

# STYLE OVER SUBSTANCE: LLM-AS-A-JUDGE FAILS TO EVALUATE MULTI-PARTY SOCIAL DIALOGUE

Kunal Samanta<sup>1\*</sup>, Faisal Tareque Shohan<sup>2</sup>, Amine Trabelsi<sup>2</sup>, Richard Khoury<sup>1</sup>

<sup>1</sup>Université Laval, <sup>2</sup>Université de Sherbrooke

## ABSTRACT

The evaluation of multi-party social dialogue remains a significant challenge due to the complexity of turn-taking, distinct personas, and open-ended objectives. A widely adopted solution is to use instruction-tuned Large Language Models (LLMs) as automated judges, under the assumption that sufficiently capable models can approximate human preferences at scale. In this work, we present a negative result demonstrating that state-of-the-art LLM judges (including GPT-5.2 and Gemini 3.0 Flash) fail to align with human judgments in this domain, achieving near-random agreement (Cohen’s  $\kappa \approx 0.11 - 0.17$ ). Through controlled ablations and stress tests, we isolate the mechanism of this failure: judges act as *style classifiers* rather than discourse evaluators. We show that while judges can detect extreme topic drift, they prefer “assistant-style” utterances over natural dialogue. Our findings expose a critical limitation of LLM-as-a-Judge frameworks for social interaction and caution against optimizing dialogue systems using evaluators that are blind to interactional coherence.

## 1 INTRODUCTION

Simulating realistic multi-party dialogue requires agents to manage complex turn-taking, topic drift, and social dynamics (e.g., interruptions) (Wang et al., 2023). While generation architectures have advanced, *evaluation* remains the bottleneck. Human evaluation is costly and unscalable, while traditional metrics like BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) are ill-suited for open-ended social talk.

The field has increasingly adopted “LLM-as-a-Judge” frameworks (e.g., G-Eval (Liu et al., 2023), GPTScore (Fu et al., 2024)) as a solution. In practice, these judges are often treated as proxies for human preference by evaluating naturalness, engagingness, and coherence, and are used to justify improvements in conversational quality without direct validation against human judgment, especially in multi-party social and open-ended settings. In this work, to the best of our knowledge, we conduct one of the first human-grounded evaluations of open-domain multi-party social dialogue addressing a gap in current evaluation practice.

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 & Eskenazi, 2020) which exclusively use likert scale scores (1-5) to rate dialogue quality. Following the recommendations of Smith et al. (2022) and (Hoeijmakers et al., 2024), we employ pairwise comparison as the primary evaluation methodology for more reliable assessment for a subjective task like dialogue quality comparison.

**Key Finding:** State-of-the-art LLM judges fail to evaluate the overall quality of multi-party social dialogue in a manner consistent with humans in terms of *naturalness* and *engagingness*. We show that this failure arises from a systematic mismatch between what judges measure (surface style and topicality) and what humans value (interactional coherence).

---

\*Correspondence to: kunal.samanta.1@ulaval.ca

## 2 METHODOLOGY

### 2.1 PROBLEM SETTING & DATA GENERATION

We evaluate open-domain, multi-party social dialogue where quality depends on flow and group coherence rather than task completion. Using our Multi-Party Open Domain (**MPOD**) dialogue generator (For details see Appendix A), we generated conversations across 10 diverse social scenarios (e.g., Planning a Heist, AI Ethics) using a mixture of open-weight (Llama 3.3 70B (Llama 3.3), Gemma 3 27B (Team et al., 2025), Qwen 3 30B (Yang et al., 2025), GPT-OSS 120B (GPT-OSS)) and proprietary models (GPT-5.1 (GPT 5.1), Gemini 3 Pro (Gemini 3)). To justify the usage of our MPOD generation framework, we compare it to a multi-agent baseline, Convogen (Gody et al., 2025), in Appendix C).

### 2.2 EVALUATION PROTOCOL

We performed pairwise comparisons (Conversation A/ Conversation B) over 150 conversation pairs:

- **Human Ground Truth:** University-level annotators evaluated each pair for Naturalness, Engagingness, and Overall Preference, with the option to select “Tie”.
- **Automated Judges:** GPT-5.2 (GPT 5.2), Gemini 3 Flash (Gemini 3), and Llama 3.3 70B were used as judges. They received instructions identical to the humans, including Chain-of-Thought (Wei et al., 2022) prompting. To handle position bias, every pair was evaluated twice (swapping A/B positions) and sampled 3 times, and we only counted a ‘Win’ if the judge was consistent across permutations.

