

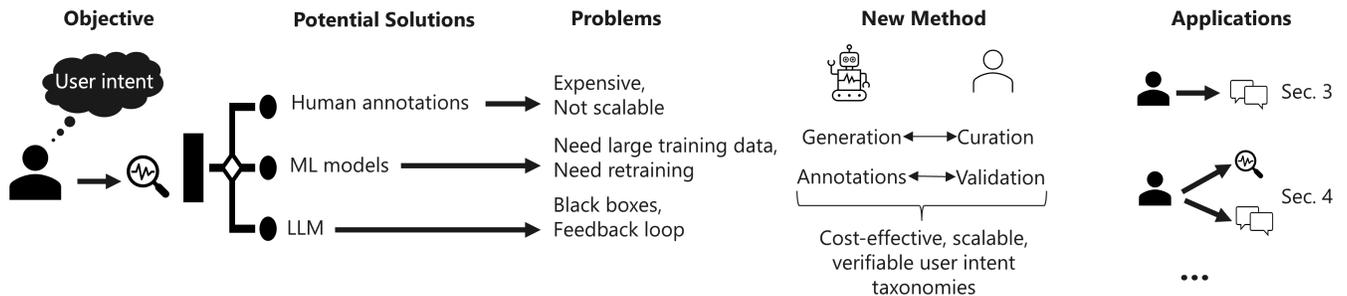
# Using Large Language Models to Generate, Validate, and Apply User Intent Taxonomies

Chirag Shah<sup>1†‡</sup>, Ryen W. White<sup>2†</sup>, Reid Andersen<sup>2</sup>, Georg Buscher<sup>2</sup>, Scott Counts<sup>2</sup>, Sarkar Snigdha Sarathi Das<sup>3‡</sup>, Ali Montazer<sup>4‡</sup>, Sathish Manivannan<sup>2</sup>, Jennifer Neville<sup>2</sup>, Xiaochuan Ni<sup>2</sup>, Nagu Rangan<sup>2</sup>, Tara Safavi<sup>2</sup>, Siddharth Suri<sup>2</sup>, Mengting Wan<sup>2</sup>, Leijie Wang<sup>1‡</sup>, Longqi Yang<sup>2</sup>

<sup>1</sup>University of Washington, <sup>2</sup>Microsoft, <sup>3</sup>Pennsylvania State University, <sup>4</sup>University of Massachusetts Amherst

†Corresponding authors: chirags@uw.edu, ryenw@microsoft.com

‡Work done while working at Microsoft, USA



**Figure 1: Detecting user intent in search situations is very important, but the potential solutions have different sets of problems. We propose a new method involving a unique collaboration between LLMs and humans to leverage best of both worlds.**

## ABSTRACT

Log data can reveal valuable information about how users interact with web search services, what they want, and how satisfied they are. However, analyzing user intents in log data is not easy, especially for new forms of web search such as AI-driven chat. To understand user intents from log data, we need a way to label them with meaningful categories that capture their diversity and dynamics. Existing methods rely on manual or ML-based labeling, which are either expensive or inflexible for large and changing datasets. We propose a novel solution using large language models (LLMs), which can generate rich and relevant concepts, descriptions, and examples for user intents. However, using LLMs to generate a user intent taxonomy and apply it to do log analysis can be problematic for two main reasons: such a taxonomy is not externally validated, and there may be an undesirable feedback loop. To overcome these issues, we propose a new methodology with human experts and assessors to verify the quality of the LLM-generated taxonomy. We also present an end-to-end pipeline that uses an LLM with human-in-the-loop to produce, refine, and use labels for user intent analysis in log data. Our method offers a scalable and adaptable way to analyze user intents in web-scale log data with minimal human effort. We demonstrate its effectiveness by uncovering new insights into user intents from search and chat logs from Bing.

This article is distributed under Creative Commons Attribution-NonCommercial- NoDerivatives License 4.0 (CC BY-NC-ND).

## KEYWORDS

User intents, Large language models, Taxonomies, Log data

## 1 INTRODUCTION

Understanding the purpose or the task behind a user’s request in an information access context is highly desired for a search or a recommender system to be able to provide the most relevant and meaningful results [54]. However, extracting user intents from log data is extremely difficult due to two main reasons: fluidity in what user intents are or can be, and how these intents can be identified using log data that may not include enough context. Additionally, in the case of emerging modalities such as AI-driven chat, users’ understanding, usage, and behaviors are rapidly evolving that call for on-demand, task-focused labels and taxonomies. We need new methods to identify, extract, and apply user intents in various information access systems, especially those with emerging modalities.

