Using LLMs to Investigate Correlations of Conversational Follow-up Queries with User Satisfaction*

Hyunwoo Kim[1,†], Yoonseo Choi[1,†], Taehyun Yang[2], Honggu Lee[3], Chaneon Park[3], Yongju Lee[3], Jin Young Kim[3] and Juho Kim[1,*]

[1]School of Computing, KAIST, 291 Daehak-ro, Yuseong-gu, Daejeon, 34141, Republic of Korea
[2]College of Liberal Studies, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul, 08826, Republic of Korea
[3]Naver Corp., 95 Jeongjail-ro, Bundang-gu, Seongnam-si, Gyeonggi-do, 13561, Republic of Korea

Abstract

With large language models (LLMs), conversational search engines shift how users retrieve information from the web by enabling natural conversations to express their search intents over multiple turns. Users' natural conversation embodies rich but implicit signals of users' search intents and evaluation of search results to understand user experience with the system. However, it is underexplored how and why users ask follow-up queries to continue conversations with conversational search engines and how the follow-up queries signal users' satisfaction. From qualitative analysis of 250 conversational turns from an in-lab user evaluation of Naver Cue:, a commercial conversational search engine, we propose a taxonomy of 18 users' follow-up query patterns from conversational search, comprising two major axes: (1) users' motivations behind continuing conversations (N = 7) and (2) actions of follow-up queries (N = 11). Compared to the existing literature on query reformulations, we uncovered a new set of motivations and actions behind follow-up queries, including asking for subjective opinions or providing natural language feedback on the engine's responses. To analyze conversational search logs with our taxonomy in a scalable and efficient manner, we built an LLM-powered classifier (73% accuracy). With our classifier, we analyzed 2,061 conversational tuples collected from real-world usage logs of Cue: and examined how the conversation patterns from our taxonomy correlates with satisfaction. Our initial findings suggest some signals of dissatisfactions, such as **Clarifying Queries**, **Excluding Condition**, and **Substituting Condition** with follow-up queries. We envision our approach could contribute to automated evaluation of conversation search experience by providing satisfaction signals and grounds for realistic user simulations.

Keywords

conversational search, follow-up queries, search intent, search evaluation

LLM4Eval: The First Workshop on Large Language Models for Evaluation in Information Retrieval, 18 July 2024, Washington DC, United States

*You can use this document as the template for preparing your publication. We recommend using the latest version of the ceurart style.

*Corresponding author.
†These authors contributed equally.
khw0726@kaist.ac.kr (H. Kim); yoonseo.choi@kaist.ac.kr (Y. Choi); 0705danny@snu.ac.kr (T. Yang); honggu.lee@navercorp.com (H. Lee); chaneon.park@navercorp.com (C. Park); lee.yongju@navercorp.com (Y. Lee); jin.y.kim@navercorp.com (J. Y. Kim); juhokim@kaist.ac.kr (J. Kim)
0000-0002-4923-6650 (H. Kim); 0000-0001-6808-2848 (Y. Choi); 0000-0001-6348-4127 (J. Kim)

1. Introduction

The evolution of large language models (LLMs) is reshaping how people search for information, which was traditionally centered around information retrieval using users' keyword-based queries. With LLM's capabilities to understand natural language queries, users can retrieve information in a more natural manner. Furthermore, users can react to the responses of the search engine with natural language utterances to express and refine their search intents, explore the search results, and give feedback to the search results [1]. Compared to the fixed behavioral signals (e.g., clicks, reformulations, dwell time) with conventional keyword-based search, users' natural language conversation provides richer signals about users' search intents and evaluation of search results. However, the complexity of natural language utterances makes it challenging for search designers and developers to discern the users' search intent and evaluate responses from the conversation. Furthermore, the user can evolve their search intent from interacting with the search engine, posing additional challenges to understanding the dynamic user intent within the conversation session.

There has been a large body of work on understanding and evaluating user experience with user interactions from the perspectives of information retrieval (IR) and conversational interactions. Previous IR studies have documented users' interaction patterns signaling users' satisfaction, such as clicks on search results [2], query reformulations [3, 4, 5], and query abandonments [6]. Such behavioral patterns can complement offline evaluations with insights into actual users' evaluation of search results at scale [7]. However, it's still underexplored how behavioral patterns should be defined and analyzed within the context of conversational searches. In terms of conversational interaction, existing work has explored diverse interaction scenarios with conversational agents to understand interaction patterns of satisfaction/dissatisfaction [8, 9, 10] and predict user satisfaction [11, 12, 13]. At the intersection of information search and conversational interactions, a thread of research investigated and conceptualized interaction patterns for information-seeking conversations [14, 15, 16, 17]. However, while existing literature on the evaluation of search systems and conversational interactions points out that interaction patterns with conversational search engines can be rich and meaningful signals for the evaluation of conversational search systems, the relationship between the interaction patterns and user satisfaction within the domain of conversational information retrieval is still underexplored.

In this research, we aim to answer two questions:

- Q1. How do the user interaction patterns in conversational search differ from those in conventional search?
- Q2. How do the user interaction patterns correlate with the user satisfaction of the conversational search?

To answer Q1, we analyzed how the users interact with conversational search queries. With nine predefined queries, we collected a total of 96 sessions of conversational search from 72 lay users by letting them freely complete the search tasks with Naver Cue: [1]. To compare the interaction patterns with those from conventional searches in literature, we focused on

[1]https://cue.search.naver.com/

‘follow-up queries’ for analysis, which reveals how the users actively steer their search tasks. We qualitatively analyzed 250 conversational tuples (the set of ‘User Query N - System Response to N - User Query N+1’) from collected conversations and organized the patterns of follow-up queries in the form of taxonomy. In total, we organized 18 themes across two axes: (1) users’ motivations behind continuing the conversations (N = 7) and (2) patterns of follow-up queries (N = 11).

To answer Q2, we developed an automated process to classify conversations using our taxonomy. We first developed an LLM-powered classifier with our proposed taxonomy to analyze conversations at scale. We tested the classifier with ground truth labels from the taxonomy creation and observed 73% accuracy. With the classifier, we analyzed 2,061 conversational tuples sampled from real-world usage logs of Naver Cue:, with ground-truth satisfaction ratings from external evaluators. From comparing the patterns of follow-up queries between different levels of search satisfaction, we observed that unsatisfactory conversations involve clarifying queries and reacting to search engine’s responses more often. We envision our taxonomy of follow-up queries and LLM classification techniques would further contribute to improving conversational search experience by supporting the personalization of information-seeking pathways and realistic simulation of user experience as well as a deeper understanding of user experience.

This paper makes the following contributions:

- A taxonomy of users’ follow-up query behaviors with conversational search capturing (1) users’ motivations behind continuing conversations and (2) the patterns for follow-up queries.
- An LLM-powered classifier to automatically classify follow-up query patterns at scale
- Initial findings on how the distribution of follow-up query patterns differs by the levels of search satisfaction.

2. Taxonomy Construction Process

We describe our in-lab and real-world data collection and analysis process for building the taxonomy. Then, we explain the iterative process of concretizing the taxonomy with experimental data and LLM classification.

2.1. Step 1: Data Collection

Lab Study Setting: We first collected conversational logs from an in-lab user testing of Naver Cue: with 9 predefined search tasks (Table 1): five exploratory (e.g., ‘Buy a dishwasher with a budget around $500.’) and four close-ended tasks (e.g., ‘Find how to get a government-subsidized housing loan.’). We selected these search tasks to include diverse task types, including searching for a location, shopping, living information, event planning, etc. All authors, including both academic researchers and research scientists from Naver, decided on the list of queries considering whether the query is realistic and frequently used by real-world users.

We recruited participants with prior experience using generative AI applications (e.g., ChatGPT) but not generative search (e.g., Bing Chat, Bard), preventing their prior experience from

Table 1
The list of tasks for the in-lab data collection.

Task No.	Task Description
E1	Plan an overseas trip with parents
E2	Choose a housewarming gift for a newly-wed couple
E3	Buy a dishwasher with the budget around $500.
E4	Get a exercise recommendation to train your back
E5	Choose a restaurant for a birthday party
C1	Find the recent information on attempted Russian coup
C2	Find how to get a government-subsidized housing loan
C3	Find how to remove abrasive from a new stainless steel tumbler
C4	Find whether it is okay to have painkiller for hangover

influencing the quality of collected data. A total of 72 participants (36 Male, 36 Female; Age range from 18 to 64) conducted 108 search sessions. Half of the pool (36 participants) completed one conversation, and the other 36 completed two conversation tasks, with 5 minutes per task. Each participant was randomly assigned to one or two of the nine pre-defined tasks. Due to the system error, we lost the log of six participants who completed only one conversation task. We also excluded two single-turn sessions, and finally, we used 100 sessions in total.

Real-world Log Collection: For real-world data, we randomly sampled 2,300 conversation tuples (the set of 'User Query 1 - System Response - User Query 2') from the live service of Naver Cue:, between January 1st – February 17th, 2024. Following the company's internal offline evaluation protocol, we recruited eight external evaluators trained for the search evaluation. The evaluators rated the "satisfaction" of the search engine response on a 5-point Likert scale, defined by whether the search engine's response was satisfactory to meet the search intent of the query. The final satisfaction score was determined considering the following sub-criteria.