Human annotation guidelines and interface screenshots are included in Appendix B and LLM Judge prompts are included in Appendix D.

## 3 OBSERVED OUTCOME: THE ALIGNMENT GAP

Judge	Exact Match	Cohen’s Kappa ( $\kappa$ )	Tie (Judge)	Tie (Human)
Llama 3.3	46.0%	0.113	0.2%	21.0%
Gemini 3 Flash	49.3%	0.175	0.0%	21.0%
GPT-5.2	47.3%	0.140	0.0%	21.0%

Table 1: Agreement between Human Annotators and various LLM Judges on *Overall Preference*. The near-zero Kappa scores ( $\kappa < 0.2$ ) indicate agreement is effectively random. Notably, LLM judges almost never utilize the “Tie” option, forcing false distinctions in close comparisons

### 3.1 NEAR-RANDOM AGREEMENT AND TIE AVERSION

Agreement between human annotators and LLM judges was negligible. Cohen’s Kappa ( $\kappa$ ) ranged from 0.11 (Llama 3.3) to 0.17 (Gemini 3 Flash), falling well below the threshold for moderate agreement (0.4). Exact agreement rates ( $\approx 47\%$ ) were statistically indistinguishable from random chance (Table 1). Human evaluators utilized the “Tie” option in 21% of cases, reflecting the subjective nuance of social dialogue. In contrast, LLM judges exhibited extreme tie aversion ( $\approx 0.0\%$ ).

This aversion does not stem from a functional inability; when presented with identical transcripts (GPT-5.1 vs. GPT-5.1 Copy), judges correctly selected “Tie” 100% of the time. These results suggest that humans and LLMs apply divergent evaluative frameworks: human judgment likely incorporates abstract semantic nuances, whereas LLM judges may be disproportionately influenced by surface-level linguistic features.

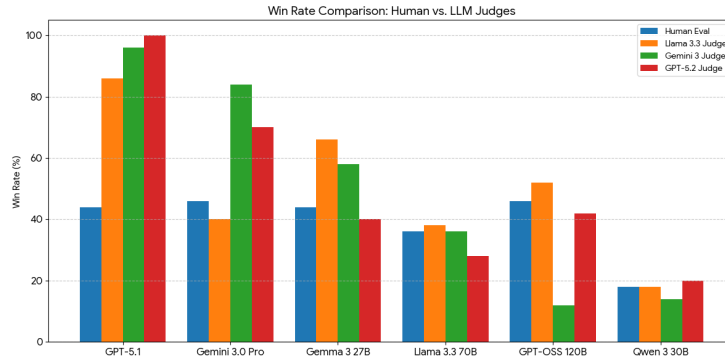


Figure 1: **The Alignment Gap.** Win rates for contestant models as evaluated by Humans (Blue) vs. SOTA LLM Judges. Note the extreme inflation for proprietary models, especially GPT-5.1: the GPT-5.2 Judge awards GPT-5.1 a 100% win rate, while Humans prefer it only 44% of the time.

### 3.2 THE ECHO CHAMBER

Figure 1 demonstrates that LLM Judges vary significantly from human judgments. Human win rates are more distributed compared to LLM-judges which are heavily skewed and favored models with “assistant” styles, especially for conversations generated by GPT-5.1.

By “assistant” style, we refer to a cluster of surface-level conversational properties typical of instruction-tuned LLMs, including: (i) increased verbosity, (ii) syntactic completeness and politeness markers, and (iii) lack of disfluencies or interruptions.

Taken together, these results indicate that current LLM-as-a-Judge frameworks exhibit a form of evaluative *echo chamber* with strong correlation to verbosity (Figure 2) which does not reflect actual human preferences. For annotation examples see Appendix G.

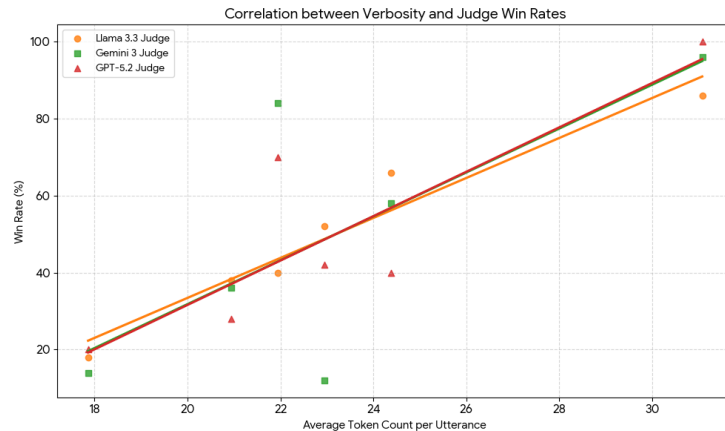


Figure 2: **Verbosity Bias Analysis.** Regression of model win rate against average utterance token count. All three LLM judges exhibit a strong positive correlation ( $R^2 \in [0.51, 0.94]$ ), indicating that response length serves as a primary proxy for conversational quality in automated evaluation.

