TIDES: A Longitudinal Bilingual Dataset for Modeling Multi-Party Social Dynamics

Heechan Lee[1,*], **Jeonggyu Kang**[1,*], **Junho Myung**[1], **Jaywoong Jeong**[1], **Juho Kim**[1,2], **Joseph Seering**[1]
KAIST[1], SkillBench[2]
{hclee99, jeonggyumarkk, seering}@kaist.ac.kr

Abstract

Group conversations are fundamental to human collaboration, yet standard large language models (LLMs) still struggle with the complexities of multi-party interaction. This challenge persists in part because existing group conversation datasets are often limited to short-term lab settings with contrived tasks, failing to capture the long-term social dynamics of real-world teams. To bridge this gap, we introduce TIDES, a high-resolution longitudinal dataset tracking 12 university project teams over a full semester. Comprising **75,971** utterances in both English and Korean from in-person meetings, TIDES provides a naturalistic record of teams working on self-managed projects. Our socio-structural annotations—covering interaction types, emergent roles, and development stages—allow for modeling of team evolution over months. Experiments show that fine-tuning on TIDES improves next-speaker prediction by 13.8 percentage points over a bigram baseline (64.53%) and yields performance comparable to strong proprietary zero-shot models. The model also comes within 2.1 percentage points of the published state of the art on the AMI Meeting Corpus while using approximately 42% less training data. However, human evaluations suggest that better next-speaker prediction does not necessarily yield more natural or coherent utterances, as fine-tuned models were generally less preferred than vanilla models. This potential mismatch motivates further study of how structural modeling can support natural multi-party generation.

tides.cstlab.org github.com/cstl-kaist/TIDES_dataset

1 Introduction

Group conversations are fundamental to human collaboration, yet standard large language models (LLMs), despite their rapid progress in dyadic settings, continue to struggle when deployed in multi-party interactions (Tan et al., 2023; Wei et al., 2023). Unlike dyadic settings, successful group collaboration requires complex social coordination: understanding latent social dynamics (Zhou et al., 2025; Gu et al., 2022) and managing turn-taking across multiple speakers (Ekstedt & Skantze, 2020; Hilgert & Niehues, 2025; Castillo-López et al., 2025). This challenge is further compounded by the fact that social relationships and team dynamics are not static, but continuously evolve over prolonged interactions (Fang et al., 2025). Developing agents capable of naturalistic multi-party interaction therefore necessitates high-quality, longitudinal resources that capture authentic social dynamics as they naturally unfold.

However, existing multi-party datasets remain insufficient for capturing the social dynamics of collaboration as they unfold in the wild. Many widely-used datasets rely on lab-based observations (Carletta et al., 2005; Karadzhov et al., 2023), scripted media (Poria et al., 2019; Chen et al., 2020; Zhu et al., 2021), or synthesized conversations (Kirstein et al., 2025; Jang

[*]Equal contribution.

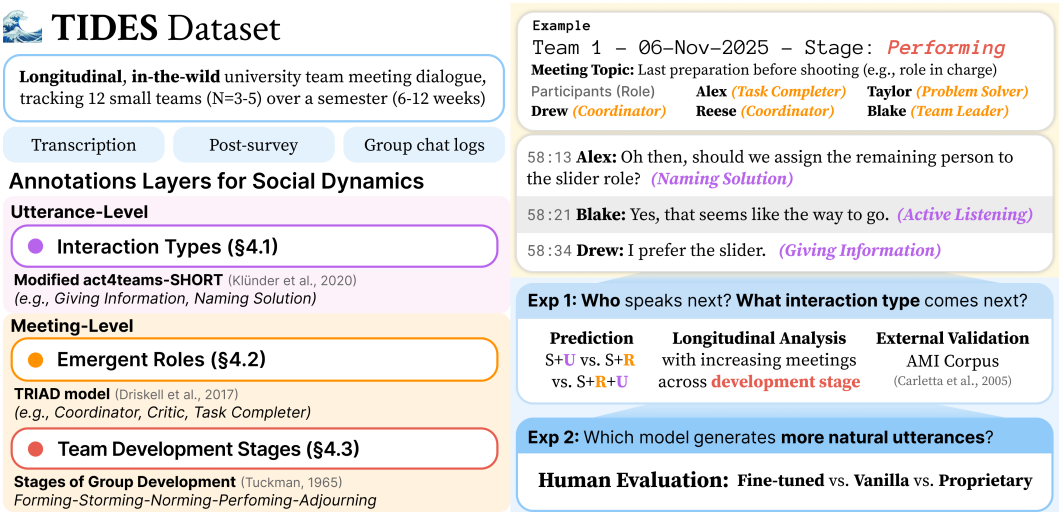

Figure 1: Overview of TIDES. TIDES is a longitudinal, in-the-wild bilingual dataset of university team project meetings, comprising 12 teams, 88 dated meetings represented in 104 transcript files, and 75,971 utterances collected over 6–12 weeks. Beyond transcripts, it provides layered annotations for social dynamics, including utterance-level interaction types, meeting-level emergent roles, and team development stages. The figure also shows an annotated example and the tasks enabled by these annotations.

et al., 2023), and while such settings can approximate multi-party interaction, they often miss the real stakes, shared history, and evolving interpersonal dependencies that characterize authentic teamwork. These social conditions matter because latent social dynamics—such as influence, alignment, tension, and role occupancy—do not simply appear within a single isolated exchange; they emerge through repeated collaboration as team members negotiate responsibility, expertise, and participation over time. However, even naturalistic meeting corpora (Carletta et al., 2005; Shriberg et al., 2004; Van Segbroeck et al., 2020) mostly capture only short-term episodes, limiting researchers' abilities to study how such dynamics accumulate and shift across a team's lifespan. Prior work has also centered on predefined or assigned roles (Jurgens et al., 2023), despite real teams more often exhibiting emergent functional roles that arise from ongoing interaction and changing task demands (Benne & Sheats, 1948).

To begin to bridge this gap, we introduce TIDES, **T**eam **I**nteraction and **D**ynamic **E**mergent **S**ocial-roles, a longitudinal bilingual dataset tracking real-world university project teams over a full semester. University courses provide a naturalistic setting where teams form organically, collaborate over weeks toward shared goals, and operate under real stakes (i.e., academic grading). We tracked 12 teams over a semester across a diverse range of university classes, recording their actual project meetings to collect a total of **75,971** utterances from 88 dated meetings (about 110 hours), represented in 104 transcript files because some longer meetings were split into multiple parts. To capture not only what teams said but also how their latent social dynamics evolved, participants completed post-meeting surveys after each session, reporting collaboration satisfaction and their perceptions of each member's emergent role. The resulting dataset pairs privacy-preserved transcripts with longitudinal social annotations, including meeting-level emergent roles, utterance-level interaction types, team development stages, and silence gaps (annotated examples in Appendix C).

The longitudinal nature of TIDES, in combination with the social annotations, enables us to take steps toward a vision of more socially-aware LLMs for group conversation, addressing four initial questions: (1) *Can models learn turn-taking and interaction patterns from longitudinal team data?* We evaluate next-speaker and next-intention prediction under three context conditions that vary the balance between utterance content and social structure. (2) *How quickly does performance adapt to a specific team as more meetings are observed?* We train on

incrementally more meetings from individual teams, measuring when prediction performance plateaus. (3) *Do these patterns generalize to unseen meeting corpora?* We evaluate on the AMI Meeting Corpus, directly comparing against published state-of-the-art performance. (4) *Does structural understanding translate to generating utterances that humans find natural?* We compare fine-tuned, vanilla, and proprietary models through automatic metrics and a human evaluation on Prolific.

Experiments show that models fine-tuned on TIDES improve prediction of team-level turn-taking patterns. For next-speaker prediction, the primary fine-tuned condition reaches 64.53%, a 13.8 percentage-point improvement over the bigram baseline and performance comparable to strong proprietary zero-shot models. In chronological single-team analyses, most observed gains occur within the first three to four meetings, which we treat as descriptive evidence of rapid team-specific adaptation (§5.2). Through transfer to the AMI Corpus (Carletta et al., 2005), our model achieves comparable performance to the published state of the art while using approximately 42% less training data. However, preference ratings from human evaluators suggest that improvements in conversational structure prediction do not necessarily lead to more human-preferred utterance generation. Although the fine-tuned models outperform the baselines in predicting the correct next speaker, human judges tend to prefer the vanilla models in terms of naturalness and coherence across multiple dimensions. These findings indicate a potential mismatch between optimizing for conversational structure and producing content that users perceive as natural, motivating further investigation into how LLM agents can generate more natural utterances and what kinds of naturalness users expect from them.

While our experiments primarily evaluate local prediction and generation within short context windows, the longitudinal annotations in TIDES provide the foundation for modeling how team dynamics evolve across meetings; initial evidence, such as the correlation between development stage and prediction accuracy (Appendix K), begins to support this direction.

2 Related Work

Multi-Party Dialogue Datasets. Prior work has introduced a range of multi-party meeting and discussion datasets, including the AMI and ICSI Meeting Corpora (Carletta et al., 2005; Shriberg et al., 2004), which provide foundational recordings of real meetings. More recent resources such as MeetingBank (Hu et al., 2023), QMSum (Zhong et al., 2021), and ELITR (Nedoluzhko et al., 2022) have expanded the scale and utility of meeting corpora, particularly for summarization and automatic minuting, while datasets like DeliData (Karadzhov et al., 2023) focus on deliberative problem-solving interactions and MPDD (Chen et al., 2020) and FAME (Kirstein et al., 2025) provide controlled or synthetic settings for analyzing interpersonal dynamics. However, most existing resources are limited in at least one critical dimension: most capture short-term interactions rather than longitudinal collaboration, some are drawn from scripted or institutional settings that do not fully reflect authentic teamwork, and others rely on assigned, fictional, or synthetic roles rather than emergent team dynamics. In contrast, TIDES captures real-world collaborative behavior *in the wild*, at scale and over a longer time span, enabling the study of how social roles, interaction patterns, and team development evolve throughout the teamwork.

Socially-Aware Dialogue Modeling. Understanding group dynamics computationally requires more than just dialogue transcripts. It demands annotations of the social structures that shape interaction. Organizational science has long studied how teams develop over time, from Tuckman's stage model (Tuckman, 1965) to emergent role theories (Kauffeld et al., 2018), yet these frameworks have seen limited adoption in NLP, due to the lack of suitably annotated data. Existing benchmarks for social intelligence rely on dyadic or scripted conversations (Zhou et al., 2024; Zhan et al., 2023; Rashid & Blanco, 2018; Tigunova et al., 2021; Jia et al., 2021; Jurgens et al., 2023), which do not capture the longitudinal, multi-party dynamics that characterize real teamwork. TIDES addresses this with socio-structural annotations—interaction types, emergent roles, and development stages. These annotations

enable models to tackle core aspects of group dynamics, such as predicting who speaks next and what type of contribution they make.

Conversational Agents in Group Settings. From an HCI perspective, researchers have investigated how conversational agents can support cooperation and facilitate participation in group discussions (Claggett et al., 2025; Kim et al., 2020; Houde et al., 2025), with a complementary line equipping agents with internal reasoning about *when and why* to speak (Liu et al., 2025). Such approaches assume that richer reasoning about group structure will lead to more natural contributions. However, systematic human evaluation of this assumption in multi-party settings remains limited, as prior studies often rely on structural metrics or LLM-based judges and are typically conducted in simulated or short-term settings. By combining longitudinal social annotations with human evaluation, TIDES enables the relationship between structural understanding and natural utterance generation to be examined directly.

