Estimating Relationships between Participants in Multi-Party Chat Corpus

Akane Fukushige[1], Koji Inoue[1], Keiko Ochi[1],
Tatsuya Kawahara[1], Sanae Yamashita[2], Ryuichiro Higashinaka[2]

[1]Graduate School of Informatics, Kyoto University,
[2]Graduate School of Informatics, Nagoya University
Correspondence: fukushige.akane.52z@st.kyoto-u.ac.jp

Abstract

While most existing dialogue studies focus on dyadic (one-on-one) interactions, research on multi-party dialogues has gained increasing importance. One key challenge in multi-party dialogues is identifying and interpreting the relationships between participants. This study focuses on multi-party chat corpus and aims to estimate participant pairs with specific relationships, such as family and acquaintances. We evaluated the performance of large language models (LLMs) in estimating these relationships, comparing them with a logistic regression model that uses interpretable textual features, including the number of turns and the frequency of honorific expressions. The results show that even advanced LLMs struggle with social relationship estimation, performing worse than a simple heuristic-based approach. This finding highlights the need for further improvement in enabling LLMs to naturally capture social relationships in multi-party dialogues.

1 Introduction

In multi-party dialogues, humans naturally infer interpersonal relationships and degrees of intimacy among participants, adapting their linguistic choices and social behaviors accordingly. Recognition of these interpersonal relationships plays a crucial role in facilitating smooth and effective communication. However, modeling these relationships is significantly more complex than in dyadic (one-on-one) settings (Ishizaki and Kato, 1998; Clark, 1982; Novick et al., 1970). Multi-party conversations involve intricate phenomena such as sub-dialogues, shifting listener roles, and unacknowledged utterances, making the automatic estimation of social relationships a considerable challenge.

Despite its importance, most research on computational relationship recognition has focused on dyadic interactions, such as estimating rapport (Nishihara et al., 2008) and intimacy (Matsumoto et al., 2018). These methods, often designed for specific dyadic or scripted contexts, are not directly applicable to the complex, spontaneous nature of multi-party chat. Furthermore, while large language models (LLMs) like GPT have been applied to various multi-party tasks (e.g., addressee recognition, response generation), their ability to robustly infer nuanced social relationships remains limited and not well-understood (Tan et al., 2023).

This study addresses this gap by focusing on the estimation of interpersonal relationships within a Japanese multi-party chat corpus. Our primary objective is to identify participant pairs with pre-existing relationships (specifically, **family** and **acquaintances**) and distinguish them from participants meeting for the first time. In multi-party dialogues involving three or more participants, it is often the case for some relationships to have already been established. We also assume a setting in which a system joins an ongoing dialogue between two persons. In these cases, it is important for the system to estimate relationships within the party. We also explore the task of assessing relationship **depth** based on conversation history. To achieve this, we investigate the efficacy of two distinct approaches: (1) a simple, interpretable logistic regression model using explicit textual features (e.g., number of turns, frequency of honorifics, and use of mention tags), and (2) a recent LLM, GPT-4o (OpenAI, 2024), using zero-shot and few-shot prompting.

Our main contribution is a comparative analysis that reveals the limitations of current LLMs in this social reasoning task. The results demonstrate that the simple interpretable logistic regression model, leveraging heuristic features, significantly outperforms GPT-4o in identifying both the type of relationship and the specific pairs involved. We find that LLMs tend to misinterpret frequent

Proceedings of the 16th International Workshop on Spoken Dialogue Systems Technology, pages 379–390
February 26–March 1, 2026. ©2026 Association for Computational Linguistics

interaction or empathy as a sign of a pre-existing relationship, particularly struggling with ‘Family’ dialogues where participants may interact *less* than acquaintances. These findings highlight a critical area for improvement: enabling LLMs to capture the subtle, and sometimes counter-intuitive, social relationships inherent in multi-party communication.

2 Multi-Party Chat Corpus

This section provides an overview and examples of the corpus used in this study and the preprocessing that was applied.

2.1 Corpus Overview

The multi-party chat corpus used in this study was developed by Tsuda et al. (2025) and consists of text-based three-party dialogues in Japanese. The participants engaged in approximately 100-turn text-based chats in an online meeting space. Here, a unit that ends with a line break is counted as a separate turn. Each dialogue was terminated at a natural topic boundary after it exceeded 100 turns.

The dialogues are broadly categorized into three types based on the relationships among the three participants: dialogues among first-timers (meeting for the first time), dialogues involving two family members and one first-timer, and dialogues involving two acquaintances and one first-timer. The first type will be referred to as "First-time dialogue," the second as "Family dialogue," and the third as "Acquaintance dialogue." The participants consist of six family pairs (12 participants), a group of 16 mutual acquaintances, and 115 participants who were complete first-timers. Each turn is annotated with the speaker, the utterance content, and, when the speaker wants to, a mention tag (@name) explicitly indicating the intended addressee. Each group of participants engaged in five dialogues, except for a small subset of the family dialogue; each dialogue was conducted independently, and the discussion topics were not shared across dialogues.

The corpus contains 1,000 First-time dialogues, 500 Family dialogues, and 500 Acquaintance dialogues. An example of a Family dialogue is presented in Table 1. From this example, we can easily infer the relationships among the participants: Speaker A is Speaker B’s mother.