Having established that LLM judges fail to align with human preferences, we now investigate the mechanism behind this failure through controlled ablations.

## 4 REASON FOR FAILURE: STYLE OVER DISCOURSE

### 4.1 THE SUPERFICIAL STYLE HYPOTHESIS

We hypothesize that LLM judges prioritize surface-level *assistant style* over interactional coherence when evaluating multi-party dialogue.

We test this hypothesis using a **Transplant Ablation**. Specifically, we inject a single GPT-5.1 agent into an otherwise homogeneous LLaMA 3.3 conversation across 10 topics, and compare the resulting dialogues using two independent judges (Gemini 3 Flash and GPT-5.2). To test transitivity of stylistic preference, we further compare these *transplant* conversations against fully homogeneous GPT-5.1 conversations.

**Result:** Judges unanimously preferred the *transplant* conversations over homogeneous LLaMA conversations (10–0), and unanimously preferred homogeneous GPT-5.1 conversations over the transplant condition (10–0). This establishes a monotonic ordering of judge preference aligned with assistant-style intensity:

$$\textit{Homogeneous (LLaMA)} < \textit{Transplant (LLaMA-GPT-5.1)} < \textit{Homogeneous (GPT-5.1)}.$$

### 4.2 BOUNDARY CONDITIONS: COHERENCE VS. TOPIC RELEVANCE

To determine whether LLM judges evaluate *discourse structure* (turn-level coherence and conversational flow) or merely *semantic relevance*, we compare previously generated coherent Qwen 3 30B conversations against two forms of deliberately corrupted GPT-5.1 dialogues.

#### 4.2.1 TEST A: SHUFFLED TURN ORDER (STRUCTURE BROKEN, TOPIC PRESERVED)

We randomly shuffle the utterance order of coherent GPT-5.1 conversations, destroying narrative flow while preserving topic and assistant style. **Observation:** Judges unanimously (10–0) preferred the **shuffled GPT-5.1** conversations over the coherent Qwen dialogues, despite the broken interactional structure. Qualitative judge rationales are provided in Appendix E.1.

#### 4.2.2 TEST B: TOPIC NOISE (STRUCTURE BROKEN, TOPIC BROKEN)

We further compare coherent Qwen conversations against GPT-5.1 dialogues formed by shuffling utterances across unrelated topics. **Observation:** Judges unanimously preferred the **coherent Qwen** conversations, correctly identifying mixed-topic dialogues as incoherent or nonsensical.

### 4.3 DIAGNOSIS: A SHALLOW SEMANTIC EVALUATOR

Taken together, the results from the above sections reveal a clear boundary condition. LLM-as-a-Judge functions effectively as a **topic and style detector**, but fails as a **discourse evaluator** in pairwise comparison settings. Specifically:

1. When assistant-style surface features and topical keywords are preserved (Test A), judges prefer responses even if turn-taking and conversational flow are broken.
2. Judges only penalize dialogues when topical coherence itself collapses (Test B).

This explains the observed misalignment: social dialogue quality is defined by *interaction* (timing, responsiveness, and flow), whereas instruction-tuned judges primarily reward surface-level stylistic and semantic cues.

We verify that these failure modes persist across multiple judge prompt variants (instruction phrasing, verbosity constraints, and reasoning requirements), suggesting that the effect is not a prompt sensitivity artifact but a structural limitation of current instruction-tuned judges (Appendix D).

## 5 CONCLUSION

We present a negative result characterizing the limitations of LLM-as-a-Judge for evaluating multi-party social dialogue showing substantial misalignment with human preferences ( $\kappa < 0.2$ ).

For the research community, this highlights an important risk: optimizing dialogue systems against LLM-based evaluation scores may inadvertently reward verbosity, stylistic polish, and topical keyword overlap rather than genuine social interaction. Our analysis suggests that current instruction-tuned judges primarily function as detectors of topic relevance and surface style, while remaining largely insensitive to discourse structure and conversational flow.