3 Data Collection

We collected authentic, longitudinal team dynamics by tracking real university project teams throughout a semester. Teams formed organically, determined project goals and milestones within the guidelines of different course projects, collaborated under real stakes (i.e., grade evaluation), and interacted repeatedly over weeks toward shared goals.

3.1 Collection Procedure

We recruited 12 student teams (total N = 50 recruited participants) enrolled in university courses at full-time Korean universities during the Fall 2025 semester. Participants were recruited through university- and student-managed announcement boards and lists. Teams ranged from 3 to 5 members, and their projects ranged in duration from 6 to 12 weeks across a variety of disciplines, including Design, Computer Science, and Industrial Engineering. Prior to the study, all participants attended an orientation session where they were briefed on data collection procedures. Throughout the semester, teams were asked to keep audio recordings of all team project meetings and to submit the recordings to the research team after each session. Following each meeting, every team member individually completed a post-meeting survey so that we could capture meeting-specific perceptions of collaboration and track how team dynamics changed over time. At the end of the semester, each team completed a final team-level survey to reflect on their overall collaboration and project outcome after their last meeting. Additionally, chat message logs regarding the project between team members were collected from all but one team, which did not use a messaging platform during the project. This experiment was approved by our institution's Institutional Review Board and we compensated each team with 500,000 KRW (approximately 325 USD).

3.2 Post-processing

Converting 110+ hours of bilingual group conversation audio into research-ready transcripts required a six-stage pipeline. We first transcribed all recordings with Whisper Large-V3 (Radford et al., 2023), then assigned speaker labels through a custom diarization pipeline combining pyannote 3.1 (Bredin, 2023; Plaquet & Bredin, 2023) VAD with ECAPA-TDNN (Desplanques et al., 2020), re-clustering to the known team size.

Next, we removed personally identifiable information using Microsoft Presidio (Microsoft, 2023) for the five English-speaking teams and a locally-run Qwen3-30B-A3B (Qwen Team, 2025) for the seven Korean-speaking teams, where no comparable off-the-shelf tool exists. All Korean transcripts were then translated to English by the same local LLM. To resolve the cross-session speaker identity problem—where the same person may receive different labels across recordings—we unified speaker labels using WeSpeaker (Wang et al., 2023) embeddings and the Hungarian algorithm (Kuhn, 1955), producing 422 consistent mappings across 12 teams. Then, at least one representative from each team validated the full transcript against the original audio recordings and corrected errors. Validators

were compensated at 35,000 KRW (approximately 23 USD) per hour of reviewed audio. The authors manually verified all data to prevent leakage of personal information, and redacted eight utterances containing offensive speech. After transcript anonymization, we refined utterance boundaries with GPT-5-mini guided by Language Development Project transcription rules (MacWhinney, 2000), yielding 1,238 splits and 309 merges across the processed corpus.

Full pipeline details, model configurations, and processing statistics are provided in Appendix F.

4 Data Annotation

After collecting and post-processing the recordings, we constructed multiple annotation layers to capture social dynamics in small teams. These annotations drew on different sources: meeting transcripts for utterance-level interaction types, post-surveys for emergent role assignment, and end-of-semester group reflection for meeting-level team development stages. As these layers capture complementary constructs from different perspectives and timescales, they should be interpreted according to their sources rather than as interchangeable measurements of a single latent social state.

4.1 Utterance-level Interaction Type

We annotated utterance-level interaction types using a modified version of `act4teams-SHORT` (Klünder et al., 2020). Derived from the original `act4teams` taxonomy (Kauffeld et al., 2018), this taxonomy was designed to capture interaction behaviors in team problem-solving and collaboration. Although we considered standard dialogue-act taxonomies such as AMI-DA (Hain et al., 2007) and MRDA (Shriberg et al., 2004), we selected an act4teams-based scheme because our goal was to characterize collaboration-oriented intentions rather than general conversational or discourse functions. However, because `act4teams-SHORT` largely omits the social activity dimension, we extended it to better capture the social signals. Specifically, drawing from the original `act4teams` taxonomy (Kauffeld et al., 2018), we separated *Counterproductivity* into *Social Negative* and *Task/Process Negative*, and readopted *Social/Humor*, *Active Listening*, and *Other/Neutral*. This resulted in our modified act4teams-SHORT scheme with a total of 15 categories.

To construct human-labeled gold data, we recruited 260 annotators through Prolific and assigned three independent annotators to each utterance. We then aggregated labels using majority voting and retained labels agreed upon by at least two of the three annotators, resulting in 5,705 human-labeled gold utterances ($\simeq$ 7.5% of all utterances). We then fine-tuned `Gemma-3-12B` on the gold data and evaluated several candidate models under a leave-one-team-out setup, in which one team was held out from fine-tuning to reduce data leakage across teams. We selected `Gemma-3-12B` for large-scale annotation because it achieved the highest macro-F1 score (0.54). In addition, two authors jointly reviewed sample outputs from the candidate models to confirm the plausibility of the resulting annotations before large-scale application. Considering the complexity of the 15-category taxonomy and the subjective nature of the task, and the moderate agreement among human annotators (Fleiss' $\kappa = 0.400$) (Wong et al., 2021), we used the fine-tuned model to annotate the remaining 70K+ utterances. Additional details on the crowdsourcing process, model selection, and analysis on gold data are provided in Appendix A.

4.2 Emergent Roles

We assigned one peer-perception-based emergent role to each participant at each meeting. In the survey, participants evaluated their teammates along three dimensions in the TRIAD model (Driskell et al., 2017)—*Dominance, Sociability, and Task Orientation*—using nine 5-point Likert items (three items per dimension). We chose this dimension-based design rather than asking participants to directly assign one of the 13 TRIAD roles, as direct role categorization is difficult for non-expert participants. Detailed survey items are provided in Appendix E.

Because our survey used a 5-point Likert scale whereas TRIAD roles are defined in a different coordinate space (7-point Likert scale), we normalized the survey scores before role assignment. We used team-specific cumulative normalization, where each participant's score at meeting m was standardized using the responses from the same team observed up to and including that meeting:

$$z_{t,m,i}^{(k)} = \frac{x_{t,m,i}^{(k)} - \mu_{t,\leq m}^{(k)}}{\sigma_{t,\leq m}^{(k)}}. \tag{1}$$

Here, $x_{t,m,i}^{(k)}$ denotes participant i's average survey score on dimension k in team t at meeting m, and $\mu_{t,\leq m}^{(k)}$ and $\sigma_{t,\leq m}^{(k)}$ denote the mean and standard deviation of that dimension computed from all responses in team t up to meeting m. We then assigned each participant the role whose TRIAD prototype, $\tilde{\mathbf{p}}_r$, was closest in Euclidean distance to the normalized score vector with both represented in the same standardized coordinate space:

$$\hat{r}_{t,m,i} = \arg\min_{r \in \mathcal{R}} \left\| \mathbf{z}_{t,m,i} - \tilde{\mathbf{p}}_r \right\|_2. \tag{2}$$

$\hat{r}_{t,m,i}$ denotes the emergent role assigned to participant i in team t at meeting m. To support alternative analyses beyond discrete role labels, TIDES also includes the raw scores on the three dimensions for each participant at each meeting.

4.3 Team Development Stage Annotation

To annotate how teams evolved over time, we assigned each meeting a stage from Tuckman's team development model: Forming, Storming, Norming, Performing, or Adjourning (Tuckman, 1965). During the final post-data collection meeting, team members gathered and reviewed previous meetings prepared by the research team and discussed which development stage best characterized the team stage of each meeting. The agreed-upon stage was then recorded as a meeting-level label.

4.4 Annotated Dataset Summary

After post-processing and annotation, the dataset contains 75,971 utterances from 88 dated meetings, represented in 104 transcript files, involving 50 recruited participants across 12 student teams. Team sizes ranged from 3 to 5 members; 7 teams primarily communicated in Korean and 5 in English. Using post-meeting peer-perception surveys, we additionally assigned 352 emergent-role labels, with one label for each participant at each meeting for which survey responses were available.

The annotated interaction types were skewed toward a small number of frequent collaborative behaviors. *Giving Information* was the most common category (35.29%), followed by *Active Listening* (13.09%) and *Linking Solutions* (12.19%), suggesting that the meetings were dominated by task-relevant information sharing, solution-building, and responsive discussion. Less frequent categories such as *Task/Process Negative* (0.85%), *Social Negative* (0.61%), and *Linking & Connecting* (0.11%) appeared only rarely. For emergent roles, *Problem Solver* (17.61%), *Coordinator* (16.76%), and *Critic* (15.91%) were the most common, indicating that task-oriented and coordination-oriented roles appeared more often than socially disruptive or off-task roles. Detailed team-level statistics and full label distributions are provided in Appendix B; annotated transcript excerpts are shown in Appendix C.

5 Experiments

We evaluate TIDES through two experiments, described below.

5.1 Experimental Setup

Common configuration. Unless stated otherwise, we fine-tune with LoRA (rank 16, $\alpha = 32$) on Gemma-3-12B-IT, trained on 32,243 samples (Teams 1, 2, 3, 5, 8, 9, 10, 11, and 12), validated

on 1,857 samples (Team 4), and tested on 5,094 samples (Teams 6&7) with 5-turn context windows. Teams 6 and 7 together comprise approximately 13% of all utterances, represent 3- and 4-member configurations, and span the full Tuckman lifecycle; holding them out also keeps the three largest teams in training. Team 4, the smallest team, is used for validation to minimize the amount of training data withheld. Fine-tuned models use completion-only loss; inference uses single-token logit scoring for open-source models and generative decoding (temperature 0) for proprietary APIs. All results are single runs. We additionally validate with Llama-3.1-8B-Instruct and benchmark four proprietary models (GPT-5.4, GPT-5.4-mini, Opus 4.6, Sonnet 4.6) in zero-shot. All Korean transcripts were translated to English during post-processing (§3.2); language labels (Korean/English) throughout refer to the language originally spoken during meetings.

We design two experiments:

1. **Prediction and analysis.** To probe whether models can learn group-level turn-taking patterns from TIDES, we evaluate next-speaker prediction (3–5 classes per team) and next-intention prediction (14 substantive classes, excluding *Other/Neutral*) under three context conditions: **SU** (Speaker + Utterance), **SRU** (+ annotated role and Tuckman stage), and **S+R** (structure only—speaker IDs, roles, and stages, with all utterance text removed). To investigate how much team-specific data is required—a practical consideration for deploying such models to new teams—we further analyze *data efficiency* by training single-team models with chronologically increasing meetings on Team 10 (Korean, 4 members) and Team 5 (English, 5 members). To test whether TIDES-trained models generalize beyond our corpus, we conduct *external validation* on the AMI Meeting Corpus (Carletta et al., 2005) (138 meetings, 4 English speakers each) using Llama-3.1-8B for direct comparison with Hilgert & Niehues (2025), who report 47.85% using AMI plus MultiLIGHT data.
2. **Generation.** To examine whether structural understanding translates to realistic utterance generation, we compare three conditions: **FT-Plain** (fine-tuned, plain generation: "SPEAKER: utterance"), **FT-Reason** (fine-tuned with social-cue reasoning: the model generates an explicit role/intention prediction before producing the utterance), and **Vanilla** (Gemma-3-12B without fine-tuning), plus four proprietary models. Each generates 100 continuations for each generation length (1, 3, 5 turns). We evaluate with automatic metrics (speaker accuracy, semantic similarity based on `all-MiniLM-L6-v2`), a human evaluation on Prolific (N=66 evaluators, 1,142 pairwise judgments, £4.50/evaluator), and LLM-as-judge evaluation.