Table 1: Example of chat corpus (Example from the Family dialogue subset (translated from Japanese). “@” indicates a mention tag.

Speaker	Utterance
A	Did you have breakfast this morning?
B	@A Yes, I did!
C	@A I had soba!
B	@A The green onions were spicy in mom’s natto rolls.
A	@C Looks great for the morning!
C	@B Natto rolls!
A	I had to make 6 natto rolls. For three people.
C	That sounds like a lot of effort!

2.2 Preprocessing

For the corpus used in this study, we prepared three types of datasets, as shown below, by applying processing related to mention tags. Since the criteria for assigning mention tags can vary across participants, relying on human annotation alone may lead to inconsistencies. To address this issue, we prepared two versions of the corpus: one with all mentions removed and another with mentions automatically estimated.

- Original data
- Data without mention tags (by removing them)
- Data with estimated mention tags (by predicting them)

First, we conducted experiments using the original corpus data, as shown in Table 1. Second, we created a version of the corpus with the mention tags removed. Third, we created a version with automatic mention tags assignment for all turns using GPT. Specifically, we provided GPT-4o with a sequence of 10 turns, and for the final turn, we asked it to estimate the mention tag as either “@A”, “@B”, “@C”, or “@all”. To obtain stable outputs, a few-shot prompt was used. This process was applied to all turns, resulting in the creation of a chat corpus with mention tags for all turns. A number of studies have been conducted on the addressee recognition (AR) task (e.g. Le et al., 2019; Li and Zhao, 2023; Tan et al., 2023), and according to Tan et al. (2023), the accuracy for GPT-4 in the AR task is 82.5%. For the corpus used in this study, the accuracy was 65.2%. A large difference in performance is that the dataset used by Tan et al (Tan et al., 2023). was from the Ubuntu IRC (Hu et al., 2019), which mainly consists of questions and answers, and is different from the casual conversation

Table 2: Input and Output Example (Acquaintance dialogue, R: Relationship, RP: Relational Pair, R and P: Relationship and Pair, RD: Relationship Depth).

Input	
Speaker	Utterance
A	I've been immersed in baseball with my kids.
B	That's nice!
C	Sounds great!
C	You even play catch when you go home during the week, right?

Task	Correct Output Example
R	Acquaintance dialogue
RP	A and C
R and P	Acquaintance: A and C
RD	1

used in this study.

3 Task Definition

To systematically evaluate a model's ability to estimate interpersonal relationships from dialogue, we define four distinct tasks. These tasks are designed to assess performance across multiple dimensions of social reasoning: from the general classification of a dialogue's social context (i.e., whether it contains a pre-existing relationship) to the specific identification of the related pair, the type of relationship, and finally, the depth of the relationship established over time. This section details the objective and input-output format for each task. All tasks are evaluated using accuracy. Table 2 provides a concrete example of the input dialogue and the expected output for each of the defined tasks.

3.1 Relationship Identification Task (R)

The relationship identification task is defined as a three-class identification task aimed at determining the dialogue type based on participant relationships, as mentioned in Section 2.1: First-time, Family, and Acquaintance dialogues.

3.2 Relational Pair Identification Task (RP)

The relational pair identification task focuses on Family and Acquaintance dialogues. This task identifies which two of the three participants have a relationship (either the family pair in Family dialogues or the acquaintance pair in Acquaintance dialogues). Here, the task is performed for given dialogues consisting of two family or acquaintance participants and one first-timer.

3.3 Relationship and Pair Identification Task (R and P)

This task is a combination of the two tasks mentioned above, that is to identify the two participants with a relationship in Family and Acquaintance dialogues, and simultaneously determine whether they are a family pair or an acquaintance pair. The simultaneous estimation of both the relationship and the pair will facilitate its application to dialogue systems.

3.4 Relationship Depth Assessment Task (RD)

In the relationship depth assessment task, we focus on data from the first and fifth dialogues with the same participants, and identify whether the dialogue is the first or fifth one. Each group of participants was engaged in five or more dialogue sessions. Hayashi et al. (2023) define rapport as the feeling of connection and harmony with the other person, showing that rapport increases as the number of conversations grows. Therefore, a higher rapport, a deeper relationship, and the depth of the relationship are expected to emerge in the fifth session compared to the first session.

4 Method and Analysis

In this section, we describe the methods used to estimate interpersonal relationships from the chat corpus. We first detail an interpretable baseline model, a logistic regression classifier, including the specific Dialogue features selected for the task. We then present a detailed statistical analysis of these features to validate their effectiveness and to uncover the distinct interaction patterns that characterize each relationship type.

4.1 Logistic Regression-based Approach

We performed logistic regression using the Dialogue features extracted from the sentences. We standardized all features using z-score normalization (mean = 0, standard deviation = 1). We trained an $l2$-regularized logistic regression classifier with C=1.0 (inverse regularization strength). To evaluate the model, we employed 10-fold cross-validation. Logistic regression is used as an interpretable baseline, positioned as a means to demonstrate the performance gap between simple feature-based models and LLMs.

Table 3: Mean values (and standard deviations) per participant for each dialogue type and feature.

