 + Logistic Regression	0.750	0.469	0.704	0.638	0.669
MLLM (Gemini 2.5 Flash, no training data)					
5-shot (fixed exemplars)	0.853	0.683	0.881	0.727	0.796
Majority vote (3 methods)	**0.893**	**0.772**	**0.906**	**0.813**	**0.857**

Note: All supervised baselines use grouped 5-fold cross-validation stratified by label, with meetings as groups to prevent leakage. ECAPA-TDNN embeddings are 192-dimensional speaker verification vectors from a model pretrained on VoxCeleb; Wav2Vec2 embeddings are 768-dimensional mean-pooled hidden states from a model pretrained on LibriSpeech. Neither embedding model was fine-tuned on our data.

model trained on VoxCeleb, and Wav2Vec2 [Baevski et al. 2020], a 768-dimensional self-supervised speech model trained on LibriSpeech. For each speaker, we extract embeddings from up to three clips (first, middle, last by start time) and average across clips. All supervised methods use grouped 5-fold cross-validation with meetings as groups to prevent leakage. Appendix H reports the leading feature-importance rankings.

The speaker embedding baselines are informative about what audio representations capture. ECAPA-TDNN, a model designed for speaker *verification* (distinguishing individual identities), achieves only 77.4%—better than chance but far below the MLLM. This indicates that speaker role is not simply a function of voice identity: officials and public participants span diverse vocal characteristics.

Wav2Vec2, a general-purpose speech representation model, performs similarly (75.0%), suggesting that self-supervised speech features also do not capture the higher-order cues that distinguish roles.

Feature importance analysis of the prosodic classifiers reveals that temporal position (`first_appearance`, `rel_entry`) dominates acoustic features, consistent with our finding that public participants systematically enter meetings later than officials (Section 6; top rankings in Appendix H). This suggests that prosodic features capture low-level acoustic properties that overlap substantially between officials and public participants—both are adults speaking into microphones in the same room—while the MLLM can additionally leverage higher-order cues: procedural vocabulary spoken with practiced fluency, turn-taking rhythm with the chair, and contextual signals that are audible but not captured by summary statistics of the acoustic signal.

Cross-city generalization. A leave-one-city-out (LOCO) evaluation on the same $N = 1{,}081$-speaker feature-extraction subset reveals a sharp contrast in generalizability. For the prosodic Gradient Boosting classifier, training on four cities and testing on the held-out city yields a macro-average accuracy of 70.0%, compared to 80.3% under within-city cross-validation—a 10-point drop. Performance varies widely by city (from 61.2% in New Brunswick to 80.6% in Jersey City), reflecting the extent to which temporal and acoustic features are city-specific. The MLLM majority-vote ensemble, by contrast, achieves a macro-average accuracy of 89.5% across the same five LOCO folds, with a narrow range from 87.6% (Asbury Park) to 91.8% (Brockton). The MLLM's ability to generalize without retraining is a key advantage for scaling to new cities: it requires no labeled data from the target jurisdiction, whereas a supervised prosodic classifier trained on other cities would suffer substantial degradation. We caution, however, that this is generalization across *similar* jurisdictions: all five cities are urban or suburban, in Democratic-leaning states, and hold meetings in a broadly comparable format. The held-out city in each LOCO fold differs in council size and acoustics but not in basic institutional type. Whether the pipeline transfers to systematically different contexts—rural town meetings, partisan-charged hearings, or jurisdictions with very different public-comment rules—is not established here and is a priority for future validation (Section 7).

C Data Collection Details

C.1 Audio Acquisition

Meeting recordings were obtained from official city YouTube channels using `yt-dlp`, an open-source command-line tool for downloading media from YouTube and other platforms. For each of the five study cities, we identified the official government YouTube channel and downloaded all available city council meeting recordings posted between January 2017 and December 2023. The download command extracted the best available audio stream directly (without re-encoding video), producing MP3 files at the source bitrate (typically 128–192 kbps). Meeting titles, dates, and URLs were logged in a metadata CSV for provenance tracking. Table 7 lists the five-city collection frame.

Table 7: Five-city collection frame. Counts are recordings downloaded per city before file-level preprocessing (Table 8).

City	State	Meetings	Years
Brockton	MA	503	2017–2023
Asbury Park	NJ	372	2017–2023
Jersey City	NJ	301	2017–2023
New Brunswick	NJ	277	2017–2023
Inglewood	CA	154	2017–2023
Total		**1,607**	

C.2 Audio Format Conversion

Raw MP3 files were converted to 16 kHz mono WAV format using `librosa` prior to diarization, as the `pyannote.audio` pipeline expects single-channel input at 16 kHz. Original MP3 files were retained on Dropbox as the canonical source; WAV conversion was performed on-the-fly during batch processing to minimize local storage requirements. Individual recordings range from approximately 30 minutes to over 5 hours in duration.

C.3 Storage and Reproducibility

The data pipeline is organized into four tiers:

(1) **Raw audio** (MP3 files, ~2 TB total) is stored on Dropbox and is not included in the code repository due to file size.
(2) **Diarization output** (segment-level CSVs with timestamps and prosodic features, ~940 MB) is stored locally.
(3) **Sampled audio clips** (short WAV files, one per speaker segment) are generated on-the-fly for classification and are treated as ephemeral intermediate files.
(4) **Classification results** (speaker-level labels, small CSVs) are the final research data and are version-controlled in the project repository.

All processing scripts are included in the project repository. Diarization and classification outputs are linked to source recordings by filename, enabling end-to-end replication given access to the original YouTube URLs.

C.4 Classification Model and API Settings

Model and interface. All three production classification strategies (5-shot, Pass 3 bootstrapping, and window 2b) use Google `gemini-2.5-flash` through the `google-genai` Python client, calling `client.models.generate_content`. Audio clips are transmitted with the Gemini Files API (`client.files.upload`). Supervised and modality baselines use different backbones (`gemini-2.5-flash-lite` for some ablations, `whisper-large-v3` for ASR, `Qwen2-Audio-7B`, ECAPA-TDNN, and Wav2Vec2), reported where used.

Hardware. Diarization (`pyannote.audio` 3.1) ran locally on a single NVIDIA GeForce RTX 5090 (32 GB GDDR7)—well above the model's ~2–4 GB VRAM requirement—at roughly 15–30 minutes per meeting. Classification is entirely API-based (Gemini) and requires no local GPU.

Generation settings and determinism. Every call sets `temperature` = 0.0; we set no custom `top_p`, `top_k`, or `seed`, and use the API's default safety settings (no custom `HarmCategory` thresholds). Temperature 0 makes each classification deterministic given identical input, so repeated runs reproduce the same labels and the bootstrapping loop converges to a fixed point rather than drifting on run-to-run randomness. Upstream clip selection is seeded (`seed` = 42) with a 30 s-per-speaker cap in weighted sampling.

Failures, retries, and unparsable responses. Each API call is retried up to three times (four for the ASR baseline) with exponential back-off; rate-limit errors (HTTP 429 / `RESOURCE_EXHAUSTED`) trigger back-off-and-retry rather than dropping the speaker. The text response is lowercased and scanned for "official" or "citizen"; if neither token is present—or the response is empty or blocked by a safety filter—the output is labeled `unknown`, and a call that still errors after all retries is recorded as an `error`. Two buckets are worth separating.

