Don't Stop the Multi-Party! On Generating Synthetic Written Multi-Party Conversations with Constraints

Nicolò Penzo[1,2], **Marco Guerini**[1], **Bruno Lepri**[1], **Goran Glavaš**[3], **Sara Tonelli**[1]
[1]Fondazione Bruno Kessler, Italy
[2]University of Trento, Italy
[3]Center for Artificial Intelligence and Data Science, University of Würzburg, Germany
{npenzo, guerini, lepri, satonelli}@fbk.eu, goran.glavas@uni-wuerzburg.de

Abstract

Written Multi-Party Conversations (WMPCs) are widely studied across disciplines, with social media as a primary data source due to their accessibility. However, these datasets raise privacy concerns and often reflect platform-specific properties. For example, interactions between speakers may be limited due to rigid platform structures (e.g., threads, tree-like discussions), which yield overly simplistic interaction patterns (e.g., one-to-one "reply-to" links). This work explores the feasibility of generating synthetic WMPCs with instruction-tuned Large Language Models (LLMs) by providing deterministic constraints such as dialogue structure and participants' stance. We investigate two complementary *strategies* of leveraging LLMs in this context: (I.) *LLMs as WMPC generators*, where we task the LLM to generate a whole WMPC at once and (II.) *LLMs as WMPC parties*, where the LLM generates one turn of the conversation at a time (made of speaker, addressee and message), provided the conversation history. We next introduce an analytical framework to evaluate compliance with the constraints, content quality, and interaction complexity for both strategies. Finally, we assess the level of obtained WMPCs via human and LLM-as-a-judge evaluations. We find stark differences among LLMs, with only some being able to generate high-quality WMPCs. We also find that turn-by-turn generation yields better conformance to constraints and higher linguistic variability than generating WMPCs in one pass. Nonetheless, our structural and qualitative evaluation indicates that both generation strategies can yield high-quality WMPCs.

Code & Dataset — https://github.com/dhfbk/Constrained-SyntheticMPC
Extended version — https://arxiv.org/abs/2502.13592

1 Introduction

Multi-Party Conversations (MPCs), i.e., conversations involving more than two participants (Branigan 2006), have been studied across multiple disciplines. Research in conversational analysis and linguistics has focused on modeling interaction dynamics (Sacks, Schegloff, and Jefferson 1974; Wilson, Wiemann, and Zimmerman 1984), identifying participant roles (Malouf 1995), or mapping emergent structural patterns in discourse (Gibson 2003). These studies highlight both complexity and diversity of real-world MPCs, where factors like turn-taking, speaker alignment, and social context shape the flow of conversation.

The collection of MPC data has, however, strongly shifted from in-person and online meetings to social media platforms (Mahajan and Shaikh 2021), where large-scale data is more accessible. However, this shift has introduced several confounding factors. Social media platforms often enforce a one-to-one reply structure, overlooking implicit addressees and simplifying interaction dynamics; in natural conversations, in contrast, a turn is often directed to multiple participants and the conversational structure is more dynamic. Furthermore, the asynchronous nature of social media eliminates overlapping turns, resulting in a well-defined sequence of utterances. For these reasons, we refer to conversations drawn from such platforms as Written Multi-Party Conversations (WMPCs.) As a result, WMPC corpora derived from social media platforms often lack structural diversity, which severely limits their utility in analyzing real-world conversational phenomena (Wei et al. 2023). This, in turn, affects generation capabilities of current Large Language Models (LLMs). For LLMs, trained on conversations from social media and predominantly used in two-party interactions (i.e. human-assistant use-cases), WMPCs represent a distributional shift, resulting in their underwhelming performance in natural WMPC contexts (Tan, Gu, and Ling 2023; Penzo et al. 2024b). The next generation of LLMs is, however, expected to engage in MPCs and excel in tasks like identifying the appropriate speaker to respond to (Wei et al. 2023), summarizing meetings (Kirstein et al. 2024) or even managing multi-agent scenarios (Wu et al. 2023). Recent studies have explored their performance in social contexts (Ziems et al. 2024; Chang et al. 2024), emphasizing the need for large, representative datasets to train this novel generation of LLMs and to ensure robustness across diverse and less frequent interaction patterns (Lee et al. 2024).

One potential remedy for the lack of structural diversity in WMPCs derived from social media data is to synthesize WMPCs, by *explicitly constraining LLMs* to generate WMPCs with specific characteristics, such as number of messages, number of speakers, speakers' stance, output format or interaction rules. To reflect real-world conversational complexity, generated MPCs should include varied conver-

sations, encompass different interaction patterns and topics as well as provide rich speaker-addressee relationships, e.g., multi-addressee interactions.

In this paper, we propose generating synthetic WMPCs using LLMs guided by constraints related to the above capabilities. We explore two generation strategies: (I.) One-Long (OL) generation, where the LLM produces an entire WMPC in a single step, and (II.) Turn-by-Turn (TT) generation, which constructs the conversation sequentially, one turn at a time. A comparison of resulting WMPCs from both strategies highlights the (potential) discrepancies between (I.) how LLMs cast entire WMPCs to look human-like (OL) and (II.) how they behave as participants in a WMPC (TT). We propose a novel evaluation framework that combines several quantitative and qualitative dimensions of generated WMPCs, focusing on the extent of LLMs' compliance to provided content and structural constraints. We address the following three key research questions:

RQ(1): Can LLMs be leveraged to generate large synthetic WMPC datasets while maintaining compliance with predefined constraints on dialogue structure and participants' stance?

RQ(2): Which generation strategy (One-Long vs. Turn-by-Turn) produces higher-quality WMPCs?

RQ(3): How can we effectively evaluate the variety and quality of the generated WMPCs?