Traditional qualitative methods such as coding and thematic analysis are time-consuming and require human expertise [8]. On the other hand, existing quantitative methods such as metrics and statistics may not capture the nuances and diversity of user intents and experiences [34]. Large language models (LLMs) have become quite capable of generating coherent texts from various inputs using deep neural networks [10]. But can they be useful in a reliable, verifiable ways to conduct such research?

There are several attempts in the recent months to use LLMs in various applications ranging from ranking and recommendations to content generation and evaluation (e.g., [6, 20, 36]). However, many

of these works often lack rigor and reliability since LLMs are used as black boxes without a meaningful understanding of their inner workings or there are feedback loops with weak or non-existent validation for the method. Simply focusing on the promising results without sufficient support of scientific rigor in the methodology can lead to misleading and even dangerous outcomes. We believe that while LLMs have shown great promise for aiding us in various informational tasks, their usage must be done with responsibility and sufficient validation. This leads us to the following research questions (RQs):

- (1) Can we use LLMs to reliably generate taxonomies for analyzing user intents in logs?
- (2) Can an LLM correctly apply a user intent taxonomy to annotate logs?
- (3) Are there situations where an LLM can do better than human annotators? Or can we do things with LLMs that we could not do before beyond just reducing efforts and increasing efficiency?

To address these RQs, we investigated if/how LLMs can help in creating an end-to-end solution for developing user intent taxonomies from AI chat logs. We picked GPT-4, which is the latest version of the Generative Pre-trained Transformer model that can produce texts across multiple domains and tasks [39], since it is one of the leading LLMs currently available and we wanted to test the potential of the most advanced models. In the process, we devised a new methodology for employing GPT-4 as a collaborator in an iterative qualitative analysis process that leverages its ability to generate summaries, questions, and categories from chat transcripts.

The primary purpose here is to develop and validate a scientific method for using LLMs in understanding user intents. To that end, our contributions are three-fold: (1) a novel method for using an LLM for end-to-end user intent analysis from log data that combines human interpretation and machine generation to reduce effort and increase comprehensiveness; (2) a new data-driven taxonomy of user intents for analyzing chat and other user interaction logs; and (3) new insights about how and why users are using search and chat modalities for their information access needs.

## 2 RELATED WORK

We focus on generating applicable user intent taxonomies from logs, but we are not the first ones to study user intents, build taxonomies, or use such taxonomies to generate insights from logs. Therefore, before we dive into our novel contributions, viz., using LLMs for generating and using user intents taxonomies with scientific rigor, it is important to briefly review some of the related work.

### 2.1 Taxonomy Generation, Validation, and Use

Taxonomies are hierarchical classifications of concepts, terms, or entities. They can facilitate information seeking, retrieval, or behavior by providing structure, organization, and navigation for users and systems [12, 13, 51]. However, generating and validating taxonomies is a challenging task that requires balancing multiple criteria such as coverage, coherence, consistency, granularity, usability, and adaptability, e.g., [28, 33, 37, 43]. Moreover, different domains and contexts may have different requirements and preferences for taxonomy design and evaluation [31]. Taxonomies can be generated manually through an iterative process and the research

community has developed tools to generate taxonomies automatically from document collections using methods such as clustering [52]. In this paper, we are the first to leverage the power of LLMs to automatically generate taxonomies within the context of search, focused specifically on conversational search; we also validate that LLM-based methodology with human assessors. More importantly, we provide a method for other researchers to do the same for their specific needs.

### 2.2 Use of LLMs in Research

The emergence of LLMs has unlocked many opportunities for rapid research advances. LLMs have been used to enable scientific discovery [25], with remarkable progress in areas such as medicine [42] and finance [4]. Early language models, such as BERT, and various natural language processing methods have been used to auto-code qualitative data [1, 23], although not at a near-human level. More recent work by Byun et al. [11] has shown that LLM-driven AI is capable of qualitative analysis and can generate nuanced results that are comparable to human researchers.

This progress has stirred discussions about the role of LLMs in quantitative and qualitative research. Some view these as tools to enrich human understanding, while others perceive them as threats to core scientific values [5]. Watkins [48] proposes considerations that can be applied by researchers to guide the use of LLMs in their workflows. Bano et al. [5] predict that we will see LLMs redefining the interplay between humans and AI in the qualitative research process. LLMs can derive insights from unstructured data and collect and analyze massive amounts of data from diverse sources. However, in their current form they do not possess the intuitions, nuances, and sensitivity to culture and context that humans have. A shift toward greater human-AI teaming for qualitative analysis is more likely in the short term given current LLM capabilities than a complete LLM-centric transformation in how qualitative research is performed [19].