Dialogue Type	Participants	#Utterances	#Honorifics	#Questions
First-time	First-timer	34.4 (8.0)	20.6 (7.9)	3.2 (2.7)
Family	First-timer	39.8 (7.2)	22.9 (7.4)	7.2 (4.5)
Family	Family	32.6 (7.8)	16.5 (6.6)	3.0 (2.4)
Acquaintance	First-timer	27.9 (7.9)	13.1 (7.3)	4.0 (3.0)
Acquaintance	Acquaintance	39.3 (10.8)	8.5 (5.6)	4.4 (3.2)

Table 4: Mean values of mention-related the Dialogue features.

Speaker	Mentioned person	#Mention tags	#Mention tags /w honorifics	#Mention tags /w questions
First-timer	First-timer	4.0	2.9	0.5
First-timer	Family	5.8	3.7	1.3
Family	First-timer	5.2	4.1	0.9
Family	Family	2.7	0.3	0.4
First-timer	Acquaintance	5.4	2.9	0.9
Acquaintance	First-timer	8.0	4.0	1.2
Acquaintance	Acquaintance	9.1	0.4	1.2

Following Matsumoto et al. (2018), we investigated "Dialogue features" hypothesized to reflect social relationships as:

- Number of turns per participant
- Number of honorifics per participant
- Number of questions per participant
- Number of mention tags per participant
- Number of mention tags with honorifics per participant
- Number of mention tags with questions per participant

The number of honorific expressions was measured using a dictionary-based pattern matching approach, in which common Japanese polite endings such as "desu" and "masu" were detected and each counted as one instance. Similarly, the number of questions was measured using a rule-based method, counting each occurrence of a question mark ("?") as one instance. For the mention-related features, we measured three types of interactions *between each pair* of participants, based on the assumption that identifying relationships would be easier by referring to the addressee of honorifics and questions:

- Number of mention tags used from each participant to each other participant
- Number of mention tags with honorifics from each participant to each other participant
- Number of mention tags with questions from each participant to each other participant

Table 5: t-test results (two-tailed). Asterisks indicate significance: * for p < 0.05 and ** for p < 0.01. The t-value is bolded when Family > First-timer or Acquaintance > First-timer (Acq: acquaintance).

Participants	Dialogue Features	t-value
First-timer and Family (@Family dialogue)	#Turns	17.8**
	#Questions	19.0**
	#Honorifics	16.4**
	#Mention tags	17.4**
	#Mention tags w/ honorifics	31.0**
	#Mention tags w/ questions	14.7**
First-timer and Acquaintance (@Acq dialogue)	#Turns	**23.1****
	#Questions	**2.4***
	#Honorifics	12.4**
	#Mention tags	**12.8****
	#Mention tags w/ honorifics	27.9**
	#Mention tags w/ questions	**3.1****

All features were measured by absolute counts per dialogue, noting that all dialogue sessions consist of approximately 100 turns. Note that since the number of mention tags could not be measured in the dataset where mention tags were removed, we did not use any mention-related features.

4.2 Statistical Analysis

To validate the effectiveness of the features used for the logistic regression model, we conducted a statistical analysis of the dataset. Our goal was to confirm that these "Dialogue features" (Matsumoto et al., 2018)–including the number of turns, honorifics, questions, and mention tags–exhibit statistically significant and distinct patterns across the different relationship types.

The results of this analysis are presented in Table 3 (for participant-level features) and Table 4 (for pair-wise, mention-related features). To test the statistical significance of these observations, we performed t-tests comparing the mean differences between the first-timer and the family members (in Family dialogues), and between the first-timer and the acquaintances (in Acquaintance dialogues), with the results shown in Table 5.

The results reveal statistically significant differences ($p < 0.05$) between the participant types. Key findings include: (1) In Family dialogues, family members had significantly fewer turns, used fewer honorifics, and asked fewer questions compared to the first-timer. (2) In Acquaintance dialogues, acquaintances had significantly *more* turns and used more mention tags than the first-timer.

This analysis confirms that distinct interaction patterns emerge based on the relationship context. As illustrated in Figure 1, conversations in Family dialogues tended to evolve around the first-timer, with fewer direct exchanges between the family pair. Conversely, in Acquaintance dialogues, the two acquaintances often engaged more actively with each other. These statistically validated patterns provide useful information for our logistic regression model, demonstrating that the selected features are indeed indicative of the underlying social relationships.

5 Evaluations

We evaluated the performance on each task defined in Section 3 by comparing our interpretable baseline against a state-of-the-art LLM. This section details the experimental setup for both models and presents the results for each of the four tasks.

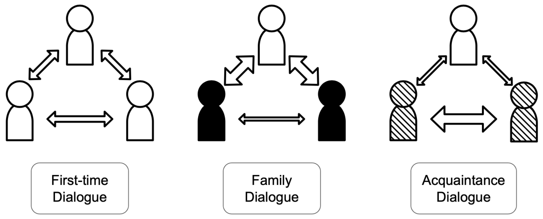

Figure 1: Interaction patterns in First-timer, Family, and Acquaintance dialogues. The white person represents the first-timer, the black person represents a family member, and the striped person represents an acquaintance. Arrow thickness indicates the frequency of interaction between each pair.

For the logistic regression experiments, we trained and evaluated the model (described in Section 4) using three distinct data preparations to understand the impact of mention tags: (1) the original data with human-annotated mention tags, (2) data with all mention tags removed, and (3) data with automatically estimated mention tags. For comparison, we also conducted ablation studies using the original data, removing either all honorific-related features or all question-related features to assess their respective contributions. The final trained regression coefficients for these models are reported in Appendix A.