5.2 Results

Experiment 1a: Speaker prediction. As a reference point, the bigram baseline—which predicts the most frequent next speaker given the current speaker—achieves 50.75%. Fine-tuning on TIDES yields large gains (Table 1): Gemma FT-SU reaches 64.53%, +13.8 pp over the bigram baseline. Adding role labels helps minimally (FT-SRU 64.82%, Δ = +0.3 pp), but removing text entirely still matches performance (FT-S+R 65.74%), indicating that structural features alone are sufficient for predicting who speaks next. This may reflect that utterance text introduces team-specific lexical patterns (e.g., project topics, jargon) that do not transfer to held-out teams, whereas structural features—speaker order, annotated roles, and development stages—capture more generalizable turn-taking regularities. Fine-tuning outperforms all proprietary zero-shot models, with Gemma FT-SU (64.53%) exceeding Opus 4.6 (63.25%) and GPT-5.4 (61.97%). Llama-3.1-8B replicates all patterns with comparable accuracy (Appendix G). A longitudinal signal is also evident. Speaker prediction accuracy increases from 57.8% in Forming-stage meetings to 70.3% in Performing-stage meetings (+12.5 pp; Appendix K), consistent with the theoretical expectation that established teams develop more predictable patterns.

Experiment 1b: Intention prediction. Intention prediction is a substantially harder task, likely due both to the larger number of classes and the subjective nature of categorization (Table 1). The best model (FT-SU, 32.96%) only marginally exceeds the majority baseline (30.05%), and unlike speaker prediction, text is necessary (FT-S+R drops to 31.17%). No

			Speaker		Intention	
Model	**Context**	**Type**	**Acc.**	**Δ Bigram**	**Acc.**	**Δ Maj.**
Baselines						
Random	—	—	30.22	—	7.14	—
Bigram	—	—	50.75	(ref)	—	—
Majority	—	—	—	—	30.05	(ref)
Zero-shot / Vanilla						
Gemma-3-12B	SU	vanilla	43.56	−7.2	25.48	−4.6
GPT-5.4	*SU*	*zero-shot*	*61.97*	*+11.2*	*25.68*	*−4.4*
Opus 4.6	*SU*	*zero-shot*	*63.25*	*+12.5*	*26.68*	*−3.4*
Fine-tuned (Gemma-3-12B-IT + LoRA)						
Gemma-3-12B	SU	FT	64.53	+13.8	**32.96**	+2.9
Gemma-3-12B	SRU	FT	64.82	+14.1	32.53	+2.5
Gemma-3-12B	S+R	FT	**65.74**	+15.0	31.17	+1.1

Full results including Llama-3.1-8B, additional proprietary models, and few-shot baselines in Appendix G. Majority baseline for intention = "Giving Information" (30.05%).

Table 1: Next-speaker and next-intention prediction on the TIDES test set (5,094 samples, Teams 6&7). Context conditions: SU = Speaker + Utterance; SRU = + Role + Tuckman stage; S+R = structure only (no text). **Bold** = best. *Italic* = proprietary zero-shot.

Team	**Meeting prefix**	**Train** N	**Acc. (%)**	**Ref. (%)**
Team 10 (Korean, n=4)	1/3/4/6	276/795/1,073/1,811	50.87/55.37/62.62/65.82	67.09
Team 5 (English, n=5)	1/3/5/7	203/956/1,648/2,570	50.00/63.84/64.81/64.59	68.27

Table 2: Chronological single-team adaptation with Gemma-3-12B. Meeting prefixes, training sizes, and accuracies are listed in order. Remaining meetings form the test set at each prefix, so results are descriptive rather than fixed-test.

Source	**Training condition**	**Training data**	**Acc. (%)**
H&N	Zero-shot Llama 8B	—	34.88
H&N	Zero-shot Llama 70B	—	35.81
Ours	TIDES only (merged)	32K TIDES	35.33
Ours	TIDES only (unmerged)	62K TIDES	40.04
Ours	Balanced TIDES + AMI	124K total	**45.79**
Ours	Unbalanced TIDES + AMI	170K total	30.02
H&N	Fine-tuned Llama 8B (in-domain)	AMI + MultiLIGHT	47.85

Table 3: AMI external validation (12,515 test samples; context window 8). H&N denotes Hilgert & Niehues (2025). "Unmerged" preserves the original TIDES utterance boundaries; the balanced mix uses approximately 42% fewer examples than the H&N in-domain result (Appendix I).

zero-shot model—proprietary or open-source—surpasses the majority baseline, consistent with the task's class imbalance and annotation ambiguity.

Experiment 1c: Data efficiency. Extending the prediction analysis to single-team settings, observed accuracy rises most sharply within the first three to four meetings (Table 2). For Team 10, accuracy increases from 50.87% with one meeting to 62.62% with four; for Team 5, it increases from 50.00% with one meeting to 63.84% with three and then remains near 65%. Because the held-out set consists of the remaining meetings and therefore changes and shrinks as training meetings are added, these rows are not directly comparable as a fixed-test learning curve. We interpret them as descriptive evidence that a small amount of team-specific data can recover much of the performance observed later in the sequence, a pattern also seen with Llama-3.1-8B (Appendix I).

Experiment 1d: External validation. To test generalization beyond TIDES, we evaluate on AMI. A balanced TIDES+AMI mix reaches **45.79%**, within 2.06 pp of published SOTA

Model	Preferred (%)	Naturalness	Coherence	Speaker consistency
FT-Plain vs. Vanilla (407 judgments; tie = 69.5%)				
FT-Plain	9.3	3.07	2.78	3.05
Vanilla	**21.1*****	**3.99*****	**4.10*****	**3.93*****
FT-Reason vs. Vanilla (347 judgments; tie = 65.4%)				
FT-Reason	12.7	3.14	2.82	3.04
Vanilla	**21.9****	**3.98*****	**4.06*****	**3.85*****
FT-Plain vs. FT-Reason (388 judgments; tie = 53.4%)				
FT-Plain	21.4	3.21	3.08	3.13
FT-Reason	**25.3**	**3.42**	**3.40*****	**3.41*****

Preferred = percentage selecting that model; the remainder are ties. Ratings are mean 1–5 scores. Stars denote within-pair tests: $^{***}p < .001$, $^{**}p < .01$ after Bonferroni correction.

Table 4: Human evaluation (66 evaluators; 1,142 quality-controlled judgments). FT-Plain = direct fine-tuning; FT-Reason = social-cue fine-tuning; Vanilla = no fine-tuning.

with ∼42% less training data (Table 3). The TIDES-only transfer model (35.33%) is close to Hilgert & Niehues' zero-shot result (34.88%), suggesting that some learned turn-taking regularities transfer across corpora. However, the full unbalanced mix (170K, trained for 1 epoch) degrades to 30.02%, suggesting that data balance matters more than volume for cross-corpus transfer.

Experiment 2a: Generation: automatic metrics. FT-Reason achieves the highest 1-turn speaker accuracy (57.0% vs. FT-Plain 49.0%, Vanilla 14.0%), and both FT conditions produce far fewer invalid speakers than Vanilla (152–178 vs. 786 at 10 turns), suggesting that fine-tuning on TIDES improves structural coherence in generation. However, proprietary models outperform all fine-tuned models: Opus 4.6 reaches 62.0% at 1-turn with zero invalid speakers, and degrades more slowly (39.0% at 5-turn vs. FT-Reason's 29.2%). Proprietary models also show higher semantic similarity to ground truth (0.36–0.44 vs. 0.28–0.36; Table 20). Unlike prediction, where fine-tuning closes the gap with proprietary models, generation quality appears to benefit more from model scale.

Experiment 2b: Generation: human evaluation. Despite the results above, human evaluators reveal a contrasting pattern (Table 4): although most judgments involving Vanilla were ties (65.4–69.5%), directional preferences **significantly favor Vanilla over fine-tuned outputs**. FT-Plain wins only 9.3% of judgments vs. Vanilla's 21.1% ($p < .001$); FT-Reason wins 12.7% vs. 21.9% ($p = .004$). Vanilla scores 0.8–1.3 Likert points higher on naturalness, coherence, and speaker consistency (d = 0.46–0.72; all p-values remain significant after Bonferroni correction). FT outputs, while contextually grounded, exhibit surface artifacts (truncation, missing punctuation), whereas Vanilla generates polished but often off-topic continuations; however, the gap persisted even after we post-processed all FT outputs to correct capitalization, punctuation, and truncated endings (Appendix L). This suggests that the difference extends beyond formatting: **improvements in structural modeling may not directly transfer to natural conversation generation**, as accurately predicting *who* speaks next does not necessarily produce utterances that humans perceive as realistic. We note, however, that our evaluators were not members of the recorded teams, and fine-tuning also adapts models to team-specific register and project-specific language; the current evaluation therefore cannot fully separate a structure–content mismatch from a domain-familiarity effect, and we treat this result as an open question.

6 Limitations and Future Work

TIDES captures authentic social dynamics in Korean- and English-speaking team collaboration over the course of a semester, and therefore offers substantial potential for research directions that remain underexplored.

Alternative Utterance-level Annotation. Currently, TIDES adopts a modified version of `act4teams-SHORT` to capture the collaborative intent of each utterance. Although the categories in `act4teams-SHORT` are well suited to representing utterances in collaborative settings, the relatively large number of categories and the inherently subjective nature of the annotation task contribute to only moderate inter-rater agreement. In future work, we aim to extend the dataset with annotations based on more general dialogue-act taxonomies, such as AMI-DA (Hain et al., 2007) and MRDA (Shriberg et al., 2004). Such annotations would improve the comparability of TIDES with prior work and facilitate its use in a broader range of downstream tasks.

Longitudinal Analysis. In experiment 1c (§ 5.2), we examined how much team-specific conversational data was needed for a model to adequately capture a team's interaction dynamics and predict the next speaker. However, TIDES also includes longitudinal measures of collaboration satisfaction, emergent roles, and team development stages collected over the course of a semester, providing opportunities to track and analyze how a team's social dynamics evolve over time. Accordingly, in future work, we aim to use TIDES to evaluate models' longitudinal capabilities, including predicting changes in members' emergent roles, transitions between team development stages, and changes in collaboration satisfaction from conversational histories.

Bilingual Data. TIDES includes both Korean and English data and provides English translations for teams that primarily communicated in Korean. We also examined how the language composition of the training data affected next-speaker prediction, as described in Appendix J. In future work, we aim to extend this analysis by examining whether the evolution of within-team social dynamics differs across teams with different primary languages or cultural contexts.

7 Conclusion

We introduced TIDES, a longitudinal bilingual dataset tracking 12 university project teams over a full semester, with socio-structural annotations including emergent roles, interaction types, and team development stages. Through a series of experiments, we demonstrated that fine-tuning on TIDES substantially improves next-speaker prediction—outperforming proprietary large-scale models—and that only a few hours of well-structured meeting data is sufficient for models to capture a team's interaction dynamics, with learned patterns transferring competitively to the AMI Meeting Corpus.

These findings indicate a potential mismatch between predicting conversational structure and producing content that users perceive as natural.

Our experiments primarily evaluated local prediction within short context windows. Modeling how team dynamics accumulate and shift across a team's full lifespan—for example, predicting role transitions, detecting phase shifts in real time, or adapting to evolving team norms—remains an important open direction. The longitudinal structure and social annotations in TIDES open the door to such research, offering a resource where cross-meeting dynamics, role trajectories, and team development can be studied as they naturally unfold.

