Can Large Language Models Understand Dynamic Joint Decision-Making Processes? Evidence on Destination Choice for Leisure Activities

Sung-Yoo Lim*[1], Koki Sato[2], Kiyoshi Takami[3], Giancarlos Parady[4], Eui-Jin Kim[5]

[1] Graduate Student, Department of Transportation Systems Engineering, Ajou University, Republic of Korea

[2] Graduate Student, Department of Urban Engineering, the University of Tokyo, Japan

[3] Associate Professor, Department of Urban Engineering, the University of Tokyo, Japan

[4] Lecturer, Department of Urban Engineering, the University of Tokyo, Japan

[5] Assistant Professor, Department of Transportation Systems Engineering, Ajou University, Republic of Korea

SHORT SUMMARY

This study examines the potential of large language models (LLMs) to understand joint travel decisions related to social activities, using group chat data from messaging platforms like WhatsApp. These decisions involve dynamic negotiations, considering the preferences and constraints of each participant. To fully understand the decision-making process, it is necessary to infer nuanced and implicit information from the social and cultural context of each generation and country. Specifically, decision-making factors must be represented in a structured format, requiring extensive human-labeled annotations. A customized prompt, based on chain-of-thought reasoning, is designed to first identify individual/clique characteristics and then trace how travel decisions evolve through iterative negotiations, capturing both explicit and implicit factors from Japanese and Korean data. We quantitatively evaluate the performance of LLMs in structuring the data, identify optimal prompting methods, and recognize situations where LLMs struggle. These findings highlight the potential of LLM-based analysis for enhancing context-rich activity-based modeling.

Keywords: Activity-based modelling, Group chat data, Joint travel decisions, Large language models, Social networks, Leisure travel behavior

1 INTRODUCTION

The growth of social networks and social media has reshaped lifestyles and travel behaviors, increasing the complexity of how individuals engage in activities and make destination choices across multiple locations (Bifulco et al., 2010). Specifically, joint travel decisions related to social activities are made in coordination with clique members (Puhe et al., 2021). This coordination involves a dynamic social negotiation process, considering the preferences and constraints of each member. The complexity of joint travel decisions necessitates more detailed datasets to capture the nuances of activity destination choices. Group chats on messaging platforms such as WhatsApp provide empirical data that reflect this complex negotiation process. This study focuses on the joint destination choice of dining locations for social activities, collecting and analyzing relevant group chat data from Japan and South Korea.

To fully understand the decision-making process in these group chats, it is essential to infer both nuanced and implicit information based on the social and cultural context of each country and generation (e.g., Generation Z, Baby Boomers). Additionally, to quantify this process, decision-making factors in the group chat data must be represented in a structured format. This involves identifying not only the choice outcome but also the decision-making process itself, such as the alternatives within the choice set, the individual and clique characteristics that may influence the process, and the discussion behind the choice, including each individual's preferences and constraints related to alternative-specific attributes (e.g., restaurant service quality, accessibility, time-space constraints). Previous research has structured these factors through careful examination of each chat by someone with a strong understanding of the social and cultural context (Parady et al., 2023). However, this process requires extensive human-labeled annotations, leading to significant time and cost.

This study explores the potential of large language models (LLMs) to understand group chats in the context of joint destination choice. LLMs have shown great potential for extracting both implicit and explicit information from unstructured text, considering the social and cultural context (Tao et al., 2024). The capability of LLMs highly depends on the prompts (i.e., the input text provided to the model to elicit a response). Prompt engineering is the process of designing and optimizing these prompts to achieve desired outputs from LLMs.

We design LLM prompts to structure unstructured group chat data using various prompting methods. These methods guide the LLM in identifying both explicit and implicit factors from Japanese and Korean group chat data. The proposed prompts, based on chain-of-thought reasoning, first identify static attributes, such as individual and clique characteristics, and then trace dynamic attributes that reveal how joint decisions evolve through iterative negotiations. We quantitatively evaluate the performance of LLMs in structuring group chat data, comparing different prompting methods to identify optimal designs and recognize situations where LLMs struggle to understand the context. These findings highlight the potential of LLM-based analysis for understanding activity-based travel decisions, thereby enhancing context-rich activity-based modeling.

2 BACKGROUND

LLMs and prompt engineering

Despite their remarkable capabilities, LLMs have an inherent limitation: by design, they generate responses based on learned statistical patterns over massive textual corpora, often aligning with *fast thinking* (Kahneman, 2011), which is intuitive and relies on quick heuristic judgments. These *fast thinking* outputs may sound plausible but lack careful, iterative reasoning.

Prompt engineering techniques, such as chain-of-thought reasoning, aims to explicitly guide an LLM to emulate *slow thinking*, which is deliberate and relies on logical and analytical judgments. By carefully structuring the prompts, we can encourage the LLM to produce more thoughtful responses and extract deeper insights from group chat data. This approach is especially important for capturing social and emotional contexts embedded in joint decision with social networks.