For the GPT-4o experiments, we evaluated its performance using three prompting strategies to test its reasoning capabilities under different conditions:

- **Zero-shot**: Provided only the task description.
- **Few-shot (FS)**: Provided the task description along with several representative examples of inputs and correct outputs.
- **Few-shot + Statistics-aware (FS+ST)**: An enhanced FS prompt that explicitly incorporated the key statistical properties identified in our analysis (Section 4.2). Specifically, we informed the model that: honorifics are rarely used among family members or acquaintances; utterances between family members are infrequent; and utterances between acquaintances are more frequent.

5.1 Relationship Identification Task

The following is the prompt for the GPT-based method for the relationship identification task.

> Analyze a conversation between three people and output in one line an estimate of whether it includes two family members, or two people who are not family members but who know each other, or whether no one is family or knows each other.
> The output format should be “family” only if

Table 6: Results of Relationship Identification Task (FS: Few-Shot prompt, ST: Statistics-aware prompt, M: Mention, EM: Estimated Mention).

Method	Accuracy
GPT-4o w/ M	0.41
GPT-4o w/o M	0.41
GPT-4o w/ EM	0.33
GPT-4o FS w/ M	0.65
GPT-4o FS w/o M	0.53
GPT-4o FS w/ EM	0.56
GPT-4o FS+ST w/ M	0.65
GPT-4o FS+ST w/o M	0.60
GPT-4o FS+ST w/ EM	0.58
Logistic w/ M	**0.80**
Logistic w/o M	0.79
Logistic w/ EM	0.75
Logistic w/ M w/o honorifics	0.62
Logistic w/ M w/o questions	0.78

> family pairs are included. If acquaintance pairs are presumed to be included, output only "acquaintances". If neither family nor acquaintances are presumed, output only "no".

In order to estimate the relationship, we employ logistic regression with three class categories: First-time dialogue, Family dialogue, and Acquaintance dialogue.

The results are presented in Table 6. The results indicate that the logistic regression model achieved the highest performance on the original data. Although the logistic regression models generally outperformed GPT, in the case where honorific-related features were removed, the performance got close to that of GPT. When the question-related features were removed, there was no significant decrease in accuracy. In the zero-shot prompts, the accuracy decreased when using the estimated mention tags, whereas this decline was not observed in the few-shot prompts or the logistic regression model. In the prompts augmented with statistical information, a slight improvement was observed only when using the dialogues without mention tags or with estimated mention tags.

5.2 Relational Pair Identification Task

The following is the prompt for the GPT-based method for the relational pair identification task. The term "family" was replaced with "acquaintance" in the experiments involving Acquaintance dialogues.

> Analyze the conversation and estimate which two of the three are the family pair.

Table 7: Relational Pair Identification Task (Accuracy, FS: Few-Shot prompt, ST: Statistics-aware prompt, M: Mention, EM: Estimated Mention).

Method	Family	Acquaintance
GPT-4o w/ M	0.44	0.72
GPT-4o w/o M	0.44	0.70
GPT-4o w/ EM	0.35	0.61
GPT-4o FS w/ M	0.64	0.70
GPT-4o FS w/o M	0.66	0.69
GPT-4o FS w/ EM	0.51	0.67
GPT-4o FS+ST w/ M	0.69	0.68
GPT-4o FS+ST w/o M	0.59	0.65
GPT-4o FS+ST w/ EM	0.59	0.69
Logistic w/ M	0.96	**0.97**
Logistic w/o M	0.86	0.90
Logistic w/ EM	0.96	0.95
Logistic w/ M w/o honorifics	0.76	0.75
Logistic w/ M w/o questions	**0.97**	**0.97**

Table 8: Reasons for GPT's Family Pair Identification (C: Correct, IC: Incorrect, FS: Few-Shot prompt)

Reason	C	IC	C (FS)	IC (FS)
Calling by Name or Relationship	27	41	24	15
Empathy	1	36	5	20
Frequent Interactions and Questions	0	28	0	12
Shared Topics	82	6	106	15
Others	20	9	35	18

> The output format should be only "A and B", for example, if you think that A and B are a family pair.

Furthermore, for analysis only, we added 'Explain the reason for your estimation' to the prompt for a randomly sampled 210 of Family dialogues.

Logistic regression was employed with three classification targets: A and B, A and C, and B and C. The results for the Family and Acquaintance dialogues are presented in Table 7, showing the percentage of correctly identified pairs. A summary of the output reasons, including the inference process, is provided in Table 8.

According to Table 7, in the Family dialogues, the logistic regression model excluding the question-related features on the original dataset achieved the highest accuracy, while in the Acquaintance dialogues, the logistic regression model using the original dataset and the model excluding the question-related features on the original dataset achieved the best performance. However, when the honorific features were removed, the accuracy of the logistic regression model dropped significantly. In the logistic regression model, unlike the Relationship Identification Task, the performance decreased when using the dialogues without

Table 9: Example of dialogue where GPT made an error (Family dialogue)

	Utterance
A	That's why when I go to a big store, I end up taking my time looking around.
B	@*A* That's so true! When you have kids with you, you can't really take your time.
B	I quickly go while they're at school!
A	Yeah, definitely hard to take it slow with kids.
A	That's a good idea.
C	@*B* It's true, you can't really take your time.

