

Who Spoke Next Is Not Who Should Have: What Zero-Shot Failure on Multi-Party Turn-Taking Benchmarks Measures, Why the Human Reference Sits Beside the Models, and What Would Show That Knowing When to Speak Must Be Trained

Pranay Mahendrakar

SONYTECH · ORCID 0009-0003-7224-029X

pranaymahendrakar@sonytech.in · mahendrakarpranay@gmail.com · +91 6361723454

Abstract. An assistant that joins a group conversation has to decide, at every pause, whether to take the floor. Two 2026 benchmarks report that language models prompted zero-shot make this decision badly and that fine-tuning repairs it, and both conclude that knowing when to speak is learned rather than emergent. Other 2025-2026 work reports the opposite pattern on neighbouring tasks: untuned models beat human subjects at predicting the next speaker in meetings, and the predictive uncertainty of a model with no task training locates speaking opportunities better than either prompting or fine-tuning. In this analysis the author reviews and synthesises published results to show that the disagreement is about three different quantities that one speak-or-silent score mixes: discrimination (whether a system speaks more often where speaking is appropriate), operating point (how often it speaks at all, which depends on costs and a role the prompt never states), and the target (whose behaviour the label records). Re-weighting the one published human reference to the benchmark's own category mix, the author finds that human annotators would score about 57 percent balanced accuracy on the Friends portion, below the best zero-shot model's 60.5. What fine-tuning visibly adds is a shift in operating point and agreement with the corpus on the one category where humans and corpus disagree; whether it also adds discrimination, the published results do not identify. The paper gives a decision procedure for what a speak-or-silent result licenses and specifies the score-based, cost-stated and panel-labelled measurements that would show whether the capability is missing. The argument covers text-transcript evaluation most firmly; for full-duplex speech models the evidence is thinner and is treated separately.

Keywords: turn-taking; multi-party conversation; spoken dialogue systems; large language models; operating point; human label variation; emergent abilities; benchmark validity

SONYTECH ORCID: 0009-0003-7224-029X

1. Introduction

A spoken or text assistant placed in a meeting, a family chat or a classroom faces a decision that a two-party chatbot never meets: most of what is said is not said to it. Treating every detected pause as an invitation, as Bhagtani et al. (2026) put it, works tolerably in dyadic dialogue; with three or more people the same policy is disruptive, because pauses are many and most of them belong to someone else. The question of who speaks next, and when, has been studied in conversation analysis since Sacks, Schegloff and Jefferson (1974), in spoken dialogue systems for two decades (Raux and Eskenazi, 2012; Skantze, 2021), and in multi-party settings specifically (Traum, 2004; Bohus and Horvitz, 2010; Ouchi and Tsuboi, 2016; Gu, Tao and Ling, 2022). Large language models have moved it from a component problem to a product question, because the same model that writes the answer is now also asked to decide whether to say anything (Mao et al., 2024; Hu et al., 2025; Zhu and Callison-Burch, 2025).

Two 2026 benchmarks put the question directly. Bhagtani et al. (2026) built more than 120,000 decision points from three multi-party corpora - the AMI meetings

(Carletta et al., 2006), transcripts of the television series Friends, and SPGISpeech financial calls - and asked eight language models, prompted zero-shot, whether a named participant should speak at each one. Balanced accuracy ranged from 42.6 to 64.5 percent; LoRA fine-tuning with balanced category sampling raised it by up to 23 points; and the authors concluded that "context-aware turn-taking is not an emergent capability; it must be explicitly trained." Nama et al. (2026) built a synthetic dataset of 215,000 examples, found zero-shot macro F1 between 0.206 and 0.569 across six models, found that supervised fine-tuning lifted every model that did not collapse to between 0.697 and 0.747, and wrote that "intervention timing is a learned capability, not an emergent property of pretraining scale." Earlier work points the same way: Wei et al. (2023) described deciding when to talk in a group as one of two skills "that pairwise-trained models appear to lack", and Umair, Sarathy and de Ruiter (2024) found that language models could not locate the points inside a turn where a listener could speak.

Other results point the other way. Fukuda et al. (2026) found that text-only language models, with no training on the target domain, predicted the next speaker in AMI meetings better than both supervised models and human subjects who could see and hear the meeting. Umair and de Ruiter (2026) found that a measure of semantic

uncertainty, computed by sampling continuations from an 8-billion-parameter model with no task training, located speaking opportunities with an F0.5 of 0.545, against 0.22 for the best prompted model and 0.27 for the best fine-tuned one. Khan et al. (2026) closed most of a prompted meeting delegate's failure to speak - from 51.4 percent of opportunities missed to 2.5 percent - with an architecture that keeps track of the meeting's state around a prompted model, not by fine-tuning it. And the same benchmark that reported zero-shot failure also reported, on its Friends portion, three human annotators whose balanced accuracy of 60 to 66 percent overlaps the best zero-shot model's.

This paper argues that both sides read their numbers correctly and are answering different questions. A speak-or-silent accuracy at a set of decision points is a function of three things. The first is discrimination: whether the system's propensity to speak is higher at points where speaking is appropriate than at points where it is not. The second is the operating point: how readily the system speaks overall, which for any decision under uncertainty depends on the relative cost of interrupting and of missing a turn, and on the role the system has been given (Raux and Eskenazi, 2012; Mitra et al., 2026). The third is the target: whose behaviour the label records. In the 2026 benchmarks the label is either what one human in a corpus did next or what a language model wrote when it generated the conversation. A system can discriminate well and still score near chance if its operating point is far from the one the metric rewards; it can score well by matching a corpus convention that human judges do not share; and it can fail a target that no observer could hit, because in multi-party talk the next speaker is often whoever starts first (Sacks, Schegloff and Jefferson, 1974; Lerner, 2003).

The conclusion "must be explicitly trained" needs all three separated. It needs evidence that discrimination is low zero-shot, not only that the default operating point is wrong; that training raises discrimination rather than only moving the operating point; and that what training converges to is appropriate participation rather than one corpus's habits. The published record supplies the first only for small open models on text transcripts, supplies the second weakly, and on the one benchmark with a human reference contradicts the third.

The contribution is a separation and its consequences, stated as six claims.

1. Zero-shot speak-or-silent tables in this literature are mostly tables of operating points. In Bhagtani et al.'s own results, by the author's arithmetic, the five open models under ten billion parameters speak at 64 to 90 percent of the AMI and SPGI decision points labelled Silent, while Gemini 3.1 Pro speaks at a fifth to a quarter of them and is silence-leaning on those two corpora but, counting points rather than averaging categories,

speak-leaning on Friends; in Nama et al.'s results two models from one family sit at opposite corners, one speaking almost everywhere and the other never (Sections 5.1 and 5.4).

2. On the only benchmark portion with a human reference, the human annotators and the best zero-shot model are at the same level. Reweighted to the Friends test set's own category mix, the three annotators would score 56.5, 55.4 and 59.0 percent balanced accuracy (mean 57.0), against 60.5 for Gemini 3.1 Pro zero-shot; per category, Gemini's zero-shot accuracy is above the annotators' mean on three of four categories and 0.15 points below it on the fourth (Section 5.2).
3. Fine-tuning changes the operating point and teaches one contested convention. On Friends, the category the benchmark calls referenced-but-not-addressed carries a Silent label; the human annotators said Speak about 72 percent of the time on it, and five of the six fine-tuned models came to say Silent 47 to 78 percent of the time, exceeding every annotator by 17 to 54 points (Section 5.3).
4. Where the label is fixed by addressing, untuned models discriminate. A transcript-conditioned frontier model responds to 94.7 percent of requests that name the assistant and 30.4 percent of matched requests that do not (Ma et al., 2026); a prompted model's next-speaker accuracy is 0.697 when the current speaker selects the next one and 0.121, against a chance rate of 0.127, when participants self-select (Liu et al., 2024). The failures concentrate where conversation analysis says the outcome is not determined by the text (Section 5.6).
5. The signal for speaking opportunities is present in the predictive distribution of a model with no task training and absent from its answers to a prompt. The semantic-uncertainty result (Umair and de Ruiter, 2026) is a readout result, and it means that "not emergent" and "not elicited by this prompt" are different findings (Section 5.7).
6. The label is one realisation of an option, and the field already has the tools for that. Human label variation, consensus over listeners and cost-stated decision rules are established in neighbouring literatures (Plank, 2022; Pavlick and Kwiatkowski, 2019; Fawcett, 2006). A benchmark that scores against the single observed next speaker measures agreement with a race outcome (Section 7).

Section 2 states the scope. Section 3 defines the three quantities and the decision rule. Section 4 summarises what conversation analysis says the label is. Section 5 re-reads the evidence, Section 6 examines what "emergent" and "must be trained" would have to mean for the claim

to be testable, and Section 7 treats the label. Section 8 gives the decision procedure, Section 9 the measurements that would settle the open parts, and Sections 10 to 12 take objections, state what is not known, and conclude.

1.1 Relation to the author's earlier papers

Two earlier papers by the author use a related separation. One argued that explanations given to human decision makers more often move their threshold for accepting advice than improve their discrimination; another argued that an agent's decision not to act covers two triggers with different cost terms. The present paper borrows the separation of discrimination from operating point, which is standard signal detection, but its question is different: whether a conversational capability is missing from language models, and whether the labels used to say so record the thing they name. Neither earlier paper treats turn-taking, multi-party conversation or the validity of next-speaker labels.

2. Scope, and what this paper does not claim

The paper is about the decision whether to take the floor in conversations with three or more participants, as evaluated on text transcripts and, more briefly, on audio for full-duplex speech models. Every number below is a published value or the author's arithmetic on published values, and the arithmetic is labelled as the author's where it appears.

Five limits are stated up front. First, the central re-analysis in Section 5 rests on one unreviewed preprint's tables (Bhagtani et al., 2026) and on a human evaluation of 360 items by three annotators within it. The re-weighting in Section 5.2 inherits that sample's error, which the author estimates at roughly two points of balanced accuracy, and it is presented as a reason to doubt a conclusion, not as a new estimate of human performance. Second, the per-category figures in Section 5 are category-averaged rates, which weight the four categories equally; they are a coarse view of discrimination at one operating point and are not a substitute for the score-based curves that none of the papers report. Third, several of the benchmarks discussed use labels written by a language model (Nama et al., 2026; Ma et al., 2026; in part, Mitra et al., 2026), and some results come from scripted television dialogue; the paper says where this matters. Fourth, the paper does not argue that small open models handle multi-party turn-taking well zero-shot. On AMI and SPGI they do not, by any reading in Section 5. Fifth, the paper takes no position on how often an assistant should speak in any particular product; it argues that the answer is a stated cost and role, and that a benchmark which leaves it unstated measures the model's guess about it.