Acknowledgments

This work was supported by the KAIST C2 (Creative & Challenging) and UP Projects. This work was also supported by Institute of Information & communications Technology Planning & Evaluation (IITP) under the Leading Generative AI Human Resources Development (IITP-2026-RS-2026-25546560) grant funded by the Korea government (MSIT). We sincerely thank all participants for contributing their meeting data throughout the semester, our Prolific annotators and evaluators, and CSTL and KIXLAB members for insightful discussions and invaluable feedback.

Ethics Statement

This study was conducted under approval from our institute's Institutional Review Board (IRB). All participants provided informed consent prior to enrollment and were fully briefed on data collection procedures, including audio recording, during an orientation session. Participation was voluntary, and participants were free to withdraw at any time.

To protect participant privacy, only transcripts are made publicly available. Prior to any analysis or release, transcripts were processed through a multi-stage anonymization pipeline. Personally identifiable information—names, locations, and organizational references—was removed using Microsoft Presidio for English-speaking teams and a locally run Qwen3-30B-A3B model for Korean-speaking teams. All pseudonyms are gender-neutral. We compensated each team 500,000 KRW (approximately 325 USD) for their participation. Also, we compensated each participant 35,000 KRW per hour of reviewed audio if they validated the quality of transcripts.

For annotation on Prolific, annotators were compensated at 8 GBP (approximately 11 USD) per task (100 utterances), and for human evaluators, each evaluator was compensated at 4.5 GBP (approximately 6 USD) per task. Both rates exceeded the platform's recommended rate. There was no discrimination in the recruitment of annotators based on any demographic characteristics.

References

Kenneth D. Benne and Paul Sheats. Functional roles of group members. *Journal of Social Issues*, 4(2):41–49, 1948.

Hervé Bredin. pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In *Interspeech*, pp. 1983–1987, 2023.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Maël Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska Masson, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre D. Wellner. The ami meeting corpus: A pre-announcement. In *Machine Learning for Multimodal Interaction*, 2005. URL https://api.semanticscholar.org/CorpusID:6118869.

Galo Castillo-López, Gael de Chalendar, and Nasredine Semmar. A survey of recent advances on turn-taking modeling in spoken dialogue systems. In Maria Ines Torres, Yuki Matsuda, Zoraida Callejas, Arantza del Pozo, and Luis Fernando D'Haro (eds.), *Proceedings of the 15th International Workshop on Spoken Dialogue Systems Technology*, pp. 254–271, Bilbao, Spain, May 2025. Association for Computational Linguistics. ISBN 979-8-89176-248-0. URL https://aclanthology.org/2025.iwsds-1.27/.

Yi-Ting Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. MPDD: A multi-party dialogue dataset for analysis of emotions and interpersonal relationships. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 610–614, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.76/.

Elijah L. Claggett, Robert E. Kraut, and Hirokazu Shirado. Relational AI: Facilitating intergroup cooperation with socially aware conversational support. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–22. Association for Computing Machinery, 2025. doi: 10.1145/3706598.3713757.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification. In *Interspeech*, pp. 3830–3834, 2020.

Tripp Driskell, James E. Driskell, C. Shawn Burke, and Eduardo Salas. Team roles: A review and integration. *Small Group Research*, 48:482 – 511, 2017. URL https://api.semanticscholar.org/CorpusID:149296953.

Erik Ekstedt and Gabriel Skantze. TurnGPT: a transformer-based language model for predicting turn-taking in spoken dialog. In Trevor Cohn, Yulan He, and Yang Liu (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 2981–2990, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.268. URL https://aclanthology.org/2020.findings-emnlp.268/.

Shitao Fang, Xingyu Liu, Takeo Igarashi, and Koji Yatani. Unraveling multiparty conversations: From human interaction mechanisms to conversational agent challenges and persona design. *Int. J. Hum. Comput. Stud.*, 208:103719, 2025. URL https://api.semanticscholar.org/CorpusID:284082422.

Jia-Chen Gu, Chongyang Tao, and Zhenhua Ling. Who says what to whom: A survey of multi-party conversations. In *International Joint Conference on Artificial Intelligence*, 2022. URL https://api.semanticscholar.org/CorpusID:250637571.

Thomas Hain, Lukas Burget, John Dines, Giulia Garau, Martin Karafiat, David van Leeuwen, Mike Lincoln, and Vincent Wan. The 2007 ami (da) system for meeting transcription. In *International Evaluation Workshop on Rich Transcription*, pp. 414–428. Springer, 2007.

Lukas Hilgert and Jan Niehues. Next speaker prediction for multi-speaker dialogue with large language models. In Mourad Abbas, Tariq Yousef, and Lukas Galke (eds.), *Proceedings of the 8th International Conference on Natural Language and Speech Processing (ICNLSP-2025)*, pp. 60–71, Southern Denmark University, Odense, Denmark, August 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.icnlsp-1.7/.

Stephanie Houde, Kristina Brimijoin, Michael Muller, Steven I. Ross, Dario Andres Silva Moran, Gabriel Enrique Gonzalez, Siya Kunde, Morgan A. Foreman, and Justin D. Weisz. Controlling AI agent participation in group conversations: A human-centered approach. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, pp. 390–408. Association for Computing Machinery, 2025. doi: 10.1145/3708359.3712089.

Yebowen Hu, Timothy Ganter, Hanieh Deilamsalehy, Franck Dernoncourt, Hassan Foroosh, and Fei Liu. MeetingBank: A benchmark dataset for meeting summarization. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16409–16423, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.906. URL https://aclanthology.org/2023.acl-long.906/.

Jihyoung Jang, Minseong Boo, and Hyounghun Kim. Conversation chronicles: Towards diverse temporal and relational dynamics in multi-session conversations. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 13584–13606, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.838. URL https://aclanthology.org/2023.emnlp-main.838/.

Qi Jia, Hongru Huang, and Kenny Q. Zhu. Ddrel: A new dataset for interpersonal relation classification in dyadic dialogues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 13125–13133, 2021.

David Jurgens, Agrima Seth, Jackson Sargent, Athena Aghighi, and Michael Geraci. Your spouse needs professional help: Determining the contextual appropriateness of messages through modeling social relationships. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10994–11013, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.616. URL https://aclanthology.org/2023.acl-long.616/.

Georgi Karadzhov, Tom Stafford, and Andreas Vlachos. Delidata: A dataset for deliberation in multi-party problem solving. *Proceedings of the ACM on Human-Computer Interaction*, 7:1 – 25, 2023. URL https://api.semanticscholar.org/CorpusID:236975941.

Simone Kauffeld, Nale Lehmann-Willenbrock, and Annika L. Meinecke. The advanced interaction analysis for teams (act4teams) coding scheme. In Elisabeth Brauner, Margarete Boos, and Michaela Kolbe (eds.), *The Cambridge Handbook of Group Interaction Analysis*, pp. 422–431. Cambridge University Press, 2018. doi: 10.1017/9781316286302.022.

Soomin Kim, Jinsu Eun, Changhoon Oh, Bongwon Suh, and Joonhwan Lee. Bot in the bunch: Facilitating group chat discussion by improving efficiency and participation with a chatbot. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–13. Association for Computing Machinery, 2020. doi: 10.1145/3313831.3376785.

Frederic Kirstein, Muneeb Khan, Jan Philip Wahle, Terry Ruas, and Bela Gipp. You need to MIMIC to get FAME: Solving meeting transcript scarcity with multi-agent conversations. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 11482–11525, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.599. URL https://aclanthology.org/2025.findings-acl.599/.

Jil Klünder, Nils Prenner, Ann-Kathrin Windmann, Marek Stess, Michael Nolting, Fabian Kortum, Lisa Handke, Kurt Schneider, and Simone Kauffeld. Do you just discuss or do you solve? meeting analysis in a software project at early stages. In *Proceedings of the IEEE/ACM 42nd International Conference on Software Engineering Workshops*, pp. 557–562, 2020.

Heng-Yu Ku, Hungwei Tseng, and Chatchada Akarasriworn. Collaboration factors, teamwork satisfaction, and student attitudes toward online collaborative learning. *Comput. Hum. Behav.*, 29:922–929, 2013. URL https://api.semanticscholar.org/CorpusID:29039396.

Harold W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1–2):83–97, 1955.

Xingyu Bruce Liu, Shitao Fang, Weiyan Shi, Chien-Sheng Wu, Takeo Igarashi, and Xiang 'Anthony' Chen. Proactive conversational agents with inner thoughts. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–19. Association for Computing Machinery, 2025. doi: 10.1145/3706598.3713760.

Brian MacWhinney. *The CHILDES Project: Tools for analyzing talk*. Lawrence Erlbaum Associates, 3rd edition, 2000.

Microsoft. Microsoft Presidio: Data protection and de-identification SDK, 2023. URL https://github.com/microsoft/presidio.

Anna Nedoluzhko, Muskaan Singh, Marie Hledíková, Tirthankar Ghosal, and Ondřej Bojar. ELITR minuting corpus: A novel dataset for automatic minuting from multi-party meetings in English and Czech. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 3174–3182, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.340/.

Alexis Plaquet and Hervé Bredin. Powerset multi-class cross entropy loss for neural speaker diarization. In *Interspeech*, pp. 3222–3226, 2023.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings*

of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 527–536, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1050. URL https://aclanthology.org/P19-1050/.

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pp. 28492–28518. PMLR, 2023.

Farzana Rashid and Eduardo Blanco. Characterizing interactions and relationships between people. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4395–4404, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1470. URL https://aclanthology.org/D18-1470/.

Elizabeth Shriberg, Raj Dhillon, Sonali Bhagat, Jeremy Ang, and Hannah Carvey. The ICSI meeting recorder dialog act (MRDA) corpus. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*, pp. 97–100, Cambridge, Massachusetts, USA, April 30 - May 1 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-2319/.

Chao-Hong Tan, Jia-Chen Gu, and Zhen-Hua Ling. Is ChatGPT a good multi-party conversation solver? In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 4905–4915, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.326. URL https://aclanthology.org/2023.findings-emnlp.326/.

Anna Tigunova, Paramita Mirza, Andrew Yates, and Gerhard Weikum. PRIDE: Predicting Relationships in Conversations. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 4636–4650, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.380. URL https://aclanthology.org/2021.emnlp-main.380/.

Bruce W. Tuckman. Developmental sequence in small groups. *Psychological bulletin*, 63: 384–399, 1965. URL https://api.semanticscholar.org/CorpusID:10356275.

Maarten Van Segbroeck, Ahmed Zaid, Ksenia Kutsenko, Cirenia Huerta, Tinh Nguyen, Xuewen Luo, Björn Hoffmeister, Jan Trmal, Maurizio Omologo, and Roland Maas. DiPCo — Dinner Party Corpus. In *Proceedings of Interspeech 2020*, pp. 434–436, 2020. doi: 10.21437/Interspeech.2020-2800.

Hongji Wang, Chengdong Liang, Shuai Wang, Zhengyang Chen, Binbin Zhang, Xu Xiang, Yanlei Deng, and Yanmin Qian. Wespeaker: A research and production oriented speaker embedding learning toolkit. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10096626.

Jimmy Wei, Kurt Shuster, Arthur Szlam, Jason Weston, Jack Urbanek, and Mojtaba Komeili. Multi-party chat: Conversational agents in group settings with humans and models. *ArXiv*, abs/2304.13835, 2023. URL https://api.semanticscholar.org/CorpusID:258352487.

Ka Wong, Praveen Paritosh, and Lora Aroyo. Cross-replication reliability - an empirical approach to interpreting inter-rater reliability. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 7053–7065, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.548. URL https://aclanthology.org/2021.acl-long.548/.