In information retrieval (IR), LLMs have been shown to be effective in supporting humans in judging document relevance, an activity that is central to search engine design and evaluation [20]. They have also been recently used for synthetic dataset generation to support IR research [27] and richer user modeling to support IR experimentation [32]. Other applications of LLMs in IR have also been discussed [3]. In this paper, we will show that working directly with humans, LLMs have the potential to support two other critical activities in search engine research and development: intent understanding and generating intent taxonomies.

### 2.3 Log Analysis and Insights

Search log analysis has been used extensively to gain insights about search interactions, including queries, search engine result page (SERP) clicks, and post-SERP interactions [18, 44, 49]. Analyzing log data has historically been a highly interactive process: researchers first write scripts to extract data, they then analyze that data manually using data science tools and methods, and (optionally) human annotators label data to improve understanding about patterns and trends and generate training data for ML models, e.g., [2].

More recently, there has been a focus on user engagement with conversational search systems [22]. Researchers and practitioners

have sought to understand user intents and behaviors in the context of chat-based systems [38, 47]. Chat is becoming an increasingly popular modality for information seeking, especially in domains where users have complex or exploratory queries, need guidance or clarification, or prefer a conversational style of interaction. Datasets of chat logs have also been created and released to the community to promote research in this emerging area [40].

ML models can help support researchers in performing data analysis [24]. Recently, from applications in other domains, we have seen that LLMs may have the potential to play a supportive role in the analysis of text data, providing insights and annotations that expedite experiments and reduce human effort [45]. In this paper, for the first time, we introduce methods for using LLMs collaborating with researchers to derive insights and taxonomies from log data.

## 2.4 Understanding User Intents in Search

Intent has been a well explored area in IR and various approaches have been proposed for intent understanding and intent representation [30, 53]. Taxonomies of search intent can help systems better understand user intent. Several search intent taxonomies have been proposed, e.g., [9, 26, 29, 41]. These have been generated iteratively via manual inspection of search log data. Intent taxonomies developed from search logs include: navigation, information, transactions, browsing, and resource finding. We expect chat interactions to exhibit new intents (e.g., creation) compared to traditional search. Automatic generation of query taxonomies has also been attempted [15, 16] using query clustering to derive a taxonomy from existing query data and categorization to assign new queries to the taxonomy. Taxonomies have also been used to represent intent in question-answering [7, 14]. Xie [50] derived an empirically based classification of search intents that motivate different search behaviors. Mitsui et al. [35] developed a set of information-seeking intentions based on that classification and studied differences in intentions as a function of the search task. Building on this previous work, we leverage LLMs to generate intent taxonomies with humans-in-the-loop and evaluate the performance of LLMs in assigning search activities to intent categories.

## 3 METHODOLOGY FOR GENERATING AND VALIDATING A USER INTENT TAXONOMY

In this section, we describe a new methodology that we developed and tested for employing LLMs to generate a user intent taxonomy that can be used to generate insights and construct hypotheses from AI chat logs.

Let us begin with a problem scenario. We have access to log data from user interactions with an AI-driven chat system (e.g., Bing Chat). This data primarily includes user requests and AI responses in natural language. We can analyze this data in a number of ways, answering questions about what the users are doing (topics, domains). But if we want to understand their intents, we need a set of labels or a taxonomy of intents. In practice, one looks for relevant literature for an existing taxonomy, but if an appropriate taxonomy does not exist, one needs to create it. This can be done by taking an existing taxonomy and modifying it to fit the data or the task (top-down approach), or by building it fresh using the available

data (bottom-up approach). Following this, one needs to validate the taxonomy to ensure it meets several criteria for a good taxonomy. Finally, the newly generated taxonomy could be applied to a specific task to generate the desired insights from the data.

Given that we were interested in analyzing AI chat logs, a seemingly newer type of modality, we found it to be desirable to adopt a bottom-up approach. In this approach, one typically analyzes available data to generate codes or labels, leading to a classification scheme or a taxonomy. As detailed in [9, 26, 29, 41], this process could involve one or more researchers and a considerable effort. We wanted to use an LLM to build such a taxonomy using relevant data and instructions. However, given that we do not have enough knowledge about how an LLM creates or links various concepts from given data, we needed a way to validate LLM's generation and fine-tune it as needed. For that, we used two researchers with many years of experience in doing qualitative analysis and building taxonomies. These researchers guided the taxonomy generation process with the LLM and two human assessors. Once the taxonomy was built using training data and validated, it was applied to annotating test data. In short, our methodology uses LLM as the backbone of taxonomy generation and application with humans in the loop for curation and validation.

