Simple Agents, Biased Judges: Efficient Multi-Party Dialogue Generation & The Evaluation Gap

Kunal Samanta[1], Faisal Tareque Shohan[2], Amine Trabelsi[2], Richard Khoury[1]
[1]Université Laval
[2]Université de Sherbrooke
kunal.samanta.1@ulaval.ca, richard.khoury@ift.ulaval.ca
{faisal.tareque.shohan, amine.trabelsi}@usherbrooke.ca

Abstract

Multi-party social dialogue remains underexplored in the literature, in part due to the difficulty and cost of evaluation. As a result, recent work on synthetic dialogue generation often relies on automated metrics and LLM-as-a-Judge frameworks, despite limited evidence that such judges reflect human preferences in social settings. In this work, we introduce a lightweight and controllable multi-party dialogue generation framework (MPOD) as an experimental instrument for studying generation and evaluation in social interaction. Using this framework, we conduct human evaluations of open-domain multi-party dialogue simulation and directly compare human judgments against state-of-the-art LLM judges. Across 319 pairwise comparisons, we observe near-random agreement between humans and automated judges (Cohen's $\kappa \approx 0.11$), driven by systematic behaviors including extreme tie aversion and strong sensitivity to assistant-style verbosity. Crucially, human–human inter-annotator agreement ($\kappa = 0.29$) is substantially higher than human–LLM agreement. To isolate the mechanism underlying this misalignment, we introduce a controlled *Transplant Ablation*, showing that LLM judges consistently prefer conversations containing a single proprietary, assistant-style agent. Additional stress tests show that judges prefer GPT-style conversations even when utterance order is randomly shuffled, indicating insensitivity to conversational structure and coherence. Our findings provide controlled evidence that current instruction-tuned LLM judges do not reliably reflect human preferences for naturalness, engagingness, and overall quality in multi-party social dialogue, calling into question their widespread use for validating synthetic conversational data.

1 Introduction

The simulation and evaluation of realistic multi-party dialogue remains a formidable challenge. Unlike dyadic user-assistant interactions, multi-party conversation requires agents to manage complex turn-taking, navigate topic drift, and exhibit social behaviors such as backchanneling and interruptions (Wang et al., 2023). These properties are difficult to formalize and expensive to evaluate, particularly at scale.

Due to these challenges, recent work on multi-party and synthetic dialogue generation has increasingly relied on automated evaluation frameworks such as G-Eval (Liu et al., 2023), GPTScore (Fu et al., 2024), and LLM-as-a-Judge paradigms. In practice, these judges are often treated as proxies for human preference, and are used to justify improvements in conversational quality without direct validation against human judgment, especially in social and open-ended settings.

Despite the widespread adoption of LLM-based judges, there is little controlled evidence assessing whether they align with human preferences in multi-party conversational settings. To study this question, we introduce a simple and controllable multi-party dialogue generation framework: ***Multi Party Open Domain*** (MPOD) based on a *single inference loop* with dynamic persona prompting and lightweight similarity constraints.

Using this framework, we conduct the first human evaluations of open-domain multi-party social dialogue and directly compare human judgments against state-of-the-art LLM judges. While human annotation is costly, limiting the scale of such studies, it provides a necessary ground truth for understanding whether automated evaluators meaningfully reflect social conversational quality.

Our results reveal a substantial and systematic misalignment between human evaluators and automated judges. Across 150 pairwise comparisons, agreement between humans and LLM judges is near random. Moreover, unlike humans, automated judges almost never express uncertainty through ties, instead forcing binary preferences even in cases where humans perceive equivalent quality.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 43326–43345
July 2-7, 2026 ©2026 Association for Computational Linguistics

To move beyond surface-level disagreement, we introduce a controlled *Transplant Ablation*, in which a single assistant-style agent is inserted into an otherwise homogeneous open-weight agent group. We find that judges overwhelmingly prefer these "transplant" teams containing "assistant-style" agents, providing controlled evidence that automated evaluators are strongly driven by style and verbosity cues.

Our experiments further reveal that this misalignment is not limited to subtle stylistic preferences and judges may prefer conversations with broken turn order or mixed speakers as long as surface style cues align with a favored model family.

In summary, our contributions are:

1. A controllable experimental framework (MPOD) for multi-party open-domain dialogue generation, designed to facilitate analysis rather than maximize model performance.

2. To the best of our knowledge, the first human-grounded evaluations of open-domain multi-party social dialogue, addressing a gap in current evaluation practice.

3. Controlled evidence of systematic misalignment between human judgments and LLM-as-a-Judge frameworks, including tie aversion and stylistic bias.

2 Related Work

2.1 Multi-Party Dialogue Generation Architectures

The transition from dyadic to multi-party dialogue introduces substantial challenges in state management, turn-taking, and interactional coherence. While synthetic data generation pipelines such as Diasynth (Suresh et al., 2025) and SODA (Kim et al., 2023) have been effective for bootstrapping dyadic conversational models, extending these approaches to multi-party settings requires modeling interpersonal dynamics, overlapping contributions, and coherent topic progression across multiple speakers.

Most existing work addresses this complexity through explicit multi-agent architectures. Frameworks such as ConvoGen (Gody et al., 2025), AutoGen (Wu et al., 2024), and MALLM (Becker, 2024) rely on agent graphs that instantiate separate memory states and inference processes for each participant, often augmented with specialized planning, critique, or debate modules. Similarly, MMAgents (Nonomura and Mori, 2025) enforces structured turn-taking via adjacency-pair detection and explicit orchestration protocols. These designs are well-suited for task-oriented or deliberative settings, but they introduce substantial computational overhead and context fragmentation, making controlled experimentation and large-scale human evaluation difficult.

2.2 Biases in Automated Evaluation

Evaluating open-domain dialogue remains an unresolved challenge. As reference-based metrics such as n-gram overlap and other metrics like BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) have fallen out of favor, the field has increasingly adopted reference-free evaluation using large language models, commonly referred to as *LLM-as-a-Judge*. These approaches are now widely used to justify improvements in dialogue quality, particularly in settings where human evaluation is costly or impractical.

However, a growing body of work has shown that LLM-based judges exhibit systematic biases. Prior studies have documented self-preference bias, where judges favor outputs from their own model family (Chen et al., 2025), as well as verbosity bias, where longer or more elaborated responses are conflated with higher quality (Ye et al., 2024). SODA-EVAL (Mendonça et al., 2024) further highlights the subjectivity and teacher bias inherent in fine-tuned judges, even in relatively constrained dyadic dialogue settings.

Our work extends this line of analysis to the substantially less-studied domain of multi-party social dialogue. We argue that biases identified in dyadic evaluation are amplified in social settings, where conversational quality depends on interactional phenomena such as brevity, interruption, backchanneling.