Haolan Zhan, Zhuang Li, Yufei Wang, Linhao Luo, Tao Feng, Xiaoxi Kang, Yuncheng Hua, Lizhen Qu, Lay-Ki Soon, Suraj Sharma, Ingrid Zukerman, Zhaleh Semnani-Azad, and Gholamreza Haffari. SocialDial: A benchmark for socially-aware dialogue systems. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2712–2722. Association for Computing Machinery, 2023. doi: 10.1145/3539618.3591877.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. QMSum: A new benchmark for query-based multi-domain meeting summarization. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5905–5921, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.472. URL https://aclanthology.org/2021.naacl-main.472/.

Jinfeng Zhou, Yuxuan Chen, Yihan Shi, Xuanming Zhang, Leqi Lei, Yi Feng, Zexuan Xiong, Miao Yan, Xunzhi Wang, Yaru Cao, Jianing Yin, Shuai Wang, Quanyu Dai, Zhenhua Dong, Hongning Wang, and Minlie Huang. SocialEval: Evaluating social intelligence of large language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 30958–31012, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1496. URL https://aclanthology.org/2025.acl-long.1496/.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. SOTOPIA: Interactive evaluation for social intelligence in language agents. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=mM7VurbA4r.

Chenguang Zhu, Yang Liu, Jie Mei, and Michael Zeng. MediaSum: A large-scale media interview dataset for dialogue summarization. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5927–5934, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.474. URL https://aclanthology.org/2021.naacl-main.474/.

A Details for Utterance-level Interaction Type Annotation

A.1 Construction of Human-labeled Gold Data via Prolific

For utterance-level interaction type annotation, we recruited 260 annotators through the crowdsourcing platform Prolific. Annotators accessed our custom-built annotation interface (Fig. 2), where they first provided their Prolific ID and then proceeded to the task. Before starting the main annotation task, annotators were instructed to carefully review the definitions of the 15 categories in our modified `act4teams-SHORT` scheme. To ensure sufficient understanding of the coding scheme, annotators were required to complete a tutorial quiz and answer all questions correctly before proceeding. Each annotator was assigned a total of 100 utterances, organized into 10 windows of 10 target utterances each. To help annotators interpret each utterance in context, we additionally provided the 10 preceding utterances and 10 following utterances for each window. After completing all 10 windows, annotators received approximately 11 USD in compensation, subject to a quality check by the research team.

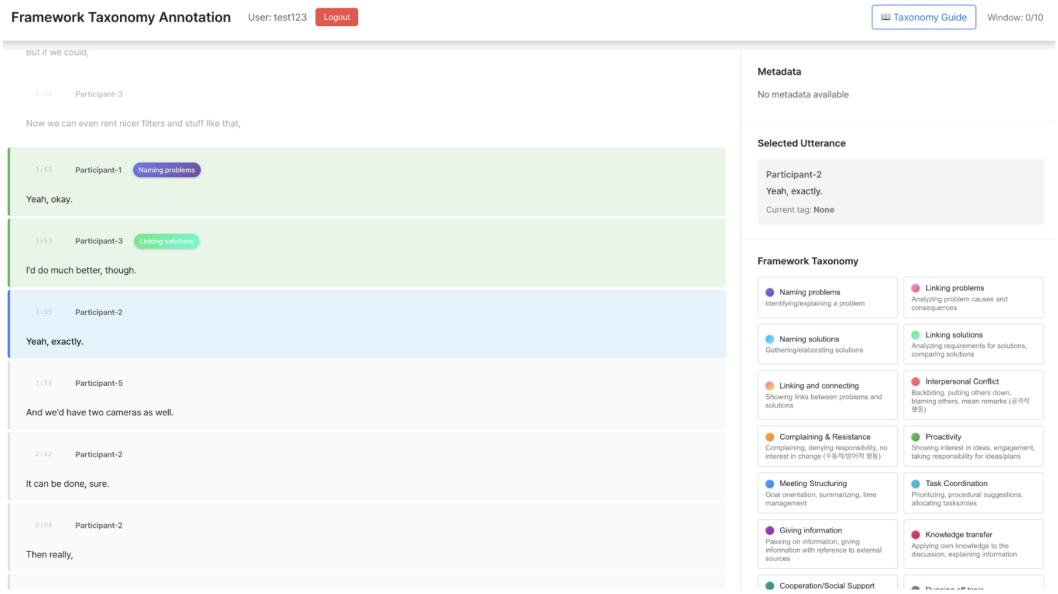

Figure 2: Custom annotation interface used for utterance-level interaction type annotation on Prolific. Annotators reviewed each target utterance with surrounding conversational context and selected one of the 15 modified `act4teams-SHORT` categories.

Interaction Type	Count	Ratio (%)
Giving Information	1,881	32.97
Active Listening	956	16.76
Linking Solutions	395	6.92
Other / Neutral	372	6.52
Structuring	362	6.35
Naming Solutions	339	5.94
Social / Humor	296	5.19
Proactivity	240	4.21
Naming Problems	217	3.80
Knowledge Transfer	171	3.00
Cooperation	162	2.84
Linking Problems	135	2.37
Social Negative	66	1.16
Task/Process Negative	62	1.09
Linking & Connecting	51	0.89

Table 5: Distribution of utterance-level interaction types in the human-labeled gold data.

A.2 Interaction Types in Human-labeled Gold Data

Table 5 summarizes the distribution of utterance-level interaction types in the human-labeled gold data. The distribution is imbalanced, with *Giving Information* and *Active Listening* accounting for a large proportion of the annotations, while several categories such as *Social Negative*, *Task/Process Negative*, and *Linking & Connecting* appear relatively infrequently.

A.3 Utterance-Level Interaction Type Annotation

We compared several base and fine-tuned models on the utterance-level interaction classification task using a leave-one-team-out validation setup (Table 6). Although fine-tuned `Qwen3-14B` achieved the highest accuracy and Cohen's κ, we selected the fine-tuned `Gemma-3-12B` for large-scale annotation because it achieved the best macro-F1 score (0.540),

which we considered the most appropriate primary metric for our imbalanced multi-class setting. The Fleiss' κ among human annotators was 0.400, indicating moderate agreement on this challenging 15-class annotation task. The released corpus marks each utterance as either human-labeled gold or model-produced silver through the `annotation_source` field. The final release contains 5,705 gold utterances and 70,266 silver utterances. Because the 15-way distribution is highly imbalanced, the silver layer should not be interpreted as uniformly reliable across categories; users can restrict analyses to the gold subset when higher-confidence supervision is required.

Model	Acc.	κ	F1
Llama3.1-8B (Base)	0.382	0.286	0.231
Llama3.1-8B (FT)	0.687	<u>0.606</u>	<u>0.528</u>
Qwen3-14B (Base)	0.488	0.397	0.313
Qwen3-14B (FT)	**0.691**	**0.612**	0.527
Gemma-3-12B (Base)	0.570	0.477	0.367
Gemma-3-12B (FT)	<u>0.687</u>	0.605	**0.540**

Table 6: Mean validation performance for utterance-level interaction type classification (leave-one-team-out). Best is bold; second-best is underlined.

B Annotated Dataset Statistics

(a) Utterance-level interaction type distribution

Utterance Type	Count	Ratio (%)
Giving Information	26,810	35.29
Active Listening	9,946	13.09
Linking Solutions	9,260	12.19
Naming Solutions	4,917	6.47
Other / Neutral	4,703	6.19
Structuring	4,643	6.11
Proactivity	3,912	5.15
Naming Problems	3,157	4.16
Social / Humor	2,510	3.30
Linking Problems	2,233	2.94
Cooperation	1,812	2.39
Knowledge Transfer	878	1.16
Task/Process Negative	642	0.85
Social Negative	465	0.61
Linking & Connecting	83	0.11
Total	75,971	100.00

(b) Emergent role distribution

Role	Count	Ratio (%)
Problem Solver	62	17.61
Coordinator	59	16.76
Critic	56	15.91
Negative	38	10.80
Task Motivator	32	9.09
Team Leader	32	9.09
Evaluator	16	4.55
Attention Seeker	13	3.69
Task Completer	11	3.13
Teamwork Support	11	3.13
Follower	10	2.84
Power Seeker	7	1.99
Social	5	1.42
Total	352	100.00

Table 7: Distribution of utterance-level interaction types and emergent roles in the dataset.

The amount of conversation varied substantially across teams, ranging from 2,788 to 14,058 utterances per team, with an average of approximately 6,331 utterances per team.

C Transcript Examples

Tables 9 and 10 show representative excerpts illustrating the full annotation layers of TIDES: pseudonymized speaker identity, emergent role, interaction type, and team development stage (in caption). Table 9 additionally shows Korean original utterances alongside their English translations.

Team	# Utterances	# Meetings
Team_3	14,058	8
Team_9	11,713	7
Team_2	11,630	13
Team_7	6,594	5
Team_5	6,168	10
Team_10	4,482	7
Team_12	4,000	12
Team_1	3,861	5
Team_11	3,852	6
Team_6	3,448	6
Team_8	3,377	5
Team_4	2,788	4
Total	75,971	88

Table 8: Team-level statistics of the released transcript corpus. Meetings are counted by unique recording date; sessions split into multiple parts count as one meeting.

Speaker	**Role**	**Korean (PII-removed)**	**English (translated)**	**Type**
Alex (A)	Task Compl.	아 그러면 이제 두 분 중에 남는 분을 슬레이터로 맡기면 되는 건가요?	Oh then, should we assign the remaining person to the slider role?	Naming Sol.
Blake (E)	Team Leader	네, 그래야 할 것 같아요.	Yes, that seems like the way to go.	Active List.
Alex (A)	Task Compl.	저야 당연히 그러면 좋습니다.	I'd be happy with that, of course.	Cooperation
Blake (E)	Team Leader	아니면 Drew님은 어떤 거 하는 게 편하세요?	Or, Drew, which one would be more convenient for you?	Structuring
Drew (C)	Coordinator	어 저는 슬레이터가 더 좋아요.	I prefer the slider.	Giving Info.
Blake (E)	Team Leader	그러면 맡길까요?	Then shall we assign it to you?	Proactivity

Table 9: Transcript excerpt from Team 1 (Korean → English, meeting 2025-11-06). Team development stage: *Performing*. The team is assigning filming roles for a video production project.

Speaker	**Role**	**Utterance**	**Type**
Riley (E)	Critic	I guess there is some random Academic Institute branch of the node pixel.	Giving Info.
Jordan (C)	Critic	Maybe you should make a copy of our stuff.	Naming Sol.
Morgan (A)	Negative	The LED is so bright that it can be fixed, but the hole is made here and there.	Linking Sol.
Morgan (A)	Negative	But to put several LEDs in a row,	Linking Sol.
Dakota (D)	Prob. Solver	I can't make the boards bend.	Naming Prob.
Dakota (D)	Prob. Solver	I think it's better to use glue gun or bond to connect them with LED lines.	Naming Sol.

Table 10: Transcript excerpt from Team 5 (English, meeting 2025-11-18). Team development stage: *Norming*. The team is building a physical LED prototype for an HCI course project.

D Meeting Metadata and Survey Examples

Table 11 shows the meeting metadata for Team 5, illustrating how team development stages progress over the semester. Table 12 shows post-meeting satisfaction scores and peer-evaluated emergent roles for the same team's first meeting.

Date	Stage	Topic (summary)
2025-10-28	Forming	Brainstormed 5 ideas, narrowed to 3
2025-11-03	Forming	Finalized topic, planned implementation
2025-11-10	Storming	Decided on circuits over Arduino
2025-11-18	Norming	Realized neopixel implementation difficulty
2025-12-02	Performing	Redistributed work after rack assembly
2025-12-06	Performing	Bug fixing and integration
2025-12-10	Performing	Button working, display planning
2025-12-12	Performing	Product finalized and functional
2025-12-13	Adjourning	Filmed demo video
2025-12-15	Adjourning	Presentation role division

Table 11: Meeting metadata for Team 5 (English, 5 members). Stages were assigned retrospectively by team consensus via Tuckman self-assessment.