Analysis framework for joint travel decisions

In joint travel decisions, individuals negotiate preferences and constraints based on individual and clique characteristics. Cultural and social contexts (e.g., within-clique hierarchy defined as age or grade in Japan and South Korea) add complexity through indirect suggestions, honorific nuances, and implicit relational signals. This study proposes a two-stage prompts: first, to identify individual and clique characteristics, and second, to capture the decision-making process, including each participant's alternatives, preferences, and constraints related to those alternatives.

The proposed prompting method is a deliberate process of designing how an LLM "thinks" to elicit deeper insights into joint decision-making processes, which were previously only identifiable through human-labeled annotation.

3 DATA AND METHOD

Data Description

We use group chat data in Japanese and Korean to evaluate whether an LLM can effectively process culturally specific conversations. The x-GDP dataset from Parady et al. (2023) comprises 217 LINE messenger group chats documenting real-time restaurant selection negotiations near the University of Tokyo campuses. This dataset was collected under IRB-approved protocols, with groups of 3-5 university students coordinating restaurant choices in controlled experimental settings. Each conversation captures the complete decision-making sequence, from initial proposals to preference negotiations and final restaurant selection, verified with photographic evidence and transaction records. From the x-GDP dataset, we randomly sampled 20 Japanese group chats to construct the Japanese dataset. The Korean dataset follows a similar approach, comprising 10 KakaoTalk messenger group chats documenting joint restaurant selection near Ajou University campuses. For more detailed information about the data, please refer to Parady et al. (2023).

Overview and Rationale

Our methodology transforms unstructured raw group chat data into structured tabular data using GPT-4o. To ensure reproducible results, we set the LLM's temperature parameter, which controls the randomness of the model's output, to 0 and apply two prompt engineering techniques, as detailed in Figure 4:

- Role Prompting: Directs the LLM to analyze conversations as a domain expert familiar with the cultural and linguistic contexts of Korean and Japanese group chats.
- Prompt Chaining: Decomposes the analysis into sequential steps to reduce error propagation.

As shown in Figure 1, this approach converts free-flowing chat conversations into structured tabular data that captures decision-making factors, including individual and clique characteristics, choice sets and outcomes, as well as each participant's preferences and constraints for the alternatives. The structured format enables systematic analysis of how group decisions emerge from conversations while preserving both explicit statements and implicit contextual factors. The bolded red text in

Figure 1 represents the structured outputs defined in the prompts, enabling systematic structuring of the data.

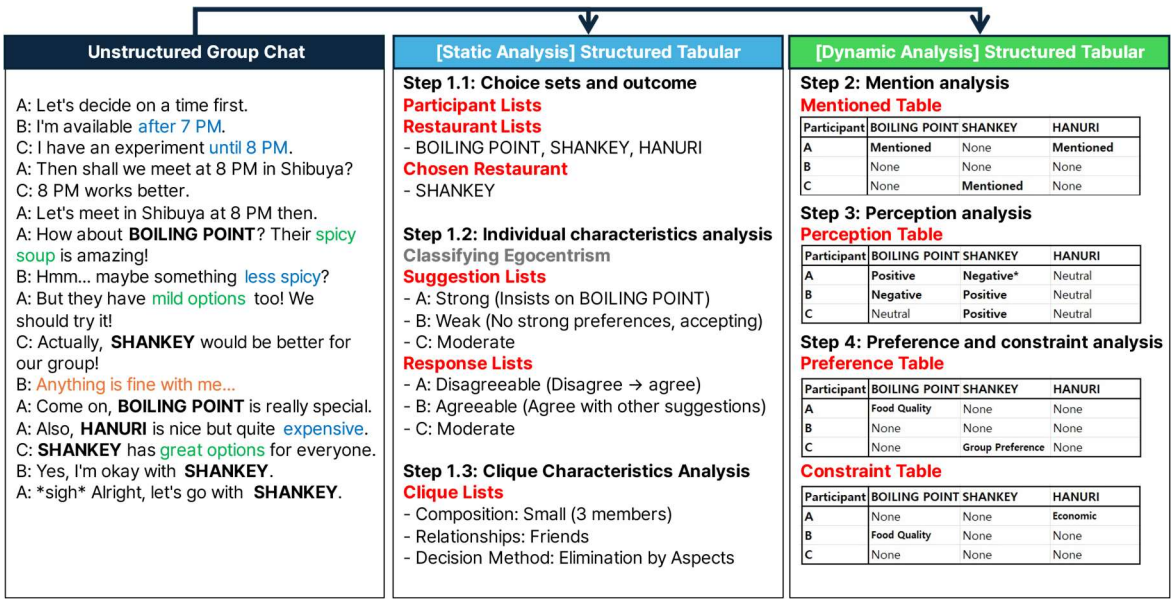

Figure 1: From Unstructured Group Chat to Structured Tabular Data

Two-stage Prompts

Figure 2 illustrates the proposed two-stage prompts. The static analysis (Step 1) captures static attributes, including individual and clique characteristics, choice sets, and outcomes. The dynamic analysis (Steps 2–4) captures dynamic attributes, such as the mentioned alternatives (Mention Analysis), each participant's perceptions of those alternatives (Perception Analysis), and each participant's preferences and constraints for alternative-specific factors (Factor Analysis).