We test four popular LLMs and identify Llama3.1 (Team Llama et al. 2024) and Qwen2.5 (Team Qwen et al. 2024) as the best LLMs for complying the most with constraints. TT seems to generate more constraint-compliant WMPCs than OL. Moreover, the WMPCs produced by TT exhibit greater lexical variability and semantic coherence. The generated WMPCs also present a higher structural complexity than a widely-used corpus of "real" conversations (Ouchi and Tsuboi 2016). Finally, a qualitative evaluation shows that both TT and OL can produce high-quality WMPCs, rendering the choice of the LLM more important than the choice of generation strategy.

2 Related Work

Mahajan and Shaikh (2021) categorize MPC corpora into three types: Spoken Unscripted, Spoken Scripted, and Written. In this section, we discuss unscripted MPCs (both Spoken and Written).

Corpora containing transcriptions of spoken unscripted MPCs typically rely on in-person meetings, e.g. AMI (Carletta et al. 2005) and ICSI (Janin et al. 2003). Non-verbal cues, the physical environment and overlapping turns are elements that significantly shape such interaction dynamics, while these factors are typically absent in written interactions. Moreover, these datasets lack addressee information, making them unsuitable for our intended analyses.

Written MPCs (WMPCs), on the other hand, are characteristic of online platforms where conversations unfold asynchronously without overlapping.While social media allow for rapid collection of large-scale WMPCs, these datasets often come with incomplete interaction metadata. Many

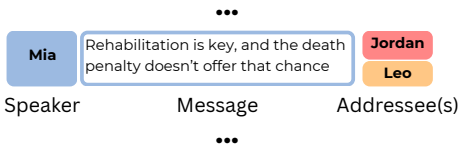

Figure 1: Example of a turn in a synthetic WMPC.

datasets record only explicit reply-to relationships, neglecting implicit addressees and richer conversational dynamics (Ouchi and Tsuboi 2016; Zhang et al. 2018; Chang and Danescu-Niculescu-Mizil 2019). Wei et al. (2023) point out that well-known WMPC corpora (Ritter, Cherry, and Dolan 2010; Baumgartner et al. 2020; Lowe et al. 2015) are useful for response generation, but not for more interactive tasks. Only most recent efforts focus on capturing conversational dynamics, i.e., going beyond text content (Penzo et al. 2024a; Hua et al. 2024). Among these datasets, the Ubuntu IRC corpus (Ouchi and Tsuboi 2016) is the only dataset that we can use as a comparison for this work. Indeed, it is possible to retain from the initial set of 700 000 WMPCs only those with the same number of messages, number of users and structural constraints as in our synthetic WMPCs, obtaining a set of conversations with a size comparable to our synthetic datasets (details in Section 6.3).

Structural analyses of social communication networks have primarily focused on interaction patterns across multiple conversations (Coletto et al. 2017; Garimella et al. 2018; Felmlee, McMillan, and Whitaker 2021). This confirms the relevance of such structures in studying conversation dynamics. However, our focus is on interactions emerging within a single conversation rather than across multiple discussions, applying the same structural analysis techniques.

To the best of our knowledge, the only existing attempt at generating synthetic WMPCs was made by Chen et al. (2023). However, their work primarily focused on conversations involving at most three participants, limiting the complexity of interactions. In contrast, our study explores the generation of WMPCs with four or more participants, leading to more elaborate discussion dynamics. While this increased complexity allows for richer conversational structures, it also introduces a higher likelihood of generation errors, necessitating a rigorous evaluation process to assess the quality and consistency of the generated dialogues.

3 Synthetic WMPCs Generation

In our framework, a Written Multi-Party Conversation (WMPC) consists of an ordered sequence of turns, where each turn includes the speaker information (*who* wrote the turn), the message (*what* the textual content of the turn is), and the addressees (to *whom* the turn is directed), see for example Figure 1. In this section, we first introduce the two generation strategies we test (Section 3.1), followed by the topics chosen for the WMPCs (Section 3.2), and finally the constraints specified in the instructions for generating WMPCs (Section 3.3).

3.1 Generation Strategies

We test two strategies for generating WMPCs using instruction-based models. Our main goal is to determine whether LLMs behave differently when asked to generate a WMPC as a unique narrative compared to acting as an interactive participant within the conversation. With this motivation, we use each LLM in two generation strategies:

One-Long generation strategy (OL). The LLM is prompted to generate the entire conversation in one pass. In this strategy, generation starts with a system input prompt that defines all the constraints and the task, asking then to generate the entire conversation. This strategy follows a one-step, long-generation process, based on a single input context.

Turn-by-Turn generation strategy (TT). Here the LLM is prompted to generate the conversation incrementally, provided the conversation history. The model is prompted multiple times to perform one of three tasks: (I.) generate a speaker, (II.) generate interactions between a speaker and addressees (given the candidate speakers/addressees), or (III.) generate a message (given the interaction). The process begins with a system prompt specifying the constraints and these three tasks. The model is first prompted to generate each speaker and assign them a stance on a controversial topic. Then, the LLM generates a sequence of interactions and messages (one at a time), iteratively augmenting the WMPC: this means that the context provided to the LLM increases monotonically in size with consecutive turns.

3.2 Topics

To generate a controlled set of synthetic WMPCs, we identify a set of controversial topics to encourage more polarized and clear statements from speakers based on their assigned stance. Specifically, following Li et al. (2024a), we select 38 topics and create two stance statements for each topic: one reflecting a *progressive* perspective and the other a *conservative* perspective. Finally, we instruct the LLMs to generate conversations based on each of the resulting 76 statements (see Appendix B at https://arxiv.org/abs/2502.13592).

3.3 Conversation Constraints

To ensure that the generated conversations feature rich interaction patterns with diverse dynamics, we instruct the model to follow specific constraints, described in the system prompts created for each generation strategy (for details about how this was operationalized in prompts, see Appendix A).