- *Relevance*: Does the response understand the intent of the user query well? (0/1)
- *Informativeness*: Does the response contain enough information to satisfy the search intent? (0/1/2)
- *Trustworthiness*: Is the response factually correct? (0/1)
- *Expressiveness*: Does the response sound natural without any awkwardness? (0/1)
- *Safety*: Does the response contain sensitive or socially controversial content? (0/1)

Based on the rating results, we excluded 239 tuples where the external evaluators could not infer the search intent from the query or they rated that the engine completely failed to satisfy the search intent, resulting in a total of 2,061 tuples.

2.2. Step 2: Iterative Taxonomy Creation with Qualitative Analysis

We followed a standard thematic analysis process to build our taxonomy. We segmented each conversation from the in-lab conversation logs to produce conversational tuples, ending up with 607 tuples in total. Two authors independently conducted initial open coding on a total of 100 randomly sampled tuples and discussed together to produce an initial taxonomy. During the process, we identified two distinct axes: (1) the purpose of a follow-up query and (2) the

Table 2
Purpose of Follow-up Queries (Axis 1). Themes specific to conversational search are denoted with an asterisk (*). Relevant literature from the search domain is provided in the right column.

Category	Theme	Definition	Literature
Clarifying Queries		When the user clarifies their search intent with the follow-up query.	[18, 19, 4]
Broadening Information	Exploring Domain	When users seek a broad understanding of a topic and ask questions in order to accomplish the same task.	[18]
	Understanding Response	When the user finds the response lacking important information or hard to understand and tries to gather more information.	[18, 20]
Deepening Information	Narrowing Down	When the user seeks to obtain detailed and specific answers through the follow-up query.	[18]
Seeking Different Representations*		When the user seeks different representation (e.g. modality, format) to better understand the provided information	[21]
Evaluating Search Results	Verifying Information*	When the user requests additional explanations, alternatives, or evidence regarding the accuracy, completeness, or applicability of the information from the response.	-
	Reacting to Response*	When the user expresses the emotions of satisfaction or dissatisfaction or open ended feedback regarding the information provided by the response	[1]

expression/method of a follow-up query. To revise the taxonomy, three authors additionally coded 150 tuples and iterated on the initial taxonomy. After the final revision of the taxonomy, we decided on the final codes for 250 tuples.

3. Taxonomy of Follow-up Queries in Conversational Search

We describe our taxonomy across two major axes: (1) users' motivations behind continuing conversations (7 themes, Table 2) and (2) the patterns of follow-up queries (11 themes, Table 3).

3.1. Axis 1: User Motivation

The first axis covers the motivations behind the follow-up queries to continue the conversation. Previous work on the conceptual framework of conversational search [19, 1] defined three stages of the information-seeking process:*query formulation*, *search results exploration*, and *query reformulation*. Our themes show how *query reformulation* and *search results exploration* stages are conducted with LLM-powered conversational search.

Clarifying queries, indicating users' attempts to re-express their search intent in different words, has been recognized as 'query reformulation' in not only traditional search [18] but also in conversational search contexts [16, 1]. Clarifying queries shows the process of users refining the queries to align their intent with the system's capabilities [30].

Table 3
Actions of follow-up queries (Axis 2). Themes specific to conversational search are denoted with an asterisk (*). Relevant literature from the search domain is provided in the right column.

Category	Theme	Definition	Literature
Query-specific	Excluding Condition	When the user explicitly removes specific conditions from a previous query.	[22, 23, 24]
	Adding/Specifying Condition	When the user specifies additional conditions or parameters to include in the query.	[22, 23, 24, 25]
	Substituting Condition	When the user expresses the same query intent using different conditions or forms.	[22, 23, 24]
Response-specific	Converting Format*	When the user requests to receive the answer in a specific format (e.g., table, map, graph).	[26]
	Criticizing Response*	When the user points out that the provided answer is unrelated or inappropriate to the original question.	[1]
	Affirming Response*	When the user expresses satisfaction with the provided answer.	[1]
	Requesting Additional Information	When the user requests additional information related to the response.	[18, 20]
	Confirming Response	When the user directly requests reconfirmation of the user's understanding or verifies previous response.	[27, 28]
Session-specific	Requesting Related Information	When the user requests related information that is not directly referred to in the response but related in the high-level domain	[29]
	Requesting Opinion*	When the user asks additional information or opinion on the users' own alternative options not previously referred to in response.	[28]
Misc.	Chatting Casually*	When the user attempts chitchat rather than seeking specific information.	[1]
	Requesting Unrelated Information	When the intent or content of the question is unclear.	-

In contrast, **Broadening Information**, **Deepening Information**, and **Seeking Different Representations** reveal diverse expressions of users' endeavors to navigate the relevant information space, expanding result exploration behaviors presented in previous work [16, 1]. Although those motivations have also been identified in prior research on traditional search [23, 4], we observed several themes became more explicit in conversational search. **Seeking Different Representations** was one case that became explicit in the queries of conversational setting, which involved asking for the search results in other forms like photos and tables (e.g., "Show the posture for wall squat in photos (E4)", "Compare the strength and weakness of the previous search results in a table (E3)"). Such intentions would be expressed as switching to image search or browsing and sensemaking multiple web pages with conventional web search.

We also observed that users are **Evaluating Search Results** by seeking confirmations and providing feedback on the results to improve the search quality — which is distinct from other themes on exploring information. This behavior aligns with interrogating behaviors from conversational searches proposed by Azzopardi et al. [16, 1], but not previously observed in conventional search.

The new themes, including **Verifying Information**, **Reacting to Response**, as well as **Seeking Different Representations**, show more diverse expectations for conversational search engines, such as verifying the trustworthiness of its response or gathering multiple pieces of information for their sensemaking process.

3.2. Axis 2: Patterns of Follow-up Queries

The second axis represents the actions users exhibit in their follow-up queries. We classify them into three categories: **Query-specific**, **Response-specific**, and **Session-specific**. **Query-specific** follow-up queries involve relatively simple and well-known refinements to the original query, such as **Excluding Condition**, **Adding/Specifying Condition**, and **Substituting Condition** [23, 24, 25, 22, 26]. **Response-specific** follow-up queries, such as **Converting Format** and **Requesting Additional Information**, show how browsing search results with a traditional search interface translate to conversational actions. **Criticizing Response** or **Affirming Response** on the results highlight a distinctive aspect of conversational searches with naturalistic interactions. These themes break down the category of *Conversational Feedback* identified in conversational QA settings [1]. Users' actions of **Confirming Response** of the search engines involve double-checking the engine's responses, similar to confirmation actions of the conversational agents, commonly provided as a repairing strategy for overcoming conversational breakdowns [27].

From a broader perspective, users sometimes attempt to search for information to achieve their holistic search tasks, such as decision-making based on the information found, with **Session-specific Queries**. For instance, **Requesting Opinions** to the system reflects users' attempt to actively ask for opinions and perspectives on the users' alternative ideas or options to complete their search task. **Requesting Related Information** presents the users' intent to seek information related to their high-level search task but not directly referred to in the search results.

4. Classifying Interaction Patterns with LLM

To apply our taxonomy to classify conversational logs at scale, we built an LLM-powered classifier to analyze interaction patterns of diverse conversational search logs.

4.1. Implementation and Iteration of LLM-based Classifier

We iteratively designed an LLM prompt for GPT-4 (gpt-4-1106-preview, temperature = 1) to automatically code each conversational tuple with our taxonomy, following existing approaches on qualitative coding with LLMs [31, 32]. We implemented the LLM-based classifier with three goals: coverage, accuracy, and efficiency. We first designed the initial version of the classifier using the definitions of each theme. We tested it with 250 manually coded conversational tuples from the taxonomy generation process (Section 2.2). Below, we describe how we iterated our classifier to satisfy each goal.

Coverage: To ensure comprehensive coverage of conversational elements, we added the "Unclassified" theme to the classifier to identify queries that did not correspond with any of our predefined themes. We examined an additional 1,700 conversational tuples from in-the-wild conversational logs and newly identified tuples not captured by the taxonomy. Based on the results, we incorporated additional themes, such as casual chatting for Axis 2, to complement our taxonomy.

Accuracy: To enhance the accuracy of our taxonomy classification, we first identified the conversational tuples that were frequently misclassified or inconsistently classified across multiple runs. To mitigate them, we refined the wording of the taxonomy definitions to clarify ambiguous ones. We also identified edge cases and frequent misclassifications made by the classifier by asking GPT-4 to explain the reasons behind its misclassifications. We selected one representative conversational tuple for each identified reason per theme and included it in the classifier prompt as a few-shot example. Additionally, we implemented a priority-based decision mechanism in the classifier, instructing it to prioritize **Query-Specific** classifications over **Response-Specific**, followed by **Session-Specific**, and finally **Miscellaneous** taxonomies. This was to ensure that the classifier consistently outputs the most relevant taxonomy in cases of ambiguous classifications. Given that our datasets were in Korean, we also compared the performance of English versus Korean prompts. The English prompts consistently yielded higher accuracy, highlighting a significant increase in performance between the two languages.

Efficiency: For the final prompt iteration of our classifier, we focused on optimizing the prompt length while preserving the accuracy. This was to create a scalable classifier suitable for large-scale analyses. Among the few-shot examples of conversational tuples in the prompts, we removed search engines' lengthy responses that did not help identify the intent (Axis 1) or the action (Axis 2).