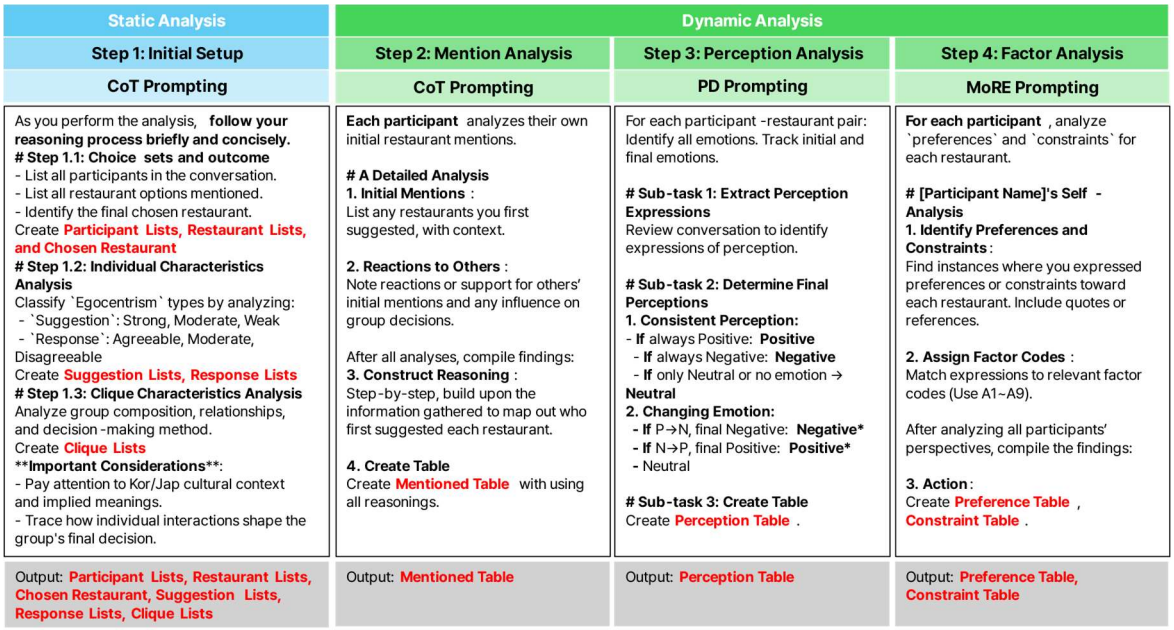

Figure 2: The proposed two-stage prompts summarizing Steps 1-4 and their outputs.

Static Analysis (Step 1)

The Step 1 in the static analysis establishes key information about the conversation, including:

- Participants (e.g., names or identifiers),
- Restaurant List and Chosen Restaurant (any options mentioned),
- Individual and Clique Characteristics (e.g., individual attributes or group dynamics).

Defining these elements prevents confusion in subsequent stages and ensures that references to participants or restaurants remain consistent throughout the analysis.

Dynamic Analysis (Steps 2–4)

Once the static attributes are in place, the dynamic analysis examines how the group's decision evolves over time. Specifically:

- Step 2 tracks the emergence of new restaurants (i.e., alternatives) mentioned, identifying which participant introduces each idea.
- Step 3 observes shifts in each participant's perceptions or emotional views on each restaurant, capturing changes in attitudes as the conversation progresses.
- Step 4 identifies each participant's preferences and constraints for each restaurant's attributes (e.g., budget, time availability, personal tastes) that influence each participant's final position.

By structuring the analysis into distinct steps, the model can more accurately parse the conversation, following the logical flow of proposals, responses, and eventual consensus.

Ground truth labels

Ground truth labels for the data were created by researchers with a strong understanding of the cultural and social context of university students in Japan and South Korea. We carefully reviewed each conversation to ensure accurate labeling, following a structured framework. **Figure 3** provides an overview of all variables used in labeling:

- Static Analysis variables: participant lists, restaurant lists, the chosen restaurant, individual suggestion and response lists, and clique lists with composition, relationships, and decision-making methods.
- Dynamic Analysis variables: the Mentioned Table for tracking newly introduced alternatives, the Perception Table for emotional views on those alternatives, and the Preference Table and Constraint Table for alternative-specific factors such as budget, time constraints, or group dynamics.