mention tags, while it improved when using the dialogues with estimated mention tags. This suggests that mention-related features have a strong impact on identifying relationship pairs, and that predicted mention tags with low accuracy were effective to some extent. In the zero-shot prompting, GPT performed better on Acquaintance dialogues than on Family dialogues. However, in the few-shot prompting, the performance on Family dialogues improved, reducing the gap between the two types of dialogue. In the prompts augmented with statistical information, no consistent improvement was observed, as the performance varied depending on the method. According to Table 8, in the zero-shot prompting, incorrect predictions were often made by empathy or frequent interactions. Table 9 presents an example where GPT made an error in pair estimation: in this case, although the correct answer was B and C, GPT incorrectly inferred that A and B formed the family pair, reasoning that they were empathizing with each other over a topic related to children. This suggests that LLMs tend to interpret close communication—such as frequent exchanges—as indicative of a close relationship. As discussed in the analysis in Section 4, Acquaintance dialogues contain more exchanges between the acquaintances themselves, which may explain why GPT produced better results for Acquaintance dialogues than for Family dialogues. However, with few-shot prompting, fewer incorrect predictions were attributed to factors such as calling by name, empathy, or frequent interactions.

5.3 Relationship and Pair Identification Task

The following is the prompt for the GPT-based method for the Relationship and Pair Identification task.

Table 10: Relationship and Pair Identification Task (FS: Few-Shot prompt, ST: Statistics-aware prompt, M: Mention, EM: Estimated Mention).

Method	Accuracy
GPT-4o w/ M	0.34
GPT-4o w/o M	0.34
GPT-4o w/ EM	0.22
GPT-4o FS w/ M	0.40
GPT-4o FS w/o M	0.44
GPT-4o FS w/ EM	0.33
GPT-4o FS+ST w/ M	0.44
GPT-4o FS+ST w/o M	0.45
GPT-4o FS+ST w/ EM	0.40
Logistic w/ M	**0.92**
Logistic w/o M	0.76
Logistic w/ EM	0.87
Logistic w/ M w/o honorifics	0.64
Logistic w/ M w/o questions	0.91

> Analyze a conversation between three people and estimate which two of the three are a related pair and what kind of relationship they have and output in one line.
> The output format should only be "Family: A and B" if family pairs are included. If the pair is not a family but an acquaintance, output only "Acquaintance: A and B".

In the logistic regression method, relationship and pair identification was performed using two classes for relationship type (family or acquaintance) and three classes for pair combinations, resulting in a logistic regression model with six classification categories. The experimental results are presented in Table 10. The table shows the percentage of correct answers where both the relationship type and the specific pair were correctly identified.

It shows that the logistic regression model using the original data achieved the highest percentage of correct answers. However, when the honorific-related features were removed, the accuracy of the logistic regression model dropped significantly, whereas this decline was not observed when the question-related features were removed. In the logistic regression model, the performance decreased when using the dialogues without mention tags, while it improved when using the dialogues with estimated mention tags. This is likely because, while the accuracy using the dialogues with estimated mention tags declined in the Relationship Identification task, the improvement in the accu-

Table 11: Relationship Depth Assessment Task (Accuracy, FS: Few-Shot prompt, M: Mention, H: honorifics, Q: questions).

Method	First-timer	Family	Acquaintance
GPT-4o w/ M	0.46	0.53	0.51
GPT-4o w/o M	0.51	0.54	0.54
GPT-4o w/ EM	0.50	0.50	0.58
GPT-4o FS w/ M	0.53	0.77	**0.70**
GPT-4o FS w/o M	0.53	**0.79**	**0.70**
GPT-4o FS w/ EM	0.52	0.76	0.65
Logistic w/ M	0.54	0.60	0.53
Logistic w/o M	**0.60**	0.64	0.58
Logistic w/ EM	0.54	0.53	0.54
Logistic w/ M w/o H	0.45	0.45	0.49
Logistic w/ M w/o Q	0.56	0.55	0.55

racy in the Relational Pair Identification task was more substantial. However, in GPT, the accuracy decreased when using the data with estimated mention tags, while it improved when using the data without mention tags. This suggests that GPT may not effectively utilize mention tags in its predictions.

5.4 Relationship Depth Assessment Task

The following is the prompt for the GPT-based method for the relationship depth assessment task.

> Analyze the conversation and output “1” or “5” for the dialogue, whether it is the first or fifth dialogue. The first and fifth dialogues data are given. The output format should be “numeric” only.

We performed the logistic regression as a binary classification task that predicts whether a dialogue is the first or the fifth session for the same participant group. The experimental results for First-time dialogues, Family dialogues, and Acquaintance dialogues are shown in Table 11.

According to Table 11, the logistic regression model achieved the highest accuracy in First-time dialogues, whereas GPT showed the highest accuracy in both Family and Acquaintance dialogues. In this task, the overall performance was low, even though it was a binary classification problem, and regardless of whether mention tags were present or not. The effect of mention tags tends to vary greatly depending on the individual, and it is likely that the mention-related features did not change significantly between the first and fifth dialogues. In this task, the decrease in accuracy caused by excluding the honorifics-related features was smaller compared to other tasks. It is possible that even by the fifth conversation, the relationship had not deepened significantly enough to be effectively captured by the model.