The outline of our methodology is shown in Figure 2. Here, we used GPT-4 as the LLM for generating the taxonomy (Phase 1), engaged human assessors to validate that taxonomy (Phase 2), and then employed both GPT-4 and human assessors to apply it (Phase 3). Through the phases of validation and application, we evaluated not only the generated taxonomy (RQ1), but also GPT-4's ability and potential to perform such research-based tasks reasonably and reliably (RQ2). Following are the details.

### 3.1 Data

We took a random sample of 1,149 fully anonymized conversations from May-June 2023, available through Bing Chat. Each conversation contained one or more turns of user request and AI response. We ensured that these conversations were in English, however, some of the user requests had non-English words mixed in. We do not believe this impacted any text processing by GPT-4 for our purposes. We used 1,000 conversations for training (building a user intent taxonomy) and set aside the rest for validation and testing.

### 3.2 Evaluating taxonomy

We first start by describing how a taxonomy should be evaluated. This will inform how we generate the taxonomy using GPT-4 (how we provide prompts), how we validate and revise it, as well as how we measure the effectiveness of the taxonomy for developing insights from logs. Using the relevant literature (some of which is summarized in the previous section) concerning taxonomy generation and validation, we consolidated the following criteria, taken from Raad and Cruz [37], with appropriate modifications.

- **Comprehensiveness:** All the data should be reliably classified using this taxonomy.
- **Consistency:** The taxonomy does not include or allow for any contradictions.

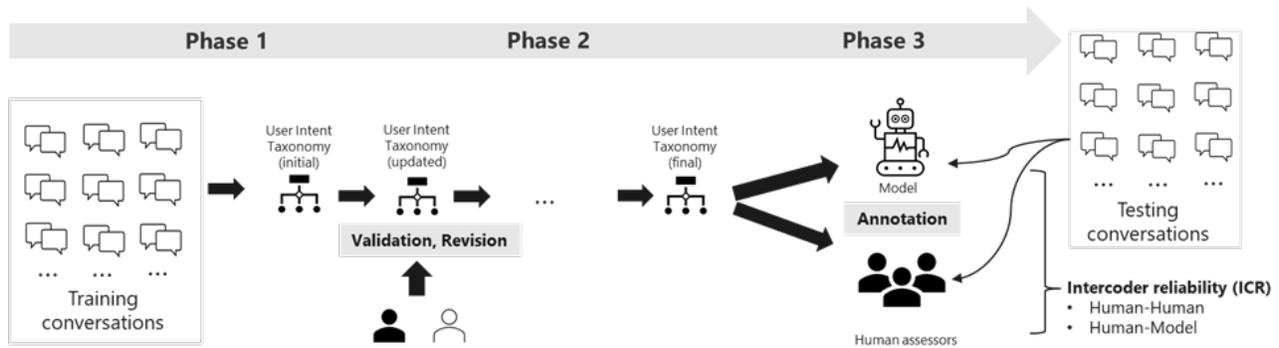


Figure 2: Three phases of user intent taxonomy generation, validation, and application.

- **Clarity:** The taxonomy should communicate the intended meaning of the defined terms. Definitions should be objective and independent of the context.
- **Accuracy:** The definitions, descriptions of classes, properties, and individuals in a taxonomy should be correct.
- **Conciseness:** The taxonomy should not include any irrelevant elements with regards to the user intents in AI chat.

Table 1 reports how we planned to evaluate or ensure the generated taxonomy meets these criteria.

### 3.3 Phase 1: Taxonomy Generation

Considering how the taxonomy should be evaluated (see Table 1), we constructed a detailed prompt for GPT-4 for generating the first version of taxonomy. We made a few design choices here, including the depth of the taxonomy (single level) and the number of categories (4-6). We asked the LLM to generate labels, descriptions, and examples for these categories. The full prompt can be seen in Appendix B.

There were variations in how different versions described the same category. For instance, ‘Learning’ had slightly different meaning and definition in each version of the taxonomy, but generally included concepts and examples of understanding and explanation. Two researchers (two of the co-authors) discussed these three versions and decided to create a consolidated version of the taxonomy, which is shown in Table 2.

### 3.4 Phase 2: Taxonomy Validation

Next, we provided the taxonomy in Table 2 to two human coders along with 10 segments of conversation. They coded them independently, after which we compared their labels. They only agreed 3 out of 10 times. We repeated the whole process with a new sample of 10 segments. It got better, but still had a high level of disagreement (60%). We, therefore, had another round of discussions and deliberations.