3. Three quantities under one score

3.1 Definitions

Fix a set of decision points D , each a moment at which a target participant T could take the floor. A system S returns, at each point, either a decision (Speak or Silent) or, if the interface allows it, a score $s(d)$ that orders points by how strongly S favours speaking.

Discrimination is the degree to which S 's propensity to speak is higher at points where speaking is appropriate than at points where it is not. With scores it is summarised by the area under the receiver operating characteristic curve; with decisions only, a single point on that curve is observed, and the best available summary is the difference between the true positive rate and the false positive rate at that point (Youden's J), which can be low because the point sits in a corner even when the curve through it is good (Fawcett, 2006).

Operating point is the point on that curve where S actually decides: how often it speaks. For a decision under uncertainty, the point that minimises expected cost is set by the ratio of the cost of a false Speak (an interruption, an intrusion on talk addressed to someone else) to the cost of a false Silent (a missed request, an assistant that ignores being asked), together with the base rate of appropriate turns. Raux and Eskenazi (2012) built a turn-taking controller for task-oriented spoken dialogue systems on exactly this view, framing turn-taking as decision making under uncertainty and pairing learned estimates with an explicit cost structure. Nama et al. (2026) state the same asymmetry in their metrics and in their reinforcement-learning reward, and Umair and de Ruiter (2026) choose a precision-weighted F score because a false entry into someone's turn costs more than a delayed one.

Target is what the label at each point records. In the benchmarks discussed here it is one of: whether T produced the next utterance in a recorded human conversation (Bhagtani et al., 2026; Fukuda et al., 2026); whether a language model generating a synthetic conversation had the assistant speak at that point (Nama et al., 2026; Ma et al., 2026; the synthetic RolePlayConv set of Mitra et al., 2026, whose other evaluation set, NOTSOFAR-1, is real meetings with one speaker designated as the assistant); or whether the proportion of a listener panel that responded at that point exceeded a threshold (Umair, Sarathy and de Ruiter, 2024; Umair and de Ruiter, 2026). These are three different quantities, and only the last is an estimate of where speaking is appropriate for listeners in general rather than for one participant.

3.2 The decision rule and what a prompt leaves out

The rule a well-specified turn-taking policy implements can be written in one line. If $p(d)$ is the probability that speaking at d is appropriate given the context and T 's role, c_I is the cost of an inappropriate entry and c_M the

cost of a missed turn, the expected-cost-minimising policy speaks when $p(d)$ exceeds $c_I / (c_I + c_M)$. The cost ratio is not a property of the conversation. It is a property of the deployment: a note-taking assistant in a board meeting and a facilitator in a classroom discussion should not share it. Mitra et al. (2026) make the role explicit as an input, motivated by the observation that what users expect of a conversational agent ranges from passive listening to active facilitation.

A zero-shot prompt of the kind used in the 2026 benchmarks states the task and asks for a binary decision. The prompt of Bhagtani et al. (2026) also states decision rules - including that the target should stay silent when "Someone mentions the target speaker in third person without expecting a response" - a qualitative tie-break ("When uncertain and the target speaker is a BYSTANDER $\rightarrow$ prefer SILENT.") and a statement that both errors matter ("False interruptions are bad, but failing to respond when involved also breaks the interaction."); its decisions are read at temperature zero. Neither benchmark's prompt states a numeric cost ratio, the base rate, or what the assistant is there to do beyond being a participant. The system's operating point is therefore set only partly by such qualitative rules and otherwise by whatever its instruction tuning and its reading of the prompt make it, and the literature on prompted classification predicts that this will vary widely and arbitrarily across models and prompt formats: language models carry biases toward particular answers that move few-shot accuracy from near chance to near the best available (Zhao et al., 2021), and meaning-preserving changes in prompt formatting alone moved accuracy by up to 76 points in one open model (Sclar et al., 2023). The abstention literature treats a model's declining to respond as a capability in its own right, with its own methods, benchmarks and metrics (Wen et al., 2024), which is a reason not to assume that an instruction-tuned model's default readiness to respond is set anywhere useful. None of this shows that a prompted model's operating point is wrong for the deployment; it shows that a zero-shot binary score is the product of an operating point that no prompt set as a threshold and a discrimination nobody measured.

3.3 What a balanced accuracy near 50 percent can and cannot show

Balanced accuracy is the mean of the true positive and true negative rates. A system that speaks at every point scores 50 percent, as does one that is silent everywhere and one that guesses at random. A system whose scores rank points perfectly but whose threshold is set so low that it speaks at 95 percent of Silent points also scores close to 50 percent. A balanced accuracy near 50 therefore excludes good discrimination only if the operating point is known to be near the middle of the curve; if the system is in a corner, it is uninformative about discrimination. The per-category rates that Bhagtani et al. (2026) publish make it possible to see

which corner each system is in, which Section 5.1 does.

4. What the label is: the conversation-analytic account

The founding description of turn-taking (Sacks, Schegloff and Jefferson, 1974) distinguishes two ways a next speaker is chosen. Either the current speaker selects the next one - by name, by gaze, by directing a question - or nobody is selected, and then any participant may self-select, with the first to start acquiring the turn; if nobody does, the current speaker may continue. Points at which a change of speaker becomes possible are transition-relevance places, and Umair and de Ruiter (2026) restate the consequence in the phrase that they are "opportunities, not obligations": at such a place a listener may speak, may produce only a minimal response, or may let the speaker go on.

Three consequences matter for benchmark labels. First, where the current speaker selects the next, the label "T spoke next" is close to a fact about the text: it records compliance with an allocation that the transcript usually shows. Lerner (2003) catalogues the means of such selection, from address terms and gaze to tacit forms that rely on circumstance and content, and notes that names in particular often do more than address. Second, where participants self-select, the label records who started first among those entitled to speak. A participant who was entitled, about to speak and beaten by half a second is labelled Silent. The overlap that results when two self-selectors start together is itself organised and resolved by the participants (Schegloff, 2000), which means the transcript keeps the winner and loses the attempt. Third, utterances in a group are addressed to some participants and designed to be understood by others: Clark and Carlson (1982) argue that a speaker in multi-party talk performs, besides the act directed at the addressees, an act of informing all participants, and Goffman's (1979) account of footing distinguishes addressed recipients from unaddressed ones and bystanders. Being mentioned is not being addressed, which is the distinction Bhagtani et al. (2026) build their S2 category on, but being mentioned often makes a response relevant, and whether a mentioned participant takes it up is again a choice.

Timing and prosody carry part of what decides these choices. Turn transitions across ten languages show a common pattern of short response latencies (Stivers et al., 2009), and the speed of turn exchange implies that listeners plan their responses while the current turn is still under way (Levinson and Torreira, 2015). Listeners project the end of a turn from its lexical and syntactic content more than from its intonation (de Ruiter, Mitterer and Enfield, 2006), but prosodic and other cues contribute to both turn-yielding and backchannel timing (Gravano and Hirschberg, 2011; Ward and Tsukahara, 2000; Ekstedt and Skantze, 2022b), pauses within a turn and gaps between turns are different events (Heldner and Edlund, 2010), and in multi-party settings gaze and

body orientation help allocate the floor (Bohus and Horvitz, 2010). A transcript that keeps the words and the speaker labels drops most of this.

Taken together, the account predicts a specific shape for the difficulty of a speak-or-silent benchmark built from transcripts. Points where the next speaker was selected by explicit address should be nearly determined by the text and easy for any competent reader. Points where nobody was selected should be only partly determined, by the content and by each participant's stake in it, and the rest decided by timing and by who started first. A label read off the transcript at those points has an irreducible error rate for any observer of the transcript, and the error rate is not small. Section 5 checks this prediction against the published results; the model results are consistent with it; no study reviewed here measures human observers separately at self-selection points.

5. The evidence, re-read

Table 1 collects the published values the argument rests on, one row per cited result. Sections 5.1 to 5.3 re-read the benchmark that states the strongest form of the claim; Sections 5.4 to 5.8 bring in the rest.

5.1 The zero-shot table is a table of operating points

Bhagatani et al. (2026) build their benchmark by turning every utterance boundary in a transcript into one decision point per participant who was not speaking. The label is Speak if that participant produced the next utterance and Silent otherwise, and each point is assigned to one of four categories: explicitly addressed by name or role (I1, Speak), an active participant not referred to but expected to respond (I2, Speak), a bystander to an exchange between others (S1, Silent), and mentioned but not addressed (S2, Silent). The Friends portion has 8,970 points: 1,114 in I1, 2,632 in I2, 639 in S1 and 4,585 in S2. AMI has 11,900 and SPGI 99,290. Each model sees a system prompt describing the task and answers at temperature zero.

The authors read the zero-shot table as a capability failure and note that "Most models exhibit a strong Speak bias". The per-category rates show how strong. Treating the I1 and I2 accuracies as the rate of speaking at Speak points, and one minus the S1 and S2 accuracies as the rate of speaking at Silent points, and weighting by the category counts, the author computes that the five open models under ten billion parameters speak at 64 to 90 percent of the Silent points in AMI and SPGI. On Friends all eight models speak at more than half of the Silent points, from 59.5 percent for Gemini 3.1 Pro to 90.0 percent for Qwen3-4B. Gemini behaves differently on the other two corpora: it speaks at 21.7 percent of Silent points in AMI and 23.1 percent in SPGI, but also at only 42.8 and 51.6 percent of Speak points, so its errors there are mostly misses. (The same weighting reproduces every reported balanced accuracy in the paper's zero-

shot table to within 0.6 points, which is the check that it is the paper's metric.)

A system in the upper-right corner of the receiver operating characteristic plane scores near 50 percent balanced accuracy whatever its discrimination, as Section 3.3 noted. The quantity the table can speak to is the gap between the two rates at the operating point each system chose. Averaging over categories, so that each category counts once, that gap is between -2.2 and 8.7 points for the six open models on AMI, between -11.3 and 5.0 points on SPGI, and between 9.8 and 29.7 points on Friends. For Gemini 3.1 Pro it is 17.3, 35.0 and 42.0 points on AMI, SPGI and Friends; for GPT-5.2 it is 15.4, 20.9 and 28.9.