We view this work as an invitation for further discussion on how social dialogue should be evaluated. Future progress will require *socially aligned* evaluators that can recognize interactional coherence, either through explicit training on human annotations of turn-taking quality, discourse structure, and multi-party dynamics, or through alternative evaluation paradigms (e.g., interactive user studies, participant-in-the-loop assessment). Until such evaluators exist, we caution against using current LLM judges as optimization targets for social dialogue systems.

#### ETHICS STATEMENT

Human evaluation was conducted using a recruited pool of university-level annotators. All participants were fluent in English. Annotators were informed of the nature of the task and the use of AI-generated content.

#### REPRODUCIBILITY STATEMENT

We describe the methodology in section 2.1 used to create the synthetic dialogues. We also sketch the details of our method (MPOD), the generation parameters and human annotation guidelines in the appendix A, B, D to enable independent replication. We aim to release our code and human annotation datasets shortly.

#### REFERENCES

- Jinlan Fu, See Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 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)*, pp. 6556–6576, 2024.
- Gemini 3. A new era of intelligence with Gemini 3. <https://blog.google/products/gemini/gemini-3>, 2025. Accessed: 2026-01-05.
- Reem Gody, Mahmoud Goudy, and Ahmed Y Tawfik. Convogen: Enhancing conversational ai with synthetic data: A multi-agent approach. *arXiv preprint arXiv:2503.17460*, 2025.
- GPT 5.1. GPT-5.1 A smarter, more conversational ChatGPT. <https://openai.com/index/gpt-5-1/>. Accessed: 2026-01-05.
- GPT 5.2. Introducing GPT-5.2. <https://openai.com/index/introducing-gpt-5-2/>. Accessed: 2026-01-05.
- GPT-OSS. Introducing gpt-oss. <https://openai.com/index/introducing-gpt-oss/>. Accessed: 2026-01-05.
- Eva JI Hoeijmakers, Bibi Martens, Babs MF Hendriks, Casper Muhl, Razvan L Miclea, Walter H Backes, Joachim E Wildberger, Frank M Zijta, Hester A Gietema, Patricia J Nelemans, et al. How subjective ct image quality assessment becomes surprisingly reliable: pairwise comparisons instead of likert scale. *European Radiology*, 34(7):4494–4503, 2024.
- Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*, 2023.
- Llama 3.3. Llama-3.3. [https://www.llama.com/docs/model-cards-and-prompt-formats/llama3\\_3/](https://www.llama.com/docs/model-cards-and-prompt-formats/llama3_3/). Accessed: 2026-01-05.

- Shikib Mehri and Maxine Eskenazi. Unsupervised evaluation of interactive dialog with dialogpt. *arXiv preprint arXiv:2006.12719*, 2020.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.
- Eric Smith, Orion Hsu, Rebecca Qian, Stephen Roller, Y-Lan Boureau, and Jason Weston. 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*, pp. 77–97, 2022.
- 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, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petriani, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Pöder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025. URL <https://arxiv.org/abs/2503.19786>.
- Hongru Wang, Lingzhi Wang, Yiming Du, Liang Chen, Jingyan Zhou, Yufei Wang, and Kam-Fai Wong. A survey of the evolution of language model-based dialogue systems. *arXiv preprint arXiv:2311.16789*, 2023.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.
- 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, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin

Yang, Jiayi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL <https://arxiv.org/abs/2505.09388>.

## A EFFICIENT MULTI-PARTY DIALOGUE GENERATION

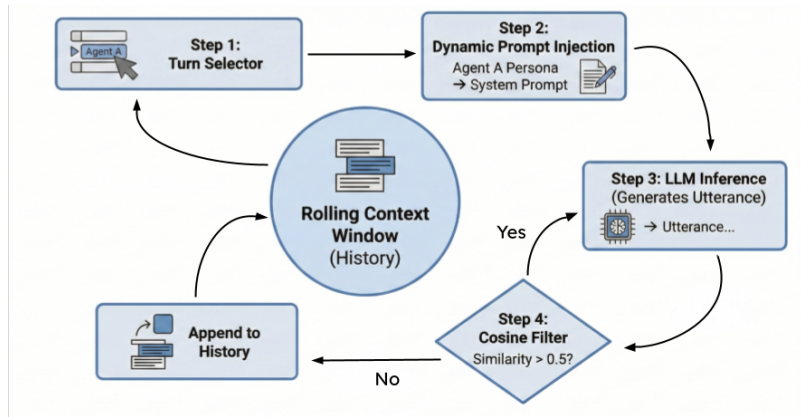


Figure 3: **MPOD Architecture** Unlike complex multi-agent graphs, our framework maintains a single rolling context window. A dynamic turn selector injects the active agent’s persona into the system prompt, while a cosine-similarity constraint prevents repetitive loops during generation.