Satisfaction Item	P1	P2	P3	P4	P5	Avg
Clear collaborative patterns	3	3	3	4	5	3.6
Members know their roles	3	5	5	5	5	4.6
Clear goals and working norms	4	5	5	5	5	4.8
Timely responses	4	5	4	5	5	4.6
Frequent communication	5	5	4	5	5	4.8

Member	G1 (Task)	G2 (Social)	G3 (Domin.)	Assigned Role
Morgan (A)	3.58	4.58	4.58	Teamwork Support
Casey (B)	4.33	4.89	4.89	Coordinator
Jordan (C)	4.44	4.11	4.22	Power Seeker
Dakota (D)	3.78	4.78	4.56	Problem Solver
Riley (E)	3.67	4.33	4.11	Critic

Table 12: Post-meeting satisfaction survey and emergent roles for Team 5 (meeting 2025-10-28, stage: *Forming*). Satisfaction items are rated 1–5; roles are derived from peer-evaluation scores across three TRIAD dimensions.

E Full Post-Survey Question Set

To capture the evolution of group dynamics throughout the project, we administered a post-survey immediately after each meeting, as well as a final survey at the end of the project. The post-survey consisted of three parts. First, participants provided a brief reflection summarizing important decisions, turning points, or notable events from the meeting. Second, they rated their satisfaction with team collaboration using questionnaire items adapted from prior work on collaboration and team processes Ku et al. (2013). Third, to capture emergent roles within the team, participants evaluated each of their teammates using nine peer-assessment items. These items reflect the three dimensions of the TRIAD model—Dominance, Sociability, and Task Orientation—with three questions for each dimension (Driskell et al., 2017). At the end of the semester, we additionally administered a final survey to assess participants' overall satisfaction and perceived quality of the project outcome. We also asked each team to collaboratively reflect on their meeting history and classify each meeting according to Tuckman's Team Development Model—Forming, Storming, Norming, Performing, and Adjourning—along with a rationale for each classification (Tuckman, 1965).

Below, we provide the full set of survey questions used in our study.

E.1 Daily Meeting Reflection

- Please share the important decision or turning point during the conversation of today's meeting.

E.2 Team Collaboration Satisfaction

Note: All items are measured on a 5-point Likert scale (1 = Strongly Disagree to 5 = Strongly Agree).

1. My team develops clear collaborative patterns to increase team learning efficiency.
2. My team members clearly know their roles during the collaboration.
3. My team has an efficient way to track the edition of documents.
4. My team sets clear goals and establishes working norms.
5. My team members reply to all responses in a timely manner.
6. My team members communicate with each other frequently.
7. I trust each team member can complete his/her work on time.
8. My team is receiving feedback from each other.
9. Communicating with team members regularly helps me to understand the team project better.
10. My team members encourage open communication with each other.
11. My team members communicate in a courteous tone.

E.3 Emergent Role Survey (Peer Evaluation)

Note: These questions evaluate the behavior of each team member on a 5-point frequency scale (1 = Never to 5 = Always).

1. Did this team member actively lead group discussions?
2. Did this team member clearly present the team's activities or direction?
3. Did this team member assert their opinions in decision-making and influence the team?
4. Did this team member create a positive and comfortable atmosphere in the team?
5. Did this team member respect and support other team members' feelings and opinions?
6. Did this team member help mediate or facilitate smooth communication during conflicts?
7. Did this team member help the team stay focused on its shared goals and avoid distractions?
8. Did this team member fulfill their assigned role responsibly and on time?
9. Did this team member show a diligent and meticulous attitude to improve task quality?

E.4 Final Survey (Individual Assessment)

1. How satisfied are you overall with the project results?
2. Do you think the project has achieved its intended outcome?
3. How do you rate the quality of the project results?

E.5 Final Survey (Team Reflection)

- **Group Discussion Task:** Looking back at all previous meetings, discuss with your team members and assign each meeting to the corresponding stage of Tuckman's Team Development Model (Forming, Storming, Norming, Performing, Adjourning). Please provide the rationale for your classifications.

F Post-processing Pipeline Details

This appendix provides full technical details for the six-stage post-processing pipeline described in Section 3.2. The pipeline converts raw audio into privacy-preserved, speaker-identified transcripts through the following stages: (1) ASR, (2) speaker diarization, (3) PII removal, (4) translation, (5) cross-session speaker unification, and (6) utterance boundary refinement.

F.1 Stage 1: ASR

We transcribed all audio files using Whisper Large-V3 (Radford et al., 2023) with float16 precision on a single RTX A6000 GPU. Table 13 summarizes the configuration. Audio files were first preprocessed with FFmpeg to remove silence regions using the filter `silenceremove=1:0:-50dB` and resampled to 16 kHz mono. We loaded Whisper Large-V3 via the HuggingFace `transformers` pipeline with hallucination reduction enabled through a repetition penalty of 1.1 and `no_repeat_ngram_size` of 3.

Parameter	Value
Model	`openai/whisper-large-v3`
Device / Precision	CUDA / float16
Batch size	24
Temperature	0.2
Entropy threshold	2.8
Sample rate	16,000 Hz
Repetition penalty	1.1
No-repeat n-gram size	3
Return timestamps	Segment-level

Table 13: ASR configuration for Whisper Large-V3.

F.2 Stage 2: Speaker Diarization

Raw Whisper output does not distinguish speakers. We developed a *Smart Pipeline* that combines pyannote 3.1 (Bredin, 2023; Plaquet & Bredin, 2023) for voice activity detection with ECAPA-TDNN (Desplanques et al., 2020) speaker embeddings (192-dim) and Median Absolute Deviation outlier filtering, then re-clusters to the known team size N, ensuring correct speaker counts by construction.

Smart Pipeline. The pipeline proceeds in five steps:

1. **VAD:** pyannote 3.1 (`pyannote/speaker-diarization-3.1`) segments the audio into speech and non-speech regions.
2. **Segment extraction:** Speech regions are extracted as individual audio segments.
3. **Embedding extraction:** Each segment is encoded with ECAPA-TDNN (`speechbrain/spkrec-ecapa-voxceleb`), producing a 192-dimensional speaker embedding. Segments shorter than 0.5 s are zero-padded.
4. **MAD filtering:** Speaker-level embeddings are computed by averaging segment embeddings per speaker, with Median Absolute Deviation filtering to remove outlier segments.
5. **Re-clustering:** Embeddings are re-clustered to the known team size N using multiple methods (KMeans, Agglomerative with ward/complete/average linkage, Spectral clustering), and the best result is selected.

F.3 Stage 3: PII Removal

English and Korean require fundamentally different PII removal approaches. For English, we used Microsoft Presidio (Microsoft, 2023), a rule-based detection framework with entity

linking and consistent pseudonymization. For Korean, we employed Qwen3-30B-A3B-Instruct-2507 (Qwen Team, 2025) running locally as a context-aware detector; an earlier regex-based approach had yielded a 91% false-positive rate, which the LLM-based pipeline substantially reduced, though 375 manual corrections were still needed. All replacements use gender-neutral pseudonyms (e.g., Alex, Jordan, Taylor).

English Pipeline. For the five English/mixed teams (329,537 segments), we used Microsoft Presidio with three custom components:

- **EntityLinker:** Maps name variants (e.g., "Mary," "MJ," "Jane") to canonical forms.
- **ConsistentAnonymizer:** Ensures the same entity always maps to the same pseudonym across the entire dataset.
- **PresidioPIIMaskerV2:** Wraps the Presidio `AnalyzerEngine` with custom Korean phone number and ID recognizers.

Detected entity types include: PERSON, LOCATION, ORGANIZATION, EMAIL, PHONE_NUMBER, URL, and DATE_TIME.

Korean Pipeline. For the seven Korean teams (48,697 segments), we employed Qwen3-30B-A3B-Instruct-2507 running locally (`bfloat16`, `device_map=auto`) in a three-stage process:

1. **Detection:** The LLM identifies PERSON, LOCATION, and ORGANIZATION entities directly from Korean text, generating an `entity_mapping.json`.
2. **Replacement:** Detected entities are replaced using regex patterns with Korean-aware word boundary matching (lookbehind for Hangul/ASCII, lookahead for ASCII only, since Korean particles attach directly to names).
3. **Translation:** PII-replaced Korean text is translated to English by the same Qwen3 model, running as a subprocess to ensure GPU memory cleanup between files.

Pseudonym Pools.

- **Person:** 32 gender-neutral English names (Alex, Jordan, Taylor, Morgan, Casey, Riley, Quinn, Avery, Parker, Drew, Reese, Jamie, Sage, River, Phoenix, Blake, Charlie, Emerson, Hayden, Skyler, Dakota, Finley, Rowan, Ellis, Cameron, Peyton, Logan, Spencer, Bailey, Kendall, Harper, Addison).
- **Location:** Coded patterns (Building-A, Room-101, Campus-North, City-A, Lab-Alpha, etc.).
- **Organization:** Coded patterns (Tech-Lab, Company-A, Institute-Alpha, etc.).

Quality Assurance. After automated PII removal, we ran a PII leak checker across all anonymized files (TXT, CSV, XLSX formats) to detect residual personal information. The Korean pipeline required many manual corrections, primarily due to common Korean names that are also everyday words (e.g., some names are identical to common pronouns).

F.4 Stage 4: Translation

To unify the dataset into a single language for downstream modeling, we translated all Korean transcripts to English using Qwen3-30B-A3B-Instruct-2507, running entirely on local machines.

All Korean transcripts were translated to English using Qwen3-30B-A3B-Instruct-2507 (`bfloat16`, `device_map=auto`). Translation was performed per-file in isolated subprocesses to ensure complete GPU memory release between files. Segments without Korean characters were automatically skipped via a `has_korean()` check. For segments exceeding 200 characters, individual (rather than batch) translation was used to maintain quality. Pseudonym preservation was validated post-translation using the `validate_pseudonyms()` function, which verifies that all pseudonyms present in the source appear in the translated output.

F.5 Stage 5: Cross-session Speaker Unification

Diarization assigns arbitrary speaker labels that reset across sessions, so the same individual may receive different labels in different recordings. We unified identities using WeSpeaker (Wang et al., 2023) 256-dimensional embeddings with the Hungarian algorithm (Kuhn, 1955) and a complementary *mention matrix*—leveraging the observation that speakers rarely say their own name—producing 422 speaker-file mappings with a within-speaker cosine similarity of 0.89 versus 0.34 between speakers.

Step 1: Embedding Extraction & Hungarian Matching. We used the WeSpeaker model (`pyannote/wespeaker-voxceleb-resnet34-LM`) to extract 256-dimensional speaker embeddings. For each speaker in each session, up to 5 segments were randomly sampled (minimum duration 1.5 s, capped at 30 s), and their embeddings were averaged to obtain a representative centroid. Files were processed in order of decreasing speaker count: the first file initializes the centroid pool, and subsequent files are matched against existing centroids using the Hungarian algorithm with cost matrix $C = 1 - \text{cosine\_similarity}(E, M)$, where E is the embedding matrix and M is the centroid matrix. Unmatched speakers create new centroids (e.g., `PERSON_F`).

Step 2: Label Application. The resulting CSV mapping (422 entries) was applied to update the `speaker` field in all released JSON files, converting arbitrary per-session labels (`PERSON_0`, `PERSON_1`, ...) to consistent cross-session labels (`PERSON_A`, `PERSON_B`, ...). Team 1 was mapped manually and verified independently.