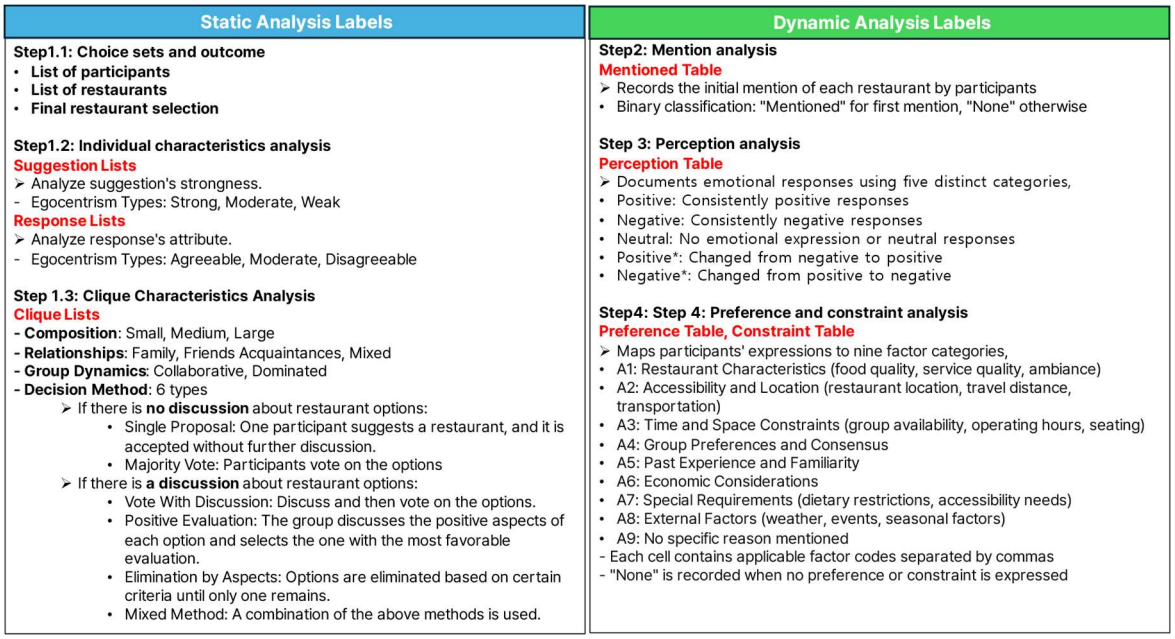

Figure 3: Labels for Static and Dynamic Analysis

Prompt Engineering Techniques

We employ multiple prompt engineering techniques tailored to the complexity of each step in our group chat data analysis. As shown in Figure 4, Static Analysis (Step 1) utilizes three prompting methods, while Dynamic Analysis (Steps 2–4) incorporates four methods. The optimal prompting methods for each analysis are identified through quantitative evaluation, as discussed in a later section.

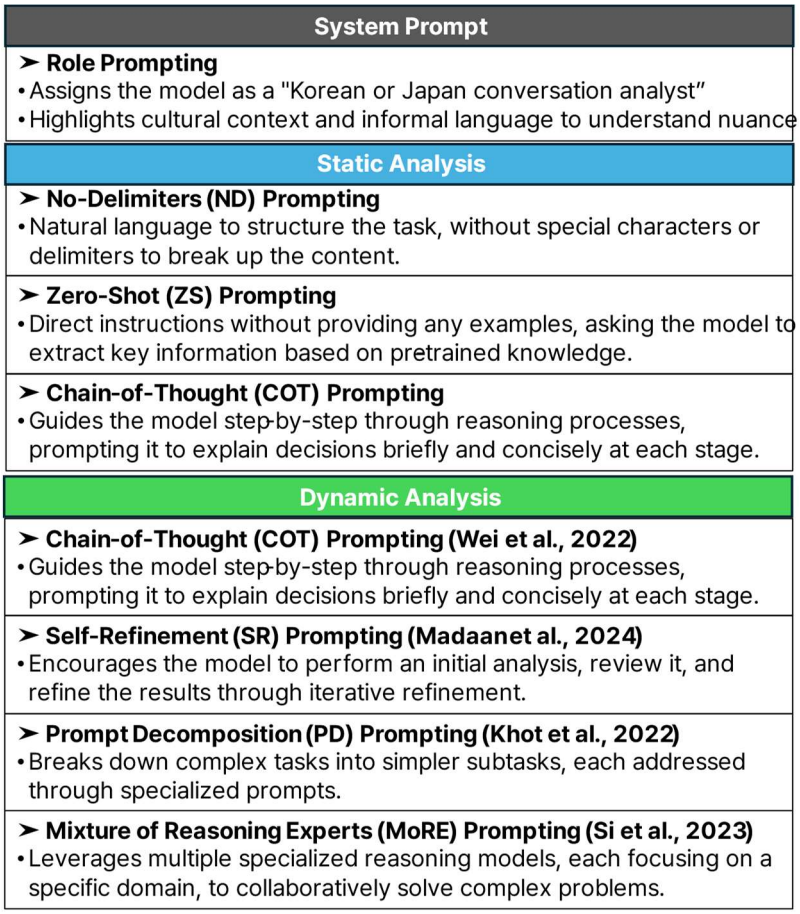

Figure 4: Overview of prompt engineering techniques at each step: ND, ZS, and CoT for Static Analysis, and CoT, SR, PD, and MoRE for Dynamic Analysis.