Unparsable (`unknown`) outputs are negligible at the speaker level: across the full production corpus they number 0 (Pass 3), 1 (5-shot), and 0 (window 2b) of ≈29,250 speakers. (The often-cited "≥95% parseable" figure is a stricter *clip*-level rate measured during prompt tuning on 402 segments, Appendix D.)

API `error` outputs (a call still failing after all retries) affect 0.33% of speakers for 5-shot and 0.37% for window 2b, but 7.4% for Pass 3. The Pass 3 rate is concentrated almost entirely in Jersey City (20.4% of its speakers, vs. 0.1–2.6% in the other four cities), reflecting a rate-limit/throughput episode during that city's large run rather than model refusals.

Effect on the evaluation denominator. The validation set is restricted to speakers with a valid label from *all three* methods; on the labeled ground truth this drops only a single speaker (Pass 3 and window 2b each error on one, 5-shot on none), taking N from 1,762 to the reported 1,761. The exclusion is therefore immaterial to the reported metrics.

Handling in the production measures. The majority-vote ensemble votes over whichever of the three methods returned a valid label for a speaker, so a speaker missing one method's label—for example a Pass 3 API error—is still classified from the other two; only speakers with no valid label from any method are dropped. The Pass 3 errors are thus almost entirely recovered, and the production corpus retains near-complete coverage (29,238 classified speakers). Because the ensemble label drives the participation measures, the Jersey City rate-limit episode does not bias the floor-share or headcount statistics.

Run provenance and scale. The production classification was run in **May 2026** (bootstrapping passes 0–3 completed 2026-05-10; the 5-shot and window 2b runs completed 2026-05-18). Across the three strategies the run issued **175,493 model calls** (117,004 for bootstrapping across passes 0–3; 29,238 for 5-shot; 29,251 for window 2b), consuming approximately **309 million input tokens** (audio tokenization dominates this count) and **168,000 output tokens** (classifications are a single word). At list Gemini 2.5 Flash prices this corresponds to an API cost on the order of **$90**; we report the token counts so the figure can be recomputed at current rates.

Artifacts. Code (diarization, sampling, prompts, classification, and analysis scripts), releasable annotations, split files, and the exemplar pool are available in a public repository; to preserve single-blind review the URL and versioned DOI are omitted here and provided on acceptance. Per-call token usage and predicted labels for every pass are stored in the `output/production_results*` CSVs, from which the counts above are reproduced.

C.5 Corpus Documentation (Data Statement)

Table 8 reports per-city descriptive statistics for the analytic corpus (referenced from Section 3), and Table 9 documents the corpus as a data statement, consolidating purpose, provenance, scale, preprocessing, partitions, leakage controls, known biases, and access terms in one place.

D MLLM Prompt Variants

Table 10 reports classification performance across the five prompt configurations we tested on the 402-segment ground truth (Gemini 2.5 Flash, temperature 0). All valid predictions are included for each variant; the N_{valid} column reports how many segments received a parseable response.

Several patterns emerge. First, prompt engineering substantially affects parse reliability: the original 4-shot prompt produced valid responses for only 89 of 402 segments (22.1%), while all revised variants achieve ≥95% validity. Second, adding exemplars consistently improves κ relative to zero-shot, though the gain is modest. Third, the 5-shot v3 configuration—which provides four official exemplars and one public-participant exemplar from previously validated ground truth with a minimal instruction prompt—achieves the best balance of accuracy (89.9%), reliability ($\kappa = 0.683$), and public recall (0.85), and is used as the baseline for all subsequent experiments.

Exemplar presentation. All audio—the five exemplars and the target speaker's query segment(s)—is supplied as *separately uploaded files* through the Gemini Files API within a single ordered prompt, *not* concatenated into one audio stream. For each exemplar we append its text label followed by its audio file ("Example i — This speaker is: {label}"), in the fixed order official, official, public participant, official, official (the single minority-class exemplar placed centrally); we then append a short instruction and the query segment(s), and finally the one-word-answer prompt. When a speaker contributes multiple query segments, all are attached to the same call with a note that they are from the same speaker, and the model returns a single label.

Why 4 officials + 1 public participant. The exemplar class balance and all shot/prompt choices were selected on the development sites (Albany and Chino), disjoint from the production evaluation. We preferred the official-heavy 4:1 ratio for two reasons. First, it approximates the base rate—officials dominate meetings—so the exemplar set does not misrepresent the prior. Second, a class-*balanced* exemplar set empirically performed worse: the balanced 4-shot v2 reaches only 85.5% accuracy versus 89.9% for 5-shot v3, because balancing pushed the model to over-predict the minority public-participant class (its recall rose to 0.93 but precision collapsed to 0.54). Retaining a single public-participant exemplar still gives the

Table 8: Corpus Descriptive Statistics by City.

City	Meetings	Speakers	Segments	Avg. Speakers/Meeting
Asbury Park	364	7,270	46,030	20.0
Brockton	499	6,605	48,899	13.2
Inglewood	141	2,346	16,869	16.6
Jersey City	299	9,143	66,515	30.6
New Brunswick	262	3,880	20,714	14.8
Total	**1,565**	**29,244**	**199,027**	**18.7**

Table 9: Corpus documentation / data statement.

Dimension	Description
Purpose	Measure public participation (speaker role and floor-time share) in U.S. local-government meetings.
Sources	Official city YouTube channels and government websites; automated scraping (Section C).
Dates	Meetings recorded 2017–2023; retrieved prior to the May 2026 classification run.
Collection frame vs. analytic sample	Frame: 1,607 downloaded recordings from the five study cities (Table 7). Analytic sample: 1,565 diarized meetings from the same cities (Table 8).
Exclusions	Corrupted downloads, files below the minimum-duration threshold, and diarization failures account for the 42-recording frame-to-analytic reduction.
Scale and units	≈2,600 hours; 199,027 speech segments; 29,244 meeting-level speaker clusters. Units: meeting, speaker cluster, segment.
Formats and preprocessing	MP3/WAV audio resampled to 16 kHz mono; segments < 0.5 s discarded; per-speaker audio preparation caps each speaker's selected speech at 30 s (Section 4).
Annotations	Speaker role (official / public participant) by three research assistants over eight rounds; disagreements resolved by majority vote (Appendix F).
Data partitions	Development (Albany, Chino; prompt tuning and diarization checks); validation ground truth (5 production cities, $N = 1{,}761$); production corpus (unlabeled, application only). See Table 4.
Leakage controls	Development cities are disjoint from production; the production corpus is held out from prompt, exemplar, pass, and ensemble selection.
Missingness	Uneven temporal coverage (2019–2022 dominant); video framing present for 45–100% of speakers by city; ASR / open-audio comparisons run on subsets with locally available source audio.
Known biases	Purposive sample of dense, urban/suburban, Democratic-leaning cities drawn from a rent-control frame; YouTube-posting requirement favors higher-capacity clerk offices. Not a representative census of U.S. local government.
Privacy	Public records of open meetings; analysis assigns de-identified labels (SPEAKER_00, …) and classifies by institutional role, not individual identity.
Access and licensing	Source audio remains public on the cities' channels; we release derived labels, metadata, and code under an open license and distribute audio as URL pointers rather than re-hosting.
Intended / prohibited uses	Intended: aggregate measurement of participation and institutional comparison. Prohibited: identifying, profiling, or re-contacting individual speakers.
Versioning and citation	Released code and annotations will be archived with a versioned DOI on acceptance.