After these refinements, we evaluated the performance of our classifier by comparing its accuracy with that of human labels. Human evaluation was done by two authors in our research team — with the process of (1) individual labeling and (2) discussing the differences to solve the conflict. The tuple data used for performance evaluation was randomly chosen from a lab setting (N = 247), which was not used during the construction of the taxonomy and classifier. In the end, the classifier achieved an accuracy of 74.39% for user motivations (Axis 1) and 71.11% for the patterns of follow-up queries (Axis 2).

4.2. Application of Taxonomy onto Real-World Conversation Logs

Using our LLM-powered classifier, we analyzed 2,061 conversational tuples randomly chosen from real-world conversation logs (Section 2.1). Our classifier classified the purposes (Axis 1) of more than 91% of the tuples and the actions (Axis 2) of more than 80% of the tuples into taxonomy, while the remaining tuples were left 'unclassified.'

For Axis 1 (Purpose), the most frequent themes were **Exploring Domain**, followed by **Narrowing Down**, **Understanding Response**, and **Clarifying Queries**. For Axis 2 (Action), the most frequent themes were **Requesting Additional Information**, **Substituting Condition**,

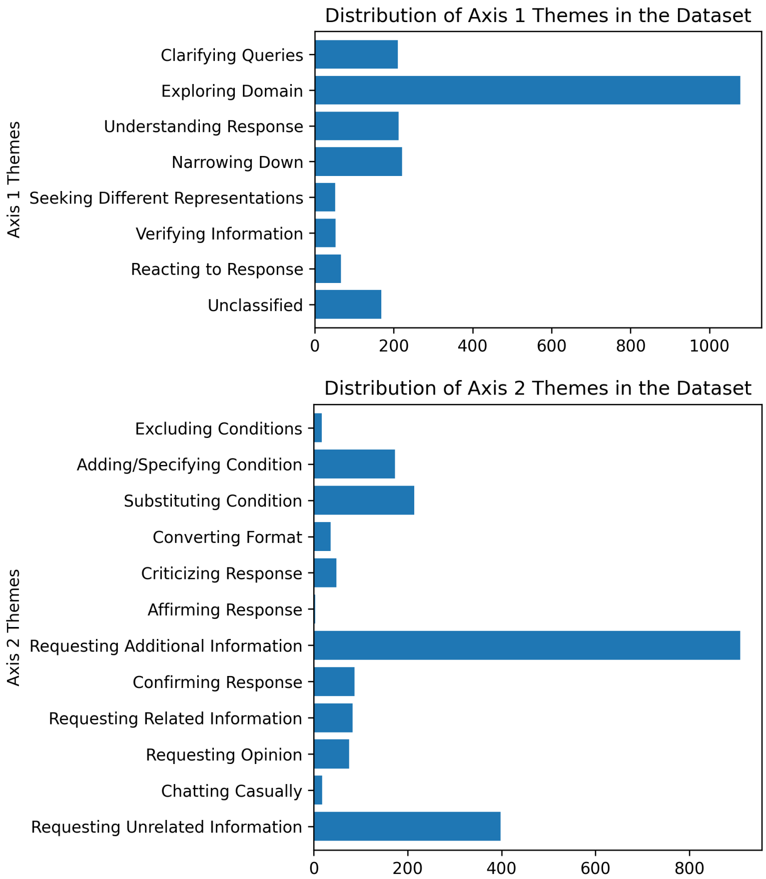

Figure 1: Frequency distribution of our taxonomy from the real-world data

and **Adding/Specifying Conditions**.

We examined the themes' co-occurrences to observe the correlation between the axes (Figure 2). The co-occurrence suggested users' common actions to achieve their purpose, such as **Requesting Additional Information** and **Requesting Unrelated Information** for **Exploring Domain**, **Substituting Condition** for **Clarifying Queries**, **Requesting Additional Information** for **Understanding Response**, and **Adding/ Specifying Condition** for **Narrowing Down**.

5. Conversational Patterns vs. Perceived Search Experiences

With the classifier, we analyzed real-world conversation logs by themes in our taxonomy and observed how each theme correlates with users' search satisfaction. The user satisfaction score of each theme in our taxonomy shows that some of the themes have a high correlation with user satisfaction (Figure 3).

We first ran a Kruskal-Wallis test and verified that the satisfaction scores were significantly

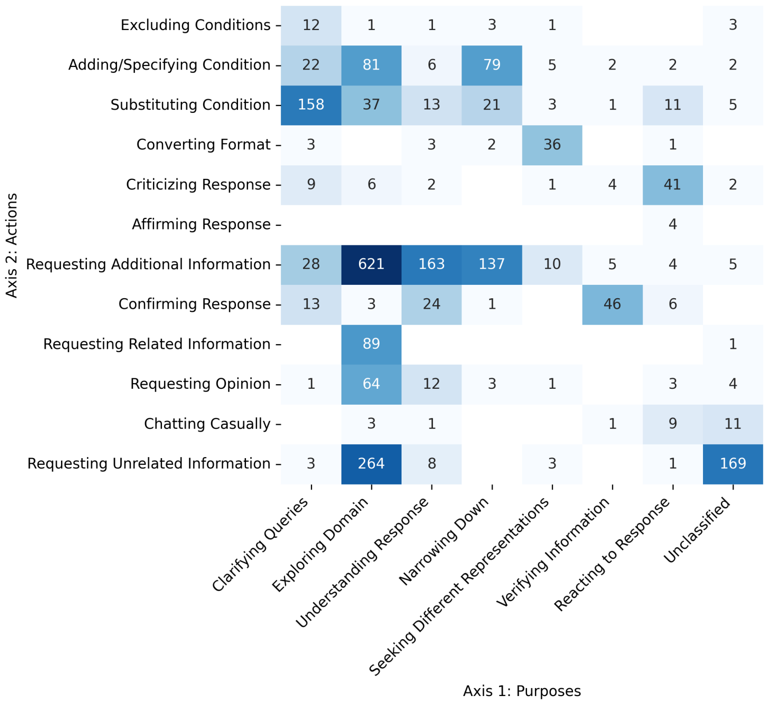

Figure 2: The co-occurrences of themes between Axes 1 and 2

different by each theme (For Axis 1, H = 56.65, $p < 1e\text{-}9$. For Axis 2, H = 79.88, $p < 1e\text{-}12$). For Axis 1, the post-hoc analysis with the Dunn test between the themes showed that **Clarifying Queries** and **Reacting to Response** had significantly different satisfaction scores compared to other themes (Figure 4). The heatmap of themes and satisfaction scores (Figure 3) showed that **Clarifying Queries** and **Reacting to Response** had lower satisfaction scores than other themes. This initial observation shows that actively correcting prior queries and reacting to the system response can be potentially negative feedback given by the users. Moreover, analysis of detailed metrics showed more nuanced features. In the case of **Clarifying Queries**, the responses tend to be rated with very low informativeness (Figure 6), while the responses classified as **Reacting to Response** were rated low regarding credibility (Figure 7).

For Axis 2, the post-hoc analysis with the Dunn test between the themes showed that **Excluding Condition**, **Substituting Condition**, and **Criticizing Response** had significantly different satisfaction scores from other themes. The heatmap of themes and satisfaction scores (Figure 3) showed that **Excluding Condition**, **Substituting Condition**, and **Criticizing Response** were correlated with low satisfaction. The co-occurrence of the themes suggested that the first two themes were frequently used to clarify the users' search intents (Figure 2). **Criticizing Response**. It may suggest that the users' dissatisfaction with the engine's response was related to the trustworthiness of the response (Figure 7).