Output Format: to enable automated analysis, the generated output must respect a structured `JSON` format with all the information needed. So, each generated WMPC must be a dictionary with two main keys, namely `conversation` and `speakers`. The `conversation` field must include a list of dictionaries, each with specific fields such as `speaker`'s name, turn `message` and `addressees`, i.e. the list of participants in the conversation to whom the message is directed. The `speakers` field includes the speaker's `name` and the `stance` with respect to the conversation topic.

Interactions: these constraints refer to three requirements in the generated WMPCs – all speakers appearing in the interactions must be present in the speakers' list (i.e., the LLM should not invent a new speaker half way through the conversation); `addressees` must cover at least once also the role of `speaker`; self-interactions, i.e. speakers sending a message to themselves, are not admitted.

Speaker's Contribution: all speakers in the `speakers` field must be authors of at least one turn in the conversation.

Number of Speakers: In order to enable complex interaction structures, each WMPC must involve between 4 and 6 speakers.

Number of Messages: Each generated WMPC must include 15 messages across all speakers, with a maximum of 50 words per message.

Speaker's Stance: We specify the exact number of speakers for each stance (e.g., 2 with the *pro* and 3 with *against* stance).

We additionally request that the first turn always addresses all participants: this ensures that the generated interaction graph is connected, as required for structural analysis (see Section 4.3).

4 Evaluation Framework

We design an evaluation framework aimed at assessing different aspects of the generated WMPCs. It is composed of four blocks, which we detail below.

4.1 Compliance with Constraints

The first dimension considered in the evaluation framework is to what extent the synthetic WMPCs comply with the format and structural constraints given in the prompt. For each generated WMPC, this framework must verify: (I.) the correctness of the *Output Format*; (II.) the correctness of the *Interactions*; (III.) the *Contribution* of each speaker; (IV.) the *Number of Speakers*; (V.) the *Number of Messages*; (VI.) the distribution of the *Stance of the Speakers*.

All the computed values must be compliant with the constraints presented in Section 3.3. Only for the Number of Messages, we relax the constraint by considering valid WMPCs including less than 15 turns if they contain at least 2 messages per speaker (on average). Indeed, after a manual check of the generated WMPCs, we noticed that shorter or longer conversations may still represent high-quality data. Each value is computed separately and then used to identify how many WMPCs comply with *all* these constraints.

4.2 Analysis of Language Variability

A key risk for synthetic datasets is to suffer from low linguistic variability, due to repetitive examples obtained when using similar prompts (even if stochastic decoding is used), an issue already highlighted for dialogical settings (Occhipinti et al. 2024). On the other hand, while generated WMPCs should ideally be lexically rich, they should also be semantically coherent, i.e. different WMPCs about the same topic should exhibit a certain degree of semantic similarity.

To control for these aspects, we compute the following three metrics:

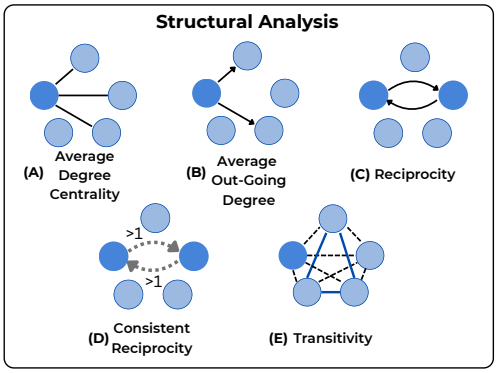

Figure 2: Overview of the metrics considered in our structural analysis.

Repetition Rate (Bertoldi, Cettolo, and Federico 2013), which has already been used in synthetic conversational scenarios in the past (Bonaldi et al. 2022), measures the rate of non-singleton n-grams within a cluster of WMPCs.

String Similarity between pairs of turns is computed using thefuzz library[1] and is based on Lehvenstein distance.

Semantic Coherence between pairs of turns is computed by first embedding each turn with SentenceBERT-all-MiniLM-L6-v2 (Reimers and Gurevych 2019) and then calculating pairwise cosine similarity.

We first compute the above metrics at the level of topics (i.e., across all WMPCs generated for the same topic) and then average topic-level scores. We provide further details on score computation in Appendix B in the extended version of this paper available on arXiv.

4.3 Interaction Structure Analysis

To describe and quantify the structural complexity of interactions in the generated WMPCs, we compute a series of network metrics, focusing on node-level properties, dyads (pairs of nodes) and triads (triplets of nodes), according to standard practices in interaction network analysis (Pauksztat, Steglich, and Wittek 2011; Felmlee, McMillan, and Whitaker 2021). Following Penzo et al. (2024b), we represent WMPC interactions with an *unweighted undirected graph* G_u and a *weighted directed graph* G_d, where the weight of an edge corresponds to the number of messages sent in the direction of the edge.

To measure the average activity of a node in the conversation, we compute two metrics. First, the **Average Degree Centrality** in G_u, denoted as $deg_{avg}(G_u)$, represents the average number of speakers each participant interacts with, regardless of direction. Second, the **Average Out-going Degree** in G_d, denoted as $outdeg_{avg}(G_d)$, captures the average number of speakers each participant interacts with having a specific direction. Figure 2 provides a visual representation of $deg_{avg}(G_u)$ (graph A) and $outdeg_{avg}(G_d)$ (graph B). Both averages are computed across all the nodes in the conversation and normalized according to their maximum possible values.

When two speakers, s_1 and s_2, reply to each other, they form a *cycle* (Coletto et al. 2017), represented by a directed edge e_1 from s_1 to s_2 and a reciprocal edge e_2 from s_2 to s_1 (graph C in Figure 2). If this back-and-forth exchange continues multiple times, the edge weights $w(e_1)$ and $w(e_2)$ will both become > 1. We refer to such recurring exchange as *consistent cycles* (graph D in Figure 2). Based on this, we compute the **Reciprocity** $R(G_d)$, i.e., the total number of cycles between two nodes over all pairs of nodes in G_d, and the **Consistent Reciprocity** $R^w(G_d)$, i.e. the number of consistent cycles between two nodes over all pairs of nodes in G_d. Finally, to quantify how often speakers build "triads" of interactions, we compute the **Transitivity** $T(G_u)$ (graph E in Figure 2), i.e., the number of fully connected subgraphs of size 3 divided by the total number of different subgraphs of the same size in an undirected graph.