Table 10: MLLM Prompt Variant Performance (Gemini 2.5 Flash, $N_{\text{total}} = 402$ segments)

Variant	Description	N_{valid}	Accuracy	κ	Cit. P	Cit. R
0-shot	No exemplars; original prompt	358	86.6%	0.554	0.63	0.65
0-shot v2	No exemplars; improved system prompt	402	88.3%	0.561	0.68	0.59
4-shot	4 random exemplars; original prompt	89	89.9%	0.682	0.65	0.87
4-shot v2	4 balanced exemplars; refined prompt	400	85.5%	0.599	0.54	0.93
5-shot v3	4 official + 1 public; minimal prompt	385	**89.9%**	**0.683**	**0.66**	**0.85**

model a reference for the minority class while keeping the official-heavy prior intact.

D.1 Prompt Text

Below we reproduce the exact text prompts used in each configuration. In all few-shot variants, labeled audio exemplars are prepended before the query clip; the text prompt follows the final audio input.

Minimal prompt (used in 0-shot, 0-shot v2, 5-shot v3, and all production runs):

```
Listen to this audio clip from a city council
meeting.

Classify the speaker as either:
- "official" (government official or staff member)
- "citizen" (member of the public)

Respond with ONLY ONE WORD: either "official"
or "citizen"

Response (one word only):
```

The 0-shot and 0-shot v2 variants differ only in the temperature parameter (default vs. 0.0), not in prompt wording.

Detailed prompt (used in 4-shot):

```
You are an expert at identifying speaker roles
in city council meetings.

LISTEN FOR THESE CLUES:

OFFICIALS typically:
- Use procedural language: "motion to approve",
"second the motion", "agenda item X"
- Refer to staff reports, legal analysis, budget
items
- Speak with formal, professional tone
- May call on speakers or manage the meeting
flow
- Direct the meeting proceedings
- Say things like "I move to...", "the chair
recognizes...", "the clerk will call roll"

CITIZENS typically:
- Address the council: "Mr. Mayor", "Council
members", "Madam Chair"
- Introduce themselves: "My name is...", "I'm a
resident of..."
- Share personal experiences or concerns
- Speak more emotionally or passionately about
issues
- Ask the council to take action on something
- Say things like "I live at...", "My neighborhood...",
"I'm here to speak about..."
- Sound like they are ADDRESSING the council,
not PART OF the council

CRITICAL DISTINCTION:
- Is this person ADDRESSING the council (speaking
TO them) → CITIZEN
- Is this person PART OF the council (managing
the meeting) → OFFICIAL

After listening, respond with ONLY ONE WORD:
- "official" if government official/staff
- "citizen" if member of public

Response (one word only):
```

Balanced prompt (used in 4-shot v2):

```
You are an expert at identifying speaker roles
in city council meetings.

LISTEN FOR THESE CLUES:

OFFICIALS typically:
- Use procedural language: "motion to approve",
"second the motion", "agenda item"
- Refer to staff reports, legal analysis, budget
items
- Manage the meeting flow: call on speakers,
call for votes
- Speak with formal, professional tone about
city business
- Say things like "I move to...", "the chair
recognizes...", "the clerk will call roll"

CITIZENS typically:
- Address the council: "Mr. Mayor", "Council
members"
- Introduce themselves: "My name is...", "I'm a
resident of..."
- Share personal experiences or concerns about
issues
- Ask the council to take action on something
- Say things like "I live at...", "My neighborhood...",
"I'm here to speak about..."

Based on what you hear, classify the speaker.

Respond with ONLY ONE WORD:
- "official" if government official/staff
- "citizen" if member of public

Response (one word only):
```

The key design evolution across variants was simplification: the detailed prompt's "CRITICAL DISTINCTION" heuristic and the balanced prompt's enumerated cues both underperformed the minimal prompt when paired with audio exemplars, suggesting that the model extracts classification-relevant information more effectively from reference audio than from textual descriptions of acoustic cues.

E Diarization Validation Protocol

Table 11 reports the pairwise diarization accuracy by city (referenced from Section 5).

E.1 Evaluation Design

We evaluated diarization quality using two complementary tasks administered to three research assistants (RAs) who worked independently.

Task 1: Speaker Pair Verification (126 development pairs from three Chino meetings, plus 360 pairs sampled across four production cities). Each item consists of two audio clips drawn from the same meeting. Some pairs come from the same model-assigned speaker label ("same-speaker" pairs) and some from different labels ("different-speaker" pairs). RAs judged whether each pair sounds like the same person, without knowing the model's assignment. This design detects two error types:

- *Merge errors:* The model grouped segments from different people under one speaker label. Detected when RAs judge a same-speaker pair as "different person."
- *Split errors:* The model assigned one person two different labels. Detected when RAs judge a different-speaker pair as "same person."

Task 2: Clip Quality Check (45 clips). RAs assessed individual clips for single-speaker presence (does the clip contain exactly one voice?), boundary quality (are segment start and end points clean?), and speech audibility (is there intelligible speech?).

E.2 Answer Key Construction

Task 1 evaluates the diarization model's assignments: a model-predicted "same-speaker" pair consists of two segments assigned the same speaker label by `pyannote`, and a predicted "different-speaker" pair consists of segments from two different labels. The majority judgment of the research assistants serves as the human reference; disagreement between that reference and the model assignment flags a potential diarization error. Pairs were sampled

Table 11: Diarization Accuracy by City (Majority Vote of 3 Raters).

City	Overall	Same-Speaker	Different-Speaker	*N* Pairs
Chino, CA	87.1%	80.2%	100.0%	124
Asbury Park, NJ	85.0%	77.5%	92.5%	80
Brockton, MA	88.8%	87.5%	90.0%	80
Inglewood, CA	91.1%	90.0%	92.3%	79
Jersey City, NJ	94.2%	93.3%	95.0%	120
Overall	**89.4%**	**85.4%**	**94.1%**	**483**

Table 12: Inter-Rater Reliability for Diarization Verification, Chino Development Meetings (Pairwise Cohen's κ)

RA Pair	Cohen's κ	Raw Agreement
RA 1 vs. RA 2	1.000	100.0%
RA 1 vs. RA 3	0.902	95.2%
RA 2 vs. RA 3	0.787	89.5%
Average	**0.897**	

to include roughly equal numbers of predicted same-speaker and different-speaker items drawn from across each meeting.