More than trying to reach a higher level of agreement, the goal here was to revise the current version of the taxonomy and develop a better understanding of how a reliable taxonomy could be generated that meets the criteria reported earlier and leads to a common and robust comprehension among the annotators. We learned that the annotators were often extrapolating why a user might have tried to do something.

That led to most divergence among them. For instance, even when all we could interpret from the data that the user asked for factual information (e.g., “Does the state of Washington have income tax?”), one of the annotators often extended that to ‘Learning’ intent. It is possible that the user was collecting such information as a part a learning task, but without additional context, it may be impossible to determine that. In such cases, it is advisable to not overextend our understanding and mark the intent based on evidence. Thus, we found it was useful to include in the taxonomy not only positive examples, but also negative examples per category to improve overall clarity. The taxonomy was further modified using negative examples for each category, and the prompt for generating a taxonomy was edited to explicitly ask for negative examples (negative examples are not listed here due to space constraints).

Once we had GPT-4 provide such examples and clarify definitions of ‘Information Retrieval’ and ‘Learning’ categories, we achieved a good match with only 20% disagreements between the two annotators. In addition, the human assessors did not find a need for any intention not covered here. Thus, the validation state of taxonomy generation was completed and we had the final version of user intent taxonomy (see Appendix C).

### 3.5 Phase 3: Taxonomy Application and Testing

We then took a different set of 124 conversations and have GPT-4 code them using the modified taxonomy generated from the above process. We also gave the same instructions to two human assessors for annotation. These instructions to humans and the prompt to GPT-4 can be found in Appendix D.

For the human coders, not a single datapoint was labeled ‘Other’. GPT-4, on the other hand, marked one out of 124 conversations as ‘Other’. This is not common in qualitative coding and further demonstrates comprehensiveness of the taxonomy.

We computed inter-coder reliability (ICR) between two human coders and found Cohen’s kappa to be 0.7620, which indicates a substantial level of agreement [17].

Next, we asked a third coder to code these 124 conversations. This allowed us to triage among the three coders. When the three annotators disagreed, we took the majority vote. If all three picked a different label, we labeled that case as ‘Other’. Finally, we computed ICR between GPT-4 labels and those generated by the majority of human coders. We computed Cohen’s kappa to be 0.7212. This also indicates a substantial level of agreement.

**Table 1: Evaluating GPT-4 generated taxonomy using human assessors.**

Criteria	How to evaluate/ensure?
Comprehensiveness	→ At the time of generation (Phase 1) by making sure the taxonomy considers as broad a scope as possible in the prompt to GPT-4. → At the time of validation (Phase 2) by looking at what proportion of instances by human assessors end up in the 'Other' category.
Consistency	→ At the time of generation (Phase 1) by making sure GPT-4 generates a taxonomy where different categories do not overlap in their meanings. → At the time of validation (Phase 2) by assessing how often the human assessors have difficulty distinguishing between two labels.
Clarity	→ At the time of generation (Phase 1) by ensuring GPT-4 provides detailed description or definition along with examples with each category. → At the time of validation (Phase 2) by eliciting from the human assessors how clear the definitions and examples are for them.
Accuracy	→ At the time of validation (Phase 2) by comparing annotations by at least two human assessors and computing inter-coder reliability (ICR).
Conciseness	→ At the time of application (Phase 3) by analyzing how well the generated and validated taxonomy serve the purpose of understanding user intents in chat logs after GPT-4 and/or human annotators use the taxonomy for annotation of test data.

**Table 2: A consolidated version of the user intent taxonomy generated by GPT-4. The examples are collected from the three versions that GPT-4 generated. Slight modifications are made in the user intent title and description using those versions. Short phrases in parentheses next to the intent name are added by researchers for clarity.**

User intent	Description	Examples
Information retrieval (looking for factual information that already exists)	The user wants to search, query, or find some information, data, or resources about a topic.	Find out the airing dates and channels of women's world cup; Search for information about a phone number; Search for corruption and unemployment statistics for a country.
Problem solving (extracting facts or answers by computing something)	The user wants to perform a mathematical or logical operation, such as a conversion, a percentage, a formula, or a function.	Compare the size of a human to a hydrogen atom and the observable universe; Compare interest rates for savings accounts; Calculate the distance between a point and a line; Convert a message from Chinese to English.
Learning (satisfying curiosity, helping learn a concept or a phenomenon)	The user wants to learn, study, or acquire new skills, concepts, or understanding about a subject. This often involves operations of calculations, comparison, and conversion.	Learn about different structural systems; Compare GPT-3 and GPT-4 versions; Explain the difference between Newtonian and non-Newtonian flow.
Content creation	The user wants to write or edit a text for a specific purpose or audience.	Write an introduction about geothermal energy; Modify a poem into different formats; Improve a report and find adverbs and connectors.
Leisure	The user wants to chat or interact with the AI or another agent about various topics or play a game with the AI or another agent.	Ask about the AI's sexual orientation and name; Listen to a romantic story; Play tennis and flirt with the user.