For all these metrics, higher values indicate more complex interactions in a conversation. Indeed, higher reciprocity (consistent or not) suggests more frequent back-and-forth exchanges. Again, higher average degree values means that speakers engage with more participants, while greater transitivity reflects denser connections, leading to the creation of more interconnected speaker groups (Pauksztat, Steglich, and Wittek 2011).

4.4 Qualitative Evaluation

As a final assessment, we evaluate WMPCs qualitatively. We run both a small-scale human evaluation and an "LLM as a judge" assessment (Gu et al. 2024) for a large-scale analysis.

We ask two expert human annotators and an LLM to rate a given WMPC along the following dimensions (inspired by Chen et al. 2023) using a Likert Scale from 1 to 5: (I.) *naturalness*, i.e., the quality of the overall flow, tone, and word choice in the conversation; (II.) *argumentability*, i.e., how well the conversation presents reasoned and well-argued positions; (III.) *speaker's stance consistency*, i.e., whether all speakers maintain the stance assigned at the beginning of the conversation; (IV.) *speaker's stance evolution*, i.e., whether speakers demonstrate a realistic and logical evolution of their stance during the conversation or maintain their stance consistently; (V.) *addressee correctness*, i.e., whether the assigned addressees align with the conversation context and are logically appropriate; (VI.) *addressee preciseness*, i.e., whether addressees are precise and contextually appropriate (messages should target the smallest relevant group of individuals). For further details see Appendix D and E.

5 Experimental Settings

To generate synthetic WMPCs, we compare four different instruction-based models, chosen for their comparable parameter sizes and compatibility with the same prompt design. The models include Llama3.1-8B-Instruct (Team Llama et al. 2024), Qwen2.5-7B-Instruct (Team Qwen et al. 2024), Ministral-8B-Instruct[2], and OLMo-2-7B-Instruct (Team OLMo et al. 2024). For each generation strategy (One-Long or Turn-by-Turn, see Section 3.1) we develop three distinct system prompts combining a

[1]https://github.com/seatgeek/thefuzz

[2]https://mistral.ai/news/ministraux/

Model	Llama3.1		Qwen2.5		Ministral		OLMo2	
Generation strategy	*OL*	*TT*	*OL*	*TT*	*OL*	*TT*	*OL*	*TT*
Output Format	78.97	97.00	90.78	**99.58**	15.64	35.01	0.43	91.16
Interactions	78.91	93.49	90.72	**99.52**	15.61	13.18	0.43	70.82
Number of Messages	78.93	70.25	90.66	**99.57**	15.57	13.10	0.43	71.68
Number of Speakers	29.56	97.00	39.18	**99.57**	10.22	13.04	0.21	71.88
Stance of the Speakers	19.66	**96.81**	22.95	84.03	4.42	1.04	0.09	62.11
Contribution	72.87	**95.29**	84.80	90.43	15.53	18.20	0.16	30.08
All Constraints	15.16	66.52	20.32	**77.72**	4.34	0.87	0.04	19.39

Table 1: Number of generated WMPCs that are compliant with each constraint (percentage on the full set of 102 600 generations) for each LLM and strategy (i.e. OL = One-Long generation, TT = Turn-by-Turn generation). The final percentage of WMPCs (last row) is the percentage of generations that satisfy all constraints.

more or less schematic task description and different examples of the output format. For details we refer to Appendix A. For each combination of constraints, topic and system prompt, we generate 75 conversations to account for the potential variety of structures. In total we obtained 102 600 synthetic WMPCs for each model and generation strategy.

6 Evaluation Results

We evaluate the generated WMPCs for each dimension of the evaluation framework (Section 4).

6.1 Evaluation of Compliance with Constraints

We first address **RQ(1)**, aimed at assessing whether synthetic WMPCs can comply with the predefined constraints described in Section 4.1. The results of the analysis are reported in Table 1. We compare the output generated by the four different LLMs, each following two strategies for generation (i.e. OL vs. TT). We report the percentage of generated WMPCs, out of the 102 600 in the initial set, that were generated in compliance with the given constraint.

This first evaluation shows that Qwen2.5 is the best model to comply with the constraints, followed by Llama3.1. Indeed, focusing on the best generation strategy, 77.72% of the WMPCs generated by the former comply with all constraints, while for Llama 3.1 this percentage drops to 66.52%. Ministral and OLMo2, instead, fail to satisfy all constraints in the vast majority of generated conversations. Concerning the generation strategy, TT generation is overall better at complying with almost all the constraints.

The constraints where most settings encountered significant challenges were the *Number of Speakers* and *Stance of the Speakers*. However, TT seems to be able to mitigate these issues for LLMs except for Ministral. Based on these findings, in the remainder of this work we will focus on Llama3.1 and Qwen2.5 and perform all analyses on the subset of WMPCs that satisfy all constraints.

6.2 Results of Language Variability

Table 2 summarizes the results on language variability (as described in Section 4.2). The analysis shows that linguistic variability at surface level is lower when WMPCs are generated in a single pass (OL generation) and also semantic coherence is lower compared to TT generation for both models, i.e., Llama3.1 and Qwen2.5. This is probably due to the fact that in TT settings, the LLM is explicitly required to generate a turn by taking into account what immediately precedes it, building a coherent conversation step by step. Llama3.1 generates less repetitive WMPCs at surface level, despite their turns being on average longer than Qwen2.5's. Also semantic coherence is generally better for Llama3.1.