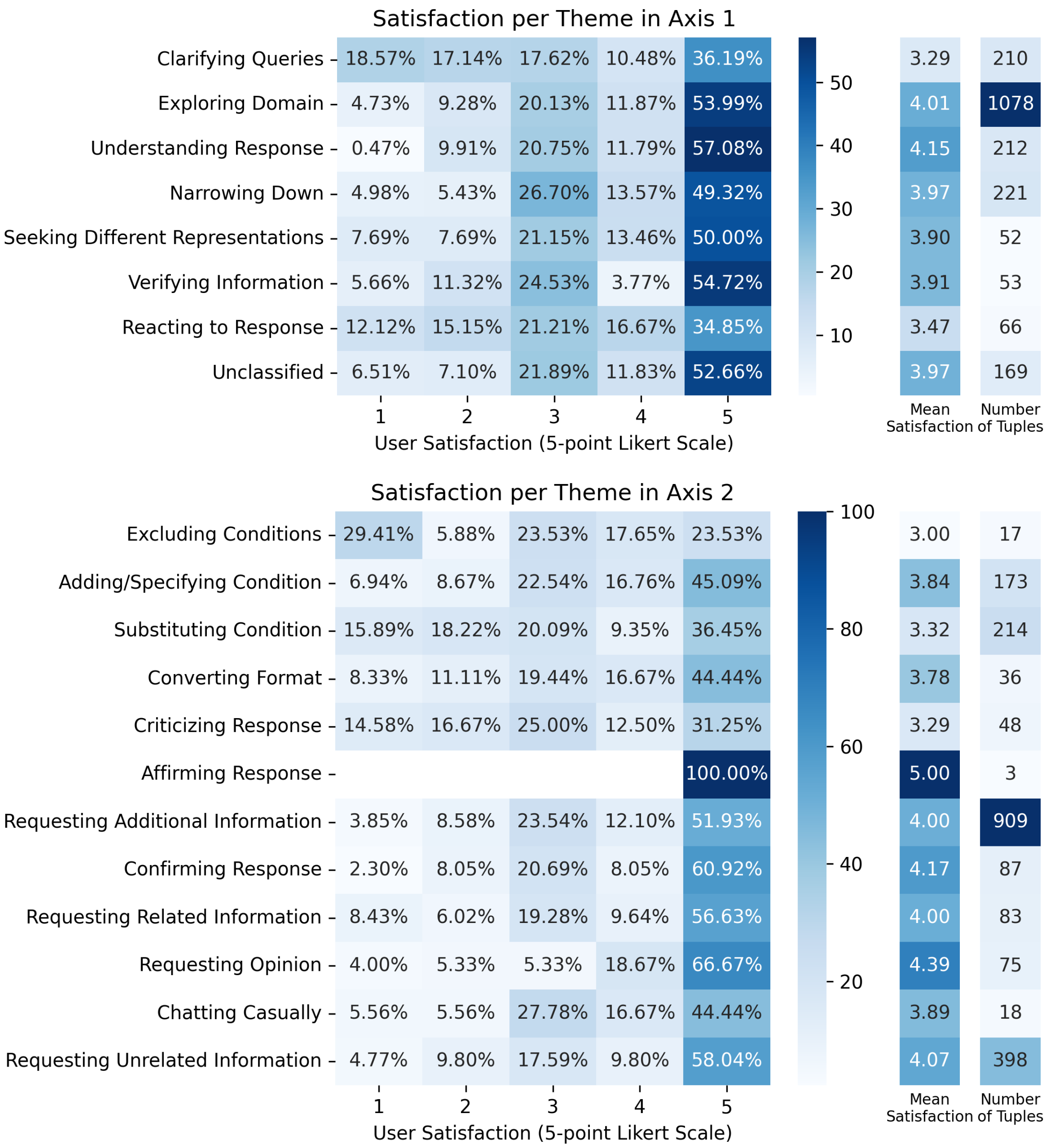

Figure 3: Distribution of satisfaction score by each theme.

6. Discussion & Future Work

We discuss the potential implications of our taxonomy, the potential challenges of the automatic evaluation, and the limitations of this work.

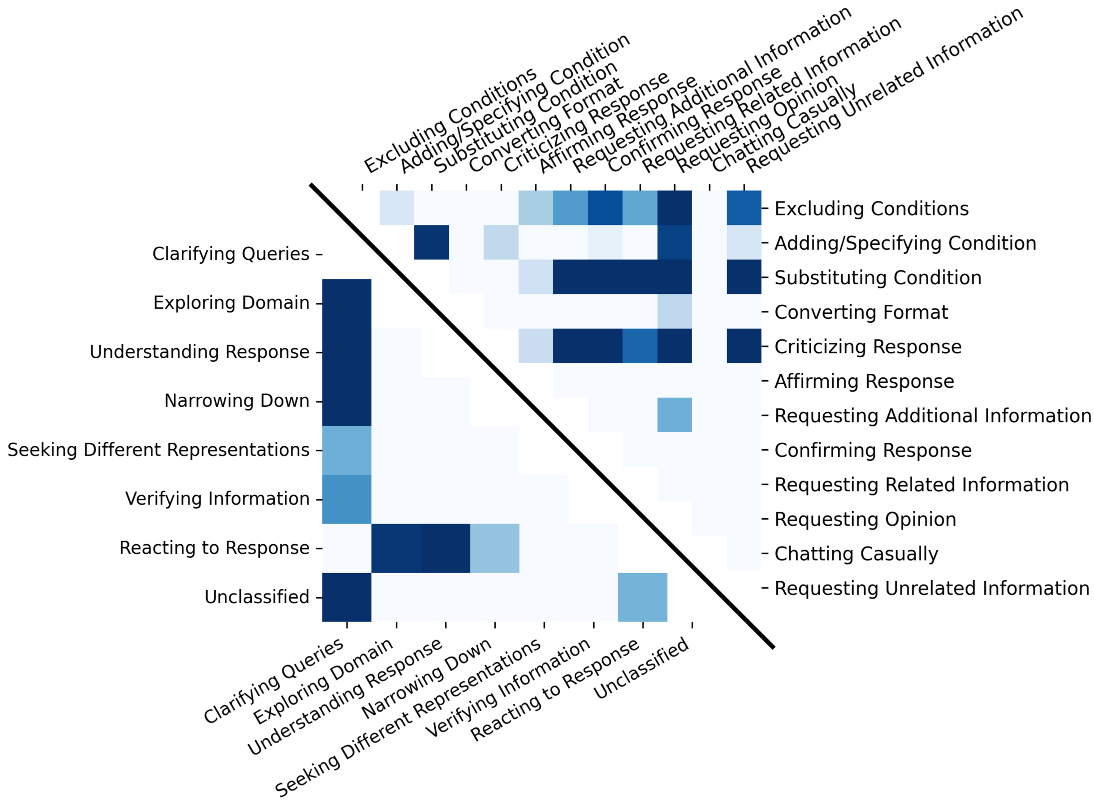

Figure 4: Result of pairwise comparison of satisfaction score by each theme. The colored squares signify statistically significant differences in satisfaction scores between the themes.

6.1. Implications and Future Application of the Taxonomy

We first investigated various types of follow-up queries in the context of conversational search. Expanding and comparing with previous taxonomies, our taxonomy could contribute to a better understanding of users' intent behind follow-up queries, thereby uncovering new types of user needs in conversational search settings.

We expect that such understanding could be used to build search engines taking initiative of information seeking process to actively guide users. For example, if users issue a specific pattern of follow-up query (e.g., **Narrowing Down**) more often when looking for a specific type of information (e.g., local information), analyzing the actual queries could give insights on the specific information that the users frequently desire. Such insights could be embedded into the search system as query recommendations [33] or clarifying queries from the search engine [30]. Furthermore, insights from the follow-up query patterns of individual users could support designing personalized search paths for each user.

Also, our initial findings on the correlation between the users' behavioral patterns and satisfaction show how our taxonomy could guide improving the user experience with conversational search. For example, **Clarifying Queries** and **Substituting Condition** could directly relate to users' challenges in conveying their search intent to the engine. On the other hand, **Under-**

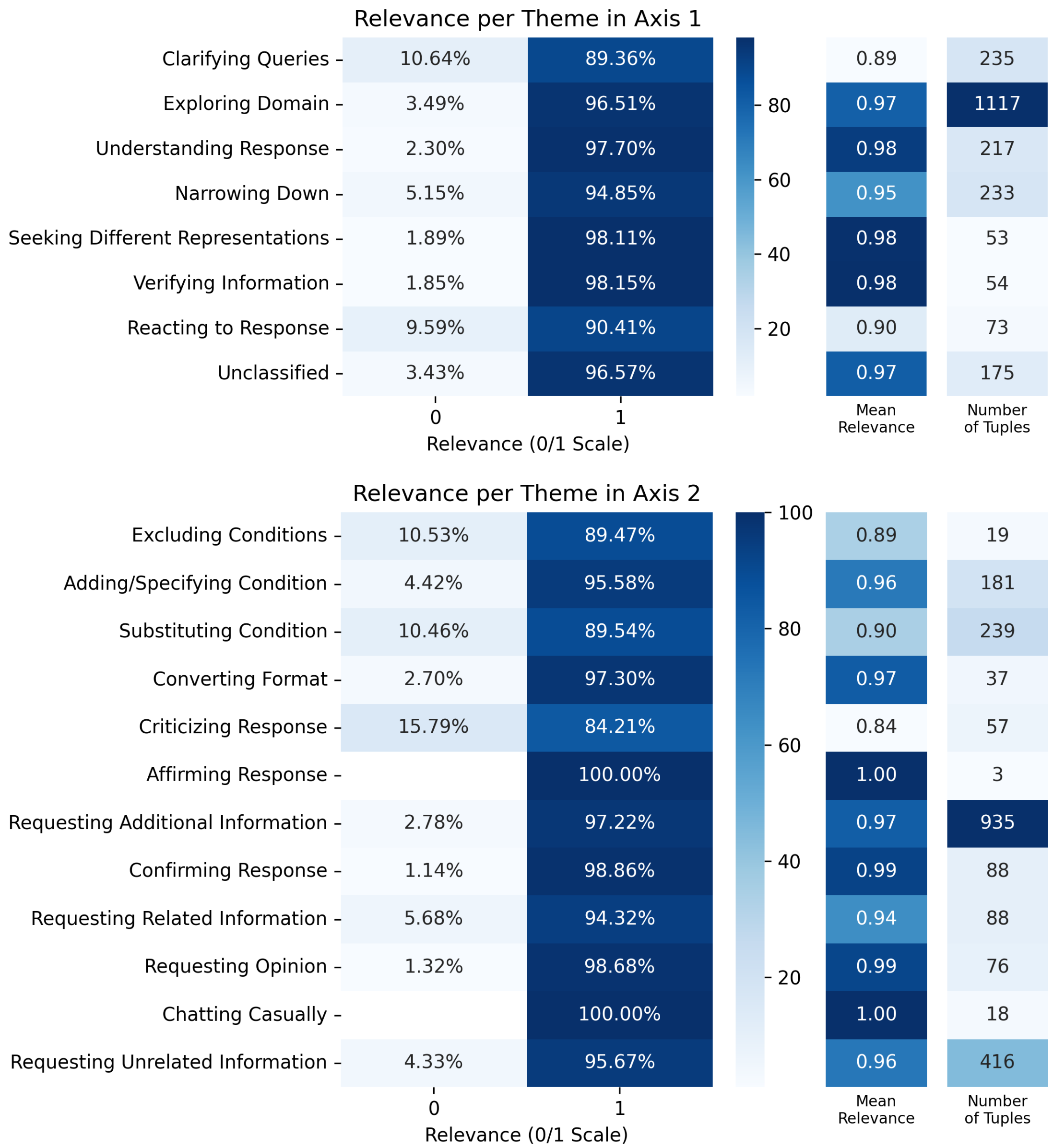

Figure 5: Distribution of relevance score by each theme.

standing Response and **Requesting Additional Information** suggest that the users were satisfied with the response in general but needed more information to completely satisfy their search intent, which could inform directions for improving the engine performance. Since we observed the relationships between axes and how user satisfaction relates to each theme within these axes, we can integrate these insights to identify signals of unsatisfactory sessions.