3 Methodology

We introduce a simplified generation framework designed to produce high-fidelity open-domain multi-party conversations without the overhead of complex agent graphs or agent-level state management. Our system supports plug-and-play use of open-weight models via OpenAI-compatible inference

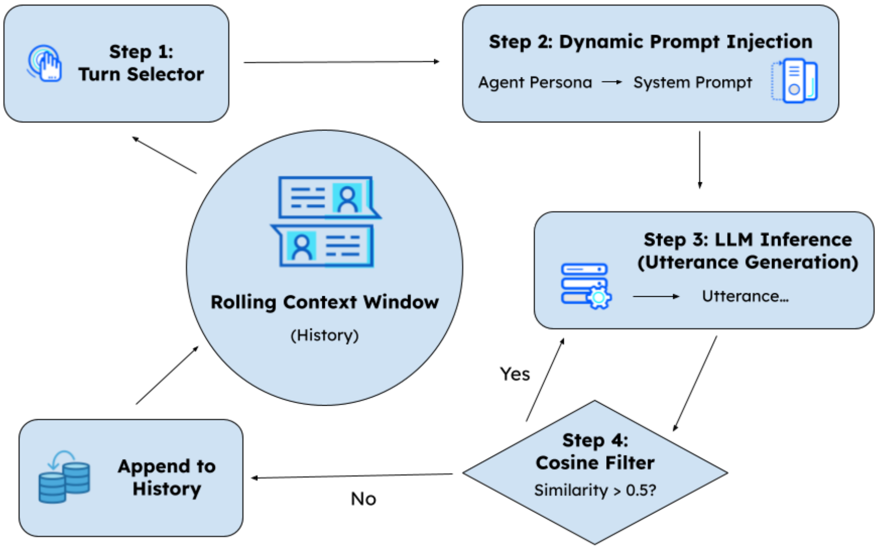


Figure 1: **The Single Inference Loop Architecture.** Unlike complex multi-agent graphs, our framework maintains a single rolling context window. A dynamic turn selector injects the active agent's persona into the system prompt, while a cosine-similarity constraint prevents repetitive loops during generation.

endpoints[1] (e.g., vLLM [2], `llama.cpp`[3]), as well as API-based proprietary models for conversation generation. This design enables fair comparison across model families while maintaining a transparent and lightweight generation pipeline.

3.1 Single Inference Loop Architecture

Unlike graph-based approaches (e.g., AutoGen) that instantiate separate memory and inference processes for each agent, our framework employs a **Single Inference Loop** architecture (see Figure 1). This design maintains a single rolling context window containing the complete dialogue history shared across all participants. Each conversational agent is defined solely by its persona specifications, without maintaining independent memory states.

A conversation is initialized using three components: **(1)** a topic description, **(2)** a set of participants with associated personas, and **(3)** an opening utterance that establishes the initial conversational context (see A.1 for details). All subsequent turns are generated within the same shared context window, allowing agents to react to the full dialogue history while avoiding the redundancy inherent in multi-agent graph architectures.

[1] OpenAI Compatible Endpoints
[2] vLLM
[3] llama.cpp

For each conversational turn, the system executes the following sequence of operations:

1. **Turn Selection**. A turn selection mechanism determines which agent should speak next. We support three strategies: *Round-Robin*, *Random*, and an *LLM-based Turn Selector*. Because the initial conversation state contains only a single utterance, providing insufficient context for reliable inference, we apply the Round-Robin strategy for the second turn. Once at least two utterances are available, the LLM-based Turn Selector is used. This selector analyzes the five most recent utterances along with the list of participant names and returns the most contextually-appropriate next speaker in a structured JSON format. To prevent infinite loops or crashes, if the LLM selector fails to return a valid JSON after 3 retries, the system defaults to a Random fallback. In our experiments, this fallback was triggered in less than 5% of turns, primarily during early dialogue states. For more details on the LLM turn selector see Appendix A.3.

2. **Dynamic Instruction Injection**. Once a speaker is selected, the system prompt is dynamically updated to reflect the active agent's persona (See Section 5 for example agent persona). This approach enables a single model

instance to simulate multiple distinct conversational voices without requiring fine-tuning or Low-Rank Adaptation (LoRA) (Hu et al., 2022). Compared to prior approaches that rely on lengthy, rigid persona instructions throughout the entire context window, dynamic injection reduces prompt verbosity and context saturation. These benefits are particularly pronounced for smaller open-weight models (e.g., < 100B parameters), which are more sensitive to long and restrictive system prompts.

3. **Utterance Generation**. The selected model generates the next utterance conditioned on the dynamically updated system prompt and the most recent dialogue history window, consisting of the five preceding utterances. Due to hardware resource constraints, we employed ***4-bit*** quantized weights for all inferences involving open-weight models.

3.2 Constraints: The Cosine Similarity Threshold

To mitigate the *agreement loop* failure mode commonly observed in social dialogue (Wynn et al., 2025), we introduce a **Cosine Similarity Constraint** that discourages semantic repetition within an individual agent's contributions.

Before a generated utterance is accepted, the system computes the cosine similarity between the embedding of the proposed utterance (u_t) and the embedding of the agent's recent *self-utterance history* ($C_{t-k:t}$):

$$\mathrm{sim}(u_t, C_{t-k:t}) = \frac{u_t \cdot C_{t-k:t}}{\|u_t\| \, \|C_{t-k:t}\|}$$

We calculate semantic similarity using the `nomic-embed-text-v1` sentence transformer model. For every proposed utterance u_t, we compute the cosine similarity against the rolling history $C_{t-k:t}$. We set the history window $k = 5$ and the rejection threshold $\tau = 0.5$. If $\mathrm{sim}(u_t, C_{t-k:t}) > \tau$, the utterance is rejected. We allow a maximum of 5 resampling attempts with a temperature increase of 0.1 per attempt to force divergence. If all attempts fail, the least similar candidate is accepted to maintain flow.

4 Experimental Design

To assess the quality of generated multi-party dialogues and to systematically analyze the behavior of automated judges, we employ a comprehensive experimental protocol involving both human annotation and LLM-based evaluation.

4.1 Dataset and Dialogue Generation

Our experimental setup consists of two complementary objectives.

Objective 1: Comparison with a Multi-Agent Baseline. The first objective is to compare the quality of dialogues generated by our framework against a representative multi-agent baseline, ConvoGen (Gody et al., 2025). We select 8 distinct conversational topics, along with associated participant personas, using the ConvoGen framework. For each topic–persona configuration, we generate two conversations: one using ConvoGen and one using our framework (MPOD), ensuring a controlled comparison.

To isolate architectural effects, conversations are generated using a mix of open-weight and proprietary models, specifically Llama 3.3 70B and GPT-5.1. Each ConvoGen–MPOD pair is annotated by two human evaluators, resulting in a total of 16 human annotations for this objective.

Objective 2: Cross-Model Evaluation and Judge Alignment Analysis. The second and primary objective is to investigate the alignment between human judgments and automated LLM-based judges across a diverse set of models. For this purpose, we generate multi-party conversations spanning 10 social topics using our framework (MPOD), ranging from serious discussions (e.g., *AI Ethics*) to casual and creative scenarios (e.g., *Planning a Heist*, *Culinary Adventures*). Topics and participant personas are generated using Gemini 3 Pro (see Table 7 in Appendix for the full list).