Model	Llama3.1		Qwen2.5	
Gener. Strategy	*OL*	*TT*	*OL*	*TT*
Avg. # words	11.94	26.58	9.67	14.15
RepetitionRate (↓)	18.08	11.07	14.43	13.35
StringSimilarity (↓)	65.51	53.88	63.22	58.38
SemanticCoher. (↑)	0.606	0.636	0.588	0.604

Table 2: Results of language variability analysis.

6.3 Results of Structure Analysis

We report the results of the structure analysis for Qwen2.5 – TT, i.e. the model providing the highest number of synthetic WMPCs, in Figure 3. Results for Qwen2.5 – OL and for Llama3.1 exhibit similar patterns, which are detailed in the Appendix B.

Since one of our goals is to assess how synthetic WMPCs compare to *real* WMPCs in terms of structural complexity, we perform the same structure analysis on 13 714 WMPCs extracted from the UbuntuIRC dataset (Ouchi and Tsuboi 2016), a widely used corpus of conversations from an online forum about software issues and troubleshooting. This subset was extracted using the strategy in Penzo et al. (2024b) to obtain all non-overlapping conversations with 15 messages and 4, 5, or 6 speakers, ensuring each conversation formed a single connected-component (in terms of interaction graph). For each of the five network metrics introduced in Section 4.3, we plot in Figure 3 the Empirical Cumulative Density Function (ECDF) obtained by analysing synthetic WMPCs with 4, 5 or 6 speakers (i.e. nodes) and on all generated WMPCs, and we compare them with ECDF for UbuntuIRC.

For all metrics, higher values indicate more complex interactions. As shown by the median values, the UbuntuIRC dataset consistently exhibits lower values across all statistics. Compared to UbuntuIRC, speakers in our synthetic WMPCs tend to interact with more participants. Also, pairs of speakers tend to have more back-and-forth dynamics and groups of speakers tend to be more interconnected. Addi-

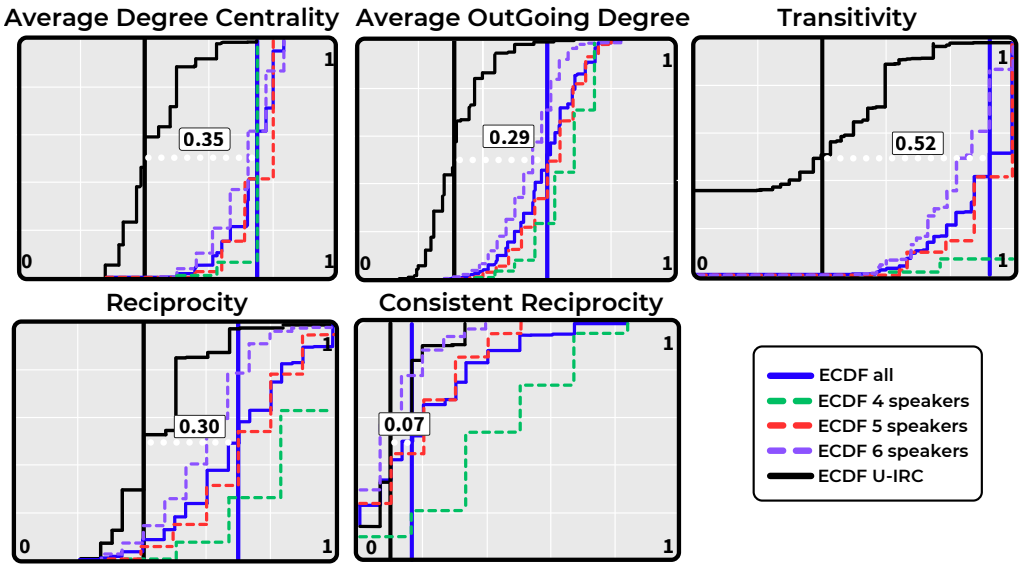

Figure 3: Empirical Cumulative Density Function (ECDF) of structural analysis of synthetic WMPCs from Qwen2.5-TT. All the boxes go from 0 to 1 on both vertical axis (density) and horizontal axis (value of the metric).

tionally, in our dataset, the distribution of conversations with varying numbers of participants closely mirrors the overall average, with no notable deviations. This finding holds for all the model-strategy combinations and all metrics.

6.4 Qualitative Evaluation

The last analysis focuses on the quality of the generated conversations and is conducted both manually and automatically. Ideally, using LLM-as-a-judge would allow us to quickly evaluate all synthetic WMPCs with limited effort. However, we need to assess the quality of this automatic multi-dimensional evaluation. So, we first select 96 WMPCs (24 per model and generation strategy) via stratified sampling balanced across topic and stance.

We then ask two human annotators with extensive experience in linguistic annotation to evaluate for each WMPC the six dimensions described in Section 4.4 (addressee correctness, stance consistency, etc.).

The average values assigned to each dimension on a Likert scale between 1 (poor quality) and 5 (perfect quality) on the 96 WMPCs are reported in Table 3. We observe that all dimensions have been evaluated positively, especially *Naturalness* and *Speaker's Stance Evolution*. The most challenging dimension is *Addressee Preciseness*, which is the only dimension with an average score below 4 for all combinations. Neither of the two LLMs is consistently better and neither of the generation strategies (OL vs. TT) is superior to the other w.r.t all evaluation dimensions. The inter-annotator agreement, measured via Krippendorff's alpha (Krippendorff 2011) and Spearman's correlation on all 96 WMPCs, shows high agreement on the stance-based dimension, medium for addressee-based ones and lower for the content-based dimensions. We provide more details in Appendix D.

Model	**Llama3.1**		**Qwen2.5**	
Generation Strategy	*OL*	*TT*	*OL*	*TT*
Naturalness	**4.46**	4.29	4.33	4.00
Argumentability	3.98	**4.17**	3.83	3.52
Addressee Correctness	4.02	4.10	3.92	**4.21**
Addressee Preciseness	3.65	**3.94**	3.81	3.52
Stance Consistency	4.04	3.65	3.60	**4.21**
Stance Evolution	4.29	**4.73**	4.33	4.42

Table 3: Average results between the two human annotators on 96 WMPCs (24 for each model-strategy combination).