6 Conclusions

In this study, we focused on a multi-party chat corpus and estimated relationships between participants using GPT-4o and logistic regression models. The analysis confirmed that First-time dialogues, Family dialogues, and Acquaintance dialogues each exhibit distinctive characteristics. The logistic regression models achieved significantly higher accuracy than GPT on many tasks, including detecting the presence of relationships and identifying specific relationship pairs. In particular, the logistic regression model showed better performance in the relationship pair identification task. However, when the honorific-related features were removed, the performance of the logistic regression model significantly decreased. GPT tends to emphasize frequent and dense communication, resulting in relatively good performance for acquaintance conversations in the pair identification task, but showing lower accuracy for family conversations. Also, GPT performed better on the relationship depth assessment task compared to the other tasks. These findings suggest that GPT is relatively capable of estimating the depth of relationships, despite its limitations in accurately identifying specific relationships.

The framework presented in this study has broader applicability. Relationships such as family and acquaintances are universal, and the method used in this study can potentially be adapted to other languages and cultural contexts. Thus, this study not only demonstrates the effectiveness of a simple, interpretable model in Japanese multi-party dialogues but also provides a generalizable framework for relationship estimation in dialogue systems. Future challenges include generalizing the model using diverse datasets, such as the Corpus of Everyday Japanese Conversation (CECJ, Koiso et al., 2022). While fine-tuning was not performed in this study, as the focus was on providing an interpretable baseline, it will likely be necessary for future improvements. Additionally, because honorific expressions are unique to the Japanese, careful adaptation would be required when applying this approach to other languages.

Acknowledgments

This work was supported by JST Moonshot R&D JPMJPS2011.

References

H. H. Clark. 1982. Hearers and speech acts. *Language*, pages 332–373.

Takato Hayashi, Ryusei Kimura, Ryo Ishii, Fumio Nihei, Atsushi Fukayama, and Shogo Okada. 2023. Ranking conversations based on rapport in first meeting conversations and friend conversations. In *SIG-SLUD*, pages 72–79.

Wenpeng Hu, Zhangming Chan, Bing Liu, Dongyan Zhao, Jinwen Ma, and Rui Yan. 2019. Gsn: A graph-structured network for multi-party dialogues. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5010–5016.

Masato Ishizaki and Tsuneaki Kato. 1998. Exploring the characteristics of multi-party dialogues. In *Association for Computational Linguistics*, page 583–589.

Hanae Koiso, Haruka Amatani, Yuichi Ishimoto, Yuriko Iseki, Yasuyuki Usuda, Wakako Kashino, Yoshiko Kawabata, Yayoi Tanaka, Yasuharu Den, Kenya Nishikawa, and Yuka Watanabe. 2022. Design and features of the corpus of everyday japanese conversation. In *NLP*, page 2008–2012.

Ran Le, Wenpeng Hu, Mingyue Shang, Zhenjun You, Lidong Bing, Dongyan Zhao, and Rui Yan. 2019. Who is speaking to whom? learning to identify utterance addressee in multi-party conversations. In *Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, page 1909–1919.

Yiyang Li and Hai Zhao. 2023. Em pre-training for multi-party dialogue response generation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 92–103.

Kazuyuki Matsumoto, Kyosuke Akita, Ren Fuji, Minoru Yoshida, and Kenji Kita. 2018. Intimacy estimation of the characters in drama scenario. *Intelligence and Information*, pages 591–604.

Yoko Nishihara, Wataru Sunayama, and Masahiko Yachida. 2008. Human friendship and hierarchical relationship estimation from utterance texts. *The Institute of Electronics, Information and Communication Engineers Transactions. Information and Systems: D*, pages 78–88.

David Novick, Lisa Walton, and Karen Ward. 1970. Contribution graphs in multiparty discourse. In *International Symposium on Spoken Dialogue (ISSD)*, pages 53–56.

OpenAI. 2024. Hello gpt-4o.

Chao-Hong Tan, Jia-Chen Gu, and Zhen-Hua Ling. 2023. Is chatgpt a good multi-party conversation solver? In *International Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Taro Tsuda, Sanae Yamashita, Koji Inoue, Tatsuya Kawahara, and Ryuichiro Higashinaka. 2025. Multi-relational multi-party chat corpus. In *NLP*.

A Logistic Regression Coefficients

In this appendix, we provide the detailed regression coefficients obtained from the logistic regression models used in our experiments. Each figure corresponds to a specific experimental setting described in Section 5.1-5.4, and lists the coefficients associated with each input feature. The coefficients indicate the relative contribution of each feature to the prediction of the target variable, with positive values representing a positive correlation and negative values representing a negative correlation. All coefficients were standardized before training to allow for comparison across features. Figures A.1-A.21 summarize the coefficients for each condition. We include these detailed values to facilitate reproducibility and to allow readers to interpret the influence of individual features on the model's decision boundaries.

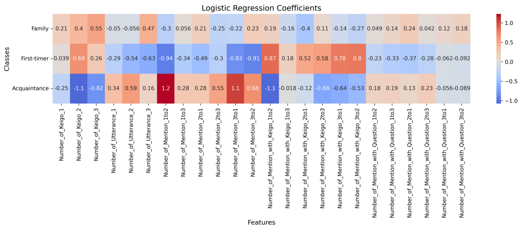

Figure A.1: Heatmap of logistic regression coefficients for the relationship identification task (with mention tags).