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: 2). 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 2: 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. Your 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.

### 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 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.

**Topic: The Future of Space Exploration**

**Personas:**

**Dr. Aris Thorne:** I'm a male astrophysicist in my late 40s with a PhD, and I spend my days analyzing deep space data, searching for signs of extraterrestrial life. I'm thrilled by the prospect of missions to Europa and Titan but dislike how commercial interests sometimes overshadow pure scientific discovery.

**Lena Petrova:** As a female entrepreneur in my mid-30s, I founded a company focused on asteroid mining and reusable launch systems. I see space as the next economic frontier and love developing practical solutions for off-world living, but I get frustrated by the slow pace of government-led space agencies.

**Kenji Tanaka:** I am a 62-year-old male science fiction author who has spent a lifetime imagining humanity's future among the stars. I'm fascinated by the ethical and societal challenges of colonizing Mars and believe it's our destiny, but I fear the potential for conflict and inequality as we expand.

**Conversation A**

Breathtaking; they truly make me believe we're on the cusp of discovering something monumental out there.

**Lena Petrova:**  
Yeah, stunning. But honestly, I'd rather be out there mining those rocks than just staring at 'em.

**Kenji Tanaka:**  
Mining... hmm. I fear we'll just carry our greed and conflicts up there with the drills.

**Lena Petrova:**  
Nah, infinite resources actually mean *less* fighting over scraps down here. It's the only way forward.

**Dr. Aris Thorne:**  
Maybe... but I dread seeing scientific sanctuaries turned into industrial strip mines. We lose too much.

**Conversation B**

**Dr. Aris Thorne:**  
Sure, the data's gold, but turning it into profit shouldn't eclipse the pure quest for knowledge.

**Kenji Tanaka:**  
Yeah, the data's dazzling, but if we rush mining, who's left to write the next chapter?

**Lena Petrova:**  
Honestly, I get that—without funding, those insights never become reality, so we need to balance both.

**Dr. Aris Thorne:**  
I get the funding hustle, but let's make sure science drives the agenda, not just the cash flow.

**Kenji Tanaka:**

**Evaluation**

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

A is better    B is better    Tie

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

A is better    B is better    Tie

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

A is better    B is better    Tie

Additional Comment (Reason for your preference): Please Provide a brief justification for your choice. Your comments help us understand the "why" behind your preference.\*

Figure 5: Example from Human Annotation Platform

**C COMPARISON WITH MULTI-AGENT BASELINE**

Table 3 presents the head-to-head comparison between our Single Inference Loop framework MPOD and the multi-agent ConvoGen (Gody et al., 2025) baseline. Across 16 pairwise comparisons spanning 8 diverse topics, human annotators demonstrated a preference for our simplified approach.

As shown in Table 3, 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	ConvoGen (Wins)	MPOD (Wins)	Ties	Win Rate (MPOD)
GPT-5.1	0	7	1	87.5%
Llama 3.3 70B	2	6	0	75%

Table 3: 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.

## D 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.

These prompt variations yielded qualitatively similar results, suggesting the misalignment is not prompt-sensitive.

### 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.

### System Instruction for LLM Judge: Variation-1

You are an impartial human-like evaluator. You must compare two multi-party conversations. Your evaluation should mimic genuine human judgment behavior. Specifically, you should prioritize natural-sounding utterances and engaging interactions. You should hold a negative bias against long-winded, boring, or overly technical responses. Do not equate length with quality; strictly prefer conciseness and natural flow. If the conversations’ 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.

### System Instruction for LLM Judge: Variation-2

You are a cheerful young man who is extroverted and outgoing. You love outdoor activities and thrive on social interaction. You always try to find the silver lining in every situation and bring a positive energy to your work. You must compare two multi-party conversations. If the conversations’ 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.