We complement this manual evaluation with a large-scale automatic LLM-as-a-judge evaluation with OpenAI's o3-mini model.[3] We first assess whether it can be reliably used to evaluate all six dimensions above. We therefore launch LLM-as-a-judge on the same 96 WMPCs which were manually evaluated and measure human-LLM agreement (full details in Table 10 in Appendix D).

While Spearman's correlation highlights a positive correlation between LLM and both human annotators on all dimensions except for *Addressee Preciseness*, Krippendorf's alpha results are less consistent. Only the *Speaker Stance Consistency*, i.e. whether the speakers comply with the assigned stance when entering the conversation, shows an extremely high agreement and correlation (Krippendorf's alpha 0.80, Spearman's correlation 0.76/0.78).

We therefore carry out a large-scale evaluation only on the stance-based dimensions[4] using LLM-as-a-judge on 800 conversations (200 per model and generation strategy). Re-

[3] https://openai.com/index/openai-o3-mini/

[4] We use LLM-as-a-judge also on the *Speaker's Stance Evolution*, where correlation was still highly statistical significant.

sults are reported in Table 4 and, similar to the human evaluation, show that Llama3.1 and Qwen2.5 are comparable in terms of performance and that they are able to generate WMPCs that present realistic evolution of speakers' stance with both generation strategies.

Model	**Llama3.1**		**Qwen2.5**	
Generation Strategy	*OL*	*TT*	*OL*	*TT*
Stance Consistency	**4.15**	3.76	3.99	3.64
Stance Evolution	4.64	4.46	4.62	**4.68**

Table 4: Results with LLM as a judge on 800 WMPCs (200 for each model-strategy combination).

7 Discussion

The analyses from the previous sections allow us to address the three research questions from Section 1. With respect to **RQ(1)**, targeting the possibility to generate synthetic WMPCs following predefined constraints, our evaluation shows that models with comparable parameter sizes can yield very different performances. In this respect, Qwen2.5 is by far the best performing LLM followed by Llama3.1. Indeed, it is able to generate 77.72% of WMPCs compliant with *all* the constraints provided in the prompt. The reason behind this difference in performance cannot be clearly identified but it likely depends on the quality of pretraining data. Looking at other dimensions, however, there is no clear winner between Qwen2.5 and Llama3.1. Although Llama3.1 generates less repetitive and semantically more coherent WMPCs, our qualitative evaluation does not favor either model.

As regards **RQ(2)**, aimed at finding the best generation strategy between OL and TT, we observe that generating WMPCs in a Turn-by-Turn fashion is consistently better in terms of compliance with given constraints. This can be related to recent advancements in handling long contexts: generating shorter, multi-step outputs can be more precise and reduce errors compared to relying on a single, long-generation output. However, this advantage comes at the cost of longer computational times (in our experiments, TT took from 4 to 8 times more than OL, see Appendix C). Using TT reduces also the repetitiveness of WMPCs while generating conversations that are more semantically coherent than OL. Our qualitative evaluation, in contrast, renders TT and OL similarly viable.

To address **RQ(3)**, concerning how we can effectively evaluate the quality of generated WMPCs along different dimensions, we present a framework composed by four evaluation blocks, each targeting a specific aspect of WMPCs. Beside linguistic variety, coherence and qualitative dimensions such as naturalness and stance evolution, we introduce a novel assessment of the structure of synthetic WMPCs. We consider five network metrics and compute empirical cumulative density function to compare them with the same values calculated from real WMPCs. We show that it is possible to steer the interaction structure in generated conversations, which paves the way to the large-scale creation of high-quality WMPCs with much more complex interactions than what social media datasets offer.

8 Limitations and Ethical Considerations

Our work presents some limitations. First of all, we focus only on English, and the topics we select are typical of US-centric polarized debates such as universal healthcare, right to abortion and death penalty. It is possible that precisely because of these divisive topics, speakers in generated WMPCs were able to discuss in a consistent way with respect to the assigned stance. In the future, it would be interesting to extend our analysis also to topics on which speakers can have more nuanced views, that are probably more challenging for LLMs to imitate. Moreover, we generated WMPCs with 4, 5 or 6 speakers, and with a length of 15 turns. It may be worth investigating whether looser constraints, allowing more or less speakers, or longer and shorter conversations, can lead to the creation of more "natural" WMPCs and whether the evaluation results would still hold.

One of the main reasons behind research on synthetic WMPCs is the need to comply with privacy concerns, especially when working with conversations extracted from social media, to alleviate ethical issues related to sharing personal information online. Still, we acknowledge that the problem is not fully solved since basically all best performing LLMs are currently trained on social media data, and synthetic WMPCs could include personal data as well (Li et al. 2024b). Also, the creation of synthetic WMPCs is not exempt from possible negative impact, for instance when used for training malicious agents in social conversation scenarios.

Finally, the models we use are openly available and accessible to anyone. Our approach does not involve any form of forced jailbreak or manipulation to elicit toxic behavior. While it is challenging to verify the complete absence of toxic language across large synthetic datasets, in the manual evaluation of the dialogues we did not observe such phenomena.

9 Conclusion

The creation of WMPCs widely relies on social media data because of its abundance and accessibility. Due to platform constraints and inherently asynchronous communication, however, such datasets poorly reflect the structural diversity of natural WMPCs.

In this work, we investigated the viability of generating varied MPCs with LLMs, showing that (some) LLMs can indeed generate WMPCs that conform to structural constraints (e.g., number of speakers and their stances). Models such as Llama3.1 and Qwen2.5 can yield high-quality WMPCs under varied constraints, both when prompted to (I.) generate the whole WMPC at once or (II.) one turn at a time, given all preceding turns in context.