Step 3: Mention Matrix Inference. As a complementary identity signal, we counted how often each speaker mentions each pseudonym using word-boundary regex matching, leveraging the observation that speakers rarely say their own name. Each assignment received a confidence tag:

- **VERIFIED:** Manual verification (Team 1 only).
- **HIGH:** 0 self-mentions and $\geq$3 total mentions by others.
- **MED:** $\leq$1 self-mention and $\geq$2 total mentions.
- **LOW:** All other cases.

This produced 49 speaker-to-real-identity mappings across 12 teams.

F.6 Stage 6: Utterance Boundary Refinement

We refined utterance boundaries using GPT-5-mini via the OpenAI API, selected for its reliable structured JSON output and applied only to already-anonymized transcripts. The model decides one of four actions per utterance—KEEP, MERGE_NEXT, MERGE_PREV, or SPLIT—guided by 10 rules from the Language Development Project (LDP) transcription guidelines (MacWhinney, 2000), using a sliding window of 8 utterances. Hard constraints prevent merging across different speakers or pauses ≥ 2 s.

Model & Configuration. We used `gpt-5-mini` via the OpenAI API (`AsyncOpenAI`) with the following settings:

LDP Rules. The system encodes 10 numbered rules from the Language Development Project transcription guidelines, plus 3 unnumbered heuristic rules (13 prompt items total). Two rules serve as hard constraints that cannot be overridden:

- **Rule 4.6:** Never merge utterances from different speakers.
- **Rule 4.8.1:** Never merge across pauses ≥ 2 seconds.

The remaining rules govern splitting and merging decisions:

Parameter	Value
Model	`gpt-5-mini`
Window size	8 utterances
Stride	6 (overlap = 2)
Pause threshold	2.0 s
Max concurrent calls	10
Max retries	3 (exponential backoff)

Table 14: Utterance boundary refinement configuration.

Rule	Category	Action	Description
4.3	Self-correction	Keep	Within-thought corrections
4.4	Abandoned thought	Split	Different-thought restart
4.5	False start	Keep	Same topic, same utterance
4.7	Multi-sentence	Split	No conjunction between sentences
4.9.1	Non-sentence tag	Merge	"ok," "right," "honey"
4.9.3	Complete tag	Split	"I know," "I don't know"
4.9.4	Perception verb	Keep	"I know it's crazy" = 1 utt.
4.10	Repetition	Keep	"no no no" preserved

Table 15: LDP rules encoded in the utterance boundary refinement system.

Processing Results. The released corpus contains 104 transcript files from 12 teams spanning 88 unique meeting dates; longer sessions were split into multiple parts. Across the processed corpus:

- Splits applied: 1,238 (including multi-sentence splits producing 2–5 segments each); Merges applied: 309
- API calls: 1,015 windows processed
- Error rate: 0% (all windows successful)

G Full Prediction Results

Table 16 shows the complete results for Experiment 1, including all proprietary models, few-shot baselines, and Llama-3.1-8B fine-tuned conditions.

H Model-Agnostic Validation

To confirm that our findings are not artifacts of a specific model, we replicate the full Experiment 1 conditions with Llama-3.1-8B-Instruct using identical LoRA configuration (rank 16, α = 32, 2 epochs, lr = 2e-4). Table 17 shows results across all six conditions.

I Learning Curve and AMI Experiment Details

Learning curve setup. For each team, we chronologically order all meetings and incrementally add them to the training set. At step k, the model trains on meetings 1 through k and is tested on all remaining meetings $k{+}1$ through N. This means that the test set changes and shrinks as k increases, so rows should not be interpreted as evaluations on an identical fixed test set; late-stage estimates also have high variance and may exceed the cross-team reference. Training uses the same LoRA configuration as Experiment 1, with the SU (Speaker + Utterance) format. We additionally evaluate the 9-team general model (trained on all 32,243 samples) on each team's full test data to establish a cross-team reference. Table 18 reports the full learning-curve results for both Gemma-3-12B and Llama-3.1-8B; the main paper (Table 2) reports Gemma only.

			Speaker		Intention	
Model	**Context**	**Type**	**Acc.**	**Δ Bigram**	**Acc.**	**Δ Maj.**
Baselines						
Random	—	—	30.22	—	7.14	—
Bigram	—	—	50.75	(ref)	—	—
Majority	—	—	—	—	30.05	(ref)
Zero-shot / Vanilla						
Llama-3.1-8B	SU	vanilla	29.07	−21.7	15.57	−14.5
Gemma-3-12B	SU	vanilla	43.56	−7.2	25.48	−4.6
GPT-5.4-mini	*SU*	*zero-shot*	*59.23*	*+8.5*	*24.22*	*−5.8*
Sonnet 4.6	*SU*	*zero-shot*	*59.70*	*+9.0*	*25.38*	*−4.7*
GPT-5.4	*SU*	*zero-shot*	*61.97*	*+11.2*	*25.68*	*−4.4*
Opus 4.6[†]	*SU*	*zero-shot*	*63.25*	*+12.5*	*26.68*	*−3.4*
Gemma-3-12B	SU	7-shot	53.85	+3.1	—	—
Fine-tuned (Gemma-3-12B-IT + LoRA)						
Gemma-3-12B	SU	FT	64.53	+13.8	**32.96**	+2.9
Gemma-3-12B	SRU	FT	64.82	+14.1	32.53	+2.5
Gemma-3-12B	S+R	FT	**65.74**	+15.0	31.17	+1.1
Fine-tuned (Llama-3.1-8B + LoRA)						
Llama-3.1-8B	SU	FT	65.17	+14.4	**33.18**	+3.1
Llama-3.1-8B	SRU	FT	**66.20**	+15.5	32.82	+2.8
Llama-3.1-8B	S+R	FT	65.65	+14.9	31.72	+1.7

[†]Speaker: 154 API errors (3.0%); Intention: 57 errors (1.1%), treated as incorrect. Retrying yields 65.04% speaker accuracy, above FT-SU (64.53%) but below the best fine-tuned condition (S+R, 65.74%). Majority baseline for intention = "Giving Information" (30.05%). 7-shot intention not evaluated.

Table 16: Full prediction results for Experiment 1, including all models and conditions. See Table 1 for the main results.

		Speaker (%)		Intention (%)	
Format	**Mode**	Gemma	Llama	Gemma	Llama
SU	Vanilla	43.56	29.07	25.48	15.57
SU	FT	64.53	65.17	32.96	33.18
SRU	Vanilla	38.69	28.01	24.48	14.96
SRU	FT	64.82	66.20	32.53	32.82
S+R	Vanilla	47.78	42.64	21.50	11.17
S+R	FT	65.74	65.65	31.17	31.72

All FT results are within 0.1–1.4 pp across models for speaker prediction, and within 0.2–0.6 pp for intention prediction, suggesting model-agnostic patterns.

Table 17: Model-agnostic validation: Gemma-3-12B vs. Llama-3.1-8B across all conditions. Same LoRA config, same data splits.

AMI data preparation. We use the AMI Meeting Corpus from HuggingFace (`edinburghcstr/ami`, IHM configuration). The official test split contains 16 meetings from 4 groups (EN2002, ES2004, IS1009, TS3003). We construct sliding-window samples with context-size 8 to match Hilgert & Niehues (2025), yielding 12,515 test samples. Speakers are anonymized to PERSON_A/B/C/D by order of first appearance in each meeting.

For the balanced TIDES+AMI training mix (124K), we use TIDES in unmerged format (62,221 samples, context 8) and downsample AMI training data from ∼108K to 62,221 samples to match. The unmerged format preserves finer utterance boundaries, that align with AMI's segmentation style. The full unbalanced mix (170,038 samples, trained for 1 epoch) uses all available data from both corpora without downsampling, but degrades accuracy to 30.02%—below both the TIDES-only transfer condition (35.33%) and Hilgert & Niehues'

		Gemma-3-12B		Llama-3.1-8B	
Mtgs	**Train** N	**Acc.**	**% ref.**	**Acc.**	**% ref.**
Team 10 (Korean, 4 members, 7 meetings)					
1	276	50.87%	75.8%	53.61%	75.6%
3	795	55.37%	82.5%	58.46%	82.4%
4	1,073	62.62%	93.3%	63.20%	89.2%
6	1,811	65.82%	98.1%	64.35%	90.8%
Team 5 (English, 5 members, 10 meetings)					
1	203	50.00%	73.2%	60.73%	99.6%
3	956	63.84%	93.5%	64.83%	106.3%
5	1,648	64.81%	94.9%	65.10%	106.8%
7	2,570	64.59%	94.6%	64.68%	106.1%

References—Gemma: 67.09% (T10), 68.27% (T5); Llama: 70.89% (T10), 60.96% (T5). Llama T5 exceeds its reference due to the small test set (187 samples).

Table 18: Full chronological single-team adaptation results for both models. % ref. is relative to each model's 9-team general model evaluated on that team; because the remaining-meeting test set changes across rows, the percentages are descriptive and are not fixed-test learning-curve estimates.

zero-shot result (34.88%)—suggesting that corpus balance is important for cross-corpus transfer.

J Training Composition and Cross-Culture Analysis

To understand how language composition in training data affects speaker prediction, we fix the test set (Teams 6&7, Korean) and vary training composition across four conditions: Korean-only, English-only, balanced (Korean downsampled to match English), and the original mixed set. Note that "Korean teams" and "English teams" refer to the language originally spoken during recorded meetings, not the actual language of the post-processed data.

Condition	**Lang.**	**Train** N	**Acc. (%)**
Vanilla	—	0	43.56
Bigram	—	—	50.75
EN-only	100% EN	8,703	57.20
Balanced	50/50 KR/EN	17,406	58.74
KR-only	100% KR	23,540	60.21
Mixed (orig.)	73/27 KR/EN	32,243	**64.53**

Table 19: Effect of training language composition on speaker prediction. Test set fixed: Teams 6&7 (Korean). Gemma-3-12B-IT + LoRA.

Table 19 shows that **data volume is the primary driver**: Mixed (64.53%, 32K samples) outperforms all alternatives, and KR-only (60.21%) surpasses EN-only (57.20%) by 3.0 pp, reflecting a same-language advantage for the Korean test set. EN-only still exceeds the bigram baseline by +6.4 pp, suggesting that cross-lingual transfer of turn-taking dynamics is substantial—conversation structure transfers across languages even without target-language data. Balanced (58.74%) slightly exceeds EN-only but falls below KR-only, indicating that subsampling from 23K to 8.7K Korean samples carries a cost not fully offset by language diversity.

	Speaker Acc. (%)			ROUGE-L	Inv.	Semantic Sim.		
Condition	1t	3t	5t	(1t)	Spkrs†	1t	3t	5t
Fine-tuned (Gemma-3-12B)								
FT-Plain (plain generation)	49.0	31.0	28.6	4.79	178	.307	.321	.345
FT-Reason (social-cue reasoning)	57.0	37.7	29.2	5.46	152	.283	.317	.360
Vanilla (no fine-tuning)	14.0	21.7	20.2	4.41	786	.319	.346	.348
Proprietary (zero-shot generation)								
GPT-5.4	51.0	42.3	37.0	—	0	.364	.432	.441
GPT-5.4-mini	60.0	46.3	37.0	—	0	.360	.424	.436
Opus 4.6	**62.0**	**46.3**	**39.0**	—	0	.379	.411	.426
Sonnet 4.6	59.0	41.7	38.2	—	0	**.396**	**.437**	**.440**

†FT/Vanilla: counted at 10-turn generation; proprietary: per-turn (all 0). ROUGE-L not computed for proprietary (different prompt format).