Overall, what we learned is that when a taxonomy is generated by GPT-4 and verified by humans, it leads to a very high amount of agreement for annotation. That speaks to the **validity** of the generated taxonomy. Also, given that GPT-4's own coding achieves a high level of ICR with human coders shows that GPT-4 can be used with high **reliability** for the annotation task.

### 3.6 Insights About and From Annotations

Now that we have demonstrated the end-to-end methodology for generating, validating, and using a taxonomy for understanding user intents in chat logs, let us consider what insights we could derive from the 124 conversation segments analyzed by annotators and GPT-4.

Table 3 presents the confusion matrix between the two annotators. We can see that Information Retrieval (IR) is the largest category, followed by Learning (LR). The greatest number of times the two annotators disagree is for IR and LR categories. This is understandable since LR always contains IR, but it may not always be easy to evaluate if an IR process extends enough to qualify as LR. As noted earlier, this was the biggest factor leading to disagreements among the annotators.

Table 4 presents the confusion matrix between human annotations (after doing triaging of three annotators) and those of GPT-4. Once again, we find that IR and LR are the largest categories and

**Table 3: Confusion matrix for user intent annotations between two human annotators. IR=Information Retrieval, PS=Problem Solving, LR=Learning, CR=Content Creation, LS=Leisure.**

		Annotator-2				
		IR	PS	LR	CR	LS
Annotator-1	IR	42	2	10	0	0
	PS	0	8	0	4	0
	LR	3	0	36	0	0
	CR	0	0	0	8	0
	LS	1	0	0	0	9

**Table 4: Confusion matrix for user intent annotations between human and GPT-4 assessments. IR=Information Retrieval, PS=Problem Solving, LR=Learning, CR=Content Creation, LS=Leisure, OT=Other.**

		GPT-4					
		IR	PS	LR	CR	LS	OT
Human	IR	46	1	5	0	0	1
	PS	0	8	2	0	0	0
	LR	12	3	26	1	0	0
	CR	0	0	0	10	0	0
	LS	0	0	3	0	5	0
	OT	1	0	0	0	0	0

also where we see the most disagreements. Specifically, several (12 out of 124) conversation segments that are marked as LR by humans are labeled as IR by GPT-4. To understand who may be better or more appropriate in picking the labels here, we examined these conversations closely. They all include IR components, but the question is – do they go far enough to indicate an LR (learning) task?

Unfortunately, we do not have the ground truth here since we do not have access to the original user who conducted the conversation. We interviewed the assessors and found that they extended their understanding of what the users were doing in those segments of conversations to what they might want to do with that information beyond the logged interactions. This often led to a conversation segment being marked as an LR instead of an IR.

GPT-4 here is strictly labeling the data without making further assumptions, which is desirable. But how consistent this LLM is while making such subjective decisions? To test this, we ran the same test data through GPT-4 four more times and measured ICR among the five sets of annotations by the LLM. We found Fleiss' kappa [21] to be 0.8516, indicating a very high level of agreement and consistency. Therefore, we believe that the labels generated by GPT-4 are better than those generated by humans in this case as they are more objectively and consistently assigned without undesirable extrapolation or assumptions that may not be well-founded, addressing RQ3.

## 4 APPLICATION OF USER INTENT PIPELINE

Now that we have developed and validated a new methodology that could construct a data and application-driven taxonomy and apply it to test data with scale, we decided to use this methodology to build an end-to-end pipeline for an application of user intents in information access context. Here, we will include both search and chat under this information access umbrella. This pipeline, as depicted in Figure 3, is built on the proposed methodology, but requires a lot less human intervention. There is still a human in the loop but given that we know how to construct and apply a reliable taxonomy using a process that is already validated, we can now use human intervention in shaping the process and doing light touch validations. This reduces the workload significantly, speeds up taxonomy generation and application, and produces trustworthy instruments, labels, and insights.