Two readings follow. For the small open models on AMI and SPGI, the speak rate barely depends on the category: a model that speaks at 88 percent of Silent points and 86 percent of Speak points, as Mistral-7B does on AMI by this averaging, is close to ignoring the context at its operating point. That is a real failure, and nothing in this paper disputes it. But the gap at a corner is a lower bound on the best gap available along the same model's curve, and without scores nobody knows how much better the same model would do with its threshold moved. The paper's test of whether the failure is "instruction neglect" - repeating the system prompt twice, which moved balanced accuracy by at most about three points - checks whether the model attended to the instruction. It does not check the operating point, because the instruction stated one only through qualitative rules and a tie-break, not as a threshold. The rules may also contribute to the Speak bias: they list nine conditions for Speak, among them "A natural backchannel or reaction is expected" and "Staying silent would unnaturally drop them from the conversation". For the frontier models the gap at the operating point is substantial on SPGI and Friends, and on AMI their errors are misses, not intrusions. Calling all eight results the same failure merges a system that cannot tell the cases apart with one that tells them apart and is set too cautiously.

5.2 The human reference, weighted like the test set

The paper reports a human evaluation on Friends, chosen because its social dialogue is the most accessible of the three corpora to non-expert annotators. Three annotators labelled 360 test items - 100 from each category except S1, which had only 60 - given the same context and target information as the models. Their balanced accuracies were 64.81, 60.31 and 66.13 percent, mean 63.75; their pairwise Cohen's kappas were 0.57, 0.38 and 0.53. The authors conclude that "even humans disagree on whether to speak in ambiguous situations" and that their best trained models "match or exceed human-level balanced accuracy." By comparison, the best zero-shot model on Friends, Gemini 3.1 Pro, scores 60.54, and the open models 50.70 to 55.92.

Table 1. Published values on deciding when to speak in multi-party settings, one row per cited result. BA is balanced accuracy (percent); F1 is macro F1; FIR and MIR are the false-intervention and missed-intervention rates. Rows marked "author's arithmetic" are computed by the author from the cited source's own tables (Section 5.2).

Source	Setting	System, condition	Reported value	Reading
Bhagtani et al. 2026	Friends, transcript	Gemini 3.1 Pro, zero-shot	BA 60.54; S2 acc 34.10	Best zero-shot
Bhagtani et al. 2026	Friends, transcript	Llama-3.1-8B, zero-shot	BA 54.21; S1 50.00, S2 16.01	Speak bias
Bhagtani et al. 2026	Friends, transcript	Llama-3.1-8B, LoRA SFT	BA 72.52; S2 acc 69.44	Above humans on S2
Bhagtani et al. 2026	Friends, 360-item subset	3 human annotators	BA 60.31-66.13; S2 24-30; kappa 0.492	Moderate agreement
Author's arithmetic	Friends, test-set mix	Same 3 annotators	BA 55.4-59.0 (mean 57.0)	Below best zero-shot
Nama et al. 2026	Synthetic, GPT-4 labels	Qwen3-4B / Qwen3-8B, zero-shot	F1 0.206, FIR 0.897 / always Silent	Opposite corners
Nama et al. 2026	Synthetic, GPT-4 labels	Llama-3.3-70B SFT; Llama-3.1-8B SFT then RL	F1 0.747, FIR 0.050 (70B SFT); MIR 0.521 → 0.186-0.218 (8B RL)	Cost moves the point
Khan et al. 2026	AMI delegation	GPT-4o prompt-only / CAPA	Opportunities missed 51.4% / 2.5%	Silence bias, no training
Ma et al. 2026	Scripted, transcript	Gemini 3.1 Pro	Responds 94.7% if named, 30.4% if not	Addressing discriminated
Liu et al. 2024	Multi-party chat corpus	GPT-4o, zero-shot	Next speaker 0.697 allocated, 0.121 self-selected (chance 0.127)	Chance where self-selected
Fukuda et al. 2026	AMI, online	Text LLMs vs humans	LLMs above humans on next speaker; humans' F1 about 60%	Hard for humans too
Wei et al. 2023	MultiLIGHT	Trained next-speaker model	54.4% vs chance 33.3%; no human-rated gain over random	Label gain, no quality gain
Umair and de Ruiter 2026	Within-turn, 118 listeners	Prompt / SFT / semantic uncertainty	F0.5 0.22 / 0.27 / 0.545	Signal in the distribution

These two sets of numbers are not computed on the same mix of items. In the full Friends data, and therefore in the test split, which the paper draws per category in the same 80/10/10 proportions, S2 makes up 4,585 of the 5,224 Silent points, or 87.8 percent; in the human subset it makes up 100 of 160, or 62.5 percent. S2 is the category on which the annotators were weakest, at 29, 30 and 24 percent accuracy. A balanced accuracy computed on the subset therefore flatters the annotators relative to models scored on the full test split.

Re-weighting each annotator's published per-category accuracies to the test split's category counts - the same computation that reproduces the models' reported scores to within 0.6 points, and that reproduces the annotators' own subset scores exactly - gives balanced accuracies of 56.5, 55.4 and 59.0 percent, mean 57.0. This is the author's arithmetic on published values. It assumes that each annotator's accuracy within a category would be the same on the full test split as on the subset, and it inherits the subset's sampling error: with 100 S2 items and accuracy near 28 percent, one binomial standard error on the S2 rate is about 4.5 points, which moves the re-weighted balanced accuracy by about 2 points.

On this footing Gemini 3.1 Pro zero-shot, at 60.54, is above all three annotators; GPT-5.2 at 55.41 and gpt-oss-20b at 55.92 are inside their range; Qwen2.5-7B at 55.00 is 0.4 points below its bottom. The per-category view, which needs no re-weighting, says the same thing for the frontier model. Gemini's zero-shot accuracies on Friends are 82.52, 80.62, 86.67 and 34.10 percent on I1,

I2, S1 and S2; the annotators' means are 82.67, 79.00, 78.33 and 27.67. The gap between the speak rates at Speak and Silent points, averaged over categories, is 42.0 points for Gemini and 33.8 for the mean annotator (36.8, 25.0 and 39.7 for the three individually). Figure 1(a) places every system in the plane of these two rates.

The conclusion this licenses is narrow and it is enough. On the one corpus where the benchmark supplies a human reference, the strongest zero-shot model does as well as the human annotators, by the paper's own metric applied to the paper's own test mix. The claim that zero-shot models "consistently fail" is true against the label and false against the human reference, and the human reference is the only evidence in the paper about what the label's ceiling is. For AMI and SPGI the paper reports no human reference at all, so the zero-shot scores there are failures against a target whose attainability is unmeasured.

Two cautions keep this from being overread. The paper chose Friends because it is the most accessible corpus for non-expert annotators, which suggests the annotators were not specialists, and three people are not a population; a larger or trained panel could do better. And Friends is scripted: who speaks next was decided by a writer, and scripted dialogue is known to differ from spontaneous talk in, among other things, how much listener feedback it contains (Pilan et al., 2024). Neither caution helps the original conclusion, which was stated about the models rather than about the corpus.

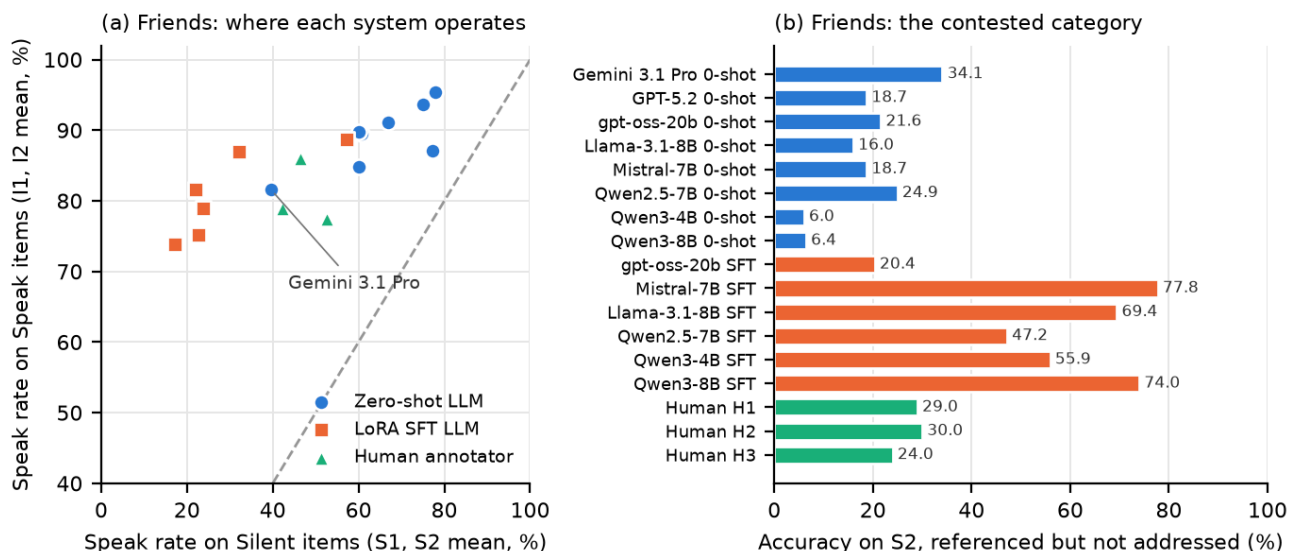


Figure 1. (a) Each system's rate of speaking at Speak-labelled points against its rate at Silent-labelled points on the Friends portion of Bhagtani et al. (2026), each rate averaged over the two categories that carry that label; the dashed line is no discrimination. This averaging differs from the count weighting of Section 5.1, so some rates differ: Gemini 3.1 Pro speaks at 59.5 percent of Silent points by count but at about 39.6 percent here, and it is speak-leaning on Friends only under the count weighting. Blue circles are zero-shot models, orange squares the LoRA fine-tuned models, green triangles the three human annotators. (b) Accuracy on S2 (mentioned but not addressed; label Silent) for the same systems. All values are taken from or computed from that paper's Tables 2, 3 and 4; the human values are on a 360-item subset.