Static Analysis (Step 1)

No-Delimiter (ND) is a baseline approach that provides raw conversation text to the LLM with minimal guidance. Zero-Shot (ZS) directly extracts specific entities (e.g., participants, restaurants) without examples or reasoning steps. Chain-of-Thought (CoT) guides the LLM through a systematic reasoning process to identify key elements in the conversation. The output of Step 1 is structured as JSON data and then used as input for Step 2. A performance evaluation identifies which technique provides the most reliable foundation for subsequent analysis.

Dynamic Analysis (Steps 2–4)

Chain-of-Thought (CoT) (Wei et al., 2022) maintains structured reasoning through complex conversational dynamics. Self-Refinement (SR) (Madaan et al., 2024) enables iterative review and correction of initial interpretations. Prompt Decomposition (PD) (Khot et al., 2022) breaks analysis into discrete subtasks for improved accuracy. Mixture of Reasoning Experts (MoRE) (Si et al., 2023) analyzes conversation from multiple participant perspectives.

Performance Evaluation

Figure 5 presents the step-by-step evaluation procedure for each stage, using ground truth labels defined in Figure 4. In **Step 1**, we compare the ND, ZS, and CoT techniques (each repeated five times) to identify participants (P^*) and restaurants (R^*). We use F1 scores as our metric to balance both precision and recall, since LLMs can either produce extra items or miss valid ones. The method with the highest average F1 score in Step 1 determines the (P^*, R^*) sets for subsequent steps.

In **Steps 2 and 3**, we also use F1 scores but expand the methods to CoT, SR, PD, and MoRE (again each repeated five times). Step 2 identifies which participant (from P^*) first mentions each restaurant (in R^*), and Step 3 tracks how perceptions of those restaurants evolve. We adopt the best-performing approach at each step to reduce error propagation.

Because **Step 4** involves multi-valued preferences and constraints for each restaurant per participant, we introduce Jaccard similarity to measure how closely the extracted factors match the ground truth. While F1 scores remain useful, Jaccard similarity better captures partial overlaps in multi-label settings, providing a more nuanced assessment of correctness.

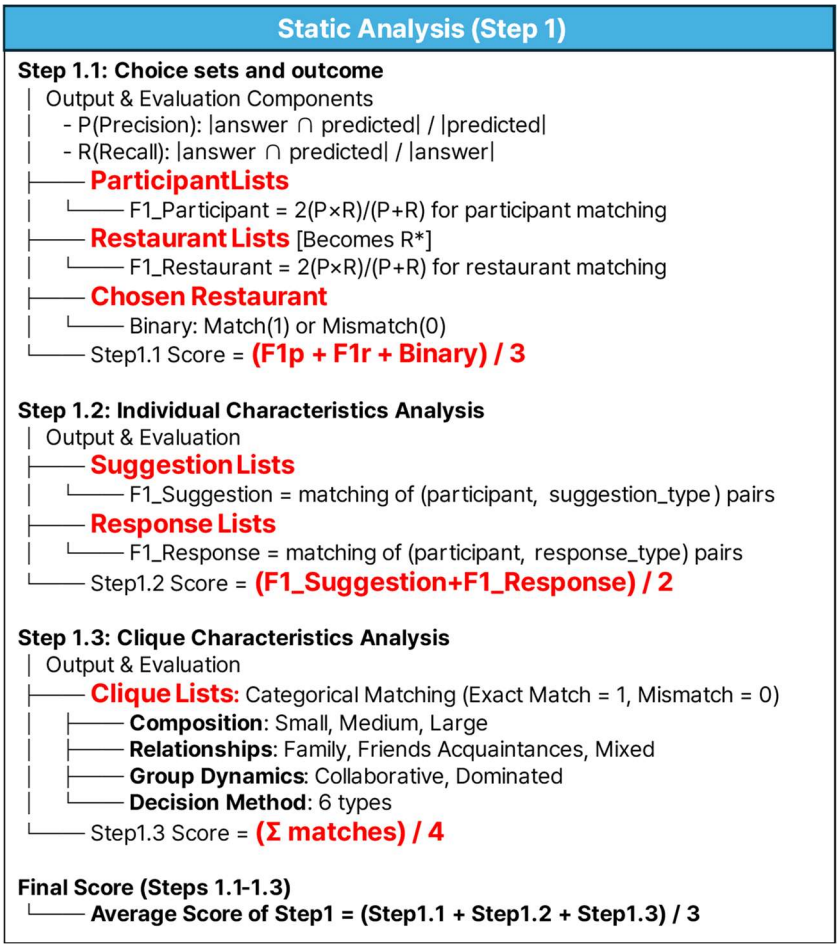

Figure 5: Evaluation metrics for static analysis (Step 1)