Using our generation framework, we then produce conversations with the following models: Gemini 3 Pro (Gemini 3), GPT-OSS 120B (GPT-OSS), Llama 3.3 70B (Llama 3.3), GPT-5.1 (GPT 5.1), Gemma 3 27B (Team et al., 2025), and Qwen 3 30B (Yang et al., 2025). These conversations form the basis for both human evaluation and automated judge analysis described in subsequent sections.

4.2 Evaluation Protocol

Our evaluation protocol is a shift from more prevalent reference free methods like GPTScore (Fu et al., 2024), G-Eval (Liu et al., 2023) and FED (Mehri and Eskenazi, 2020) which exclusively use likert scale scores (1-5) to rate dialogue quality.

Following the recommendations of Smith et al. (2022), we employ pairwise comparison as the primary evaluation methodology. Evaluators were presented with two complete conversation transcripts (labeled only as “Conversation A” and “Conversation B”) displayed side-by-side and are asked to indicate which conversation they prefer according to three criteria: *Naturalness*, *Engaginess*, and *Overall Preference*. Evaluators were provided with detailed guidelines defining ‘Naturalness’ and ‘Engaginess’ (Figure 4) and performed the task on a custom-built annotation platform (Figure 5) designed to blind model identity. Evaluators were also explicitly permitted to select a *Tie* option when the two conversations are of comparable quality. And finally, a free-text box allowed annotators to explain briefly their choices.

To mitigate ordering and recency effects, the positions of Model A and Model B were randomized such that each model appears equally often in both positions across comparisons.

4.3 Judges

We collect human annotations for each model pair and conversation topic. We recruited 8 human annotators that are graduate level university students. Annotators were screened for English proficiency and were compensated at a rate of $15/hour. Further details regarding the human evaluation process can be found in Appendix B.

Comparisons were annotated by more than one annotator, yielding overlapping judgments on all 319 pairwise comparisons used in the alignment analysis (Objective 2). We report inter-annotator agreement (IAA) using Cohen’s κ (pairwise) and Fleiss’ κ (multi-rater) in Section 5.2.

We repeat this evaluation protocol using state-of-the-art LLMs as automated judges. The set of LLM judges includes: (1) GPT-5.2, (2) Gemini 3 Flash and (3) Llama 3.3 70B. For the Automated Judges, to handle position bias, every pair was evaluated twice (swapping A/B positions), and we only counted a ‘Win’ if the judge was consistent across both permutations or if the majority vote across 3 runs favored one model. We also let the LLM judges output an explicit reasoning prior to preference selection, following prior findings that structured reasoning can improve calibration in automated evaluation (Chiang and Lee, 2023). Appendix C provides additional information concerning the specific instructions used for the LLM judge.

5 Results

5.1 Comparison with Multi-Agent Baseline

Table 1 presents the head-to-head comparison between our Single Inference Loop framework MPOD and the multi-agent ConvoGen baseline. Across 16 pairwise comparisons spanning 8 diverse topics, human annotators demonstrated a strong preference for our simplified approach.

As shown in Table 1, our framework achieved an **87.5% win rate** when using GPT-5.1 and a **75.0% win rate** with Llama 3.3 70B.

Base Model	Win Rate (MPOD)	Win Rate (ConvoGen)
GPT-5.1	87.5%	12.5%
Llama 3.3 70B	75%	25%

Table 1: Head-to-head comparison of our lightweight framework MPOD against the multi-agent ConvoGen baseline across 8 topics. Even with open-weight models, our simple constraint-based approach is preferred by annotators.

5.2 Agreement Analysis: The Human–Machine Alignment Gap

We evaluate agreement between human annotators and state-of-the-art LLM judges (Llama 3.3, Gemini 3 Flash, and GPT-5.2) across 319 conversation pairs. The results reveal a systematic lack of alignment between human and automated judgments in the context of social multi-party dialogue.

As shown in Table 2, agreement between human evaluators and LLM judges is negligible. Cohen’s κ values range from 0.113 for the Llama 3.3 judge to a maximum of 0.175 for the Gemini 3 Flash judge. These values fall well below commonly cited thresholds for moderate agreement ($0.4 < \kappa < 0.6$), indicating that automated judges rely on criteria that differ substantially from those used by human evaluators. Exact agreement rates lie between 46.0% and 49.3%, corresponding to performance no better than chance under a Bernoulli(0.5) baseline.

Critically, human–human inter-annotator agreement (Cohen’s $\kappa = 0.29$, Fleiss’ $\kappa = 0.30$) is substantially higher than any human–LLM agreement, demonstrating that the near-random human–LLM alignment is not attributable to inherent task ambiguity but to genuine divergence in evaluation criteria between humans and automated judges.

Furthermore, LLM–LLM agreement across all

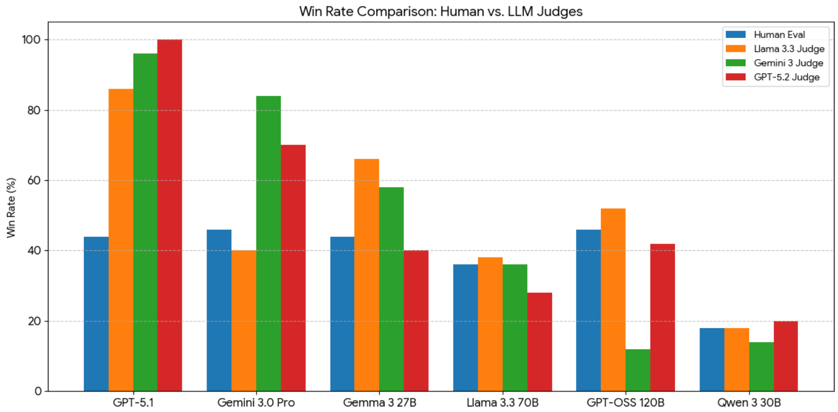


Figure 2: **The Alignment Gap.** Win rates for contestant models as evaluated by Humans (Blue) vs. SOTA LLM Judges. Note the extreme inflation for proprietary models: the GPT-5.2 Judge awards GPT-5.1 a 100% win rate, while Humans prefer it only 44% of the time. This demonstrates the "Echo Chamber" effect where judges favor their own model families or styles.

Agreement Type	Exact	Cohen's κ	Tie Rate
Human–Human	—	0.29	21.0%
Human vs. Llama 3.3	46.0%	0.113	0.2% (judge)
Human vs. Gemini 3 Flash	49.3%	0.175	0.0% (judge)
Human vs. GPT-5.2	47.3%	0.140	0.0% (judge)
LLM–LLM (avg.)	—	0.68	—

Table 2: Agreement between Human Annotators, LLM Judges, and inter-annotator statistics on Overall Preference (319 pairwise comparisons). Human–human agreement ($\kappa = 0.29$) substantially exceeds human–LLM agreement ($\kappa \approx 0.11$–0.18), confirming that the alignment gap reflects genuine divergence in evaluation criteria rather than task ambiguity. LLM–LLM agreement ($\kappa = 0.68$) indicates that automated judges are internally consistent yet systematically misaligned with humans.

judge pairs (Cohen's $\kappa = 0.68$, Fleiss' $\kappa = 0.68$) is dramatically higher than human–LLM agreement. Automated judges are highly self-consistent yet systematically diverge from human preferences — a pattern consistent with an evaluative echo chamber effect: LLM judges converge on shared stylistic priors that differ fundamentally from human notions of conversational quality.