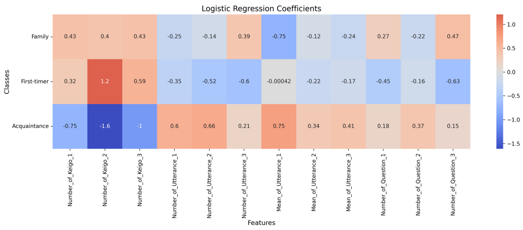

Figure A.2: Heatmap of logistic regression coefficients for the relationship identification task (without mention tags).

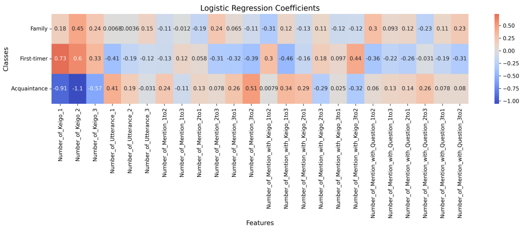

Figure A.3: Heatmap of logistic regression coefficients for the relationship identification task (with estimated mention tags).

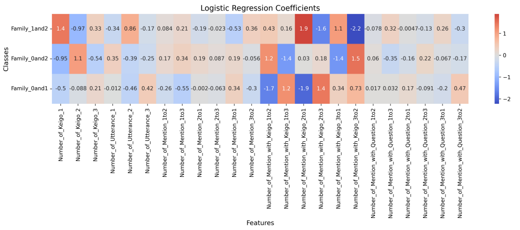

Figure A.4: Heatmap of logistic regression coefficients for the relational pair identification task (Family dialogues, with mention tags).

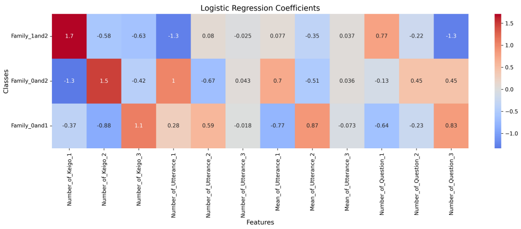

Figure A.5: Heatmap of logistic regression coefficients for the relational pair identification task (Family dialogues, without mention tags).

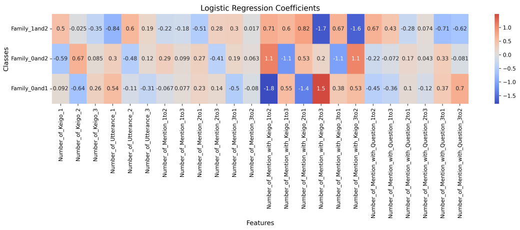

Figure A.6: Heatmap of logistic regression coefficients for the relational pair identification task (Family dialogues, with estimated mention tags).

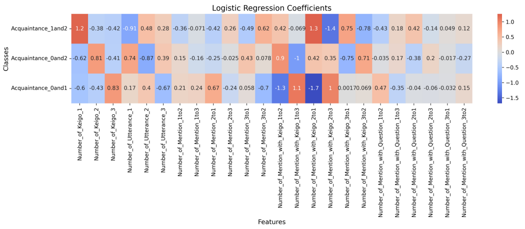

Figure A.7: Heatmap of logistic regression coefficients for the relational pair identification task (Acquaintance dialogues, with mention tags).

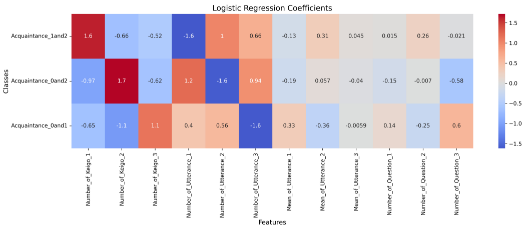

Figure A.8: Heatmap of logistic regression coefficients for the relational pair identification task (Acquaintance dialogues, without mention tags).

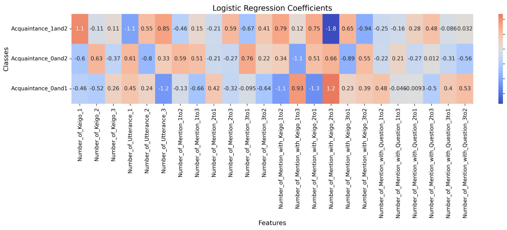

Figure A.9: Heatmap of logistic regression coefficients for the relational pair identification task (Acquaintance dialogues, with estimated mention tags).

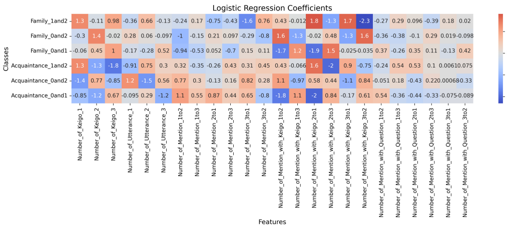

Figure A.10: Heatmap of logistic regression coefficients for the relationship and pair identification task (with mention tags).

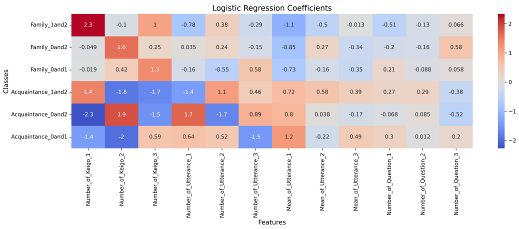

Figure A.11: Heatmap of logistic regression coefficients for the relationship and pair identification task (without mention tags).