5.3 What fine-tuning changed

The fine-tuning in Bhagtani et al. (2026) uses LoRA on the six open models with a batch sampler that draws a quarter of every batch from each of the four categories. That sampler sets the training prior to half Speak and half Silent and weights S2 no more than any other category, which matches what balanced accuracy rewards. The authors report that the largest gains came from S2 and S1, the two Silent categories.

On Friends the effect is visible in Figure 1. Averaged over categories, the five fine-tuned models other than gpt-oss-20b speak at 17.2 to 32.2 percent of Silent points, against 60.0 to 77.8 percent for the six open models zero-shot on the same averaging; their rate at Speak points falls too, from 84.9 to 95.4 percent to 73.8 to 87.0. Their gap between the two rates rises to between 52.4 and 59.6 points. gpt-oss-20b barely moves, which the authors attribute to a conflict between its built-in reasoning and the adapter's output format.

Two things happened, and the data separate them only partly. The operating point moved from the upper-right corner toward the middle; and the gap at the new operating point is larger than any zero-shot gap. Because the zero-shot gap at a corner is only a lower bound on what the same model's scores could deliver at a better threshold, the published numbers cannot say how much of the larger gap is new discrimination and how much a zero-shot model would have shown with its threshold moved. That is the measurement Section 9 asks for first.

What fine-tuning taught on S2 is a separate question, and here the human reference is direct. S2 points are

those at which the target was mentioned and someone else produced the next utterance. The annotators, shown these points, said the mentioned person should speak 70 to 76 percent of the time. Five of the fine-tuned models now say Silent on 47.2 to 77.8 percent of them. Figure 1(b) shows the result: the zero-shot models sit between 18 points below and 4 points above the annotators' range on this category, and fine-tuning moves five models away from the annotators in the direction of the label, to between 17 and 54 points above every annotator. There are two accounts of this. On one, the fine-tuned models learned a pragmatic distinction - being talked about is not being talked to - that untrained annotators miss. On the other, they learned how the Friends scripts, and the transcript-based labelling, resolve mentions: often by having someone else speak next, so that the mentioned character's not speaking is recorded as the correct decision. The conversation-analytic account in Section 4 predicts that a mention frequently makes a response relevant without obliging it, and that who then speaks is partly decided by who starts first; on that account the annotators' "Speak" is often a reasonable judgment and the label records a lost race or an authorial choice. The published evidence cannot decide between the two accounts. It can establish that the category on which fine-tuning gains most is the category on which the benchmark's label and its human reference disagree most, and that a model which beats the human reference there is agreeing with the corpus, not with people.

One further detail concerns the reasoning-trace variant. Its traces were distilled from Gemini 2.5 Flash by giving the teacher each training example together with its

correct label and asking for a one-sentence justification. Training on these traces raised Friends balanced accuracy for the model compared from 66.60 to 68.46. This is evidence that a frontier model can write a plausible reason for the corpus label when told the label. It is not evidence about whether the frontier model would have chosen that label.

5.4 A synthetic benchmark: opposite corners in one family

Nama et al. (2026) generate their data. They take question-answer examples, prompt GPT-4-family models (the paper names GPT-4o-mini and GPT-4-turbo-preview) to add a social context and an intervention type, and have GPT-4-turbo-preview write group conversations of 2 to 6 human speakers in which the assistant intervenes or stays silent; every turn becomes a Speak or Silent example over an eight-message window. The assistant intervenes on average 1.86 times per conversation, and 86.2 percent of its interventions occur without its being explicitly addressed.

The zero-shot results are a map of operating points. Qwen3-4B intervenes almost everywhere (false-intervention rate 0.897, macro F1 0.206); Qwen3-8B, from the same family, is silent everywhere (false-intervention rate 0.000, macro F1 0.466). Llama-3.1-8B and Llama-3.3-70B intervene falsely at rates of 0.731 and 0.749 and score 0.337 and 0.325. GPT-4o-mini and GPT-4.1-mini score 0.569 and 0.499, with false-intervention rates of 0.257 and 0.435. (The text attributes 0.567 to GPT-4o in the same passage; the table gives 0.569 for GPT-4o-mini. The table is used here.) That a model which never speaks outscores three that speak readily shows how much of the zero-shot ranking is the operating point.

Supervised fine-tuning lifts all six capable models to macro F1 between 0.697 and 0.747 and pushes them to the other corner: false-intervention rates of 0.036 to 0.163, and, in the authors' summary, missed-intervention rates of 0.488 to 0.549, which they call systematically over-conservative. The two smallest models collapse to always-Silent. Reinforcement learning with a reward that penalises missed and false interventions asymmetrically then reduces the missed-intervention rate to 0.186 to 0.218 and raises recall from 0.479 to 0.78 to 0.81, with a coefficient that the authors vary to trade one error against the other. Every stage of this sequence moves the system along a trade-off by changing the training signal's implicit or explicit cost ratio. None of the reported metrics separates movement along the trade-off from improvement of it.

The target deserves the same attention as in Section 5.2. The labels in this dataset are the interventions that GPT-4-turbo-preview chose to write when it generated each conversation, and the human turns around them were written by the same model in the same pass. The dataset therefore records a prompted language model's

participation policy, made self-consistent by construction, and fine-tuning teaches it to smaller models. That is a useful thing to do. It is hard to square with the conclusion that the policy cannot come from pretraining, since a pretrained and instruction-tuned model produced it. What the result shows is that a model asked to classify a conversation it did not write does not reproduce the policy that a related model used when writing it, and that training a student on the writer's choices closes the gap. Pisan et al. (2024) add a reason for care in treating such data as a picture of real groups: dialogues generated by standard language models resemble scripted dialogue more than spontaneous conversation in how much listener feedback they contain.

5.5 Meeting delegates: the silence corner

The bias reverses when a model stands in for a person. Hu et al. (2025) benchmarked language models acting as meeting delegates on real meeting transcripts and found model-specific tendencies: GPT-4 and GPT-4o balanced active and cautious engagement, Gemini 1.5 Pro was more cautious, and Gemini 1.5 Flash and the Llama 3 models more active. Khan et al. (2026) re-implemented a transcript-only delegate on GPT-4o and found that on 137 AMI meetings it did not attempt 51.4 percent of the absent participant's talking opportunities; a reflection-based variant with a longer raw context window did not attempt 49.5 percent. Their architecture, CAPA, keeps an explicit meeting state - active topic, decisions, open questions, stances, prior coverage, and who holds the floor - and separates forecasting, deciding and phrasing into modules around the same backbone. It reduced the unattempted share to 2.5 percent, doubled credited recovery of the participant's ideas from 26.1 to 52.2 percent, and raised decision F1 from 38.1 to 63.0, with hallucination at 0.6 percent. The paper describes no fine-tuning of the backbone. Across four backbones (GPT-4o, Gemini-2.5-Pro, Llama-3.3-70B and Qwen3.6-27B) the authors report that coverage and precision trade off along a single curve, with Gemini permissive and Qwen conservative.

Three things follow for the general claim. The direction of the zero-shot bias is not a property of language models; it depends on the model and on the framing of the role, and the same model family can be too eager in one benchmark and too reticent in another. The gap between a prompt-only delegate and a competent one was closed in this case by giving the model a representation of what the meeting needs, not by training it on speak-or-silent labels. And the backbone comparison is one of the few places in this literature where an operating-point trade-off across models is reported as such. Eckhaus, Berger and Stanovsky (2025) point the same way from a different setting: in online games of Mafia with human players, an agent whose decision to speak is made by a separately prompted scheduling call - with talkative-oriented and listening-

oriented variants of that prompt - produced message timing that the authors found closely mirrors that of the human players.

User studies of prompted multi-party frameworks are consistent with this. Mao et al. (2024) describe a group-chat assistant that decides what to say, when and to whom through LLM modules and report, from case studies and user studies, that it chimed in at appropriate times in goal-oriented group discussions. Liu et al. (2024) have a prompted model generate a covert train of thoughts in parallel with the conversation and rate its motivation to voice each one; in their user study participants preferred it to a next-speaker-prediction baseline over 82 percent of the time, on measures including turn-taking appropriateness. These are evaluations of perceived quality, not of agreement with a label, and they make no claim about discrimination at fixed decision points; they show that prompted systems can participate in a way people judge appropriate.

5.6 Where the text fixes the answer, untuned models discriminate

Section 4 predicted that points where the current speaker selects the next should be easy and points where participants self-select should be hard for any observer of a transcript. Four results bear on this directly.

Ma et al. (2026) built paired scenarios in which the same request is either addressed to the assistant by name or addressed implicitly, with the answer held fixed. Given the speaker-attributed transcript and a statement of the assistant's duties, Gemini 3.1 Pro chose to respond to 94.7 percent of named requests and 30.4 percent of implicit ones, and it preserved silence on 95.4 and 99.3 percent of the turns that called for it in the two conditions. Its decisions are highly sensitive to explicit addressing, and on implicit addressing it errs toward silence, which matches its I2 accuracy on AMI and SPGI in Bhagtani et al. (2026), 39.27 and 30.31 percent.

Liu et al. (2024), in a formative study preceding their framework, asked prompted GPT models to predict the next speaker in a multi-party chat corpus in which 95 percent of turns were self-selected and 5 percent allocated by the current speaker. GPT-4o was right 69.7 percent of the time when the turn was allocated and 12.1 percent of the time when it was self-selected; random guessing scored 12.7 percent. A fine-tuned GPT-3.5 scored 26.5 percent overall. When the target for self-selected turns was redefined so that "anyone" counted as correct, the fine-tuned model's accuracy rose to 81.0 percent overall and to 76.5 percent on allocated turns; the authors write that "removing the requirement to predict specific speakers in self-selection scenarios eliminates a major source of noise and unpredictability."

Fukuda et al. (2026) compared supervised models, text-only Qwen3 models, multimodal models and human subjects on AMI under an online protocol in which

nobody sees the future. Text-only models outperformed both the supervised models and the humans at next-speaker prediction, even though the humans had audio and video; humans outperformed the models at addressee detection and at predicting whether the speaker would change. The humans' F1 on next-speaker prediction was about 60 percent; the twelve participants were non-native speakers of English, which the authors flag as a limitation; model size did not help monotonically, with the 14-billion-parameter model ahead of both the 8- and 32-billion ones; and the segments that were hard for humans - frequent speaker changes, balanced participation - were hard for the models too. Their result contrasts with Inoue et al. (2025), who found GPT-4o only marginally above chance at addressee recognition in Japanese three-party conversations from what became the TEIDAN corpus (Mori et al., 2026), where explicit addressees appeared in about 20 percent of turns; in the same study GPT-4o predicted the next speaker with 46.0 percent accuracy, below the chance rate of 50 percent.