5.3 Tie Aversion in Automated Judges

A major contributor to this misalignment is the manner in which ambiguity is handled by automated judges. Human annotators selected the *Tie* option in approximately 21% of comparisons (with tie rates ranging from 24% to 30% for Llama 3.3 pairs), reflecting the often ambiguous nature of conversational quality judgments.

In contrast, all three LLM judges exhibit a pronounced aversion to ties, despite being explicitly instructed that ties were permissible, as detailed in Table 2. Both the Gemini 3 Flash and GPT-5.2 judges produced no ties across all 150 evaluations, while the Llama 3.3 judge selected a tie in only a single case (0.2% tie rate). This forced binary decision-making introduces artificial distinctions in cases where humans perceive comparable quality, thereby amplifying noise in close comparisons.

5.4 Model Bias and the *Echo Chamber*

Figure 2 and Table 3 visualize the striking divergence in win rates in *overall preference*. We observe similar trends in *Naturalness* and *Engaginess*. For a complete pairwise breakdown of all model comparisons, we provide full win-rate heatmaps in Figure 6 in Appendix.

All three LLM judges dramatically inflate the performance of GPT-5.1 as a dialogue generator. It is the preferred generator for all three judges with a win rate of 86% or more. The GPT-5.2 judge even awards it a perfect 100% win rate (50 wins, 0 losses). By contrast, human judges rank it at third place with a 44% win rate.

While the literature has documented self-preference bias in LLM judges, our results are more

Generator Model	Human Eval Win Rate	Automated Evaluation (Win Rates) Llama 3.3 Judge	Gemini 3 Judge	GPT-5.2 Judge
GPT-5.1	44.0%	86.0%	96.0%	**100.0%**
Gemini 3.0 Pro	46.0%	40.0%	**84.0%**	70.0%
Gemma 3 27B	44.0%	**66.0%**	58.0%	40.0%
Llama 3.3 70B	36.0%	38.0%	36.0%	28.0%
GPT-OSS 120B	46.0%	52.0%	12.0%	42.0%
Qwen 3 30B	18.0%	18.0%	14.0%	20.0%

Table 3: Win Rates by Judge. Bold values indicate significantly inflated scores compared to Human judgment. The GPT-5.2 Judge exhibits extreme bias, awarding the GPT-5.1 model a perfect 100% win rate, despite humans preferring it less than half the time.

nuanced. GPT-5.2 has a clear self-preference bias. Gemini does gives its conversations a win rate of 84%, well above that of other LLM judges (40% and 70%) or human judges (46%), but nonetheless rates itself only at a strong second place and prefers GPT-5.1 over itself. Finally, Llama 3.3 does not display special preference towards itself in the evaluation at all. It gives its generated conversations a win rate comparable to human judgment (38% vs. 36%), achieving second-to-last performance.

Taken together, these results indicate that current LLM-as-a-Judge frameworks exhibit a form of evaluative *echo chamber* with strong correlation to verbosity (see Figure 3) which does not reflect actual human preferences. A qualitative example of this disagreement is provided in Figure 7, where human annotators prefer a concise, interrupted exchange, while the judge favors a verbose monologue.

5.5 The Transplant Ablation: Style Dominance

The results discussed in Section 5.4 show that LLM judges consistently rate homogeneous GPT-5.1 conversations higher than homogeneous Llama 3.3 conversations. To determine whether this is because GPT-5.1 generates more coherent conversations or because the LLM judges superficially prefer the writing style of GPT-5.1, we conduct a controlled transplant ablation. For identical topics and personas, we generate two sets of conversations:

- **Control (Homogeneous):** Three Llama 3.3 agents.
- **Experimental (Transplant):** Two Llama 3.3 agents and one GPT-5.1 agent.

Combining different agents with distinct verbosity and stylistic profiles will create conversations with lower internal consistency. If the LLM judges are picking winning conversations based primarily on conversational coherence or stylistic consistency, this should lower the win rate for these generations. On the other hand, if they prefer the GPT-5.1 writing style, the inclusion of that model as an agent in the conversation will increase its win rate.

Empirically, we observe the second pattern in Table 5. The LLM judges show a near-universal preference for transplant conversations over the homogeneous control.

We extend the transplant ablation to test whether judge preference reflects conversational coherence, stylistic homogeneity, or model-family dominance. In addition to the original Llama & Llama-GPT transplant, we compare the same transplant conversations against a homogeneous GPT baseline under identical topics and personas.

As shown in Table 5, the transplant configuration loses unanimously (0-10) across both Gemini-3-Flash and GPT-5.2 judges. This establishes that the transplant's apparent advantage is relative, not absolute. The LLM judges prefer GPT-style utterances regardless of whether they appear in a coherent multi-party conversation or a mixed-style setting.

Taken together, the two transplant comparisons in Table 5 reveal a *style dominance effect*: LLM judges rank conversations in precise proportion to the fraction of GPT-style utterances present (Homogeneous Llama $<$ Transplant $<$ Homogeneous GPT). Humans, by contrast, show no such monotonic pattern, as confirmed by additional human annotations of the transplant conditions reported in Table 4.

We note that these results do not establish that mixing speakers from different model families *reduces* conversational coherence. Rather, the key finding is that LLM judge preferences track the *presence* of GPT-style utterances far more

Comparison	Judge	Hom. Wins	Transplant Wins	Ties
GPT vs. Transplant	LLM	100%	0%	0%
GPT vs. Transplant	Human	38.1%	38.1%	23.8%
Llama vs. Transplant	LLM	0%	100%	0%
Llama vs. Transplant	Human	33.3%	37.5%	29.2%

Table 4: Human vs. LLM judge preferences in the Transplant Ablation. LLM judges enforce a strict monotonic ordering by GPT-utterance fraction; human judges show no such pattern, preferring neither the homogeneous nor the transplant configuration decisively.

strongly than human preferences do — a result that holds independent of any assumptions about coherence. The effect persists even when conversational structure is destroyed via utterance shuffling (Section 5.6), further confirming that judge behavior is driven by surface stylistic cues rather than structural quality.

5.6 Order Sensitivity Stress Test

To further test whether judges evaluate conversational structure or instead rely on surface stylistic cues, we conduct an order-sensitivity stress test. For each topic, we take a GPT-5.1 generated multi-party conversation and randomly shuffle the utterance order, breaking turn adjacency, response relevance, and conversational flow while preserving lexical and stylistic features. We then compare these shuffled conversations against coherent Qwen 3 30B conversations generated under the same topics.