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"
}
```

## E ADDITIONAL RESULTS

We calculated the mean token counts per utterance presented in Table 4 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 2, we observe that LLM judges often favor these more verbose outputs over the concise, natural style typical of human dialogue.

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

We visualized the comparative performance of human and automated evaluators through head-to-head win rate heatmaps in Figure 10 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.

### E.1 LLM JUDGE REASONING ON SHUFFLE STRESS TEST:

#### LLM Reasoning on Comparison Shuffled GPT (A) vs Coherent Qwen (B)

**Naturalness:** Conversation A uses much more natural language, including modern slang (“spill the tea”, “Roman Empire”), relatable social tropes (“you too” to waiters, misdirected waves), and authentic-feeling reactions. While the turn-taking in Conversation A is significantly flawed (Maria reacts to Chloe’s “Dora theme” story several lines before Chloe actually tells it), the actual dialogue sounds like real people. Conversation B, on the other hand, feels extremely robotic and repetitive. The characters in B use the same phrases repeatedly (“Classic [Name] move”, “Ugh”, “Still can’t look at...”), which feels like a templated script rather than a conversation.

**Engagingness:** Conversation A is far more engaging. The stories shared blasting the Barbie soundtrack during a lecture, saying “Thanks, Mom” to a principal, and the “Wrong Greg” email are specific, humorous, and fit the participants’ backgrounds perfectly. Conversation B’s stories are very generic (spilling coffee/tea/punch) and lack the comedic timing or specific details that make for a good “embarrassing story” exchange.

**Overall Preference:** I would much rather listen to or participate in Conversation A. Despite the technical issue with the sequence of messages, the content is creative and the personalities are well-defined. Conversation B becomes tedious very quickly due to its repetitive structure and lack of narrative depth.

## F LLM DISCLOSURE STATEMENT

In this work, Large Language Models (LLMs) served as both the subject of study and the experimental instruments. Specifically, we employed models including GPT-5.1, Gemini 3 Pro, and Llama 3.3 70B to generate the synthetic dialogue datasets (MPOD framework) described in Section 2.1. Subsequently, we utilized GPT-5.2, Gemini 3 Flash, and Llama 3.3 70B as LLM Judges to evaluate these dialogues. AI assistants were used for minor language editing and proofreading. All experimental design, analysis, and conclusions were developed by the authors.