Wei et al. (2023), finally, trained next-speaker models on their three-party MultiLIGHT role-play data and reached 54.4 percent accuracy against a chance rate of 33.3, and 85.0 percent when the model was also shown the next utterance. In human evaluation of the resulting agents, however, the authors could not show that the trained turn-taking model improved the conversations over a random choice of speaker, and they suggest that in open-ended conversation, as long as an utterance makes sense, the choice of who says it does little harm.

The pattern across these four is the one Section 4 predicts. Where the next speaker is allocated by explicit address, untuned models discriminate well. Over all turns, not only allocated ones, one study found a text-only model predicting the next speaker better than human subjects did, though of the two models compared with the humans only one was significantly ahead (Fukuda et al., 2026). Where nobody is allocated, the observed next speaker is close to unpredictable for prompted models and for fine-tuned ones; redefining the target to accept any entitled speaker removes most of the error; and matching the observed speaker more often does not make the conversation better by human judgment. Earlier evaluations of language models on multi-party tasks reached compatible conclusions at coarser grain: zero-shot ChatGPT was weak on a number of the five multi-party tasks evaluated and GPT-4 promising (Tan, Gu and Ling, 2023), and addressee recognition depends on structural features of the conversation in a way response selection does not, with prompt sensitivity that varies by task (Penzo et al., 2024).

5.7 Within a turn: the signal is in the distribution, not in the answer

The cleanest evidence on whether language models carry the information needed to find speaking opportunities comes from the within-turn task. Umair, Sarathy and de Ruiter (2024) collected real-time responses from 118 listeners to 55 recorded single-speaker turns: each listener produced a brief backchannel wherever they perceived a chance to speak. In the labelling that Umair and de Ruiter (2026) apply to these data, a position is an opportunity when the share of listeners responding there exceeds a threshold set for each stimulus, which makes 16.2 percent of 5,195 word-adjacent positions positive. This is the only label in the literature reviewed here that is defined over a population of listeners rather than one participant, and the 2024 paper found that prompted language models could not reproduce it.

Umair and de Ruiter (2026) re-ran the comparison. Six instruction-tuned models from 7 to 72 billion parameters, under four prompting conditions - an expert definition, the listeners' own instruction, and two variants with imagined or true continuations - stayed near chance, with balanced accuracies between 0.49 and 0.54. Fine-tuning Llama-3.1-8B on 528 to 1,526 examples gave balanced accuracies of 0.50 to 0.57, not improving monotonically with data. Then the authors sampled continuations of each prefix from the same 8-billion-parameter model without task training, measured how semantically dispersed the continuations were, and applied a rule to the trajectory of that dispersion. Across 2,304 configurations fixed before evaluation, the F0.5 averaged 0.545 and ranged from 0.385 to 0.603, against 0.22 for the best prompted model, 0.27 for the best fine-tuned one, and about 0.15 for GPT-4o in the 2024 study, a figure the 2026 authors approximate from precision and recall reported under a global threshold, not under their stimulus-specific one. Two control signals computed from the same model's token probabilities - next-token entropy and length-normalised predictive entropy - reached 0.319 and 0.328 on average, both below the weakest semantic-uncertainty configuration. Most of the variance across configurations came from the decision rule, which the authors acknowledge and test with those controls.

The finding bears on the general claim in a specific way. Prompting asks the model to say where an opportunity is; fine-tuning on a small set teaches it to say so; and both fail. Reading the opportunity off the model's own uncertainty about how the turn will continue succeeds, with no task training of the model at all. Whatever information listeners use to anticipate these opportunities is, to a useful degree, present in what a model with no task training expects to come next. This is not evidence that prompted models will succeed at multi-party floor-taking, and it comes from one dataset of single-speaker stimuli; but it is direct evidence that "the model cannot do X when asked" and "the model

lacks the information for X" are different findings, and that for at least one turn-taking judgment the second is false.

5.8 Speech: thinner evidence, and a different remedy

Full-duplex speech models listen and speak at once and so must make the floor decision continuously. The evidence on them is thinner and less directly comparable, and it does not settle the training claim either way.

Models trained on two-channel conversational audio learn turn-taking behaviour from the data rather than from labels. dGSLM, trained on 2,000 hours of two-channel telephone conversation with no text or labels, produced more naturalistic turn-taking than a cascaded text-based system (Nguyen et al., 2022); Moshi models user and system speech in parallel streams so as to handle overlap and interruption (Defossez et al., 2024); and the voice activity projection objective trains a model to predict both speakers' upcoming voice activity, with turn shifts and backchannels then predictable zero-shot (Ekstedt and Skantze, 2022a) and with the model making use of prosody without being told to (Ekstedt and Skantze, 2022b). Extensions take the objective to three-party Japanese conversation, where models trained on triadic data beat baselines but accuracy depends on conversation type (Elmers et al., 2025), to multi-party audio-visual input from a single camera (Qi and Skantze, 2026), and to a robot mediator whose expected action is often not to speak at all (Cano et al., 2026). On the text side the same idea appears as a language model fine-tuned on dialogue to predict turn shifts (Ekstedt and Skantze, 2020) and as a fusion of an acoustic model with a language model that beats either alone on turn-taking and backchannel prediction (Wang et al., 2024). Multimodal input helps in dyadic settings as well: adding synchronised video and audio to a language model improved prediction of whether to stay silent, react briefly or respond in full by up to a factor of three over strong language-model baselines (Liao et al., 2025).

Evaluations of deployed full-duplex systems find them wanting. A judge model trained on human-human turn-taking found that existing spoken dialogue systems sometimes fail to speak up, can interrupt too aggressively, and rarely backchannel (Arora et al., 2025). Full-Duplex-Bench separates pause handling - whether a model takes over the turn when a user merely pauses - from backchannelling, smooth turn-taking and handling interruptions (Lin et al., 2025). In the multi-party Duplex-MPE benchmark, none of five open full-duplex systems showed a significant difference in response rate between requests that named the assistant and matched requests that did not, with paired differences of -1.75 to +0.60 points and the smallest p-value 0.2365, while the best of them, MiniCPM-o 4.5, led on three of four capabilities (Ma et al., 2026). The authors do not read this null as a failure. They write that

it "could reflect either successful inference of the addressee from context or a failure to recognise the name as an addressing cue, leaving the model similarly inclined to speak or remain silent in both conditions." And they report that MiniCPM-o produced speech after 95.45 percent of named requests to the assistant while keeping silent on 95.4 percent of the 4,704 requests addressed to other people or devices, so that it "responds frequently when addressed and usually preserves silence when someone else is addressed" (Ma et al., 2026).

For speech, then, the evidence does not establish a missing capability any more than the text evidence does: the paired test cannot tell a system that infers the addressee from context from one that ignores names, and the best system mostly held back when someone else was addressed. The remedy the literature points to for the floor decision is in any case not supervised speak-or-silent labels. It is training on multi-party conversational data with an objective that sees the floor, which the voice-activity-projection line shows can be self-supervised, together with the cues - gaze, orientation, voice identity - that tell a listener who is being addressed (Bohus and Horvitz, 2010; Skantze, 2021). A recent survey of full-duplex models names the scarcity of synchronous conversational data among the field's main obstacles (Chen and Yu, 2025), and multi-party synchronous data is scarcer still. A transcript-conditioned frontier model in the same benchmark discriminates addressing strongly (Section 5.6), so the distinction is available to a language model that reads the transcript; where the speech systems fall short of it, the published results do not say whether perception, training data or the decision itself is responsible.

6. What "emergent" and "must be trained" would have to mean

The 2026 conclusions use a vocabulary with a history. Wei et al. (2022) defined an ability as emergent if it is absent in smaller models and present in larger ones, so that it cannot be predicted by extrapolating from small-scale performance. Schaeffer, Miranda and Koyejo (2023) argued that many apparent emergent abilities are produced by the researcher's choice of metric: discontinuous or nonlinear metrics on fixed model outputs show sharp transitions where continuous metrics show smooth improvement. Lu et al. (2023) argued that claimed emergent abilities are better explained by in-context learning, memory and linguistic knowledge than by abilities arising unbidden with scale. The 2026 papers use "not emergent" in a looser sense - absent from what pretraining and instruction tuning provide, at any of the scales tested - and that sense is the one to test.

Three things would have to be shown, and the metric matters for each.

First, the information must be absent, not merely unexpressed. A thresholded binary answer at

temperature zero is exactly the kind of discontinuous readout Schaeffer, Miranda and Koyejo (2023) warn about: a model whose probability of speaking is higher at appropriate points but everywhere above the threshold looks identical, under this readout, to one that has no idea. Section 5.7 is a case in which a continuous readout of a model with no task training succeeds where the binary answer fails. None of the multi-party benchmarks reviewed here reports a continuous readout.

Second, scale must not help. The record here is mixed and partly confounded with operating point. In Bhagtani et al. (2026) the two frontier models show larger gaps between their speak rates at Speak and Silent points than the five open models under ten billion parameters on every corpus, which is what scaling would predict; in Nama et al. (2026) Llama-3.3-70B zero-shot is no better than Llama-3.1-8B, both intervening falsely at about three-quarters of Silent points; Fukuda et al. (2026) found no monotonic size effect among 8-, 14- and 32-billion-parameter Qwen3 models, and Umair and de Ruiter (2026) found no consistent benefit from scale under prompting. Where scale appears not to help, the operating point has not been held fixed, so the absence of a scale effect on a thresholded score does not show its absence on discrimination.