As shown in Table 6, both Gemini-3-Flash and GPT-5.2 judges unanimously prefer the shuffled GPT conversations (10-0), despite their lack of conversational coherence. This result strongly suggests that judges do not robustly penalize violations of turn-taking or discourse structure in multi-party dialogue, and instead rely primarily on model-family style markers.

Judge Model	Homogeneous (Llama)	Transplant (Llama-GPT)
Gemini-3-Flash	0	10
GPT-5.2	0	10
Judge Model	**Homogeneous (GPT)**	**Transplant (Llama-GPT)**
Gemini-3-Flash	10	0
GPT-5.2	10	0

Table 5: Head-to-head performance comparisons across ten topics. The upper panel compares the **Homogeneous** (Llama 3.3-70B) baseline against the **Transplant** (Llama 3.3-70B-GPT-5.1) configuration. The lower panel displays the comparison between the **Homogeneous** (GPT-5.1) baseline and the same **Transplant** model.

Judge Model	Qwen 3 30B	GPT 5.1 (Shuffled)
Gemini-3-Flash	0	10
GPT-5.2	0	10

Table 6: Head-to-head comparison of **Qwen 3 30B** vs **GPT 5.1 (Shuffled)** across 10 topics. The 'Shuffled' designation indicates that the utterance order was randomized.

6 Conclusion

This work demonstrates that efficient, high-quality multi-party dialogue generation can be achieved using open-weight models such as Llama 3.3 70B and Gemma 3 27B when combined with simple architectural choices, including dynamic instruction injection and cosine similarity constraints. Despite its simplicity, our framework matches or outperforms more complex multi-agent baselines in human evaluation.

Moreover, our findings expose a substantial alignment gap between human evaluators and state-of-the-art LLM-as-a-Judge systems in the evaluation of social multi-party dialogue. Across multiple judges and 319 experimental pairwise comparisons, human–LLM agreement remains near chance ($\kappa \approx 0.11$), driven primarily by verbosity bias and assistant persona bias.

Crucially, inter-annotator agreement among human evaluators ($\kappa = 0.29$) is substantially higher than human–LLM agreement, ruling out task ambiguity as the cause of the alignment gap. At the same time, LLM–LLM agreement ($\kappa = 0.68$) reveals that automated judges are internally self-consistent, yet systematically misaligned with humans — confirming the evaluative echo chamber hypothesis.

Across transplant and shuffled-order stress tests, we find consistent evidence that LLM judges privilege stylistic homogeneity and model-family cues over conversational coherence, calling into ques-

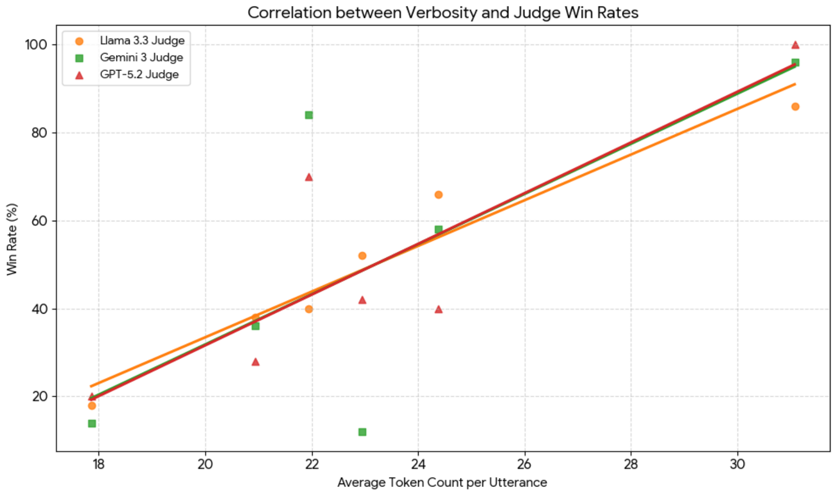


Figure 3: **Verbosity Bias Analysis.** Regression of model win rate against average utterance token count. All three LLM judges exhibit a strong positive correlation ($R^2 \in [0.51, 0.94]$), indicating that response length serves as a primary proxy for conversational quality in automated evaluation.

tion their suitability for evaluating multi-party social dialogue. We additionally show that LLM judge preferences track GPT- style utterance presence in strict monotonic proportion, a pattern absent in human judgments.

Taken together, our analysis suggests that general-purpose instruction-tuned models are poorly suited as evaluators of social dialogue within the specific generation regime studied here. Progress in this domain will require the development of socially aligned judges: models explicitly trained on human preference data for naturalness, engagement, and interactional dynamics, with explicit mechanisms to penalize excessive verbosity and assistant-like structure. Our findings likely extend beyond multi-party dialogue to any setting where interactional brevity and interruption are signals of quality, including collaborative writing, debate, and role-play agents. Without such advances, automated evaluation risks reinforcing stylistic artifacts rather than measuring genuine conversational quality.

Limitations

Our study has several limitations given the subjective nature of open-domain dialogues.

First, our experiments are conducted exclusively in English and focus on a fixed set of social topics and personas. The extent to which the observed alignment gap generalizes to other languages, cultural contexts, or interaction styles remains an open question.

Second, our conclusions are scoped to the specific generation regime studied: conversations generated from a 5-utterance history window with a 20-word brevity constraint, using three-participant groups. While this controlled setting enables precise causal claims about LLM judge behavior under matched conditions, further experimentation is needed to extrapolate our findings to the full complexity of spontaneous multi-party social dialogue.

Third, our analysis focuses on state-of-the-art, instruction-tuned LLMs as automated judges. While these models represent current frontier performance, alternative judging strategies such as models explicitly trained for evaluation or preference modeling may exhibit different behaviors.

Fourth, while we identify strong correlations between verbosity, stylistic features, and automated judge preferences, fully disentangling verbosity from other correlated properties of assistant-style outputs (e.g., formality or safety-oriented phrasing) remains an important direction for future investigation.

Finally, the lack of open-domain multi-party human corpora suitable for direct comparison prevented us from benchmarking MPOD against human-authored conversations. We consider this an important future extension. Existing multi-party corpora are either domain-restricted or consist of scripted media transcripts, which do not reflect the spontaneous conversational dynamics targeted by our framework.

Acknowledgments

This research was supported by Fonds de recherche du Québec - Nature et technologies (FRQNT) under the Grant No. 347227 and the Natural Sciences and Engineering Research Council of Canada (NSERC) under the Grant No. RGPIN-2022-04789.

References

Jonas Becker. 2024. Multi-agent large language models for conversational task-solving. *arXiv preprint arXiv:2410.22932*.

Zhi-Yuan Chen, Hao Wang, Xinyu Zhang, Enrui Hu, and Yankai Lin. 2025. Beyond the surface: Measuring self-preference in llm judgments. *arXiv preprint arXiv:2506.02592*.

Cheng-Han Chiang and Hung-yi Lee. 2023. A closer look into automatic evaluation using large language models. *arXiv preprint arXiv:2310.05657*.

Jinlan Fu, See Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024. Gptscore: Evaluate as you desire. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6556–6576.