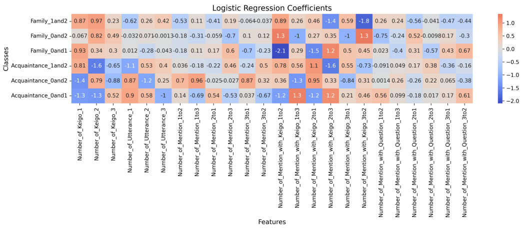

Figure A.12: Heatmap of logistic regression coefficients for the relationship and pair identification task (with estimated mention tags).

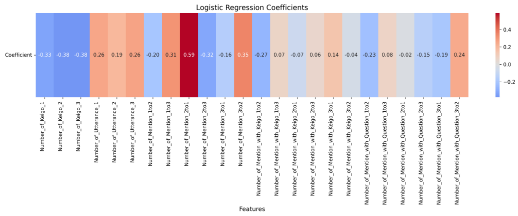

Figure A.13: Heatmap of logistic regression coefficients for the relationship depth assessment task (First-time dialogues, with mention tags).

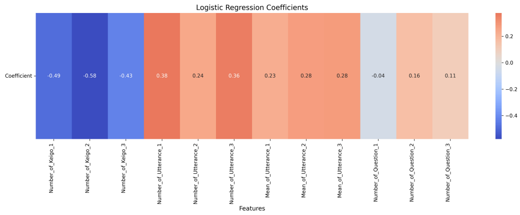

Figure A.14: Heatmap of logistic regression coefficients for the relationship depth assessment task (First-time dialogues, without mention tags).

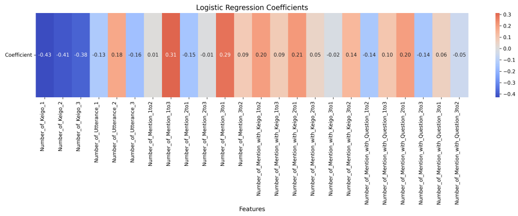

Figure A.15: Heatmap of logistic regression coefficients for the relationship depth assessment task (First-time dialogues, with estimated mention tags).

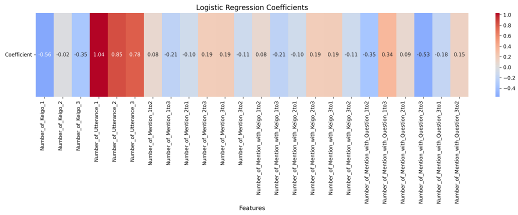

Figure A.16: Heatmap of logistic regression coefficients for the relationship depth assessment task (Family dialogues, with mention tags).

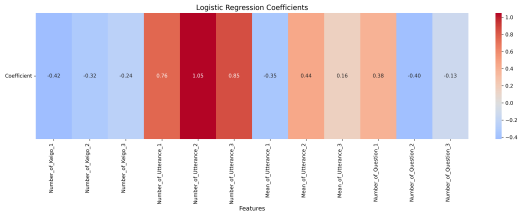

Figure A.17: Heatmap of logistic regression coefficients for the relationship depth assessment task (Family dialogues, without mention tags).

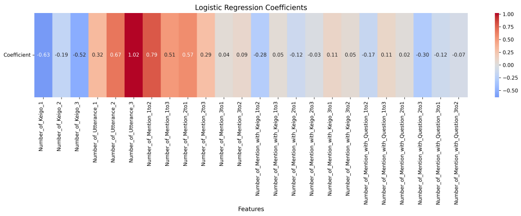

Figure A.18: Heatmap of logistic regression coefficients for the relationship depth assessment task (Family dialogues, with estimated mention tags).

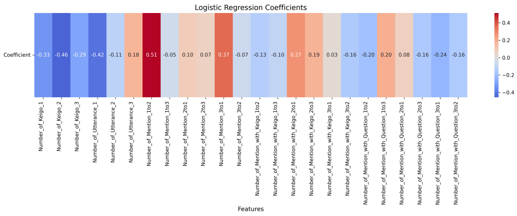

Figure A.19: Heatmap of logistic regression coefficients for the relationship depth assessment task (Acquaintance dialogues, with mention tags).

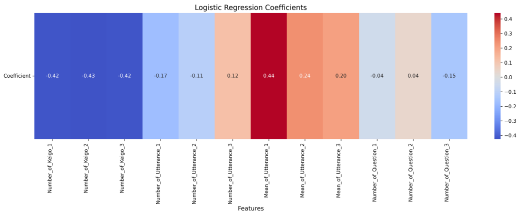

Figure A.20: Heatmap of logistic regression coefficients for the relationship depth assessment task (Acquaintance dialogues, without mention tags).

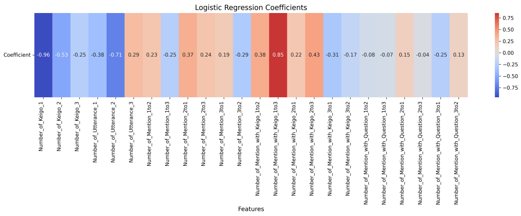

Figure A.21: Heatmap of logistic regression coefficients for the relationship depth assessment task (Acquaintance dialogues, with estimated mention tags).