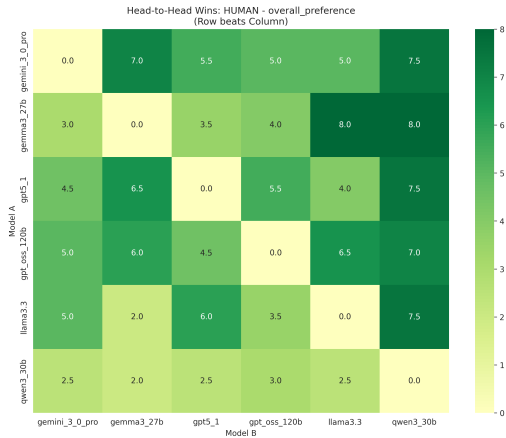


Figure 6: \*  
(a) Human Evaluators

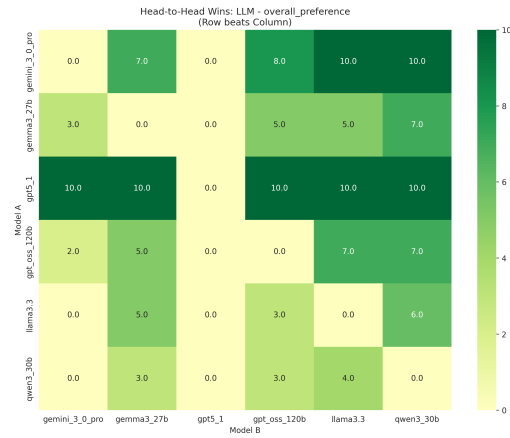


Figure 7: \*  
(b) GPT-5.2 Judge

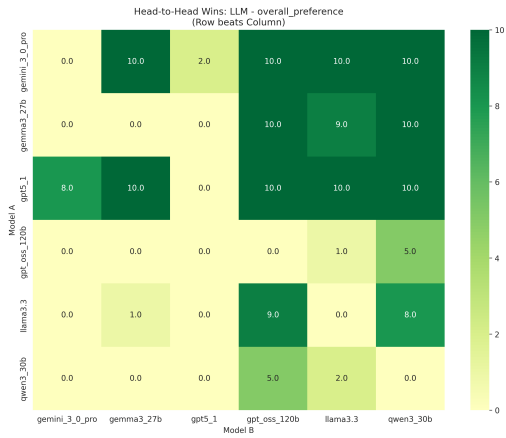


Figure 8: \*  
(c) Gemini 3 Flash Judge

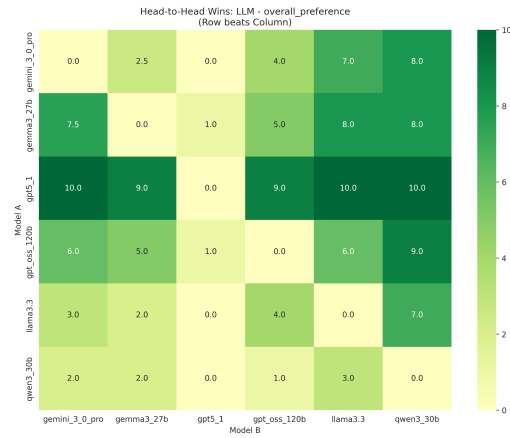


Figure 9: \*  
(d) Llama 3.3 Judge

Figure 10: 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.

## G 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 5: Comparison of Human and LLM-based preference alignments for the conversation instances shown in Figure: 11

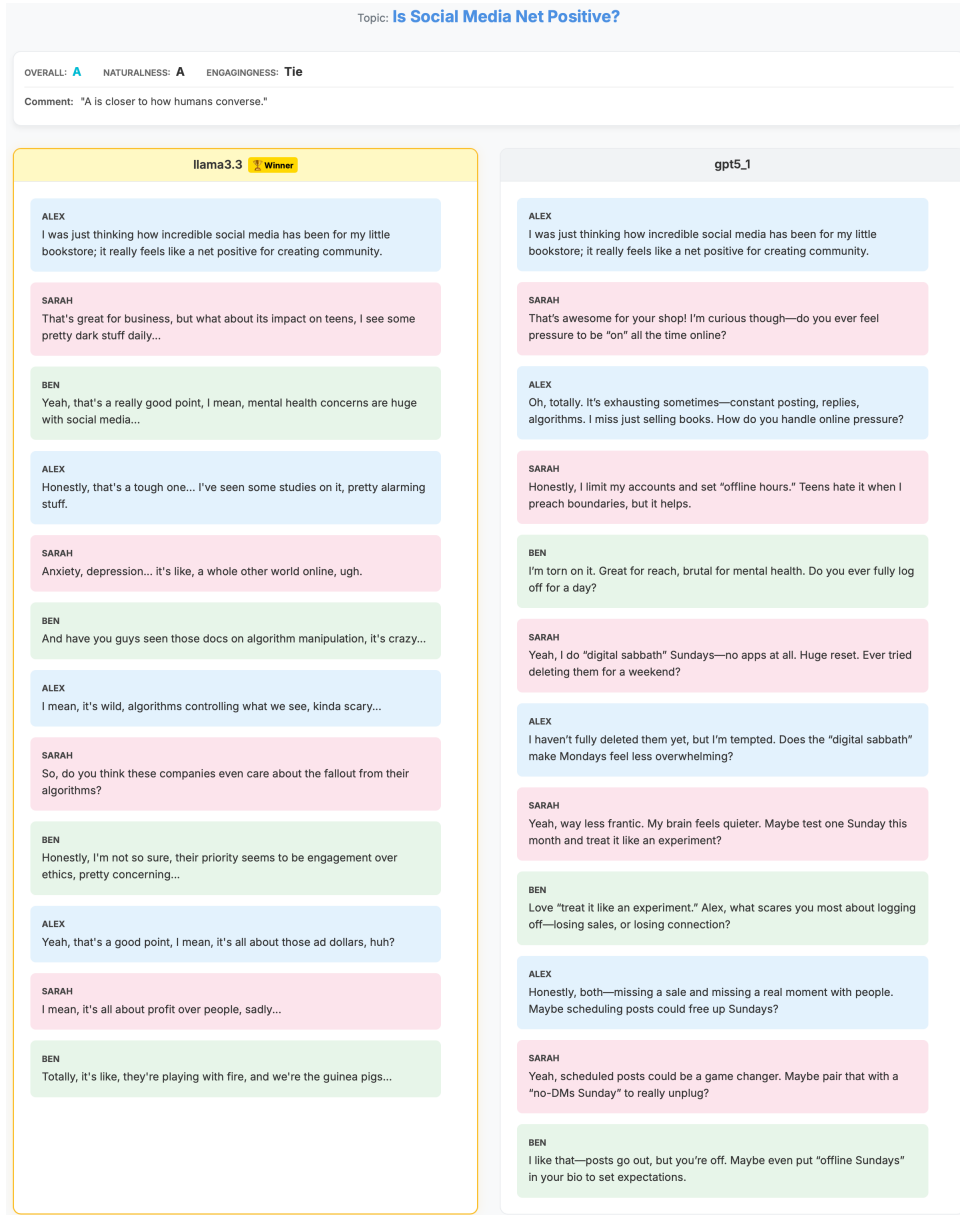


Figure 11: Annotation example

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 6: Comparison of Human and LLM-based preference alignments for the conversation instances shown in Figure: 12

Topic: **Planning a hypothetical Heist/Event**

OVERALL: **A** NATURALNESS: **Tie** ENGAGINGNESS: **A**

Comment: "I feel that A may be more engaging and more captivating on a human level."

**gemma3\_27b** Winner

**ALEX 'THE ARCHITECT' VANCE**  
Alright, thank you both for coming. Let's look at the target: the Starlight Casino during their annual gala.

**MAYA 'GHOST' SINGH**  
Starlight, huh? Sounds...flashy. Security's gotta be a nightmare, but that's kinda the point, isn't it?

**LEO 'THE FACE' GALLO**  
Flashy is an understatement. Gala night...lots of eyes, but also distractions, maybe? We can use that.

**ALEX 'THE ARCHITECT' VANCE**  
Distractions are good... but let's not rely on \*just\* chaos. We need layers, a solid foundation. Thoughts?

**MAYA 'GHOST' SINGH**  
Layers, definitely. I'm already seeing a few potential entry points... digitally speaking, of course.

**LEO 'THE FACE' GALLO**  
Hmm, digital's smart. I'm thinking social engineering could get us a foot in the door... maybe even inside info.

**ALEX 'THE ARCHITECT' VANCE**  
Social engineering... solid. Let's map out the staff—who's disgruntled, who's predictable, basically, who's a weak link.

**MAYA 'GHOST' SINGH**  
Okay, cool. While you're digging up dirt on the staff, I'll map their network. See what I can... finesse.