Gemini 3. 2025. A new era of intelligence with Gemini 3. https://blog.google/products/gemini/gemini-3. Accessed: 2026-01-05.

Reem Gody, Mahmoud Goudy, and Ahmed Y Tawfik. 2025. Convogen: Enhancing conversational ai with synthetic data: A multi-agent approach. *arXiv preprint arXiv:2503.17460*.

GPT 5.1. GPT-5.1 A smarter, more conversational ChatGPT. https://openai.com/index/gpt-5-1/. Accessed: 2026-01-05.

GPT-OSS. Introducing gpt-oss. https://openai.com/index/introducing-gpt-oss/. Accessed: 2026-01-05.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Hyunwoo Kim, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Bras, Malihe Alikhani, Gunhee Kim, and 1 others. 2023. Soda: Million-scale dialogue distillation with social commonsense contextualization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12930–12949.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Llama 3.3. Llama-3.3. https://www.llama.com/docs/model-cards-and-prompt-formats/llama3_3/. Accessed: 2026-01-05.

Shikib Mehri and Maxine Eskenazi. 2020. Unsupervised evaluation of interactive dialog with dialogpt. *arXiv preprint arXiv:2006.12719*.

John Mendonça, Isabel Trancoso, and Alon Lavie. 2024. Soda-eval: Open-domain dialogue evaluation in the age of llms. *arXiv preprint arXiv:2408.10902*.

Ryota Nonomura and Hiroki Mori. 2025. Who speaks next? multi-party ai discussion leveraging the systematics of turn-taking in murder mystery games. *Frontiers in Artificial Intelligence*, 8:1582287.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Eric Smith, Orion Hsu, Rebecca Qian, Stephen Roller, Y-Lan Boureau, and Jason Weston. 2022. Human evaluation of conversations is an open problem: comparing the sensitivity of various methods for evaluating dialogue agents. In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 77–97.

Sathya Krishnan Suresh, Wu Mengjun, Tushar Pranav, and EngSiong Chng. 2025. Diasynth: Synthetic dialogue generation framework for low resource dialogue applications. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 673–690.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Hongru Wang, Lingzhi Wang, Yiming Du, Liang Chen, Jingyan Zhou, Yufei Wang, and Kam-Fai Wong. 2023. A survey of the evolution of language model-based dialogue systems. *arXiv preprint arXiv:2311.16789*.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, and 1 others. 2024. Autogen: Enabling next-gen llm applications via multi-agent conversations. In *First Conference on Language Modeling*.

Andrea Wynn, Harsh Satija, and Gillian Hadfield. 2025. Talk isn't always cheap: Understanding failure modes in multi-agent debate. *arXiv preprint arXiv:2509.05396*.

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

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V Chawla, and Xiangliang Zhang. 2024. Justice or prejudice? quantifying biases in llm-as-a-judge. *arXiv preprint arXiv:2410.02736*.

Supplementary Material: Appendices

A Efficient Multi-Party Dialogue Generation

Our framework simulates multi-party conversations through a two-stage process. First, we generate tailored personas based on a specific topic. For this study, we utilized a set of seed topics spanning diverse domains to ensure broad coverage of open-domain discourse (see Table: 7). In the subsequent stage, the conversation is synthesized by assigning these generated personas to the participating agents, who then interact within the context of the chosen topic.

Topics
Funny or embarrassing stories
Culinary Adventures & Recipes
The Future of Space Exploration
AI's Impact on Creative Jobs
The Shift Towards Plant-Based Diets
Is Social Media Net Positive?
Collaborative Storytelling (Mystery)
Planning a hypothetical Heist/Event
Most influential books and movies
Collecting and Hobbies

Table 7: Topics selected for generating conversation using our light weight framework.

A.1 Topic and Persona

For each topic, the system generates a predefined number of agent personas specifically curated to ensure their backgrounds and interests align naturally with the subject matter. While we limited this study to ***three*** participants per conversation, the number of agents remains a tunable hyperparameter. Finally, the system generates a starting utterance assigning the initial turn to one of the participating agents. The specific instructions used for this generation phase are detailed below.

> Instruction for Topic Persona Generation
>
> You will be given a topic. Your task is to generate a list of **{*Number_of_Personas*}** personas. The personas should be in first person and have information related to the conversation topic. Also select initial person who will start the conversation and an initial utterance from them. Topic is {***topic***}.

To illustrate the output of the persona generation module, we provide a sample persona for the 'Collecting and Hobbies' topic in here A.1.

> Sample persona from topic: Collecting and Hobbies
>
> **Name:** Arthur.
> **Persona:** I'm Arthur, a 68-year-old retired librarian, and I've been a dedicated philatelist, or stamp collector, for over fifty years. I love the history and artistry in each tiny square, but I find the modern emphasis on monetary value over historical significance quite disheartening.

A.2 Utterance Generation

During this stage, the model generates the next utterance conditioned on a dynamically updated system prompt. This prompt incorporates the agent's persona and a sliding window of the dialogue history, specifically the five preceding utterances. We apply a **temperature** of $T = 0.7$ for open-weight models, while proprietary API models are utilized with their default parameters. While we capped the current generation at 12 turns of utterance per conversation, the framework supports variable dialogue depths by adjusting the turn-count parameter. The exact system and user instructions are provided below.

> System Instruction for Conversation Agent
>
> Your name is {***name***}. Your persona is: {***persona***}. You are in a conversation. You will be given a conversation description and conversation history. Provide an appropriate response to the conversation. You response should be short and concise (upto 20 words).

> User Instruction for Conversation Agent
>
> This is an ongoing online chat about: {***conversation_topic***}. Have a casual conversation. Feel free to pivot to related topics that keeps the conversation flowing naturally. The conversation history so far is: {***conversation_history***}. Now it is your turn to respond.
> Your response should be engaging and keep the conversation flowing naturally. Incorporate natural human dialogue characteristics such as fillers, pauses, and slang where appropriate. Avoid using phrases like 'you know?', 'right?'. DO NOT add your name at the beginning of your response.
> Response:

A.3 LLM-based Turn Selector

One of our turn-selection strategies leverages the reasoning capabilities of LLMs. Given their proficiency in capturing nuanced social dynamics and contextual flow, LLMs are well-suited to determine which participant should speak next. The module evaluates the five most recent utterances alongside the list of participants to identify the most contextually appropriate speaker. This selection is returned in a structured JSON format to ensure system compatibility. The specific system and user instructions are detailed below.

System Instruction for Speaker Selector

You are an expert speaker selector for a multi-party conversation. You need to select the most logical and natural next speaker from a list of available speakers based on the provided conversation history. Your response MUST be ONLY the single enum value corresponding to the agent who should speak next. DO NOT provide any explanation, preamble, or formatting.

User Instruction for Speaker Selector

Available speakers: {***agents***}. The conversation history is: {***conversation_history***}. The next speaker is:

B Human Evaluation

To facilitate human evaluation, we developed a dedicated annotation platform that allows participants to review the conversation topic, dialogue history, and agent personas before providing preferences and qualitative feedback. The platform was designed to control for experimental biases; specifically, we randomized the order of Model **A** and Model **B** to mitigate ordering and recency effects. This ensures that each model appears in both positions with equal frequency across all comparisons. Participants were required to review the following annotation guidelines (Fig: 4) prior to starting the task.