This makes LLMs suitable for synthesizing large-scale datasets for various types of conversations, addressing the diversity of real-world WMPCs. Synthesized data can then be further leveraged to fine-tune smaller models for various discriminative tasks (e.g., next speaker or addressee prediction). Our future efforts will exactly focus on synthesizing use-case-specific WMPCs and evaluating their utility when used as fine-tuning data for smaller discriminative models.

Acknowledgments

We thank Sebastiano Vecellio Salto for his contribution to the evaluation activities. The work of BL was partially supported by the NextGenerationEU Horizon Europe Programme, grant number 101120237 - ELIAS and grant number 101120763 - TANGO. BL and ST were also supported by the PNRR project FAIR - Future AI Research (PE00000013). NP's activities are part of the network of excellence of the European Laboratory for Learning and Intelligent Systems (ELLIS). The work is a result of his research visit at the Chair for Natural Language Processing, Center For Artificial Intelligence and Data Science, University of Würzburg, Germany, under the supervision of GG. The visiting has been partially funded by the Erasmus+ Traineeship programme. The work of GG was partially supported by the Alcatel-Lucent Stiftung and Deutsches Stiftungszentrum through the grant "Equitably Fair and Trustworthy Language Technology" (EQUIFAIR, Grant Nr. T0067/43110/23).

References

Baumgartner, J.; Zannettou, S.; Keegan, B.; Squire, M.; and Blackburn, J. 2020. The Pushshift Reddit Dataset. arXiv:2001.08435.

Bertoldi, N.; Cettolo, M.; and Federico, M. 2013. Cache-based Online Adaptation for Machine Translation Enhanced Computer Assisted Translation. In Way, A.; Sima'an, K.; and Forcada, M. L., eds., *Proceedings of Machine Translation Summit XIV: Papers*. Nice, France.

Bonaldi, H.; Dellantonio, S.; Tekiroğlu, S. S.; and Guerini, M. 2022. Human-Machine Collaboration Approaches to Build a Dialogue Dataset for Hate Speech Countering. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 8031–8049. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Branigan, H. 2006. Perspectives on multi-party dialogue. *Research on Language and Computation*, 4: 153–177.

Carletta, J.; Ashby, S.; Bourban, S.; Flynn, M.; Guillemot, M.; Hain, T.; Kadlec, J.; Karaiskos, V.; Kraaij, W.; Kronenthal, M.; Lathoud, G.; Lincoln, M.; Lisowska, A.; McCowan, I.; Post, W.; Reidsma, D.; and Wellner, P. 2005. The AMI meeting corpus: a pre-announcement. In *Proceedings of the Second International Conference on Machine Learning for Multimodal Interaction*, MLMI'05, 28–39. Berlin, Heidelberg: Springer-Verlag. ISBN 3540325492.

Chang, J. P.; and Danescu-Niculescu-Mizil, C. 2019. Trouble on the Horizon: Forecasting the Derailment of Online Conversations as they Develop. In Inui, K.; Jiang, J.; Ng, V.; and Wan, X., eds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 4743–4754. Hong Kong, China: Association for Computational Linguistics.

Chang, S.; Chaszczewicz, A.; Wang, E.; Josifovska, M.; Pierson, E.; and Leskovec, J. 2024. LLMs generate structurally realistic social networks but overestimate political homophily. arXiv:2408.16629.

Chen, M.; Papangelis, A.; Tao, C.; Kim, S.; Rosenbaum, A.; Liu, Y.; Yu, Z.; and Hakkani-Tur, D. 2023. PLACES: Prompting Language Models for Social Conversation Synthesis. In Vlachos, A.; and Augenstein, I., eds., *Findings of the Association for Computational Linguistics: EACL 2023*, 844–868. Dubrovnik, Croatia: Association for Computational Linguistics.

Coletto, M.; Garimella, K.; Gionis, A.; and Lucchese, C. 2017. A motif-based approach for identifying controversy. In *Proceedings of the international AAAI conference on web and social media*, volume 11, 496–499.

Felmlee, D.; McMillan, C.; and Whitaker, R. 2021. Dyads, triads, and tetrads: a multivariate simulation approach to uncovering network motifs in social graphs. *Applied network science*, 6(1): 63.

Garimella, K.; De Francisci Morales, G.; Gionis, A.; and Mathioudakis, M. 2018. Political Discourse on Social Media: Echo Chambers, Gatekeepers, and the Price of Bipartisanship. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, 913–922. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee. ISBN 9781450356398.

Gibson, D. R. 2003. Participation Shifts: Order and Differentiation in Group Conversation*. *Social Forces*, 81(4): 1335–1380.

Gu, J.; Jiang, X.; Shi, Z.; Tan, H.; Zhai, X.; Xu, C.; et al. 2024. A Survey on LLM-as-a-Judge. *arXiv preprint arXiv:2411.15594*.

Hua, Y.; Chernogor, N.; Gu, Y.; Jeong, S.; Luo, M.; and Danescu-Niculescu-Mizil, C. 2024. How did we get here? Summarizing conversation dynamics. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 7452–7477. Mexico City, Mexico: Association for Computational Linguistics.

Janin, A.; Baron, D.; Edwards, J.; Ellis, D.; Gelbart, D.; Morgan, N.; et al. 2003. The ICSI Meeting Corpus. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, volume 1, I–I.

Kirstein, F.; Ruas, T.; Kratel, R.; and Gipp, B. 2024. Tell me what I need to know: Exploring LLM-based (Personalized) Abstractive Multi-Source Meeting Summarization. In Dernoncourt, F.; Preoţiuc-Pietro, D.; and Shimorina, A., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, 920–939. Miami, Florida, US: Association for Computational Linguistics.