E.3 Inter-Rater Reliability

Table 12 reports pairwise inter-rater agreement for the Chino development meetings, where average Cohen's $\kappa = 0.897$ ("almost perfect"). Across the four production cities agreement is lower but still substantial (mean pairwise $\kappa = 0.79$, range 0.69–0.84), consistent with their noisier audio; pooling all five cities gives a mean pairwise $\kappa = 0.84$. In each case the reliability confirms that the verification task is well-defined and that RAs can distinguish speakers by voice.

E.4 Error Analysis

In the Chino development set, all five pairs flagged unanimously by all three RAs as model errors are merge errors with a consistent root cause: the audio clip contains overlapping speech from multiple people, but the model assigned the entire segment to a single speaker label. These are segmentation boundary errors rather than speaker embedding failures—the model correctly identifies speakers' voices but sometimes draws segment boundaries too broadly, capturing speaker transitions within a single segment. No split errors were detected in Chino by majority vote (0 of 45 different-speaker pairs flagged as same person). The four production cities confirm this pattern: merge errors remain the dominant failure mode (12.2% of same-speaker pairs by majority vote), while split errors—absent in Chino—appear only at a low rate (7.3% of different-speaker pairs) and are likewise concentrated in overlapping or degraded audio.

E.5 Clip Quality

Majority-vote quality assessments: 86.7% of clips contain a single speaker, and 97.8% contain audible speech. Boundary quality showed the most inter-rater variation (66.7–73.3% rated as "clean"),

Table 13: Diarization Error Rate (DER) and Jaccard Error Rate (JER) on seven densely-annotated meetings (0.25 s collar, annotator-averaged). Errors concentrate in Brockton's subcommittee audio.

Meeting	DER	JER
Jersey City (Special, Sep. 2021)	0.4%	1.9%
Asbury Park (Dec. 2018)	1.3%	3.9%
Asbury Park (Special, Mar. 2023)	3.9%	11.1%
Jersey City (Nov. 2021)	4.6%	28.0%
Inglewood (Feb. 2022)	6.8%	23.8%
Brockton (Dec. 2019)	16.8%	50.7%
Brockton (May 2020)	21.7%	39.9%
Speech-weighted	**7.6%**	**21.9%**

consistent with exact cut-off points being a more subjective judgment.

E.6 Diarization Error Rate (DER and JER)

Beyond the pairwise verification above, three research assistants densely annotated speaker turns for seven complete meetings across the production cities (five double-annotated), correcting the pyannote output in Audacity; we score against pyannote with `pyannote.metrics`. Table 13 reports per-meeting DER and JER. Errors concentrate almost entirely in Brockton (DER 17–22%), whose corpus is dominated by school-committee and subcommittee sessions with noisier, overlapping audio; the council-style meetings in the other cities fall at or below 6.8% DER (3.1% excluding Brockton). Brockton nonetheless attains among the *highest* role-classification accuracy (90.8%), indicating that boundary-level diarization errors do not always change the speaker-role decision. Speaker splits and merges can still propagate to the per-speaker classification input, however, so the role accuracy should not be interpreted as evidence that diarization errors are fully absorbed. Two caveats bound these values: annotators corrected the pyannote draft rather than labeling from scratch, so DER is a mild lower bound, and inter-annotator DER is 15.2%, reflecting the difficulty of exact boundary placement on overlapping speech.

F Ground Truth Creation

F.1 Annotation Protocol

What the RAs did. Three research assistants (RAs) independently labeled each speaker as "official," "citizen," or "other" (the last reserved for clips with no intelligible speech, pure background noise, or corrupted audio). For each speaker, the RA listened to up to three short clips (2–15 seconds each) drawn from different points in that speaker's turns; a single clip sufficed when the role was obvious, and the additional clips were available when it was not. Alongside each label the RA recorded a confidence rating (high, medium, low) and an optional free-text note (e.g., "states name and address," "sounds like roll call," "too short to tell"), which the PI used to audit borderline cases.

The decision rubric. RAs were given a written guide instructing them to rely on role-diagnostic cues rather than on the substantive content of speech: procedural and parliamentary language ("I move to approve," "is there a second," "roll call"), forms of address (officials address one another by title; members of the public introduce themselves and address the council), the audio source (dais microphone versus podium or call-in line), and the speaker's position and footprint in the meeting (officials recur throughout and appear early; members of the public typically speak once, briefly, and later). When genuinely uncertain between the two roles, RAs were instructed to apply the public-participant label (recorded as "citizen") at low confidence, since members of the public are the rarer and more consequential class to recover.

Role boundaries for ambiguous participants. The label keys on *institutional role in the meeting*, not employment status or profession. Anyone conducting or presenting the meeting's official business is coded *official*: elected members, clerks, and city staff, but also legal counsel and city attorneys, police and department heads giving reports, and *consultants, contractors, or invited experts/presenters* delivering a report at the body's request; when an official reads a resident's submitted comment aloud, the reading official's role governs. Anyone addressing the body from the floor is coded *public participant*: residents, community advocates, and business owners, in-person or call-in commenters, and any attorney, journalist, or other professional speaking on their own or a client's behalf during public comment. Ceremonial speakers (opening prayer, pledge) are coded by who delivers them—official if a staff member, public otherwise. Clips with no intelligible speech (pure noise, corrupted or too-short audio, or a speaker who cannot be made out) are coded *other* and excluded from evaluation; genuinely 50/50 cases are labeled public at low confidence. These rules were given to the RAs as worked examples and mirror the category definitions supplied to the classifier.

Training and skills required. The task does not require subject-matter expertise or transcription skill; RAs were onboarded with the written guide and a small set of worked examples. Importantly, because labeling keys on *how* and *when* a person speaks—procedural register, self-introduction, microphone source, meeting timing—rather than on understanding *what* they say, the task is robust to the features that complicate transcription: a speaker's accent, non-native or informal grammar, and partially inaudible content rarely obstruct role identification. The genuine difficulties in the task, reflected in the RAs' notes and low-confidence flags, were short or overlapping clips and the occasional ambiguous staff voice, which the confidence ratings and multi-clip design were intended to absorb. Each RA worked independently without access to the others' labels. Inter-rater reliability was computed using Fleiss' κ for all three raters and pairwise Cohen's κ; disagreements were resolved via majority vote, with three-way ties adjudicated by the PI through direct listening.

F.2 Label Propagation

The consensus label for a speaker is propagated to *all* diarization segments attributed to that speaker within a meeting. Because diarization assigns a consistent speaker ID across a recording, a single human judgment per speaker suffices to label every segment from that speaker. This label propagation step assumes that speaker identity is stable within a meeting—an assumption supported by the low rate of split errors in our diarization validation: none in the Chino development set and roughly 7% of different-speaker pairs in the production cities (Section 5).

F.3 Annotation Rounds

The ground truth was assembled across eleven annotation rounds (R1–R11), with Round 8 excluded from all analyses due to low inter-rater reliability (Fleiss' $\kappa = 0.511$).

- **Round 1** (development): 6 speakers from one Albany, NY meeting (February 3, 2020).
- **Round 2** (development + early production): 234 speakers from meetings across Chino (CA) and initial production cities.
- **Rounds 4–7** (production): Extended annotation to five production cities with progressively improved annotation guidelines, yielding 732 additional labeled speakers.
- **Rounds 9–11** (production): Further annotation across the five production cities, adding 818 labeled speakers (R9: 275, R10: 229, R11: 314).