Third, training must be necessary, and "training" must be specified. Three senses are in play. In the strongest, supervised speak-or-silent labels are required. In a weaker one, some training on multi-party conversational data, possibly self-supervised, is required. In the weakest, the model needs the policy stated - its role, the cost of each kind of error, what it is there to contribute - by prompt or by scaffold. The published record contradicts the strongest sense as a necessity claim: a prompted scaffold closed most of a delegate's silence gap (Khan et al., 2026), a prompted scheduler produced human-like message timing (Eckhaus, Berger and Stanovsky, 2025), the predictive distribution of a model with no task training located speaking opportunities better than a fine-tuned one (Umair and de Ruiter, 2026), and a self-supervised objective learns turn-taking events without labels (Ekstedt and Skantze, 2022a). It supports the weakest sense for text transcripts where addressing decides the case (Section 5.6), and for speech it points to the middle sense as a remedy without showing it to be necessary (Section 5.8). The 2026 papers' evidence - large gains from supervised fine-tuning over zero-shot prompting - is compatible with all three senses, because supervised fine-tuning supplies a stated policy, a cost ratio and in-domain data at once.

The cleanest way to put the point is as a comparison the benchmarks ran only in part. The zero-shot prompt of Bhagtani et al. (2026) already states part of the policy in words, including the S2 convention that a participant mentioned in the third person should stay silent, and zero-shot models still scored 6.03 to 34.10 percent on S2 in Friends. That counts against the simplest form of

"must be told": stating the convention did not make the models apply it. It does not settle the question, because the rule was stated qualitatively, without a threshold, next to nine rules for speaking that compete with it. Fine-tuning with a balanced sampler supplies the operating point implicitly. A zero-shot condition that states the operating point quantitatively - "interrupting costs four times as much as missing a turn," or "speak only when you are named or asked a question no one else can answer," or a threshold applied to the model's own probability of Speak - would show how much of the gain the label-free statement of the policy recovers. That comparison is the difference between "must be trained" and "must be told".

7. The label: one participant's choice, or one model's

7.1 Human label variation, applied to the floor

The general argument is settled in neighbouring areas. Plank (2022) argues that variation among human labellers often reflects disagreement, subjectivity or multiple plausible answers rather than noise, and that treating one label as ground truth distorts data, modelling and evaluation alike. Pavlick and Kwiatkowski (2019) showed that disagreements about natural language inferences persist as more ratings are collected and as raters are given more context, and that models' uncertainty does not track the structure of human disagreement. Uma et al. (2021) survey methods for training and evaluating on distributions of labels rather than single ones.

Turn-taking benchmarks meet the problem in an especially sharp form, because in the largest benchmark built from human conversation the label is not a judgment at all. In Bhagtani et al. (2026) it is an event: whether one participant produced the next utterance. Section 4 explained why that event is, at self-selection points, partly the outcome of a race. The human evaluation then measured how often three people's judgments of appropriateness matched those events, with pairwise kappas of 0.38 to 0.57. Fukuda et al. (2026) reported moderate agreement among their participants as well, and found in a pilot that participants could often not commit to a single next speaker, so they allowed sets. Liu et al. (2024) found that accepting "anyone" as the answer for self-selected turns removed most of the error. Wei et al. (2023) found that matching the observed next speaker more often did not make agents better by human judgment. Each of these is a symptom of scoring against one realisation of an option.

The within-turn work shows the alternative. Umair, Sarathy and de Ruiter (2024) collected the responses of 118 listeners and defined an opportunity as a place where enough of them responded. A multi-party analogue would present each decision point to a panel and record the proportion who judge that the target

should take the floor, with the target's role stated. The label becomes a probability; a system's score becomes a calibrated probability or a ranking; and evaluation can report both discrimination against the panel's majority and calibration against the panel's proportion. Nothing about this requires new theory. It requires that benchmark builders stop reading the label off the transcript.

7.2 Synthetic labels record the generator's policy

Three of the benchmarks discussed use labels written by a language model: Nama et al. (2026) through GPT-4-turbo-preview's generated transcripts, Ma et al. (2026) through generated scripts that supply labels and answers, and Mitra et al. (2026) through a synthetic multi-party corpus built with off-the-shelf language models and speech synthesis. For Mitra et al. the argument of this section covers only that corpus, RolePlayConv; their other evaluation set, NOTSOFAR-1, consists of real meetings in which one speaker was designated as the assistant by "a hybrid procedure that aggregates LLM-based rankings of assistant-like behavior with independent human evaluations". Each of these is legitimate for what it is: a controlled environment in which the correct behaviour is defined by construction, which is how Ma et al. (2026) can pair explicit and implicit versions of the same request. But the correct behaviour in such a benchmark is the generator's policy, and evaluating a different model against it measures agreement between two policies. Fine-tuning on it teaches the generator's policy, including its habits. Because the generator is itself a prompted language model, a result that a prompted model cannot reproduce the labels zero-shot shows that two prompted models disagree, not that prompting cannot produce a coherent participation policy. Pisan et al. (2024) give a reason to expect that such policies will look more like scripted than spontaneous talk.

7.3 The role is part of the label

Mitra et al. (2026) condition turn-taking on an explicitly assigned role and report, relative to models without role conditioning, improvements "by over 40%" in precision and "more than 70%" in recall on meeting data and on their synthetic corpus. Cano et al. (2026) design for a robot mediator whose expected action is often to wait. Zhu and Callison-Burch (2025) describe a class of "overhearing" agents that monitor a conversation and intervene only when they can help, and Deng et al. (2023) and Lu et al. (2024) treat proactive initiative as a goal-directed behaviour with its own objectives. Each of these places the same conversation under different correct policies. A benchmark that asks whether "the assistant" should speak, without saying what the assistant is for, asks a question with several right answers and scores one.

8. What a speak-or-silent result licenses: a decision procedure

Algorithm 1 states the procedure the preceding sections imply for reading a published speak-or-silent result. It asks, in order, whose behaviour the label records, whether a human reference exists on the same item mix, whether the result reports scores or only decisions, and where the system's operating point lies; and it returns the strongest claim the result supports.

Applied to Bhagtani et al. (2026), the label source is EVENT, so step 3 limits every capability claim to allocated categories and reports S2 - where the target was mentioned and someone else spoke next - as agreement with recorded events at points that are not all allocated. On Friends, step 4 records Gemini 3.1 Pro above every re-weighted annotator and GPT-5.2 and gpt-oss-20b within their range. The five fine-tuned models other than gpt-oss-20b, at 65.12 to 72.52, are also above every annotator, but much of their gain is on S2, so under step 3 much of that standing is agreement with the corpus rather than with people, and step 7 reports the gain as a change of operating point plus an unidentified change in discrimination. On AMI and SPGI there is no human reference, so step 4 records only disagreement with the label. Step 5 then puts only Mistral-7B on AMI at a corner (it speaks at 84.9 percent of Speak points and 87.2 percent of Silent points under the count weighting of Section 5.1); the other small open models fall to the branch in which the gap is a lower bound, with category-averaged gaps of -11.3 to 8.7 points, near zero or negative. That supports "these models, so prompted, do not separate the cases", which is the defensible core of the original claim. Applied to Nama et al. (2026), step 2 phrases every claim as agreement with the generator. Umair and de Ruiter (2026) come closest to step 6: their system produces a continuous signal, they report results across the whole grid of decision rules rather than one chosen threshold, and their label comes from a listener panel - the only case in this literature that combines the two, although they report precision-weighted F scores rather than a full curve.

9. Measurements that would settle the open parts

Six measurements would turn the disagreement into numbers. None requires new models, and most require only re-scoring outputs that the benchmark authors already have.

1. **Scores, not decisions.** For each model and decision point, record the probability of Speak (from token probabilities, or from repeated sampling at non-zero temperature) and report the full receiver operating characteristic curve and its area, alongside the binary result. This identifies how much of the fine-tuning gain in Section 5.3 a

moved threshold would have delivered zero-shot.

2. **A stated operating point.** Add zero-shot conditions that state the cost ratio and the assistant's role in the prompt, at two or three values, and conditions that apply a threshold to the model's probability of Speak chosen on a validation split. Compare the gain from these label-free statements with the gain from fine-tuning. This is the test of "must be trained" against "must be told" (Section 6).
3. **A human reference on the scored mix.** Collect human judgments on a sample drawn with the test split's category proportions, or report the per-category human rates so the re-weighting of Section 5.2 is exact. Use more annotators than three, and state their expertise. Without this, a zero-shot "failure" has no ceiling.
4. **A panel label at self-selection points.** For points where no one is selected, replace the observed-next-speaker label with the proportion of a judge panel who consider it appropriate for the target to speak, with the target's role stated, following the listener-panel design of Umair, Sarathy and de Ruiter (2024). Report how much of the residual error at those points disappears.
5. **Allocated versus self-selected strata.** Report every result separately for points where the next speaker was allocated by the current speaker and points where participants self-selected, as Liu et al. (2024) did for next-speaker prediction. The account in Section 4 predicts high zero-shot discrimination in the first stratum and near-chance agreement with recorded events, for everyone, in the second.
6. **Role counterfactuals.** Present the same decision points under two stated roles (for instance, note-taker and facilitator) and measure whether the model's decisions change in the direction the roles imply, following the role-conditioning design of Mitra et al. (2026) but in a zero-shot condition. A model that changes its decisions appropriately with the stated role has the pragmatic knowledge and lacked only the specification.

For full-duplex speech, the analogue of measurement 5 is the paired explicit-implicit design of Ma et al. (2026), which found no response-rate difference between named and unnamed requests for any of five systems, a null its authors read as consistent either with inferring the addressee from context or with ignoring the name; the analogue of measurement 2 is a duty statement or role delivered in audio, whose effect the same benchmark reports only as an unscored probe.