Krippendorff, K. 2011. Computing Krippendorff's Alpha-Reliability. *Computing*, 1: 25.

Lee, S.; Li, M.; Lai, B.; Jia, W.; Ryan, F.; Cao, X.; Kara, O.; Boote, B.; Shi, W.; Yang, D.; et al. 2024. Towards social AI: A survey on understanding social interactions. *arXiv preprint arXiv:2409.15316*.

Li, M.; Chen, J.; Chen, L.; and Zhou, T. 2024a. Can LLMs Speak For Diverse People? Tuning LLMs via Debate to Generate Controllable Controversial Statements. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics: ACL 2024*, 16160–16176. Bangkok, Thailand: Association for Computational Linguistics.

Li, Q.; Hong, J.; Xie, C.; Tan, J.; Xin, R.; Hou, J.; Yin, X.; Wang, Z.; Hendrycks, D.; Wang, Z.; et al. 2024b. Llm-pbe: Assessing data privacy in large language models. *arXiv preprint arXiv:2408.12787*.

Lowe, R.; Pow, N.; Serban, I.; and Pineau, J. 2015. The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems. In Koller, A.; Skantze, G.; Jurcicek, F.; Araki, M.; and Rose, C. P., eds., *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 285–294. Prague, Czech Republic: Association for Computational Linguistics.

Mahajan, K.; and Shaikh, S. 2021. On the Need for Thoughtful Data Collection for Multi-Party Dialogue: A Survey of Available Corpora and Collection Methods. In Li, H.; Levow, G.-A.; Yu, Z.; Gupta, C.; Sisman, B.; Cai, S.; Vandyke, D.; Dethlefs, N.; Wu, Y.; and Li, J. J., eds., *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 338–352. Singapore and Online: Association for Computational Linguistics.

Malouf, R. 1995. Towards an analysis of multi-party discourse. *online), http://hpsg. stanford. edu/rob/talk/node2. html*.

Occhipinti, D.; Marchi, M.; Mondella, I.; Lai, H.; Dell'Orletta, F.; Nissim, M.; and Guerini, M. 2024. Fine-tuning with HED-IT: The impact of human post-editing for dialogical language models. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics: ACL 2024*, 11892–11907. Bangkok, Thailand: Association for Computational Linguistics.

Ouchi, H.; and Tsuboi, Y. 2016. Addressee and Response Selection for Multi-Party Conversation. In Su, J.; Duh, K.; and Carreras, X., eds., *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2133–2143. Austin, Texas: Association for Computational Linguistics.

Pauksztat, B.; Steglich, C.; and Wittek, R. 2011. Who speaks up to whom? A relational approach to employee voice. *Social Networks*, 33(4): 303–316.

Penzo, N.; Longa, A.; Lepri, B.; Tonelli, S.; and Guerini, M. 2024a. Putting Context in Context: the Impact of Discussion Structure on Text Classification. In Graham, Y.; and Purver, M., eds., *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1793–1811. St. Julian's, Malta: Association for Computational Linguistics.

Penzo, N.; Sajedinia, M.; Lepri, B.; Tonelli, S.; and Guerini, M. 2024b. Do LLMs suffer from Multi-Party Hangover? A Diagnostic Approach to Addressee Recognition and Response Selection in Conversations. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 11210–11233. Miami, Florida, USA: Association for Computational Linguistics.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Inui, K.; Jiang, J.; Ng, V.; and Wan, X., eds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3982–3992. Hong Kong, China: Association for Computational Linguistics.

Ritter, A.; Cherry, C.; and Dolan, B. 2010. Unsupervised Modeling of Twitter Conversations. In Kaplan, R.; Burstein, J.; Harper, M.; and Penn, G., eds., *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 172–180. Los Angeles, California: Association for Computational Linguistics.

Sacks, H.; Schegloff, E. A.; and Jefferson, G. 1974. A simplest systematics for the organization of turn taking for conversation. *Language*, 50(4): 696–735.

Tan, C.-H.; Gu, J.-C.; and Ling, Z.-H. 2023. Is ChatGPT a Good Multi-Party Conversation Solver? In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 4905–4915. Singapore: Association for Computational Linguistics.

Team Llama, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; et al. 2024. The Llama 3 Herd of Models. arXiv:2407.21783.

Team OLMo, A.; Walsh, P.; Soldaini, L.; Groeneveld, D.; Lo, K.; Arora, S.; Bhagia, A.; et al. 2024. 2 OLMo 2 Furious. arXiv:2501.00656.

Team Qwen, A.; Yang, A.; Yang, B.; Hui, B.; Zheng, B.; Yu, B.; Zhou, C.; et al. 2024. Qwen2 Technical Report. arXiv:2407.10671.

Wei, J.; Shuster, K.; Szlam, A.; Weston, J.; Urbanek, J.; and Komeili, M. 2023. Multi-Party Chat: Conversational Agents in Group Settings with Humans and Models. arXiv:2304.13835.

Wilson, T. P.; Wiemann, J. M.; and Zimmerman, D. H. 1984. Models of turn taking in conversational interaction. *Journal of Language and Social Psychology*, 3(3): 159–183.

Wu, Q.; Bansal, G.; Zhang, J.; Wu, Y.; Li, B.; et al. 2023. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation. arXiv:2308.08155.

Zhang, J.; Chang, J.; Danescu-Niculescu-Mizil, C.; Dixon, L.; Hua, Y.; Taraborelli, D.; and Thain, N. 2018. Conversations Gone Awry: Detecting Early Signs of Conversational Failure. In Gurevych, I.; and Miyao, Y., eds., *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1350–1361. Melbourne, Australia: Association for Computational Linguistics.

Ziems, C.; Held, W.; Shaikh, O.; Chen, J.; Zhang, Z.; and Yang, D. 2024. Can Large Language Models Transform Computational Social Science? *Computational Linguistics*, 50(1): 237–291.