F.4 Inter-Rater Reliability by Round

Table 14 reports inter-rater reliability for each annotation round. Reliability improved substantially between early rounds (R2, R4: moderate agreement) and later rounds (R5–R7, R9–R11: substantial to almost perfect agreement, apart from R11 at the low end of substantial), reflecting refinement of the annotation guidelines.

F.5 Ground Truth Statistics

After excluding R8 and speakers labeled "other" or "disputed," the documented annotations comprise **1,812 consensus-labeled speakers** across the five study cities and two development sites (Albany and Chino), spanning **169 meetings**. Table 15 summarizes the distribution by site. The production validation set consists of the $N = 1{,}761$ speakers from the five study cities who received valid predictions from all three classification methods (this is the count that enters Tables 1 and 5); for the study cities the Official and Public columns below report these prediction-valid speakers, so the five-city rows sum to 1,761.

The overall public-participant rate in the production validation set is 39.3% (692/1,761). Later annotation rounds (R5–R7, R9–R11) were deliberately sampled to include more public-comment-heavy

Table 14: Inter-Rater Reliability by Annotation Round

Round	N	Fleiss' κ	Interpretation	Mean pairwise κ
R2	234	0.534	Moderate	0.539
R4	166	0.525	Moderate	0.521
R5	273	0.862	Almost perfect	0.846
R6	199	0.744	Substantial	0.745
R7	294	0.844	Almost perfect	0.845
R9	275	0.868	Almost perfect	0.868
R10	229	0.748	Substantial	0.746
R11	314	0.662	Substantial	0.664
R8 (excl.)	185	0.511	Moderate	0.529

Note: R8 is excluded from all validation analyses due to low inter-rater reliability. Mean Fleiss' κ across the eight included rounds (R2, R4–R7, R9–R11) is 0.723 (range 0.525–0.868).

Table 15: Ground Truth Summary by City and Label (R1–R11, excluding R8)

City	State	Meetings	Official	Public	N (valid)
Albany	NY	1	5	1	—
Chino	CA	3	38	7	—
Asbury Park	NJ	37	266	184	450
Brockton	MA	36	234	49	283
Inglewood	CA	35	193	125	318
Jersey City	NJ	28	252	282	534
New Brunswick	NJ	29	124	52	176
Total		**169**	**1,112**	**700**	**1,761**

Note: N (valid) counts speakers from the five study cities with valid predictions from all three classification methods. Albany and Chino are development sites excluded from production evaluation. Official and Public columns report consensus-labeled speakers (speakers labeled "other" or "disputed" are excluded).

sessions, producing a substantially more balanced evaluation set than the initial development data (15.6% public-participant rate in Chino/Albany).

G Propagation of Measurement Uncertainty

Our pipeline introduces three sources of measurement error: diarization (89.4% accuracy), MLLM classification (87.7% accuracy), and human annotation noise (Fleiss' $\kappa = 0.723$). Using predicted labels directly in downstream analyses can produce biased point estimates and invalid confidence intervals—even at 90% accuracy, plug-in estimators can exhibit severely undercovered CIs [Egami et al. 2023]. We therefore characterize how classification error affects public-participant proportions and their cross-city and temporal comparisons. A duration-weighted correction for speaking-time share requires additional analysis and is identified below as a planned refinement.

We ground our analysis in the Design-based Supervised Learning (DSL) framework of Egami et al. [2023], which provides a general theory for using predicted variables in downstream statistical analyses. Our setting maps directly onto this framework: the MLLM predictions are "surrogate labels" Q_i, the human annotations are "gold-standard labels" Y_i, and the downstream estimands are aggregate public-participant proportions and cross-group comparisons. The DSL estimator is doubly robust and provides valid confidence intervals without requiring that the classifier be correctly specified. The misclassification correction we apply below is a special case of DSL for prevalence estimation; for regressions and other downstream analyses, the full DSL framework (implemented in the `dsl` R package) applies.

We correct for classification bias using the confusion matrix from the majority-vote ensemble evaluated on the human-labeled sample ($N = 1{,}761$). Let $\text{TPR} = P(\hat{Y} = \text{public} \mid Y = \text{public}) = 538/692 = 0.777$ and $\text{FPR} = P(\hat{Y} = \text{public} \mid Y = \text{official}) = 63/1{,}069 = 0.059$. The observed (plug-in) public-participant proportion $\hat{\pi}$ relates to the reference-label proportion π by:

$$\hat{\pi} = \pi \cdot \text{TPR} + (1 - \pi) \cdot \text{FPR}. \tag{1}$$

Inverting yields the bias-corrected estimator:

$$\pi^* = \frac{\hat{\pi} - \text{FPR}}{\text{TPR} - \text{FPR}}. \tag{2}$$

Table 16 applies this correction to each city. Because $\text{TPR} - \text{FPR} = 0.719 < 1$, the denominator in Equation 2 stretches all proportions away from zero, amplifying cross-city differences rather than attenuating them.

Table 16: Observed vs. Bias-Corrected Public-Participant Proportions

City	Observed (%)	Corrected (%)	Δ (pp)
Asbury Park	29.3	32.6	+3.3
Brockton	13.4	10.4	−3.0
Inglewood	34.9	40.4	+5.5
Jersey City	41.4	49.4	+8.0
New Brunswick	28.6	31.6	+3.0
All	**29.9**	**33.4**	**+3.5**

We construct confidence intervals for π^* via the delta method. The asymptotic variance of the corrected estimator is:

$$\begin{aligned}\mathrm{Var}(\pi^*) \approx \frac{1}{(\mathrm{TPR}-\mathrm{FPR})^2}\Bigg[&\frac{\hat{\pi}(1-\hat{\pi})}{n} \\ &+ (\pi^*)^2\,\frac{\mathrm{TPR}(1-\mathrm{TPR})}{n_1} \\ &+ (1-\pi^*)^2\,\frac{\mathrm{FPR}(1-\mathrm{FPR})}{n_0}\Bigg],\end{aligned} \tag{3}$$

where n is the production sample size, and $n_1 = 692$ and $n_0 = 1{,}069$ are the human-reference public-participant and official counts. For the aggregate ($n = 29{,}238$), we obtain SE ≈ 0.011 and a 95% CI of $[31.3\%, 35.5\%]$.

The headcount-based comparisons survive correction and are amplified. The decline in public-participant prevalence from the pre-pandemic period to 2020–2021 widens from 14.9 percentage points (observed) to 20.7pp (corrected). The cross-city range in public-participant proportions expands from 28.0pp (observed: Brockton 13.4% to Jersey City 41.4%) to 39.0pp (corrected: 10.4% to 49.4%). Both differences exceed the reported aggregate CI width under the pooled correction.