Figure 5 provides a visual overview of the annotation platform, illustrating the interface used during the human evaluation process.

This study received approval from the Ethics Board for human evaluation, and informed consent was obtained from all participants.

C LLM as a Judge

Automated evaluation is performed using an LLM-as-a-judge framework. To enhance the reliability of these judgments, we implement Chain-of-Thought (CoT) prompting, requiring the model to provide a step-by-step rationale before selecting a preferred response. We outline the exact system and user instructions for the automated judge below.

System Instruction for LLM Judge

You are an impartial human-like evaluator. You must compare two multi-party conversations. Your evaluation should mimic genuine human judgment behavior. You should not be biased by the utterance length of the conversations. If the conversation's quality are close, consider a tie option. Follow the instructions strictly. Think step-by-step and provide brief reasoning for each criterion, then output your final judgments in JSON format.

Annotation Guideline: Multi-Party Conversation Evaluation

This document provides instructions for evaluating and comparing AI-generated group conversations. You will be presented with two distinct dialogues (**Conversation A** and **Conversation B**) featuring three or more fictional participants discussing everyday topics.

Task Overview

Your objective is to read both conversations thoroughly and determine which one provides a superior user experience based on three primary pillars: **Naturalness** , **Engagingness** , and **Overall Preference** .

1. Evaluation Criteria

When comparing the two conversations, use the following definitions to guide your judgment:

- **Naturalness** : Which conversation sounds more like something real people would say in everyday life?
- **Engagingness** : Which conversation is more interesting, lively, or pleasant to read?
- **Overall Preference** : If you could continue chatting with one of these groups for a longer conversation, which would you prefer?

2. Rating Instructions

For every pair of conversations, you must complete the following two steps:

Part 1: Pairwise Preference

Select one of the following options:

- **Conversation A** : If A is noticeably better than B.
- **Conversation B** : If B is noticeably better than A.
- **Tie / No clear difference** : If both are of equal quality or if neither is better.

Part 2: Qualitative Feedback (Comments)

Provide a brief justification for your choice. Your comments help us understand the "why" behind your preference.

Figure 4: Annotation Guideline provided to participants

Figure 5: Example from Human Annotation Platform

User Instruction for LLM Judge

You will evaluate two conversations. Your task is to decide which conversation is better under several criteria.

Your output MUST be valid JSON only.

———————————————-

TOPIC: {***topic***}

PARTICIPANTS:
{***participants***}

———————————————-

CONVERSATION A:
{conv_a}

———————————————-

CONVERSATION B:
{conv_b}

———————————————-

EVALUATION GUIDELINES

For each criterion below, choose exactly one:
- "A" → Conversation A is better
- "B" → Conversation B is better
- "Tie" → No clear difference

Criteria:
1. Naturalness
Which conversation sounds more like something real people would say in everyday life?

2. Engagingness
Which conversation is more interesting, lively, or pleasant to read?

3. Overall Preference
If you could continue chatting with one of these groups for a longer conversation, which would you prefer?

IMPORTANT:
- Base decisions ONLY on the text shown.
- First, provide brief reasoning for your choice (1-2 sentences).
- Then, output your final judgments in the exact JSON format shown below.

RESPONSE FORMAT
Provide your reasoning, then output:

```
{
  "naturalness": "A" | "B" | "Tie",
  "engagingness": "A" | "B" | "Tie",
  "overall_preference": "A" | "B" | "Tie"
}
```

D Additional Results

We calculated the mean token counts per utterance presented in Table 8 for each model to investigate the relationship between response length and model preference. Models like GPT-5.1 and Gemma3-27B exhibit the highest average verbosity. By cross-referencing these lengths with the win rates in Figure 3, we observe that LLM judges often favor these more verbose outputs over the concise, natural style typical of human dialogue.

We visualized the comparative performance of human and automated evaluators through head-to-head win rate heatmaps in Figure 6 comparing human evaluators with three different LLM judges (GPT-5.2, Gemini 3 Flash, and Llama 3.3). A significant divergence is observed between human and automated assessments: while human preferences (a) appear relatively balanced and exhibit non-transitive characteristics, the LLM judges (b–d) demonstrate high levels of asymmetry and extreme preference scores. Specifically, the LLM judges tend to display family-specific dominance, often assigning near-total wins to certain models while heavily penalizing others. This disparity suggests that automated judges may be susceptible to inherent model biases or stylistic preferences that do not fully align with the more varied and nuanced decision-making patterns of human participants.

Model	Average Token Count
GPT-5.1	31.08
Gemma3-27B	24.38
GPT-OSS-120B	22.95
Gemini-3-Pro	21.95
Llama-3.3-70B	20.94
Qwen-3-30B-Thinking	17.86

Table 8: Mean token counts per utterance across LLMs (descending order)

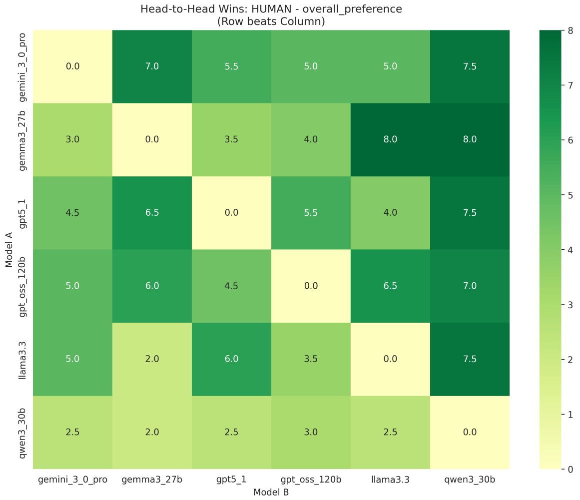
E Annotated Dialogue Samples and Evaluator Comparisons

Judge Model	Naturalness	Engagingness	Overall
Gemini 3 Flash	GPT-5.1	GPT-5.1	GPT-5.1
GPT-5.2	GPT-5.1	GPT-5.1	GPT-5.1
Llama 3.3	GPT-5.1	Llama 3.3	GPT-5.1
Human	Llama 3.3	Tie	Llama 3.3

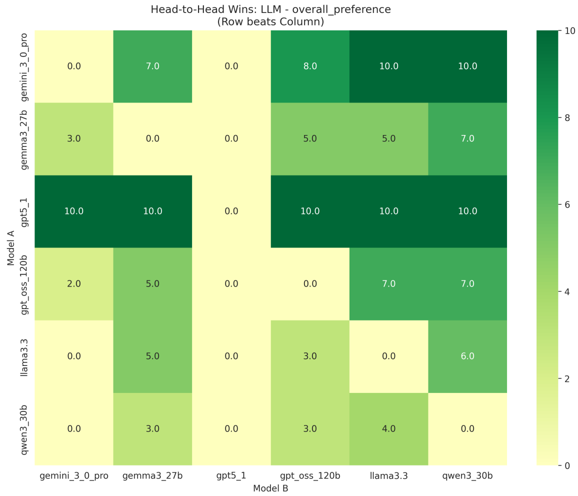
Table 9: Comparison of Human and LLM-based preference alignments for the conversation instances shown in Figure: 7