Algorithm 1. Reading a speak-or-silent result

```

Input: benchmark B (labels, categories, metric), system S,
      optional human reference H on items of B
      (H = the individual annotators, each scored separately)
1. Identify the label source of B:
   EVENT (who spoke next in a recording)
   GEN   (a generator model's choice)
   PANEL (proportion of a judge panel)
2. If label source is GEN, phrase every claim below as
   agreement with the generator's policy
3. If label source is EVENT, restrict any capability
   claim below to categories where the next speaker is
   allocated (address, question to a named person);
   at self-selection points report agreement only
4. If H is missing:
   record "S disagrees with the label of B"
   (no claim about capability; ceiling unmeasured)
   Else re-weight each annotator to the category mix on
   which S was scored, and record whether S is below every
   annotator, within their range, or above every annotator
   on the metric; continue in every case
5. If S reports only decisions (no scores):
   compute speak rates  $r_{\text{speak}}$ ,  $r_{\text{silent}}$  at Speak- and
   Silent-labelled points
   if both rates > 0.8 or both < 0.2:
     record "operating point at a corner;
             discrimination not identified"
   else gap  $J = r_{\text{speak}} - r_{\text{silent}}$  is a lower bound
     on the best gap along S's curve
6. If S reports scores:
   report AUC and the curve; compare to H at H's
   operating point; claim about discrimination allowed
7. If S improves after training, attribute the gain to
   discrimination only if step 6 holds before and after;
   otherwise report it as a change of operating point
   plus an unidentified change in discrimination
Output: the strongest claim recorded, within the limits
       set by steps 2 and 3

```

10. Objections

"The human annotators were weak, so matching them is a low bar." It may be. But the benchmark's authors chose them as the reference and used them to claim that fine-tuned models reach human level; the same reference, applied consistently, puts the best zero-shot model there too. If the reference is too weak, the claim that fine-tuned models reach human level falls with it, and the claim that zero-shot models fail is left without any reference. Either way the conclusion "not emergent" is not supported by the human comparison as published.

"Balanced accuracy is the benchmark's chosen metric, so operating point is part of the task." That is a fair position for a deployment benchmark that states its costs. Balanced accuracy weights a missed turn and an intrusion equally after correcting for base rates, which is one cost ratio among many, and the zero-shot prompt told the models only that both errors matter, not how much. A model scored on a cost it was not given and judged to lack a capability has been judged on its guess about the scoring rule. When Nama et al. (2026) did state an asymmetric cost, in a reward, the system moved along the trade-off as instructed.

"The small open models genuinely fail on AMI and SPGI; this paper concedes it." It does. The disagreement is about the inference from that failure to

the general claim that the capability is not present in pretrained models and must be trained in. Section 5 shows the inference does not survive the frontier-model results, the human reference, the addressing results and the within-turn readout.

"Fine-tuned models exceed human balanced accuracy, so training adds something real." Training adds at least two real things: an operating point matched to the metric, and agreement with the corpus on S2. Whether it adds discrimination beyond what the zero-shot model's scores would show is unmeasured (Section 5.3). Whether agreement with the corpus on S2 is a gain depends on whether the corpus or the annotators are right about mentioned-but-not-addressed participants, and the published evidence does not settle that.

"Text transcripts are the wrong medium anyway; the real problem is in audio." For deployment, probably. The text-transcript claim is still worth getting right, because it is the claim being made about language models as such, and the speech evidence in Section 5.8 points to a remedy - multi-party training data with an objective that sees the floor - that is different from supervised speak-or-silent labels.

"Synthetic data is the only way to get labels at scale." It is a way to get a generator's policy at scale. Section 7.2 does not argue against using it; it argues that results on it be stated as agreement with that policy, and that a

prompted generator's having produced the policy is evidence against the claim that prompting cannot.

11. What is not known

It is not known how well any model discriminates appropriate from inappropriate floor-taking in multi-party text when its operating point is controlled, because no benchmark reviewed here reports scores. It is not known how much of the supervised fine-tuning gains in either 2026 benchmark a stated cost or a tuned threshold would recover without training. It is not known what the human ceiling is on AMI or SPGI under the transcript protocol of Bhagtani et al. (2026), or on the Friends test mix with a larger or trained panel. It is not known whether the S2 convention that fine-tuned models learn on Friends transfers to spontaneous multi-party talk, or whether it is a property of scripted dialogue. It is not known how far full-duplex speech models condition on addressing in multi-party audio; the one benchmark that tests it found no response-rate difference between named and unnamed requests, which its authors say fits either reading, while its best system mostly kept silent when someone else was addressed. It is not known whether the semantic-uncertainty readout that locates within-turn opportunities in single-speaker stimuli extends to deciding which of several listeners should take the floor. And it is not known what cost ratio users of a group assistant actually hold, or how much it varies with the assistant's role; every benchmark discussed here either leaves it unstated or chooses it for them.

12. Conclusion

The question whether knowing when to speak in a group is an emergent capability of language models or one that must be trained has been answered "must be trained" on the strength of zero-shot scores near chance and large fine-tuning gains. The published results behind that answer mix three quantities. Zero-shot scores are dominated by operating points that no prompt set as a threshold: small models speak almost everywhere, a frontier model is cautious on two corpora and eager on a third, and two models from one family sit at opposite corners. Labels record what one participant did next, or what a generating model wrote, and at the many points where participants self-select, that event is close to unpredictable for any observer. On the one benchmark portion with a human reference, the strongest zero-shot model matches the annotators once both are scored on the same mix, and fine-tuning gains most on the category where the label and the annotators disagree.

What the evidence supports is narrower and more useful. Where the text fixes who should speak, as with explicit address, untuned frontier models discriminate. Where it does not, the target needs to be redefined before any system can be said to fail it. Small open models prompted zero-shot do not separate the cases at their default operating point. For full-duplex speech systems,

the one multi-party test cannot say whether they ignore a name or infer the addressee without it, and the remedy the speech literature points to for the floor decision is multi-party data with a self-supervised objective. And at least one turn-taking judgment that prompting and fine-tuning both miss can be read directly from the expectations of a model with no task training about what comes next.

The tests that would settle the rest are cheap: report scores, state the cost and the role, score humans on the same mix, label self-selection points with a panel, and stratify by how the next speaker was chosen. Until a benchmark does these things, a zero-shot speak-or-silent score is a measurement of a model's guess about a cost it was not given, against an event it could not have predicted.

References

- Arora, S., Lu, Z., Chiu, C.-C., Pang, R., & Watanabe, S. (2025). Talking Turns: Benchmarking Audio Foundation Models on Turn-Taking Dynamics. arXiv:2503.01174. doi:10.48550/arXiv.2503.01174. [PREPRINT]
- Bhagtani, K., Anand, M., Xu, Y. C., & Yadav, A. K. S. (2026). Speak or Stay Silent: Context-Aware Turn-Taking in Multi-Party Dialogue. arXiv:2603.11409. doi:10.48550/arXiv.2603.11409. [PREPRINT]
- Bohus, D., & Horvitz, E. (2010). Facilitating multiparty dialog with gaze, gesture, and speech. International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction (ICMI-MLMI 2010), 1-8. doi:10.1145/1891903.1891910.
- Cano, A., Perez, G., Merino, L., & Gomez, R. (2026). Multimodal Voice Activity Projection for Social Robot Mediation: Expected Behavior and Deployment Constraints. arXiv:2609.28317. doi:10.48550/arXiv.2609.28317. [PREPRINT]
- Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., et al. (2006). The AMI Meeting Corpus: A Pre-announcement. Machine Learning for Multimodal Interaction (MLMI 2005), Lecture Notes in Computer Science, 28-39. doi:10.1007/11677482_3.
- Chen, Y., & Yu, H. (2025). From Turn-Taking to Synchronous Dialogue: A Survey of Full-Duplex Spoken Language Models. arXiv:2509.14515. doi:10.48550/arXiv.2509.14515. [PREPRINT]
- Clark, H. H., & Carlson, T. B. (1982). Hearers and Speech Acts. Language, 58(2), 332-373. doi:10.2307/414102.
- de Ruiter, J. P., Mitterer, H., & Enfield, N. J. (2006). Projecting the End of a Speaker's Turn: A Cognitive Cornerstone of Conversation. Language, 82(3), 515-535. doi:10.1353/lan.2006.0130.
- Defossez, A., Mazare, L., Orsini, M., Royer, A., Perez, P., Jegou, H., Grave, E., & Zeghidour, N. (2024). Moshi: a speech-text foundation model for real-time dialogue. arXiv:2410.00037. doi:10.48550/arXiv.2410.00037. [PREPRINT]
- Deng, Y., Lei, W., Lam, W., & Chua, T.-S. (2023). A Survey on Proactive Dialogue Systems: Problems, Methods, and Prospects. IJCAI 2023 Survey Track. arXiv:2305.02750. doi:10.48550/arXiv.2305.02750.
- Eckhaus, N., Berger, U., & Stanovsky, G. (2025). Time to Talk: LLM Agents for Asynchronous Group Communication in Mafia Games. arXiv:2506.05309. doi:10.48550/arXiv.2506.05309. [PREPRINT]
- Ekstedt, E., & Skantze, G. (2020). TurnGPT: a Transformer-based Language Model for Predicting Turn-taking in Spoken Dialog. arXiv:2010.10874. doi:10.48550/arXiv.2010.10874.
- Ekstedt, E., & Skantze, G. (2022a). Voice Activity Projection: Self-supervised Learning of Turn-taking Events. Interspeech 2022.