Two features of our validation design are relevant to interpretation. First, the 87.7% accuracy is measured end-to-end on diarized speakers: human annotators labeled the same diarized audio segments that the MLLM classified, so diarization errors are represented in the confusion matrix and require no separate propagation step. Second, the inter-rater agreement ($\kappa = 0.723$) shows that the human-consensus labels are imperfect reference measurements rather than objective truth. Apparent model errors can therefore include disagreements on genuinely ambiguous speakers; we do not treat the reported accuracy as an exact estimate against error-free labels.

G.1 Transportability of the Error Rates

The correction in Equation 2 applies a single pair of error rates—TPR = 0.777 and FPR = 0.059, estimated on the labeled sample—to every city and period in the production corpus. This is valid only under a *transportability* (label-shift) assumption: conditional on a speaker's true role, the probability of misclassification does not depend on the city, the time period, or the audio conditions, so that only the class prevalence π varies across strata while (TPR, FPR) are constant. Under this assumption the stratum-specific corrected prevalence is recovered by applying the same (TPR, FPR) to each stratum's observed $\hat{\pi}$.

The assumption is substantive and can fail in two ways relevant here. First, *audio conditions shift*: virtual and hybrid meetings after early 2020 change microphone paths and introduce call-in artifacts, which could lower TPR or raise FPR in the pandemic period specifically—biasing the very COVID comparison we most care about. Second, *prevalence-dependent errors*: in extremely official-heavy meetings (e.g., Brockton), the temporal-context components can inherit the majority class, so the effective TPR may be lower where public participation is rarest. Per-city ensemble accuracy is stable (90.8% Brockton, 89.0% Inglewood, up to 0.873 public F1 in Jersey City; Section 5), which is consistent with—but does not prove—approximate transportability.

Planned refinements. Four refinements will sharpen these intervals and test the assumption directly: (i) re-estimating (TPR, FPR) *within* city and within pre/during/post-COVID period and re-running the correction stratum by stratum, to report how much the corrected COVID gap and cross-city range move; (ii) replacing the i.i.d. speaker-level variance in the delta-method expression with *meeting-clustered* (cluster-robust) standard errors, since speakers within a meeting share audio conditions and are not independent; (iii) a *duration-weighted* error analysis for ~~floor share~~ **[GS: speaking-time share]**, propagating classification error weighted by speaking time rather than headcount, since a misclassified long-speaking official distorts ~~floor share~~ **[GS: speaking-time share]** more than a brief speaker; and (iv) uncertainty bands on the principal application figures (Figures 2 and 3) from the clustered, bias-corrected estimates.

In sum, the pooled misclassification correction supports the headcount-based cross-city and temporal comparisons in Section 6, shifting public-participant prevalence estimates by 3–8 percentage points. It does not by itself correct duration-weighted speaking-time shares. The planned stratified, clustered, and duration-weighted analyses above would test the transportability assumption and extend the correction to those quantities. The DSL framework [Egami et al. 2023] provides formal guarantees—consistency, asymptotic normality, and valid confidence intervals—for subsequent regression-based downstream analyses using predicted labels when its assumptions are satisfied.

H Additional Tables and Figures

H.1 Prosodic Feature Importance

Table 17 reports the top 10 feature importance scores from the Random Forest classifier trained on the full production corpus (Section 5). Temporal features—particularly `first_appearance` (when the speaker first enters the meeting) and `rel_entry` (relative entry position)—dominate acoustic features, consistent with the structural regularity of council meetings: officials speak early during procedural business, while public participants appear later during public comment periods. Pure acoustic features (MFCCs, spectral measures) rank considerably lower, reflecting the substantial acoustic overlap between officials and public participants speaking in the same room under the same recording conditions.

Table 17: Random Forest Feature Importance (200 trees, grouped 5-fold CV, N = 27,087 speakers, 79 features)

Rank	Feature	Importance
1	First appearance (temporal)	0.073
2	Relative entry position (temporal)	0.047
3	Number of segments (temporal)	0.033
4	Total speaking duration (temporal)	0.033
5	$MFCC_5$ mean	0.023
6	Last appearance (temporal)	0.015
7	$MFCC_4$ mean	0.015
8	$MFCC_0$ mean (std)	0.014
9	$MFCC_7$ mean	0.014
10	$MFCC_0$ std (mean)	0.013

Note: Features labeled "(temporal)" are speaker-level temporal position and activity features; remaining features are segment-level acoustic features aggregated to the speaker level. The top four features are all temporal, accounting for 18.6% of total importance.

Table 18: Classification Accuracy by Meeting (Gemini 2.5 Flash)

Meeting	0-shot	5-shot	Pass 3	Window 2b	AND
Chino Sept 15	0.880	0.920	0.920	0.920	0.940
Chino Oct 20	0.893	0.907	0.913	0.920	0.933
Chino Nov 17	0.907	0.898	0.907	0.917	0.926
Albany Feb 3	0.913	0.935	0.935	0.913	0.935

H.2 Classification Accuracy by Meeting

Table 18 reports per-meeting accuracy for the four main classification methods and the best ensemble, illustrating the heterogeneity discussed in Section 5.

The November 17 Chino meeting is a consistent outlier for bootstrapping methods due to its unusually low public-participant ratio (~10%), which causes temporal neighbors to be overwhelmingly official and biases bootstrapped predictions. The AND ensemble performs most consistently across meetings.

I Audio-as-Data vs. Audio-to-Text-to-Data

A natural question is whether our audio-direct approach actually buys anything over the conventional pipeline that first transcribes speech and then classifies the text. This appendix tests that directly by holding everything fixed *except the modality*. For each human-labeled speaker we reuse the same diarized clips that the audio pipeline classifies; we then (i) transcribe each clip with OpenAI's Whisper [Radford et al. 2023] and (ii) classify the resulting transcript, either with the same MLLM operating on text (Gemini, with a prompt mirroring the audio prompt) or with a supervised text classifier (TF-IDF features and logistic regression, grouped 5-fold cross-validation by meeting). The audio arm is our production majority-vote ensemble. Because only the input representation differs, any performance gap is attributable to the transcription bottleneck rather than to the classifier or the task.

Scope and caveats. These results should be read as preliminary, and three caveats bound them. First, *coverage*: this analysis uses the 593 human-labeled speakers (public-participant rate 0.40, matching the full sample) whose source recordings were available locally at the time of analysis, not all 1,761. Second, and importantly, *ASR strength*: we transcribe with a mid-sized model (`whisper-small.en`) rather than the largest available. A stronger ASR model would help the text pipeline most precisely on the hard, noisy clips; these numbers therefore, if anything, *overstate* the audio advantage in magnitude, and the gap below should be read as an upper bound. The qualitative pattern, however, is robust to this choice. Third, we stratify by a cheap acoustic signal-to-noise (SNR) proxy computed directly from each clip, which is correlated with but not identical to transcription error.

Overall. On the common set of 593 speakers (Table 19), the audio-direct pipeline outperforms both text pipelines: +3.5 accuracy points over Whisper→Gemini and +6.7 over Whisper→TF-IDF, with markedly higher public *precision* (0.95 vs. 0.78). Transcription is therefore not free for this task: routing the signal through text loses information even when the downstream classifier is the same model.