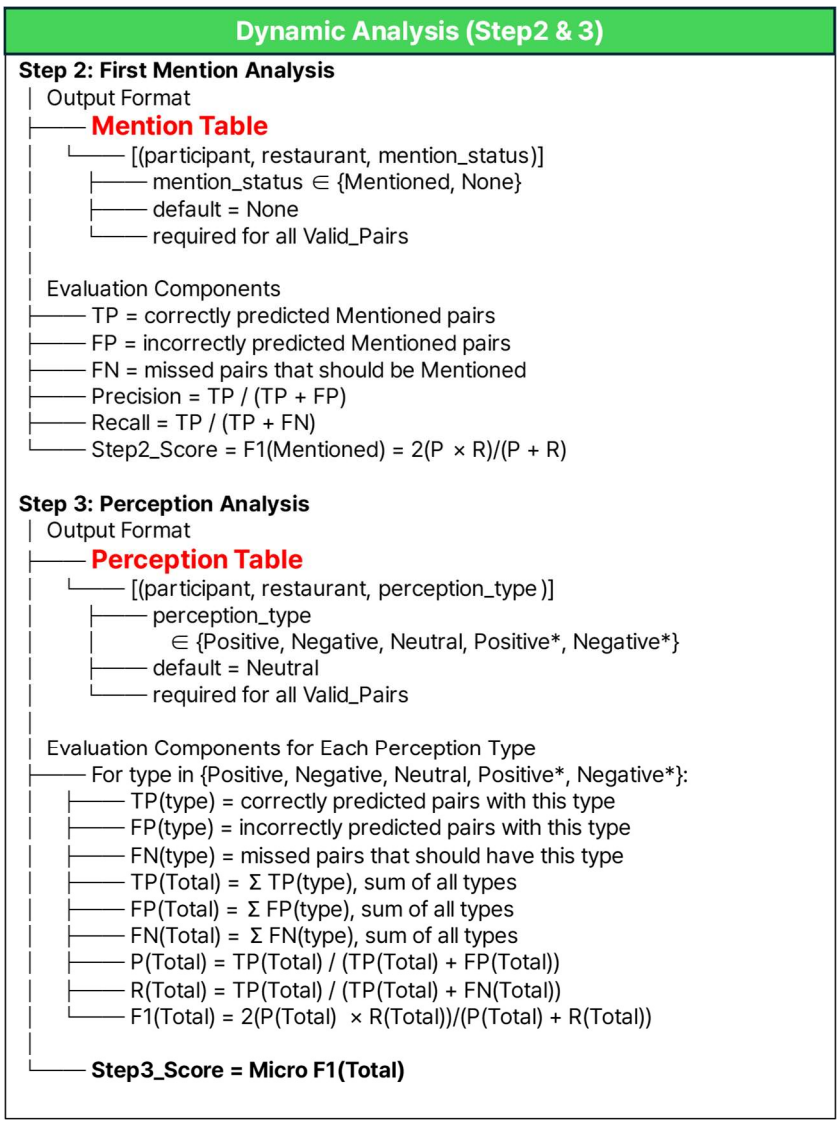

Figure 5: Evaluation metrics for static analysis (Step 2&3)