Table 20: Automatic evaluation of generated utterances (100 samples per generation length). Speaker accuracy = fraction of correctly predicted speakers averaged across all generated turns. Semantic similarity = cosine similarity of sentence embeddings (`all-MiniLM-L6-v2`) between concatenated generated and ground-truth turns. **Bold** = best per column.

K Detailed Breakdowns

Per-role analysis. Speaker prediction accuracy varies by the target speaker's emergent role. Attention Seekers are most predictable (75.4% on the test set), likely because their frequent backchanneling creates strong sequential patterns. Critics are hardest to predict (53.1%), consistent with their tendency to interject at unpredictable moments.

Tuckman stage progression. Teams in the Forming stage (Tuckman) show lower speaker prediction accuracy (57.8% on the test set) than teams in the Performing stage (70.3%), a +12.5 pp gap. This aligns with theory: established teams develop more predictable interaction patterns as they mature.

Intention class analysis. The fine-tuned model's intention predictions concentrate on Giving Information (74%) and Active Listening (21%), achieving top-3 accuracy of 62.6% but macro F1 of only 0.072. This class collapse reflects the skewed label distribution and ambiguity of the fine-grained categories, suggesting that intention prediction may require coarser targets, richer contextual signals, or task-specific architectures.

L Generation Supplementary

Automatic evaluation. Table 20 shows that FT-Reason consistently outperforms FT-Plain on structural metrics (speaker accuracy, ROUGE-L) while Vanilla produces the most invalid speakers (786 at 10-turn vs. 152–178 for FT). Degeneration rates increase with generation length: 1% at 1-turn, 9% at 3-turn, and 23% at 5-turn, motivating the 4-stage quality filtering pipeline described below.

Surface post-processing. Manual inspection revealed that FT outputs, while contextually grounded, suffered from surface-level formatting issues absent in vanilla outputs, such as missing sentence-initial capitalization, missing punctuation, lowercase "i", and truncated sentence endings. To ensure that evaluators judged content rather than formatting, we applied a two-pass surface polish to all FT utterances before evaluation. Pass 1 used GPT-5.4-mini to correct capitalization and basic punctuation (1,390/1,440 utterances modified). Pass 2 applied rule-based truncation fixes (278 trailing periods removed from incomplete sentences) followed by GPT-5.4-mini comma insertion (705 commas added). All changes were verified as surface-only. No words were added, removed, or reordered. Despite this polish, automated LLM judges still preferred Vanilla outputs (83–87% across prompt

variants), and the subsequent human evaluation confirmed the same pattern, indicating that the preference gap is not driven by surface formatting.

Quality filtering. We applied a 4-stage pipeline to select evaluation samples: (1) GPT-based defect filtering (5 defect categories, 3-run majority vote, 2,700 evaluations $\rightarrow$ 171 pass), (2) manual audit (3 removed), (3) subtle quality filtering (unnatural enumeration), and (4) uniform sampling (40 per generation length). Five-turn samples required regeneration with improved parameters (temperature 0.7$\rightarrow$0.5, repetition penalty 1.2$\rightarrow$1.3, best-of-3 selection). The final set comprises 120 samples $\times$ 3 pairs = 360 comparisons + 20 attention checks.

Human evaluation setup. We recruited evaluators through Prolific, requiring English fluency and prior experience with collaborative teamwork. The evaluation interface presented pairs of AI-generated meeting continuations (A and B, randomly assigned) alongside the original conversation context. Evaluators rated each pair on five dimensions: (1) overall preference (A much better / A slightly better / Tie / B slightly better / B much better), (2) confidence (1–5), and for each side separately, (3) naturalness (1–5), (4) coherence (1–5), and (5) speaker consistency (1–5).

Quality control used two mechanisms: attention checks (3 per evaluator, requiring identification of a clearly superior continuation) and instructed-response checks (requiring a specific Likert value). Evaluators failing $\geq$2 checks of the same type were terminated early. Of 123 total participants, 66 passed quality thresholds, yielding 1,142 quality-controlled pairwise judgments across 360 comparison items (3 pairs $\times$ 120 samples).

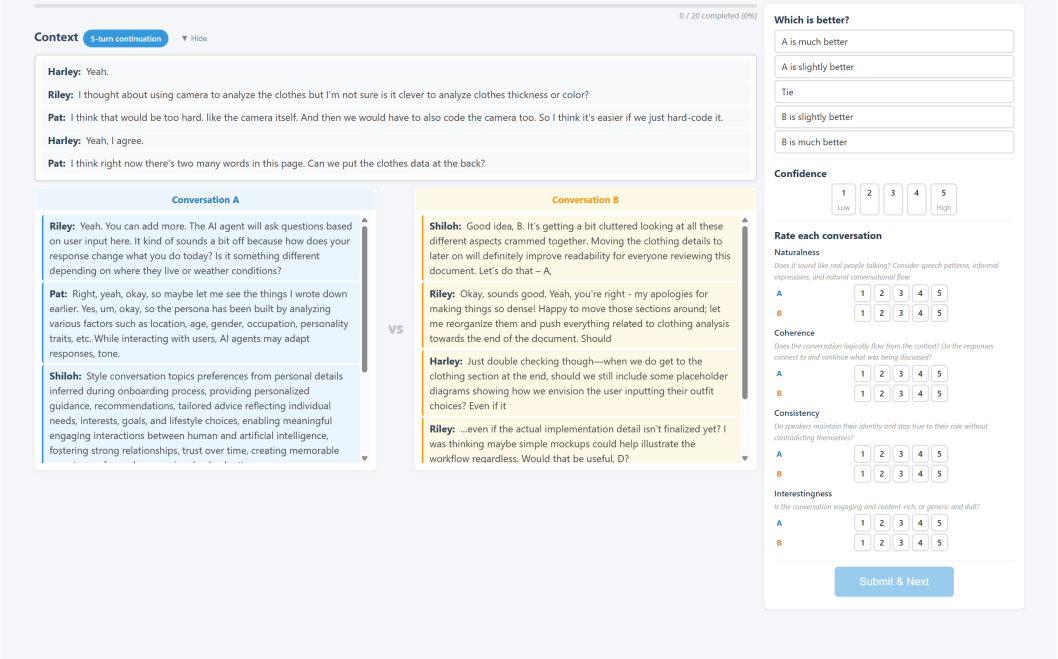

Figure 3: Human evaluation interface on Prolific. Evaluators view the original conversation context (top), two AI-generated continuations A and B (bottom left/right, randomly assigned), and rate overall preference, confidence, naturalness, coherence, speaker consistency, and interestingness on the right panel.

L.1 LLM-as-Judge Evaluation

Table 21 compares two LLM judges. GPT-5.4 strongly aligns with human preferences (Vanilla preferred in 66.7–72.5% of comparisons involving Vanilla). Opus 4.6 diverges: it prefers FT-Plain over Vanilla overall (54.2% vs. 45.0%) and FT-Reason over Vanilla (62.5% vs. 37.5%),

particularly at shorter generation lengths. At 1-turn, Opus prefers FT-Plain 72.5% of the time, but this reverses at 5-turn where Vanilla is preferred 67.5%. This suggests that Opus is more sensitive to structural coherence (where FT excels) whereas GPT prioritizes surface fluency (where Vanilla excels), highlighting that LLM-based evaluation of multi-party dialogue generation remains model-dependent.

	Win rate (%)	
Model	**GPT-5.4**	**Opus 4.6**
FT-Plain vs. Vanilla		
FT-Plain	27.5	54.2
Vanilla	72.5	45.0
FT-Reason vs. Vanilla		
FT-Reason	33.3	62.5
Vanilla	66.7	37.5
FT-Plain vs. FT-Reason		
FT-Plain	40.0	29.4
FT-Reason	60.0	51.3

GPT-5.4 aligns with human evaluators (Vanilla preferred). Opus 4.6 diverges, preferring FT at shorter horizons.

Table 21: LLM-as-judge pairwise evaluation (360 comparisons per judge). Win rate = % of comparisons in which the judge preferred that model. GPT-5.4 reports no ties; for Opus 4.6, the remainder within each comparison are ties.

Multi-turn fine-tuning failure. We attempted direct 5-turn fine-tuning (6,397 non-overlapping training samples), but 90% of generated outputs failed to parse correctly. We reverted to 1-turn fine-tuning with autoregressive multi-turn generation, which proved more reliable despite compounding errors over turns.

M Training and Inference Configuration

Table 22 summarizes the training hyperparameters for all fine-tuning experiments. All models use LoRA adapters with completion-only loss (prompt tokens masked). Inference for open-source models uses single-token logit scoring, where the prompt is fed through the model, and the candidate with the highest log-probability at the prediction position is selected. For proprietary models, we use greedy generative decoding (temperature 0, max tokens 50) via the respective APIs.

N Prompt Templates

We show the exact prompt format used for each task. Both open-source and proprietary models receive identical system and user messages.

Speaker prediction (SU format).

> **System:** *You are an expert at predicting conversational dynamics in team meetings. Given a dialogue history, predict which team member will speak next. Answer with only the speaker ID (e.g., PERSON_A).*
>
> **User:** *Below is a dialogue history from a team meeting. Predict which speaker speaks next.*
>
> *Dialogue:*
> *[Turn 1] PERSON_A: Hello. I'm Jordan.*
> *[Turn 2] PERSON_B: I'm Alex.*
> *[Turn 3] PERSON_C: I'm Drew.*
> *[Turn 4] PERSON_D: I'm Reese.*
> *[Turn 5] PERSON_A: It's being recorded, so just in case, I'll turn this one on too.*
> *Participants in this meeting: PERSON_C, PERSON_B, PERSON_A, PERSON_D*

Parameter	Gemma-3-12B	Llama-3.1-8B
LoRA rank (r)	16	16
LoRA alpha (α)	32	32
LoRA dropout	0	0
LoRA targets	q, k, v, o, gate, up, down_proj	
Trainable params	$\sim$0.56%	$\sim$0.56%
Epochs	3 (2 for Mixed)	2
Learning rate	1e-4	2e-4
LR scheduler	cosine	cosine
Warmup ratio	0.1	0.1
Batch size	2	2
Gradient accum.	8	8
Effective batch	16	16
Max seq length	4,096	4,096
Precision	bf16	bf16
Optimizer	AdamW	AdamW
Hardware	2$\times$ NVIDIA RTX A6000 (48GB)	
Wall time (Exp 1)	$\sim$18h	$\sim$8.5h

Table 22: Training configuration for all fine-tuning experiments. All use LoRA with no quantization (full bf16).

Who speaks next? Answer with only the speaker ID.
Expected output: *PERSON_B*

Intention prediction (SU format).

System: *You are an expert at predicting utterance types in team meeting conversations. Given a dialogue history, predict the utterance type of the next turn.*
[14 category definitions provided, e.g.:
- Active listening: Short utterances showing attention or agreement
- Giving Information: Providing objective facts or asking factual questions
...]
Answer with only the utterance type name.
User: *[Same dialogue format as speaker prediction, with utterance types annotated per turn, e.g.:*
[Turn 1] PERSON_A [Social / Humor]: Hello. I'm Jordan.]
Expected output: *Active listening*

Generation (FT-Plain, plain format). The model receives the 5-turn context and generates the next utterance in "`SPEAKER: utterance`" format. For FT-Reason (social-cue reasoning), the model first generates a structured chain: "`Next speaker: X / Role: Y / Intention: Z / Utterance: text`". For Vanilla and proprietary models, the same context is provided, with instructions to continue the conversation naturally.