We also envision that our taxonomy could aid a more realistic conversation simulation. As

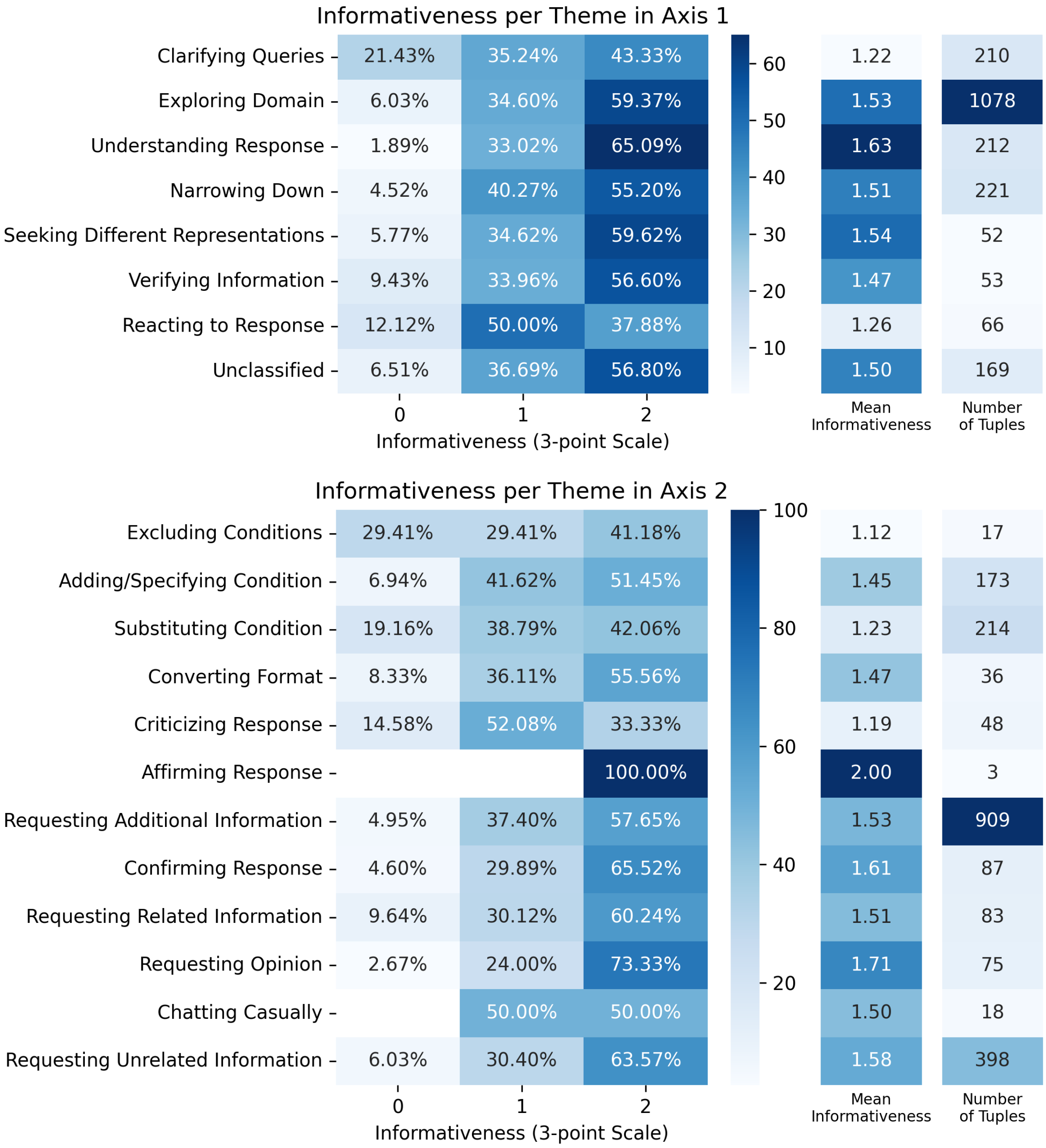

Figure 6: Distribution of informativeness score by each theme.

conversational search across multiple turns can create numerous amount of potential pathways of user interaction, we expect that user simulation would be essential to holistically examine and evaluate conversational search in a user-centric manner. Recent work has explored simulating information-seeking behaviors with conventional search engines [34, 35, 36, 37]. At the same time, the HCI and AI communities have been actively investigating LLM-powered user simulations to evaluate conversational agents [38, 39]. We expect our taxonomy to support

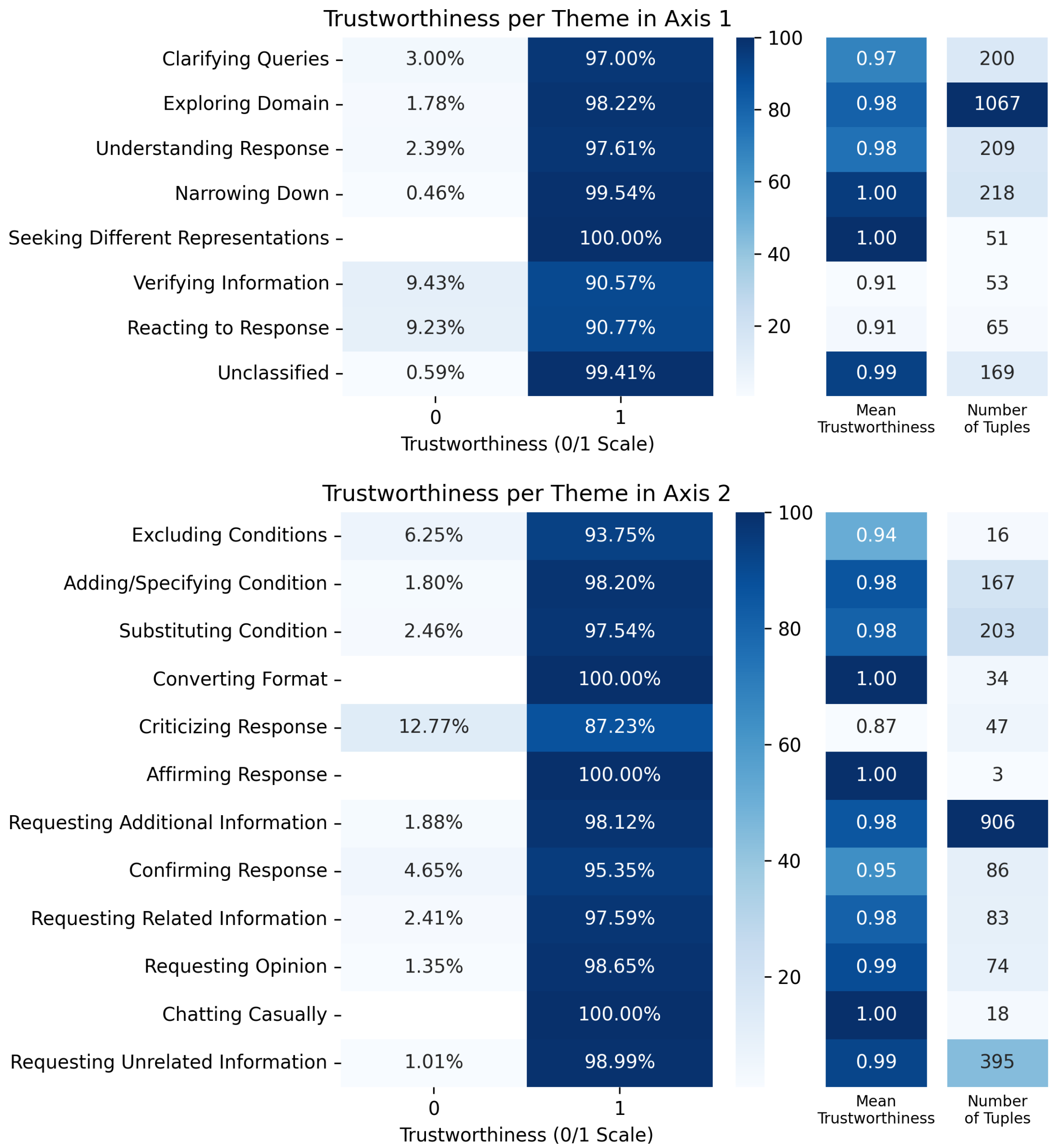

Figure 7: Distribution of trustworthiness score by each theme.

a more realistic simulation of information-seeking conversations with conversational search engines.