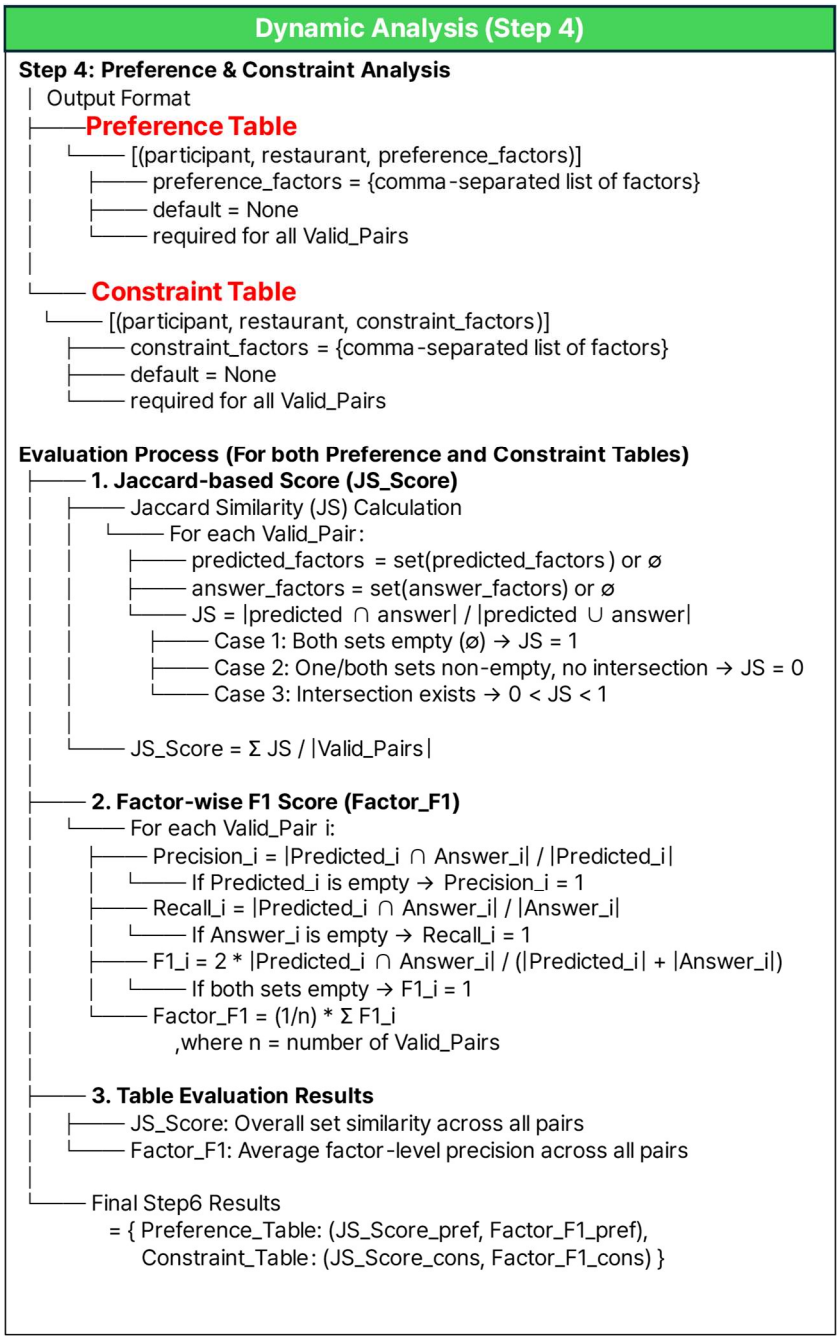

Figure 5: Evaluation metrics for static analysis (Step 4)

4 RESULTS AND DISCUSSIONS

Performance of Prompt Engineering Techniques

This section presents the performance of each prompting method at each step, evaluated using ground truth labels created by researchers with expertise in Japanese and Korean social contexts. All reported values are averaged over five runs, with standard deviations consistently below 0.02, reflecting stable performance achieved by the low temperature parameter setting.

Static Analysis Performance

Table 1 summarizes the performance of ND, ZS, and CoT in extracting static information, including participant lists, restaurant lists, the chosen restaurant, suggestion lists, and response lists, from Japanese and Korean group chat data.

In Step 1.1, all three techniques achieve perfect accuracy in identifying participants, with an F1 score of 1.00. They also demonstrate high accuracy in recognizing mentioned restaurants, with F1 scores reaching up to 0.97 for Japanese chats and 0.85–0.89 for Korean chats. For identifying the chosen restaurant (binary matching), Korean chats achieve perfect accuracy (1.00), whereas Japanese chats yield slightly lower scores (0.78–0.85).

While all three methods perform well in Step 1.1, capturing implicit elements in Step 1.2, such as individual egocentrism represented by suggestions and response lists, was more challenging. This resulted in lower scores, ranging from 0.66–0.68 for Korean data and 0.48–0.56 for Japanese data.

In Step 1.3, the three techniques achieved F1 scores ranging from 0.83 to 0.87 across the datasets from both countries. The errors in this step primarily arose from difficulties in identifying decision-making methods.

Overall, the performance of the three prompting methods in Step 1 was similar. However, significant differences were observed between the Korean and Japanese datasets, likely due to the more indirect expressions typically found in Japanese conversations compared to the explicit confirmations commonly used in Korean conversations.

Table 1. Performance in Static Analysis (Step1)

	Japanese			Korean		
	ND	**ZS**	**CoT**	**ND**	**ZS**	**CoT**
F1 (Participant lists)	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
F1 (Restaurant lists)	**0.97**	**0.97**	**0.97**	0.85	0.86	0.89
Binary matching (Chosen restaurant)	0.78	0.85	0.85	**1.00**	**1.00**	**1.00**
Step1.1_Score	0.92	0.94	0.94	0.95	0.95	0.96
F1_Suggestion	0.60	0.75	0.73	0.78	0.70	0.73
F1_Response	0.35	0.36	0.31	0.58	0.63	0.60
Step1.2_Score	0.48	0.56	0.52	0.68	0.67	0.66
Step1.3_Score	0.83	0.84	0.84	0.87	0.83	0.85
Step1 Final Score	0.74	0.78	0.76	0.83	0.82	0.83

Dynamic Analysis Performance

Step2: Mentioned Table
All four prompting methods (CoT, SR, PD, MoRE) show robust performance in identifying the participant who first mentions each restaurant (F1 ≈ 0.85–0.93). The performance is slightly higher for Korean chats (Precision: 0.91–0.92; Recall: 0.93–0.96) than for Japanese (Precision: 0.86–0.89; Recall: 0.84–0.85).

Step3: Perception Table
Perception analysis (positive, negative, neutral) demonstrates moderate performance, with F1 scores ranging from approximately 0.59 to 0.75. Korean data generally perform better (F1 up to 0.75) compared to Japanese data (F1 up to 0.65). PD achieves the highest precision and recall in Korean chats (0.75, 0.76), while Japanese conversations yield similar results across all prompting methods. The greater nuance of emotional expressions in Japanese conversations likely account for the lower scores.