Where audio earns its advantage. The overall gap masks a sharp interaction with audio quality. Figure 6 and Table 20 split the 593 speakers into quartiles of acoustic quality. On the cleanest audio (Q4), the text pipelines *match or slightly exceed* the audio pipeline—unsurprising, since speaker role is heavily lexically marked ("my name is, I live at…" vs. procedural language), and on clean speech the transcript captures it. But as audio quality degrades, the text pipelines fall off sharply while the audio pipeline holds: on the noisiest quartile (Q1, ~11 dB), accuracy drops to 0.733 (Gemini) and 0.700 (TF-IDF) against 0.820 for audio—a gap of 9–12 points. This is the expected signature of ASR error, which is concentrated on exactly the noisy, overlapping, accented, and far-microphone speech that characterizes public comment [Koenecke et al. 2020]. Because real meeting corpora are full of such clips, the audio-direct approach is more robust where it matters and avoids inheriting transcription's failure modes.

In sum, for classifying speaker roles, transcription is not a lossless intermediate step. Audio-direct classification matches the text pipeline on clean speech and is substantially more robust on the degraded audio that dominates real public-meeting recordings—while also avoiding the cost, latency, and documented racial and accent bias of ASR. A definitive version of this comparison, on all 1,761 speakers and with a larger ASR model (which would give the text arm its strongest form), is left for the final draft.

J Theoretical Foundations

This appendix provides formal grounding for three design choices in the pipeline: iterative bootstrapping as a fixed-point optimization procedure, the sliding-window classifier, and the ensemble rules as applications of the Condorcet jury theorem.

J.1 Bootstrapping as Iterative Refinement

The bootstrapping procedure (Algorithm 2) can be interpreted as a fixed-point iteration over the label space. Let $\mathcal{L} = \{0, 1\}^K$ denote

Table 19: Audio-as-Data vs. Audio-to-Text-to-Data (same $N = 593$ speakers; modality is the only difference; Whisper `small.en`)

Pipeline	N	Acc.	κ	Cit. P	Cit. R	Cit. F1
Audio-direct (MLLM ensemble)	593	**0.853**	**0.679**	**0.947**	0.672	**0.786**
Text: Whisper → Gemini	593	0.818	0.620	0.775	**0.769**	0.772
Text: Whisper → TF-IDF (sup.)	593	0.786	0.546	0.761	0.681	0.718

Table 20: Accuracy by acoustic-quality quartile ($N \approx 150$ speakers each)

SNR quartile (mean dB)	Audio-direct	Whisper→Gemini	Whisper→TF-IDF
Q1 noisiest (~11 dB)	**0.820**	0.733	0.700
Q2 (~23 dB)	**0.873**	0.812	0.813
Q3 (~34 dB)	**0.893**	0.878	0.805
Q4 cleanest (~51 dB)	0.833	**0.847**	0.820

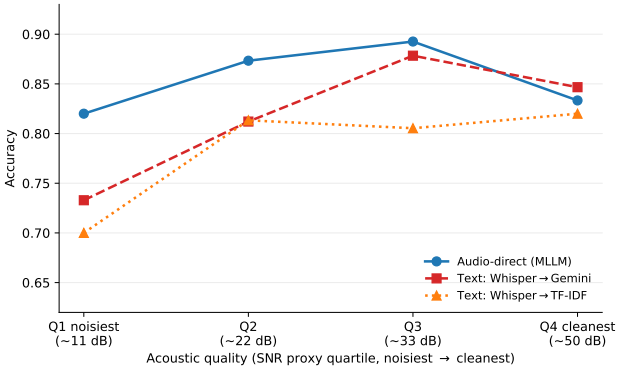

Figure 6: Classification accuracy by acoustic quality (SNR-proxy quartile, noisiest to cleanest), for the audio-direct pipeline and the two transcription-based pipelines ($N = 593$, ~150 speakers per quartile). The text pipelines match audio on clean speech (Q4) but degrade sharply on noisy speech (Q1), where ASR fails; the audio pipeline is comparatively flat.

the space of binary label assignments for all K speakers in a meeting, and let $\Phi\colon \mathcal{L} \to \mathcal{L}$ denote the update operator that, given a current label assignment $\mathbf{L}_{t-1}$, produces a new assignment $\mathbf{L}_t$ by classifying each speaker with temporal exemplars drawn from $\mathbf{L}_{t-1}$. Bootstrapping iterates $\mathbf{L}_t = \Phi(\mathbf{L}_{t-1})$ until convergence.

Convergence guarantee. Because the label space $\mathcal{L}$ is finite ($|\mathcal{L}| = 2^K$) and the classifier Φ is deterministic (temperature = 0), the sequence $\{\mathbf{L}_t\}$ must either converge to a fixed point $\mathbf{L}^* = \Phi(\mathbf{L}^*)$ or enter a cycle. In practice, we observe rapid convergence: accuracy rises over the first passes and then plateaus (84.8% → 86.4% → 86.9% → 86.7%) and the number of changed labels decreases, consistent with contraction toward a fixed point.

This structure is analogous to the expectation–maximization (EM) algorithm: Pass 0 provides an initial estimate (analogous to the E-step initialization), and subsequent passes refine labels using pseudo-labels as soft constraints (analogous to alternating E- and M-steps). Unlike EM, however, we lack a formal likelihood objective, so convergence to a global optimum is not guaranteed. The empirical monotonicity and rapid stabilization (within $T = 3$ passes) suggest that the landscape is well-behaved for this task.

J.2 Sliding-Window Classification

The window 2b classifier (Section 4.3) labels each speaker from a fixed-width window of surrounding speakers. Let $\mathbf{s} = [s_1, \ldots, s_N]$ be the meeting's unique speakers ordered by first appearance. For each interior position i the classifier reads the window $[s_{i-1}, s_i, s_{i+1}]$ (width 3, step 1) and outputs a label for the center speaker s_i; the endpoints s_1 and s_N use one-sided context (their single available neighbor). Because the sequence ranges over *unique* speakers, each speaker is the center of exactly one window and therefore receives exactly one window-2b label—no per-speaker aggregation is required. Two properties make this useful. First, consecutive windows overlap—sharing two of their three speakers—so every speaker is classified with its immediate neighbors present, which surfaces local turn-taking regularities such as the official → public → official alternation of a chair introducing commenters. Second, the same rule is applied at every position, so the classifier is position-invariant along the meeting. The "rule" is not a trained sequence model with learned weights but the pretrained MLLM applied in context; the window only governs how much local context the model sees, complementing the global context supplied by bootstrapping.

J.3 Ensemble Theory: The Condorcet Jury Theorem

The majority-vote ensemble is motivated by the Condorcet jury theorem [de Condorcet 1785]. In its classical form, the theorem states: if each of M independent voters has probability $p > 0.5$ of making the correct binary decision, then the probability that a majority votes correctly is

$$P_{\text{maj}} = \sum_{k=\lceil M/2 \rceil}^{M} \binom{M}{k} p^k (1-p)^{M-k},$$

and $P_{\text{maj}} \to 1$ as $M \to \infty$.