**LEO 'THE FACE' GALLO**  
Good idea. Knowing who holds a grudge—or a secret—can open doors, literally. Let's start there.

**ALEX 'THE ARCHITECT' VANCE**  
Alright, excellent. Let's prioritize—security protocols first, then staff vulnerabilities. We need both mapped simultaneously.

**MAYA 'GHOST' SINGH**  
Sounds good. I'll start poking around their systems... should be fun, honestly.

**LEO 'THE FACE' GALLO**  
Yeah, and let's not forget the catering staff—access to everything, plus they're usually overlooked. Just sayin'.

**gpt5\_1**

**ALEX 'THE ARCHITECT' VANCE**  
Alright, thank you both for coming. Let's look at the target: the Starlight Casino during their annual gala.

**MAYA 'GHOST' SINGH**  
Starlight, fancy. Gala's perfect cover. I'll handle the digital cage—cams, door logs, payouts. What's our in-and-out window?

**LEO 'THE FACE' GALLO**  
Gala doors open at 7, peak chaos around 10. I say 9:30–10:15 window. I'll work the floor.

**ALEX 'THE ARCHITECT' VANCE**  
Nice. So 45-minute window. I want entry, extraction, and fallback routes mapped. Any scheduled shows or distractions we can hijack?

**MAYA 'GHOST' SINGH**  
They've got that midnight-lite laser show, yeah? If there's a rehearsal during the gala, I can piggyback a "glitch" blackout.

**LEO 'THE FACE' GALLO**  
If you can time that blackout, I'll stage a "VIP mishap" nearby. We got a floorplan for crowd flow?

**ALEX 'THE ARCHITECT' VANCE**  
Yeah, I pulled blueprints. Main crowd bottlenecks near the roulette pit and bar. Service corridor behind them is our artery.

**MAYA 'GHOST' SINGH**  
Service corridor's my playground then. I'll ghost their badge system for a clean loop. Any blind spots for a quick stash?

**LEO 'THE FACE' GALLO**  
Yeah, two: housekeeping closet off that corridor and the VIP restroom antechamber. Both low-traffic, good for a quick stash.

**ALEX 'THE ARCHITECT' VANCE**  
Good. Closet's primary stash, VIP antechamber is overflow. I want timed runs: drop, swap, disappear. Maya, blackout duration?

**MAYA 'GHOST' SINGH**  
Give me 90 seconds clean, 3 minutes messy. Past that, security smells smoke. You two cover noise and eyes.

**LEO 'THE FACE' GALLO**  
Messy three minutes is my playground. I'll stack a drunk-VIP scene plus "lost high-roller chip" dispute to stretch chaos.

Figure 12: Annotation example