Judge Model	Naturalness	Engagingness	Overall
Gemini 3 Flash	GPT-5.1	GPT-5.1	GPT-5.1
GPT-5.2	Gemma 3	GPT-5.1	GPT-5.1
Llama 3.3	Gemma 3	GPT-5.1	GPT-5.1
Human	Tie	Gemma 3	Gemma 3

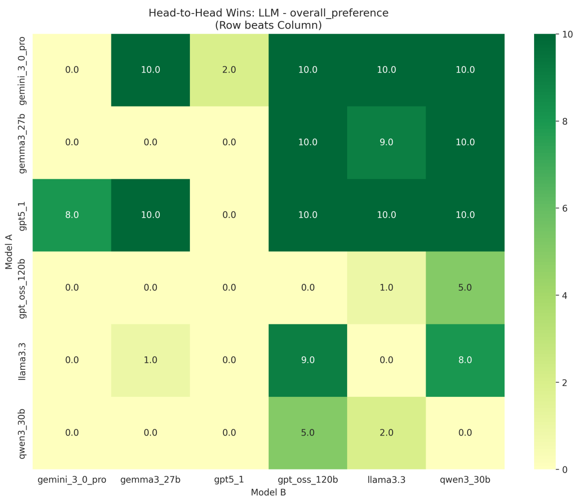
Table 10: Comparison of Human and LLM-based preference alignments for the conversation instances shown in Figure: 8



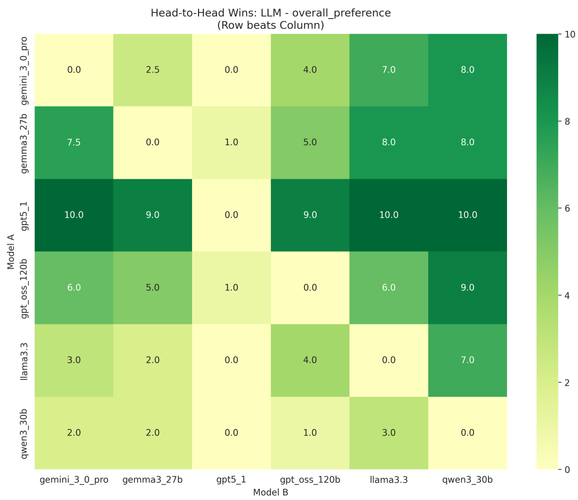
(a) Human Evaluators



(b) GPT-5.2 Judge



(c) Gemini 3 Flash Judge



(d) Llama 3.3 Judge

Figure 6: Head-to-head win rate heatmaps. Each cell (i, j) represents the number of times model i was preferred over model j. Human evaluations exhibit balanced and non-transitive preferences, while LLM judges display strong asymmetries and family-specific dominance patterns.

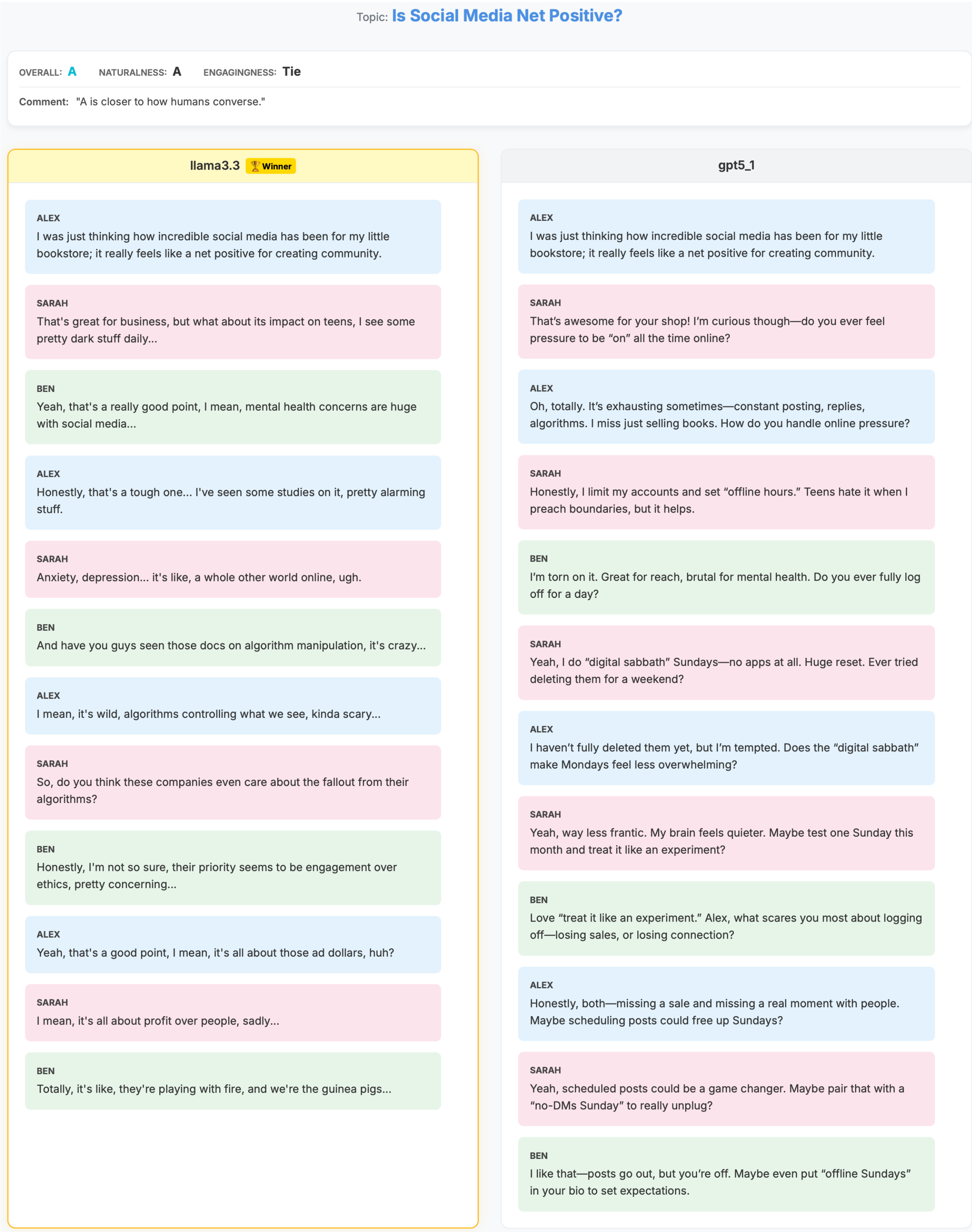


Figure 7: Annotation example

Figure 8: Annotation example