- arXiv:2205.09812. doi:10.48550/arXiv.2205.09812.
- Ekdstedt, E., & Skantze, G. (2022b). How Much Does Prosody Help Turn-taking? Investigations using Voice Activity Projection Models. *SIGDIAL 2022*. arXiv:2209.05161. doi:10.48550/arXiv.2209.05161.
- Elmers, M., Inoue, K., Lala, D., & Kawahara, T. (2025). Triadic Multi-party Voice Activity Projection for Turn-taking in Spoken Dialogue Systems. *Interspeech 2025*. arXiv:2507.07518. doi:10.48550/arXiv.2507.07518.
- Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 861-874. doi:10.1016/j.patrec.2005.10.010.
- Fukuda, R., Kano, T., Arora, S., Delcroix, M., Tawara, N., Ogawa, A., Chiba, Y., Ando, A., Chen, W., & Watanabe, S. (2026). Evaluating Large Language Models Abilities for Addressee, Turn-change, and Next Speaker Prediction in Meetings. arXiv:2606.17542. doi:10.48550/arXiv.2606.17542. [PREPRINT]
- Goffman, E. (1979). Footing. *Semiotica*, 25(1-2), 1-30. doi:10.1515/semi.1979.25.1-2.1.
- Gravano, A., & Hirschberg, J. (2011). Turn-taking cues in task-oriented dialogue. *Computer Speech and Language*, 25(3), 601-634. doi:10.1016/j.csl.2010.10.003.
- Gu, J.-C., Tao, C., & Ling, Z.-H. (2022). Who Says What to Whom: A Survey of Multi-Party Conversations. *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI 2022)*, 5486-5493. doi:10.24963/ijcai.2022/768.
- Heldner, M., & Edlund, J. (2010). Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, 38(4), 555-568. doi:10.1016/j.jwocn.2010.08.002.
- Hu, L., Yuan, S., Qin, X., Zhang, J., Lin, Q., Zhang, D., Rajmohan, S., & Zhang, Q. (2025). MEETING DELEGATE: Benchmarking LLMs on Attending Meetings on Our Behalf. arXiv:2502.04376. doi:10.48550/arXiv.2502.04376. [PREPRINT]
- Inoue, K., Lala, D., Elmers, M., Ochi, K., & Kawahara, T. (2025). An LLM Benchmark for Addressee Recognition in Multi-modal Multi-party Dialogue. *IWSDS 2025*. arXiv:2501.16643. doi:10.48550/arXiv.2501.16643.
- Khan, M., Kirstein, F., Ruas, T., & Gipp, B. (2026). Speak for Me: Giving LLMs the Situational Awareness to Participate in a Meeting. arXiv:2609.03923. doi:10.48550/arXiv.2609.03923. [PREPRINT]
- Lerner, G. H. (2003). Selecting next speaker: The context-sensitive operation of a context-free organization. *Language in Society*, 32(2), 177-201. doi:10.1017/S004740450332202X.
- Levinson, S. C., & Torreira, F. (2015). Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology*, 6. doi:10.3389/fpsyg.2015.00731.
- Liao, Z., Ouyang, Y., Lee, Y.-L., Yu, C.-P., Tsai, Y.-H., & Yin, Z. (2025). Beyond Words: Multimodal LLM Knows When to Speak. arXiv:2505.14654. doi:10.48550/arXiv.2505.14654. [PREPRINT]
- Lin, G.-T., Lian, J., Li, T., Wang, Q., Anumanchipalli, G., Liu, A. H., & Lee, H.-y. (2025). Full-Duplex-Bench: A Benchmark to Evaluate Full-duplex Spoken Dialogue Models on Turn-taking Capabilities. arXiv:2503.04721. doi:10.48550/arXiv.2503.04721. [PREPRINT]
- Liu, X. B., Fang, S., Shi, W., Wu, C.-S., Igarashi, T., & Chen, X. A. (2024). Proactive Conversational Agents with Inner Thoughts. arXiv:2501.00383. doi:10.48550/arXiv.2501.00383. [PREPRINT]
- Lu, S., Bigoulaeva, I., Sachdeva, R., Tayyar Madabushi, H., & Gurevych, I. (2023). Are Emergent Abilities in Large Language Models just In-Context Learning? *ACL 2024*. arXiv:2309.01809. doi:10.48550/arXiv.2309.01809.
- Lu, Y., Yang, S., Qian, C., Chen, G., Luo, Q., Wu, Y., Wang, H., Cong, X., Zhang, Z., Lin, Y., Liu, W., Wang, Y., et al. (2024). Proactive Agent: Shifting LLM Agents from Reactive Responses to Active Assistance. arXiv:2410.12361. doi:10.48550/arXiv.2410.12361. [PREPRINT]
- Ma, C., Feng, W., Jin, W., Dai, G., Xie, T., Ma, Y., Kang, Z., Zhao, X., Zheng, X., & Chao, F. (2026). Duplex-MPE: Benchmarking Multi-Party Interaction in Full-Duplex Dialogue. arXiv:2609.31948. doi:10.48550/arXiv.2609.31948. [PREPRINT]
- Mao, M., Ting, P., Xiang, Y., Xu, M., Chen, J., & Lin, J. (2024). Multi-User Chat Assistant (MUCA): a Framework Using LLMs to Facilitate Group Conversations. arXiv:2401.04883. doi:10.48550/arXiv.2401.04883. [PREPRINT]
- Mitra, S., Pandey, P., Jain, A., Sahith, S., & Girish, K. V. V. (2026). Adaptive Turn-Taking for Real-time Multi-Party Voice Agents. *Interspeech 2026*. arXiv:2606.13544. doi:10.48550/arXiv.2606.13544.
- Mori, T., Inoue, K., Elmers, M., Lala, D., & Kawahara, T. (2026). TEIDAN: A Multilingual Multiparty Dialogue Corpus. *ICMI Companion 2026*. arXiv:2609.00802. doi:10.48550/arXiv.2609.00802.
- Nama, V., Mendi, S., Ye, Z., & Bent, B. (2026). When2Speak: A Dataset for Temporal Participation and Turn-Taking in Multi-Party Conversations for Large Language Models. arXiv:2605.05626. doi:10.48550/arXiv.2605.05626. [PREPRINT]
- Nguyen, T. A., Kharitonov, E., Copet, J., Adi, Y., Hsu, W.-N., Elkahky, A., Tomasello, P., Algayres, R., Sagot, B., Mohamed, A., & Dupoux, E. (2022). Generative Spoken Dialogue Language Modeling. arXiv:2203.16502. doi:10.48550/arXiv.2203.16502.
- Ouchi, H., & Tsuboi, Y. (2016). Addressee and Response Selection for Multi-Party Conversation. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2133-2143. doi:10.18653/v1/D16-1231.
- Pavlick, E., & Kwiatkowski, T. (2019). Inherent Disagreements in Human Textual Inferences. *Transactions of the Association for Computational Linguistics*, 7, 677-694. doi:10.1162/tacl_a_00293.
- Penzo, N., Sajedinia, M., Lepri, B., Tonelli, S., & Guerini, M. (2024). Do LLMs suffer from Multi-Party Hangover? A Diagnostic Approach to Addressee Recognition and Response Selection in Conversations. arXiv:2409.18602. doi:10.48550/arXiv.2409.18602. [PREPRINT]
- Pilan, I., Prevot, L., Buschmeier, H., & Lison, P. (2024). Conversational Feedback in Scripted versus Spontaneous Dialogues: A Comparative Analysis. *SIGDIAL 2024*, 440-457. arXiv:2309.15656. doi:10.48550/arXiv.2309.15656.
- Plank, B. (2022). The 'Problem' of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation. *EMNLP 2022*. arXiv:2211.02570. doi:10.48550/arXiv.2211.02570.
- Qi, H., & Skantze, G. (2026). MuVAP: Multimodal Multiparty Voice Activity Projection for Turn-taking Prediction in the Wild. arXiv:2606.16731. doi:10.48550/arXiv.2606.16731. [PREPRINT]
- Raux, A., & Eskenazi, M. (2012). Optimizing the turn-taking behavior of task-oriented spoken dialog systems. *ACM Transactions on Speech and Language Processing*, 9(1), 1-23. doi:10.1145/2168748.2168749.
- Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, 50, 696-735. doi:10.2307/412243.
- Schaeffer, R., Miranda, B., & Koyejo, S. (2023). Are Emergent Abilities of Large Language Models a Mirage? arXiv:2304.15004. doi:10.48550/arXiv.2304.15004.
- Schegloff, E. A. (2000). Overlapping talk and the organization of turn-taking for conversation. *Language in Society*, 29(1), 1-63. doi:10.1017/S0047404500001019.
- Sciar, M., Choi, Y., Tsvetkov, Y., & Suhr, A. (2023). Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting. *ICLR 2024*. arXiv:2310.11324. doi:10.48550/arXiv.2310.11324.
- Skantze, G. (2021). Turn-taking in Conversational Systems and Human-Robot Interaction: A Review. *Computer Speech and Language*, 67, 101178. doi:10.1016/j.csl.2020.101178.
- Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., et al. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy*

- of Sciences, 106(26), 10587-10592. doi:10.1073/pnas.0903616106.
- Tan, C.-H., Gu, J.-C., & Ling, Z.-H. (2023). Is ChatGPT a Good Multi-Party Conversation Solver? Findings of EMNLP 2023. arXiv:2310.16301. doi:10.48550/arXiv.2310.16301.
- Traum, D. (2004). Issues in Multiparty Dialogues. *Advances in Agent Communication, Lecture Notes in Computer Science*, 201-211. doi:10.1007/978-3-540-24608-4_12.
- Uma, A., Fornaciari, T., Hovy, D., Paun, S., Plank, B., & Poesio, M. (2021). Learning from Disagreement: A Survey. *Journal of Artificial Intelligence Research*, 72, 1385-1470. doi:10.1613/jair.1.12752.
- Umair, M., & de Ruiter, J. P. (2026). Using Semantic Uncertainty to Estimate Transition Relevance in Turn-taking. arXiv:2609.10934. doi:10.48550/arXiv.2609.10934. [PREPRINT]
- Umair, M., Sarathy, V., & de Ruiter, J. P. (2024). Large Language Models Know What To Say But Not When To Speak. arXiv:2410.16044. doi:10.48550/arXiv.2410.16044.
- Wang, J., Chen, L., Khare, A., Raju, A., Dheram, P., He, D., Wu, M., Stolcke, A., & Ravichandran, V. (2024). Turn-taking and Backchannel Prediction with Acoustic and Large Language Model Fusion. ICASSP 2024. arXiv:2401.14717. doi:10.48550/arXiv.2401.14717.
- Ward, N., & Tsukahara, W. (2000). Prosodic features which cue back-channel responses in English and Japanese. *Journal of Pragmatics*, 32(8), 1177-1207. doi:10.1016/S0378-2166(99)00109-5.
- Wei, J., Shuster, K., Szlam, A., Weston, J., Urbanek, J., & Komeili, M. (2023). Multi-Party Chat: Conversational Agents in Group Settings with Humans and Models. arXiv:2304.13835. doi:10.48550/arXiv.2304.13835.
- Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., et al. (2022). Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research*. arXiv:2206.07682. doi:10.48550/arXiv.2206.07682.
- Wen, B., Yao, J., Feng, S., Xu, C., Tsvetkov, Y., Howe, B., & Wang, L. L. (2024). Know Your Limits: A Survey of Abstention in Large Language Models. *Transactions of the Association for Computational Linguistics*. arXiv:2407.18418. doi:10.48550/arXiv.2407.18418.
- Zhao, T. Z., Wallace, E., Feng, S., Klein, D., & Singh, S. (2021). Calibrate Before Use: Improving Few-Shot Performance of Language Models. ICML 2021. arXiv:2102.09690. doi:10.48550/arXiv.2102.09690.
- Zhu, A., & Callison-Burch, C. (2025). Overhearing LLM Agents: A Survey, Taxonomy, and Roadmap. arXiv:2509.16325. doi:10.48550/arXiv.2509.16325. [PREPRINT]