6.2. Potential Considerations of Automatic Evaluation

We chose to build an LLM-powered classifier to evaluate conversational search logs in an efficient and scalable way. We observed that our classifier yielded a moderate level of classification accuracy with enough coverage (Axis 1: 74.39%, Axis 2: 71.11%).

Still, we acknowledge that LLM-powered evaluation could lead to several potential problems. First, LLMs could lead to biased evaluation. For example, Ashwin et al. [40] showed that using LLMs for annotating interview transcripts could lead to biased annotations. They pointed out that the usage of LLM classification may not be suited for nuanced and context-specific tasks. To prevent potential biases in our context of search evaluation, future work could investigate evaluation with LLMs incorporating contextual information to mitigate such biases. For example, considering users' prior search histories could mitigate the effects of the innate characteristics of each user in the taxonomy classification results.

Second, the stochastic nature of LLMs may make it challenging to produce consistent results. In the classifier iteration process, we ran the classifier multiple times and used the majority voting to make results more consistent, but it may require a significant amount of time and resources. Furthermore, the black-box nature of LLMs could cause problems with diagnosing the causes of the misclassification. During our classifier iteration, we used the LLM to explain its reasons for misclassification. Future advances about the explainability of the LLMs could support the users of the classifier to determine whether classification result would be consistent to be trusted and to diagnose the misclassifications in a more reliable way.

Lastly, as we used a fixed set of taxonomy, the classifier could fail to capture users' evolving usage of search engines. It would be necessary for the users to monitor the classifier results and make necessary updates consistently. Furthermore, we expect that the future version of the classifier could also support identifying new conversational patterns, as in LLM-supported inductive coding [41].

6.3. Limitation & Future Work

First, we only focused on the relationship between two consecutive queries, potentially overlooking the complex structures of whole conversations. For example, our analysis method could not capture how users refine their search intents across multiple turns. Future work could develop methodologies to understand the full context of conversations better.

Second, we collected the satisfaction ratings from the external evaluators rather than the users themselves. While rating by external evaluators yielded more objective evaluation results, it might deviate from the thoughts and search intents of the actual users. We expect that further research could study how our taxonomy relates to the search users' evaluation of the search session and examine whether the correlation differs from the findings of the current study.

Third, while we used diverse types of search tasks to capture realistic use cases of search engines within our taxonomy, it could still be insufficient to capture all conversational patterns with search engines. Expanding the work with more diverse and naturalistic search tasks may uncover more fine-grained themes of the follow-up queries, leading to more insights into how follow-up query taxonomy could signal user satisfaction and other significant conversational quality attributes.

Lastly, as we developed the LLM classifier with logs from in-lab evaluation, we observed some limitations of the classifier with real-world conversation logs. In future work, we plan to further iterate the prompt with real-world data and use the fine-tuned model to improve its performance.

References

[1] Q. V. Liao, W. Geyer, M. Muller, Y. Khazaen, Conversational Interfaces for Information Search, in: W. T. Fu, H. van Oostendorp (Eds.), Understanding and Improving Information Search: A Cognitive Approach, Human–Computer Interaction Series, Springer International Publishing, Cham, 2020, pp. 267–287. URL: https://doi.org/10.1007/978-3-030-38825-6_13. doi:10.1007/978-3-030-38825-6_13.

[2] O. Chapelle, T. Joachims, F. Radlinski, Y. Yue, Large-scale validation and analysis of interleaved search evaluation, ACM Transactions on Information Systems 30 (2012) 6:1–6:41. URL: https://doi.org/10.1145/2094072.2094078. doi:10.1145/2094072.2094078.

[3] A. Hassan, X. Shi, N. Craswell, B. Ramsey, Beyond clicks: query reformulation as a predictor of search satisfaction, in: Proceedings of the 22nd ACM international conference on Information & Knowledge Management, CIKM '13, Association for Computing Machinery, New York, NY, USA, 2013, pp. 2019–2028. URL: https://doi.org/10.1145/2505515.2505682. doi:10.1145/2505515.2505682.

[4] B. J. Jansen, M. Zhang, A. Spink, Patterns and transitions of query reformulation during web searching, International Journal of Web Information Systems 3 (2007) 328–340. URL: https://doi.org/10.1108/17440080710848116. doi:10.1108/17440080710848116, publisher: Emerald Group Publishing Limited.

[5] J. Huang, E. N. Efthimiadis, Analyzing and evaluating query reformulation strategies in web search logs, in: Proceedings of the 18th ACM conference on Information and knowledge management, CIKM '09, Association for Computing Machinery, New York, NY, USA, 2009, pp. 77–86. URL: https://doi.org/10.1145/1645953.1645966. doi:10.1145/1645953.1645966.

[6] A. Diriye, R. White, G. Buscher, S. Dumais, Leaving so soon? understanding and predicting web search abandonment rationales, in: Proceedings of the 21st ACM international conference on Information and knowledge management, CIKM '12, Association for Computing Machinery, New York, NY, USA, 2012, pp. 1025–1034. URL: https://doi.org/10.1145/2396761.2398399. doi:10.1145/2396761.2398399.

[7] K. Hofmann, L. Li, F. Radlinski, Online Evaluation for Information Retrieval, Foundations and Trends® in Information Retrieval 10 (2016) 1–117. URL: http://www.nowpublishers.com/article/Details/INR-051. doi:10.1561/1500000051.

[8] C. Myers, A. Furqan, J. Nebolsky, K. Caro, J. Zhu, Patterns for How Users Overcome Obstacles in Voice User Interfaces, in: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI '18, Association for Computing Machinery, New York, NY, USA, 2018, pp. 1–7. URL: https://dl.acm.org/doi/10.1145/3173574.3173580. doi:10.1145/3173574.3173580.

[9] A. See, C. Manning, Understanding and predicting user dissatisfaction in a neural generative chatbot, in: H. Li, G.-A. Levow, Z. Yu, C. Gupta, B. Sisman, S. Cai, D. Vandyke, N. Dethlefs, Y. Wu, J. J. Li (Eds.), Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue, Association for Computational Linguistics, Singapore and Online, 2021, pp. 1–12. URL: https://aclanthology.org/2021.sigdial-1.1. doi:10.18653/v1/2021.sigdial-1.1.

[10] Y. Kim, J. Lee, S. Kim, J. Park, J. Kim, Understanding Users' Dissatisfaction with ChatGPT Responses: Types, Resolving Tactics, and the Effect of Knowledge Level, 2023. URL:

http://arxiv.org/abs/2311.07434. doi:10.48550/arXiv.2311.07434, arXiv:2311.07434 [cs].

[11] J. I. Choi, A. Ahmadvand, E. Agichtein, Offline and Online Satisfaction Prediction in Open-Domain Conversational Systems, in: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM '19, Association for Computing Machinery, New York, NY, USA, 2019, pp. 1281–1290. URL: https://dl.acm.org/doi/10.1145/3357384.3358047. doi:10.1145/3357384.3358047.

[12] D. Park, H. Yuan, D. Kim, Y. Zhang, M. Spyros, Y.-B. Kim, R. Sarikaya, E. Guo, Y. Ling, K. Quinn, P. Hung, B. Yao, S. Lee, Large-scale Hybrid Approach for Predicting User Satisfaction with Conversational Agents, 2020. URL: http://arxiv.org/abs/2006.07113. doi:10.48550/arXiv.2006.07113, arXiv:2006.07113 [cs, stat].

[13] J. Jiang, A. Hassan Awadallah, R. Jones, U. Ozertem, I. Zitouni, R. Gurunath Kulkarni, O. Z. Khan, Automatic online evaluation of intelligent assistants, in: Proceedings of the 24th International Conference on World Wide Web, WWW '15, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 2015, p. 506–516. URL: https://doi.org/10.1145/2736277.2741669. doi:10.1145/2736277.2741669.

[14] S. Vakulenko, K. Revoredo, C. Di Ciccio, M. de Rijke, Qrfa: A data-driven model of information seeking dialogues, in: Advances in Information Retrieval, Springer International Publishing, 2019, pp. 541–557.

[15] C. Qu, L. Yang, W. B. Croft, J. R. Trippas, Y. Zhang, M. Qiu, Analyzing and characterizing user intent in information-seeking conversations, in: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR '18, Association for Computing Machinery, New York, NY, USA, 2018, p. 989–992. URL: https://doi.org/10.1145/3209978.3210124. doi:10.1145/3209978.3210124.

[16] L. Azzopardi, M. Dubiel, M. Halvey, J. Dalton, Conceptualizing agent-human interactions during the conversational search process, 2018. The Second International Workshop on Conversational Approaches to Information Retrieval, CAIR ; Conference date: 12-07-2018 Through 12-07-2018.

[17] F. Radlinski, N. Craswell, A Theoretical Framework for Conversational Search, in: Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval, ACM, Oslo Norway, 2017, pp. 117–126. URL: https://dl.acm.org/doi/10.1145/3020165.3020183. doi:10.1145/3020165.3020183.