Step4: Preference Table and Constraint Table
Preference analysis shows moderate performance, with Jaccard Similarity (JS) scores ranging from 0.59 to 0.67, while constraint analysis achieves notably higher scores (JS: 0.75–0.80). MoRE demonstrates the strongest performance in constraint analysis (JS: 0.80), whereas CoT shows the lowest performance across both datasets. The performance gap between preference and constraint analysis is likely due to constraints being more concrete (e.g., budget, time), making them easier to extract explicitly, whereas preferences are often expressed more implicitly.

Table 2. Performance in Dynamic Analysis (Step2-4)

	Japanese				Korean				
	CoT	**SR**	**PD**	**MoRE**	**CoT**	**SR**	**PD**	**MoRE**	
Precision	0.89	0.86	0.88	0.87	0.92	0.91	0.92	0.91	**Mentioned Table**
Recall	0.84	0.84	0.85	0.85	0.96	0.95	0.96	0.93	
Step2_Score	0.86	0.85	0.86	0.86	0.93	0.93	0.93	0.92	
Precision	0.66	0.66	0.66	0.66	0.74	0.75	0.75	0.70	**Perception Table**
Recall	0.64	0.65	0.65	0.57	0.75	0.76	0.76	0.71	
Step3_Score	0.65	0.65	0.65	0.59	0.74	0.75	0.75	0.70	
JS_Score	0.61	0.60	0.61	0.59	0.65	0.67	0.65	0.66	**Preference Table**
Factor_F1	0.63	0.62	0.63	0.62	0.66	0.68	0.66	0.67	
JS_Score	0.77	0.80	0.78	0.80	0.75	0.78	0.78	0.80	**Constraint Table**
Factor_F1	0.77	0.80	0.78	0.80	0.75	0.78	0.78	0.80	

5 CONCLUSION

This study demonstrates how joint decision-making can be systematically analyzed using LLMs guided by optimal prompts. We quantitatively evaluate the effectiveness of LLMs against large-scale ground truth data across various subtasks, ranging from simple tasks like detecting participants to more complex tasks like identifying each participant's preferences and constraints for restaurant attributes, such as accessibility and quality. Through this approach, we explore whether LLMs can effectively transform unstructured group chat data into structured tabular data by capturing the dynamic negotiations involved in joint travel decisions.

Our results provide three noteworthy findings. First, LLMs can reliably capture explicit decision elements (e.g., chosen alternatives) but often struggle with more implicit, context-dependent content, underscoring the complexity of understanding dynamic social negotiation processes. Second, GPT-4o performs well with Japanese and Korean group chats when guided by carefully designed prompts, with slightly higher performance observed for Korean conversations, likely due to more direct

communication patterns compared to Japanese conversations. Third, the state-of-the-art prompting methods used in the dynamic analysis (Table 2) demonstrated consistent performance, suggesting that LLMs exhibit robust capabilities in handling dynamic negotiations, regardless of the specific prompting method employed.

Our findings suggest that GPT-4o, one of the most widely used LLMs, demonstrates some capability in structuring group chat data, showing potential to reduce the high costs associated with human-labeled annotations for interpreting group decisions. This capability could accelerate the use of large-scale chat data to better understand activity and travel behavior. However, significant room for improvement remains in capturing implicit, context-dependent content, which can be addressed through advancements in prompt engineering and further development of LLM models.

REFERENCES

Bifulco, G. N., Cartenì, A., & Papola, A. (2010). An activity-based approach for complex travel behaviour modelling. European Transport Research Review, 2, 209-221.

Kahneman, D. (2011). Thinking, fast and slow. *Farrar, Straus and Giroux*.

Khot, T., Trivedi, H., Finlayson, M., Fu, Y., Richardson, K., Clark, P., & Sabharwal, A. (2022). Decomposed prompting: A modular approach for solving complex tasks. *arXiv preprint arXiv:2210.02406*.

Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., ... & Clark, P. (2024). Self-refine: Iterative refinement with self-feedback. Advances in Neural Information Processing Systems, 36.

Parady, G., Oyama, Y. and Chikaraishi, M., 2023. Text-aided Group Decision-making Process Observation Method (x-GDP): A novel methodology for observing the joint decision-making process of travel choices. Transportation, pp.1-25.

Puhe, M., Schippl, J., Fleischer, T., & Vortisch, P. (2021). Social network approach to analyze stability and variability of travel decisions. Transportation research record, 2675(9), 398-407.

Si, C., Shi, W., Zhao, C., Zettlemoyer, L., & Boyd-Graber, J. (2023). Getting more out of mixture of language model reasoning experts. *arXiv preprint arXiv:2305.14628*.

Tao, Y., Viberg, O., Baker, R. S., & Kizilcec, R. F. (2024). Cultural bias and cultural alignment of large language models. PNAS nexus, 3(9), pgae346.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems, 35, 24824-24837.