In our setting, $M = 3$ classifiers with individual accuracies $p_1 = 0.847$ (5-shot), $p_2 = 0.867$ (Pass 3), and $p_3 = 0.852$ (window 2b). The independence assumption is violated—all three classifiers use the same underlying MLLM—but the classifiers make *complementary* errors due to different input representations (fixed exemplars vs. temporal exemplars vs. sliding window). Empirically, pairwise inter-method κ ranges from 0.725 to 0.751, and 24–30% of each method's errors are unique (not shared with either other method), confirming substantial but imperfect correlation. The empirical majority-vote accuracy (87.7%) exceeds all individual classifiers, consistent with the Condorcet prediction under weak positive correlation.

Formally, given M classifiers $f_1, \ldots, f_M$ with $f_m(k) \in \{0, 1\}$ for speaker k, the three combination rules are: *majority vote* $\hat{y}_k = \mathbf{1}\left[\sum_{m=1}^{M} f_m(k) \geq \lceil M/2 \rceil\right]$; *AND* $\hat{y}_k = \mathbf{1}\left[f_a(k) = 1 \wedge f_b(k) = 1\right]$; and *OR* $\hat{y}_k = \mathbf{1}\left[f_a(k) = 1 \vee f_b(k) = 1\right]$.

The AND and OR ensembles represent principled tradeoffs along the precision–recall frontier:

- **AND rule** (predict public participant only if *both* agree): increases the effective decision threshold, yielding high precision at the cost of recall. For two classifiers with independent false-positive rates α_a and α_b, the AND rule achieves false-positive rate $\alpha_a \cdot \alpha_b < \min(\alpha_a, \alpha_b)$.
- **OR rule** (predict public participant if *either* agrees): decreases the threshold, yielding high recall at the cost of precision. The OR rule achieves false-negative rate $\beta_a \cdot \beta_b < \min(\beta_a, \beta_b)$ under independence.

These theoretical properties are borne out empirically: the AND ensemble of 5-shot ∩ window 2b achieves the highest public precision (0.930), while the OR ensemble of Pass 3 ∪ window 2b achieves the highest public recall (0.854), with the majority vote providing the best F1 (0.832).

Dependence between window 2b and Pass 3. Window 2b conditions on Pass 3's bootstrapped neighbor labels, so the two are positively correlated and the majority vote gives Pass-3-aligned signal somewhat more weight than three independent voters would. Three considerations make the ensemble appropriate nonetheless. (i) Window 2b is not a deterministic function of Pass 3: it also reads the neighbors' *audio*, disagrees with Pass 3 at $\kappa \approx 0.73$, and contributes 24–30% unique errors, so it carries information Pass 3 does not. (ii) The 5-shot classifier uses fixed, meeting-external exemplars and no bootstrapped labels; it is therefore independent of Pass 3 and anchors the vote, so a shared Pass-3/window-2b error is overturned whenever 5-shot dissents. (iii) A leave-one-out check confirms both dependent methods contribute: on the $N = 1{,}761$ set the three-way majority reaches 87.7%, whereas dropping window 2b (voting 5-shot with Pass 3) reaches at most 86.7% and dropping Pass 3 (voting 5-shot with window 2b) at most 85.6%. Removing either method costs accuracy, so the gain is not an artifact of double-counting Pass 3. An audio-only window variant (2a) is fully Pass-3-independent but less accurate than 2b, which is why 2b is the production choice.

AND/OR pairs and their selection. Table 1 reports *all* pairwise AND and OR ensembles. The pairs we single out—5-shot ∩ window 2b for peak precision and Pass 3 ∪ window 2b for peak recall—were identified *after* inspecting validation performance, so they characterize the precision–recall frontier rather than pre-registered operating points. Only the majority-vote rule is committed to *a priori* and used in the application.

K Extended Limitations and Ethical Considerations

K.1 Scope and Inference

The five study cities were purposively selected from municipalities with accessible meeting recordings and are concentrated in urban or suburban, Democratic-leaning settings. The observed cross-city and temporal patterns therefore demonstrate the measurement pipeline rather than provide a representative portrait of U.S. local government. In particular, the post-2020 decline is descriptive: the uneven city-by-year panel does not identify the pandemic or remote-meeting policies as its cause. Transfer to rural jurisdictions, politically different settings, and institutions with different public-comment rules requires additional validation. Future work can broaden coverage through resources such as LocalView [Barari and Simko 2023] and use explicit causal designs to study how institutional rules affect public floor-time share.

K.2 Measurement Validity and Model Durability

Human-consensus labels establish agreement with trained annotators who use procedural language, microphone source, timing, and meeting position; they do not independently verify every role against rosters, agendas, minutes, or speaker cards. The operational classes also compress heterogeneous roles: consultants, invited presenters, attorneys, contractors, journalists, and unintelligible speakers may not fit neatly into either *official* or *public participant*. Independent documentary validation and a clearer policy for ambiguous roles would strengthen construct validity.

The pipeline uses a closed, hosted MLLM whose outputs may change across model versions even at temperature 0. We report the exact model identifier, run dates, settings, and per-pass behavior (Appendix C.4), but absolute performance remains version-specific. Operating on audio preserves lexical, prosodic, and turn-taking information, yet the model does not isolate which cues drive its decisions. The matched audio-versus-text analysis suggests that transcription loses information on noisy recordings, but it currently covers a subset of speakers and does not use the strongest available ASR system (Appendix I).

K.3 Public Records, Privacy, and Data Processing

All source recordings are public records of open government meetings published by the cities. The project analyzes existing recordings without intervening with speakers or collecting private records and received an institutional determination that it did not constitute human-subjects research; research-assistant annotation labels institutional roles rather than studying the annotators. Nevertheless, a public recording contains identifiable voice data, and public availability should not be treated as permission for unrestricted secondary profiling.

We assign meeting-specific codes (e.g., `SPEAKER_00`), do not identify or link speakers to voter, property, or other personal records, and do not release a voice database. Raw audio remains on the

source city channels and is distributed only through provenance URLs rather than re-hosted. Audio inputs were processed using the paid Gemini API. Under the paid-service terms in effect during the analysis, prompts, associated files, and responses were not used to improve Google's products; Files-API uploads are automatically deleted after 48 hours. These provisions reduce but do not eliminate the risks introduced by third-party processing, and the applicable terms should be rechecked before future collection or replication.

K.4 Fairness, Redistribution, and Misuse

Audio classification may perform unevenly across accents, languages, speech impairments, microphone types, remote-call systems, and recording quality. We report per-city results and sensitivity to aggregate classification error, but the annotations do not contain demographic attributes and therefore do not support subgroup fairness estimates by race, gender, age, citizenship, or accent. Claims of demographic fairness would be unwarranted, and transfer to new speech populations should be evaluated before substantive use.

The intended use is aggregate measurement of participation and comparison of institutions. We release code, aggregate measures, and ethically releasable annotations, but not raw audio or derived voice embeddings. The pipeline is not validated for identifying, profiling, ranking, surveilling, or re-contacting individual speakers, and it must not be used to infer demographic or legal attributes. Researchers extending the pipeline should preserve provenance, minimize retained voice data, document access controls, and reassess legal and ethical requirements in each jurisdiction.