### 4.1 Stepwise Process for the Full Pipeline

In the steps below, we describe this process as a full pipeline for how one could leverage LLM for analyzing log data. Through the process, we will also focus on evaluating various aspects of the generated taxonomy. These aspects include comprehensiveness, consistency, clarity, accuracy, and conciseness.

But why generate a new taxonomy if there are several existing taxonomies, including the one generated in the previous section? While one of these taxonomies could be fitting, it is desirable that we have a taxonomy that is rooted in specific data and application we have under consideration. Given the small cost of generating a taxonomy may also justify at least attempting to construct a new taxonomy and deciding if it is more fitting than anything available.

### Step-1: Identify Application and Data

The first step is to identify what kind of data we want to extract user intents from and why. Here, we are interested in understanding how users have different or overlapping intents between two modalities – search and chat. Using log data available to use from Bing Search and Bing Chat, we needed to first build a new user intent taxonomy and then apply that taxonomy to annotate log data. We extracted a random sample of fully anonymized users who had used both Bing Search and Bing Chat from May-June 2023. From those users, we extracted 2,456 queries and 15,531 chat requests they had sent to the respective services. We used 500 search queries and 500 chat requests (a total of 1,000 user inquiries to Bing) for training and set aside the rest for testing. We randomized their order, forming our training set with 1,000 data points.

### Step-2: Build/Fine-tune Taxonomy with Human-in-the-loop

To get started with the LLM (here, GPT-4), we built the initial prompt that explained what we are trying to do, what the data contains, and what are some of the criteria or constraints. For example, we indicated that we are looking for a taxonomy of user intents with no more than 5 categories and the criteria for a good taxonomy are comprehensiveness, consistency, clarity, accuracy, and conciseness as defined earlier. The full prompt is given in Appendix E. This resulted in the zero-shot version of the taxonomy with the following five categories:

- *Ask for Advice or Recommendation*: The intent to seek suggestions, opinions, or guidance from others on a specific topic or situation. Examples: recommend me a good book; how can I improve my sleep quality.
- *Create*: The intent to use AI tools or platforms to generate, edit, or manipulate information objects. Example: write me a poem about breakup; create a short report on effects of glaciers melting.
- *Information Retrieval*: The intent to find existing information or answers on the internet. Examples: what is the capital of France; where is biggest deposit of oil.
- *Learn*: The intent to acquire new knowledge or skills on a subject of interest. Examples: teach full stack web development; explain market cap rate in commercial real estate.
- *Leisure*: The intent to enjoy oneself by engaging in amusing activities such as games, jokes, stories, etc. Examples: play tic tac toe; make me laugh.

This is not very different from what we saw in the previous section. However, we noticed that the descriptions and examples associated with these labels were different and more suitable for our purpose. Even if there was evidence or intuition that an existing taxonomy would be sufficient for our purpose, given the reasonable cost for generating a new taxonomy, it may be desirable to go through these two steps to validate and revise that taxonomy with a goal to fare better along the criteria for a quality taxonomy described before.

### Step-3: Measure Taxonomy Comprehensiveness/Consistency

We now need to test how complete and consistent this taxonomy is. For that, we fed it as a prompt to the LLM and have it label each of the samples we used before separately. This time, we asked it explicitly to label anything that does not fit the provided labels as 'Other'. We found that no sample fell under this category. This

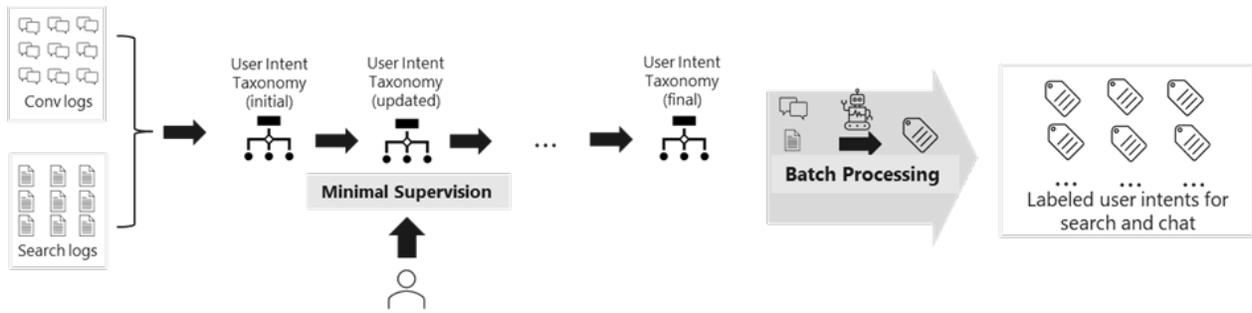


Figure 3: Using LLM in an end-to-end pipeline for generating, validating, and applying a taxonomy for user intents.

indicated that the taxonomy was comprehensive and consistent enough.