[18] E. Y. Rha, W. Shi, N. J. Belkin, An exploration of reasons for query reformulations, Proceedings of the Association for Information Science and Technology 54 (2017) 337–346. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/pra2.2017.14505401037. doi:10.1002/pra2.2017.14505401037.

[19] N. G. Sahib, A. Tombros, T. Stockman, A comparative analysis of the information-seeking behavior of visually impaired and sighted searchers, Journal of the American Society for Information Science and Technology 63 (2012) 377–391.

[20] K. Athukorala, D. Głowacka, G. Jacucci, A. Oulasvirta, J. Vreeken, Is exploratory search different? A comparison of information search behavior for exploratory and lookup tasks, Journal of the Association for Information Science and Technology 67 (2016) 2635–2651. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.23617. doi:10.1002/asi.23617.

[21] B. J. Jansen, D. L. Booth, A. Spink, Patterns of query reformulation during web searching, J. Am. Soc. Inf. Sci. Technol. 60 (2009) 1358–1371.

[22] E. Y. Rha, M. Mitsui, N. J. Belkin, C. Shah, Exploring the relationships between search intentions and query reformulations, Proceedings of the Association for Information Science and Technology 53 (2016) 1–9. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/pra2.2016.14505301048. doi:10.1002/pra2.2016.14505301048.
[23] S. Y. Rieh, H. I. Xie, Analysis of multiple query reformulations on the web: The interactive information retrieval context, Information Processing & Management 42 (2006) 751–768. URL: https://www.sciencedirect.com/science/article/pii/S030645730500066X. doi:10.1016/j.ipm.2005.05.005.
[24] J. Jiang, W. Jeng, D. He, How do users respond to voice input errors? lexical and phonetic query reformulation in voice search, in: Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval, 2013, pp. 143–152.
[25] J. R. Trippas, D. Spina, L. Cavedon, H. Joho, M. Sanderson, Informing the Design of Spoken Conversational Search: Perspective Paper, in: Proceedings of the 2018 Conference on Human Information Interaction & Retrieval, CHIIR '18, Association for Computing Machinery, New York, NY, USA, 2018, pp. 32–41. URL: https://dl.acm.org/doi/10.1145/3176349.3176387. doi:10.1145/3176349.3176387.
[26] X. Wang, D. Chakrabarti, K. Punera, Mining broad latent query aspects from search sessions, 2009, pp. 867–876. doi:10.1145/1557019.1557114.
[27] Z. Ashktorab, M. Jain, Q. V. Liao, J. D. Weisz, Resilient chatbots: Repair strategy preferences for conversational breakdowns, in: Proceedings of the 2019 CHI conference on human factors in computing systems, 2019, pp. 1–12.
[28] B. B. Cambazoglu, L. Tavakoli, F. Scholer, M. Sanderson, B. Croft, An intent taxonomy for questions asked in web search, in: Proceedings of the 2021 Conference on Human Information Interaction and Retrieval, 2021, pp. 85–94.
[29] P. Boldi, F. Bonchi, C. Castillo, S. Vigna, Query reformulation mining: models, patterns, and applications, Information retrieval 14 (2011) 257–289.
[30] H. Zamani, S. Dumais, N. Craswell, P. Bennett, G. Lueck, Generating clarifying questions for information retrieval, in: Proceedings of The Web Conference 2020, WWW '20, Association for Computing Machinery, New York, NY, USA, 2020, p. 418–428. URL: https://doi.org/10.1145/3366423.3380126. doi:10.1145/3366423.3380126.
[31] R. Chew, J. Bollenbacher, M. Wenger, J. Speer, A. Kim, Llm-assisted content analysis: Using large language models to support deductive coding, arXiv preprint arXiv:2306.14924 (2023).
[32] Z. Xiao, X. Yuan, Q. V. Liao, R. Abdelghani, P.-Y. Oudeyer, Supporting qualitative analysis with large language models: Combining codebook with gpt-3 for deductive coding, in: Companion Proceedings of the 28th International Conference on Intelligent User Interfaces, 2023, pp. 75–78.
[33] R. Baeza-Yates, C. Hurtado, M. Mendoza, Query recommendation using query logs in search engines, in: International conference on extending database technology, Springer, 2004, pp. 588–596.
[34] E. Zhang, X. Wang, P. Gong, Y. Lin, J. Mao, Usimagent: Large language models for simulating search users, 2024. arXiv:2403.09142.
[35] B. Engelmann, T. Breuer, J. I. Friese, P. Schaer, N. Fuhr, Context-driven interactive query simulations based on generative large language models, in: N. Goharian, N. Tonellotto,

Y. He, A. Lipani, G. McDonald, C. Macdonald, I. Ounis (Eds.), Advances in Information Retrieval, Springer Nature Switzerland, Cham, 2024, pp. 173–188.

[36] P. Owoicho, I. Sekulic, M. Aliannejadi, J. Dalton, F. Crestani, Exploiting simulated user feedback for conversational search: Ranking, rewriting, and beyond, in: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '23, Association for Computing Machinery, New York, NY, USA, 2023, p. 632–642. URL: https://doi.org/10.1145/3539618.3591683. doi:10.1145/3539618.3591683.

[37] I. Sekulić, M. Aliannejadi, F. Crestani, Evaluating mixed-initiative conversational search systems via user simulation, in: Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining, WSDM '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 888–896. URL: https://doi.org/10.1145/3488560.3498440. doi:10.1145/3488560.3498440.

[38] E. Svikhnushina, P. Pu, Approximating online human evaluation of social chatbots with prompting, in: S. Stoyanchev, S. Joty, D. Schlangen, O. Dusek, C. Kennington, M. Alikhani (Eds.), Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Association for Computational Linguistics, Prague, Czechia, 2023, pp. 268–281. URL: https://aclanthology.org/2023.sigdial-1.25. doi:10.18653/v1/2023.sigdial-1.25.

[39] J. Afzali, A. M. Drzewiecki, K. Balog, S. Zhang, Usersimcrs: A user simulation toolkit for evaluating conversational recommender systems, in: Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining, WSDM '23, Association for Computing Machinery, New York, NY, USA, 2023, p. 1160–1163. URL: https://doi.org/10.1145/3539597.3573029. doi:10.1145/3539597.3573029.

[40] J. Ashwin, A. Chhabra, V. Rao, Using large language models for qualitative analysis can introduce serious bias, arXiv preprint arXiv:2309.17147 (2023).

[41] J. Gao, Y. Guo, T. J.-J. Li, S. T. Perrault, Collabcoder: A gpt-powered workflow for collaborative qualitative analysis, in: Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing, CSCW '23 Companion, Association for Computing Machinery, New York, NY, USA, 2023, p. 354–357. URL: https://doi.org/10.1145/3584931.3607500. doi:10.1145/3584931.3607500.

A. Appendix

In this section, we present the prompts used for our classifier. The dataset utilized in this study is primarily in Korean therefore, the few-shot examples are original in Korean. However, parts of the classifier prompts and examples have been translated into English for clarity in this documentation. The whole code including the Korean prompt can be found at `https://github.com/kixlab/searchgpt-classifier`.

A.1. Axis 1 Classifier

```
You are a classifier analyzing an excerpt of a conversational search log, consisting of the
    ↪ initial query Q1, the engine's response R, and the follow-up query Q2. Among the
    ↪ taxonomy of the purpose of Q2 presented below, choose one that matches the best and
    ↪ only tell me the name of the taxonomy as this is scripted.

1. Clarifying Queries - Q2 clarifies the exact search intent of Q1. This occurs when Q2 is
    ↪ repeating the same question as Q1 to get a more specific answer by repeating the same
    ↪ question in Q1 with exact words or by asking the question about the same exact thing.
    ↪ If Q1 and Q2 are different, ask if the rewording of Q2 changes the meaning from Q1 or
    ↪ is just clarifying.

2. Exploring Domain - Q2 asks a question that explores the same general topic as Q1. Use
    ↪ taxonomy's definition widely.

3. Understanding Response - Q2 seeks more information necessary to understand R or to expand
    ↪ the range of information from R. This occurs when Q2 specifically asks for more
    ↪ information about information mentioned in R.

4. Narrowing Down - Q2 seeks more details or specific information that narrows down the
    ↪ question asked in Q1. This only occurs when Q2 asks for more specific information
    ↪ about something mentioned in Q1.

5. Seeking Different Representations - Q2 seeks different modalities (e.g., images, videos)
    ↪ or formats (e.g., tables) to better present the information provided in R.

6. Verifying Information - Q2 asks for more information to verify the information by asking
    ↪ for evidence, explanations, or alternatives from R

7. Reacting to Response - Q2 expresses satisfaction/dissatisfaction with R or provides
    ↪ feedback on R.

8. Unclassified - Q2 is completely irrelevant to Q1 or R.

Follow the examples as a guide to classify the queries. Assume the examples are always
    ↪ correct.

1:
Q1: What is the most preferred type of dishwasher by newlyweds?
R: Regarding dishwasher types, the most preferred product by newlyweds is a 12-person
    ↪ dishwasher. This capacity is generally suitable for a family of four. However, a
    ↪ 6-person...
Q2: Tell me the product name of the dishwasher preferred by newlyweds.
```

```
Taxonomy: Clarifying Queries

1A:
Q1: How can I apply for an additional government-supported rental loan if I already own a
    ↪ house?
R: The method to apply for an additional government-supported rental loan using your
    ↪ existing house as collateral is as follows...
Q2: I am curious about the income requirements for the government-supported rental loan.
    ↪ Tell me.

Taxonomy: Clarifying Queries

1B:
Q1: Tell me more about the Russian coup on June 24, 2023.
Q2: On June 24, 2023, tell me in detail about the military uprising led by the 'Wagner
    ↪ Group' in Russia.

Taxonomy: Clarifying Queries

2:
Q1: Is there any medicine I can take for a headache caused by a hangover?
R: There is no medicine you can take for a headache caused by a hangover. A hangover is the
    ↪ process of alcohol being metabolized...
Q2: What should I do if I accidentally took medicine for a severe hangover?

Taxonomy: Exploring Domain

2A:
Q1: Is it okay to bring toilet paper as a housewarming gift for newlyweds?
R: As for bringing toilet paper as a housewarming gift for newlyweds, it can be a good
    ↪ choice. However...
Q2: Tell me about a housewarming gift under 5,000 won that is not too expensive.

Taxonomy: Exploring Domain

2B:
Q1: So I shouldn't take headache medicine for a hangover?
Q2: Recommend a pain reliever with ibuprofen that helps with hangovers.

Taxonomy: Exploring Domain

3:
Q1: How about whiskey?
R: A whiskey product recommended as a housewarming gift for newlyweds in the 50,000 won
    ↪ range is available. Drinking it 'on the rocks' is a method of enjoying whiskey
    ↪ warmly, which can also be easily enjoyed at a newlywed's home. The product is an
    ↪ 'Oreforse' brand whiskey, which is popular for its luxury design and quality. Here
    ↪ are some recommended products:
Q2: What is 'on the rocks'?

Taxonomy: Understanding Response

4:
```

```
Q1: Choose a housewarming gift for newlyweds.
R: Suitable housewarming gifts for newlyweds include practical gifts like toilet paper, wet
    ↪ wipes, and coffee machines...
Q2: Choose a housewarming gift for newlyweds in the 50,000 won range.

Taxonomy: Narrowing Down

4A:
Q1: Tell me the product name of the dishwasher most preferred by newlyweds.
R: The dishwasher product most preferred by newlyweds is the BESPOKE dishwasher. This
    ↪ product offers various capacities and types...
Q2: How much is the BESPOKE dishwasher?

Taxonomy: Narrowing Down

4B:
Q1: What is the preferred capacity for a dishwasher?
R: The preferred capacity for a dishwasher is as follows...
Q2: Tell me the price of an LG dishwasher.

Taxonomy: Narrowing Down

5:
Q1: Heating cost fundraising poster
R: I cannot provide specific information about a heating cost fundraising poster, but
    ↪ generally, such posters include the following elements...
Q2: Show me a picture related to heating cost fundraising.

Taxonomy: Seeking Different Representations

6:
Q1: How do I remove polish from a steel tumbler, not stainless steel?
R: The method to remove polish from a steel tumbler is as follows...
Q2: Do I need to follow the steps from 1 to 6 in order?

Taxonomy: Verifying Information

6A:
Q1: Is the Wagner Group the largest army in Russia?
R: The largest army in Russia is the Russian Ground Forces, which is the largest in scale
    ↪ among the Russian military, with approximately 370,000 personnel. The Russian Ground
    ↪ Forces focus on mechanized and airborne units...
Q2: Isn't the Russian Ground Forces affiliated with the Russian government?

Taxonomy: Verifying Information

6B:
Q1: Can Tylenol help with a hangover?
R: Tylenol can help relieve headaches caused by hangovers. However, taking Tylenol after
    ↪ drinking can lead to liver damage...
Q2: Can I take a painkiller that contains acetaminophen to relieve a hangover?

Taxonomy: Verifying Information
```

```
7:
Q1: None of the three models you mentioned are in the 500,000 won range.
R: If I understand correctly, you want to know which of the three models I mentioned are not
    ↪ in the 500,000 won range...
Q2: I was talking about dishwashers, not Galaxy Tabs... Get it together, (LLM Agent Model
    ↪ name).

Taxonomy: Reacting to Response
```

A.2. Axis 2 Classifier

```
You are a classifier analyzing an excerpt of a conversational search log, consisting of the
    ↪ initial query Q1, the engine's response R, and the follow-up query Q2. Among the
    ↪ taxonomy of the purpose of Q2 presented below, choose one Taxonomy within A, B, C, or
    ↪ D. Just list the name of the Theme. This taxonomy is based on priority. Check
    ↪ chronologically, so check if the conversation fits 1 then 2, then 3 and so on... The
    ↪ lower number gets priority.

A. Query-Specific (When Q2 is related to Q1):
1. Excluding Conditions - When the user explicitly removes specific conditions from the
    ↪ previous query (Q1).
2. Adding/Specifying Condition - WWhen the user specifies additional conditions or
    ↪ parameters to include in the query
3. Substituting Condition - When the user expresses the same query intent using different
    ↪ conditions or forms.

ONLY if Q2 is not related to Q1, then:
B. Response-Specific (When Q2 is unrelated to Q1 but related to R):
4. Converting Format - When Q2 asks in a different format (e.g., table, map, graph, image).
5. Confirming Response - When the user seeks clarification or confirms information from R
    ↪ using Q2.
6. Criticizing Response - When the user expresses criticism of the provided response (R).
7. Affirming Response - When the user expresses satisfaction with the provided response (R).
8. Requesting Additional Information - When the user requests additional information related
    ↪ to the response (R).

ONLY if Q2 is not related to Q1 and R, then:
C. Session-Specific (When Q2 is unrelated to Q1 and R but broadly to the topic):
9. Requesting Opinion - When the user asks about something an opinion not previously
    ↪ referred to in response (R).
10. Requesting Related Information - When the user requests related information that is
    ↪ broadly related to the same topic.

ONLY if Q2 is not related to Q1, R, and the topic, then:
D. Miscellaneous (Other):
11. Chatting Casually - When the user attempts chitchat rather than seeking specific
    ↪ information.
12. Requesting Unrelated Information.

Follow the examples as a guide to classify the queries. Assume the examples are always
    ↪ correct.
```

```
1:
Q1: What dishes can be made with beef, enoki mushrooms, shishito peppers, onions, and bean
    ↪ sprouts?
Q2: What dishes can be made with beef, enoki mushrooms, and onions?

Taxonomy: Excluding Conditions

2:
Q1: What is the most luxurious coffee machine?
Q2: Tell me about a more expensive and better coffee machine, excluding capsule coffee
    ↪ machines.

Taxonomy: Adding/Specifying Condition

2A:
Q1: Are there any models under 500,000 KRW from large corporations?
Q2: Not robot vacuums, but dishwashers!

Taxonomy: Adding/Specifying Condition

2B:
Q1: What kinds of special housewarming gifts are there for newlyweds?
Q2: Recommend types of foreign electronic products.

Taxonomy: Adding/Specifying Condition

2C:
Q1: Tell me about the Russian coup.
Q2: Tell me more about the recent coup related to the Wagner Group.

Taxonomy: Adding/Specifying Condition

3:
Q1: Recommend a restaurant in Daejeon suitable for a 70th birthday party with a rating of at
    ↪ least 4.8.
Q2: Recommend a well-reviewed restaurant in Daejeon suitable for a 70th birthday party.

Taxonomy: Substituting Condition

3A:
Q1: Tell me more about the second coup.
Q2: Tell me more about the Russian coup on June 24, 2023.

Taxonomy: Substituting Condition

5:
Q1: Can I bring toilet paper as a housewarming gift for newlyweds?
Q2: Tell me about a housewarming gift under 5,000 won that isn't too expensive.

Taxonomy: Criticizing Response

7:
Q1: Is the Wagner Group the largest army in Russia?
```

```
R: The largest army in Russia is the Russian Ground Forces, which is the largest among the
     ↪ Russian military, with approximately 370,000 personnel...
Q2: Isn't the Russian Ground Forces affiliated with the Russian government?

Taxonomy: Confirming Response

7A:
1: Can Tylenol help with a hangover?
R: Can Tylenol help with a hangover? - A hangover involves symptoms like headache, vomiting,
     ↪ and dizziness that occur after drinking alcohol...
Q2: So, can I take a pain reliever containing acetaminophen for hangover relief?

Taxonomy: Confirming Response

7B:
Q1: Russia doesn't have a military organization other than the Wagner Group, does it?
Q2: Is the Wagner Group the largest army in Russia?

Taxonomy: Confirming Response

8:
Q1: Choose a housewarming gift for newlyweds in the 50,000 won range.
R: Here are some recommended housewarming gifts for newlyweds. 1. Handmade pottery gifts for
     ↪ newlyweds. Q2: How about whiskey?

Taxonomy: Requesting Opinion

8A:
Q1: Is it okay to exercise when I have a severe hangover?
R: Opinions vary on whether you can exercise with a severe hangover. However, some argue
     ↪ that you can still engage in some physical activity even with a severe hangover.
Q2: Can I go to a bathhouse if I have a headache from a severe hangover?

Taxonomy: Requesting Opinion

9:
Q1: I'm planning a ryokan trip, any good places?
R: Are there any good ryokan places for a trip? - Many ryokans in the Yufuin area are
     ↪ frequently searched. Some ryokans are praised for their breakfast,...
Q2: Are there other places worth visiting in Yufuin?

Taxonomy: Requesting Additional Information
```