**Step-4: Improve Taxonomy Clarity**

Next, we asked the LLM to expand each category label with more description and examples to improve its clarity. Taking the lesson from before, we also asked GPT-4 for negative examples per category, improving on the taxonomy’s clarity.

**Step-5: Measure Validity and Accuracy**

As the final step of validation and refinement, we asked the LLM to use the constructed taxonomy to label the same data that was used to generate the taxonomy. Normally, this is not a practice for testing, but here we are looking for internal validity and accuracy of the taxonomy. Recall that we had 1,000 data points (500 search queries and 500 chat requests) for training. Once the LLM labeled each of these, we took a random sample of 100 and manually checked if the assigned label follows the definition for that label as generated before. We found that the answer was ‘yes’ for 95 of these samples and that there was no sample assigned ‘Other’ category. This analysis provided us with the assurance the taxonomy was valid and accurate.

**Step-6: Perform Annotations and Measure Conciseness**

Finally, we ran our test data – 1,956 search queries and 15,031 chat requests – through GPT-4 with the final version of the taxonomy as a part of the prompt. This prompt is given in Appendix F. We found that no sample was marked with ‘Other’ label, ensuring that all the important concepts were covered. Also, no category had too few (subjective, but in our case < 2%) samples, indicating that the taxonomy was concise.

**4.2 Insights about Intents in Search vs. Chat**

The steps above demonstrated that we could create a user intent taxonomy fulfilling all the criteria for a high-quality, reliable, and robust taxonomy. We did not have ground truth for the test data to be able to compute accuracy, but given the rigor taken in building and validating the taxonomy, we feel confident in the labels generated by the LLM. If one needs additional assurance at this point, a small sample of this test data can be taken for human assessment and ICR can be computed between that assessment and the one from the LLM.

For our purposes, we decided to move on to deriving insights from this test data. Given that we had an uneven number of queries and chat requests, we normalized them around each intent category.

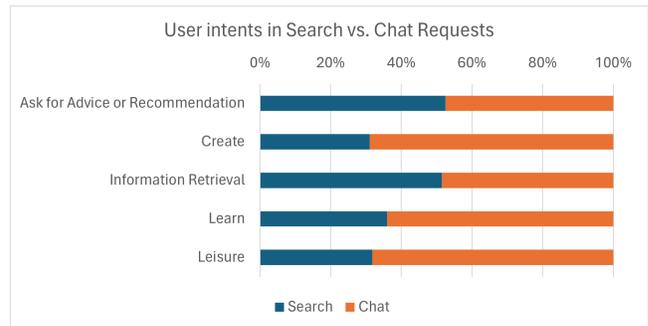


Figure 4: Comparing user intents between search and chat.

Figure 4 shows the distribution of user intents for search and chat. As shown, ‘Ask for Advice or Recommendation’ and ‘Information Retrieval’ are almost evenly distributed between search and chat, with a little bit of skew toward search. The other three categories (‘Create’, ‘Learn’, and ‘Leisure’) are heavily leaning toward chat. This requires a close examination.

First off, it is important to understand that this figure shows a view from the user intent perspective. If a user had an intent or a task related to one of the five intents considered here, where would they go – search or chat? We found that while they could use either for their ‘Information Retrieval’ or ‘Ask for Advice or Recommendation’ needs, they are favoring (with a significant tilt toward) chat for their create, learn, and leisure intents. ‘Create’ makes sense because it is more suitable to use a generative AI tool like Bing Chat for a creation task than a search engine. Of course, users are still sending search engines with create-related requests, but we hypothesize that as generative IR tools such as AI chat become more capable and known, that intent will shift more dramatically from search to chat. Similarly, ‘Leisure’ makes sense because here the user intends to have a social or open-ended conversation, which is by design supported through chat. Perhaps more interesting finding here is with respect to ‘Learn’. Learning is considered to be a higher-level goal or task in information seeking [46]. While people have used keyword-based search system for such a task, with chat-based generative IR systems, the intent fits the modality more appropriately. Through a manual inspection of some of the logs available to us, we could see that users are indeed issuing higher-level and complex requests, often associated with learning, through the chat interface. We should note this with

a caveat that our unit of analysis here is a single request from the user. It is possible that the user issued multiple queries in a given search session to accomplish their learning task. Even then, it is interesting to learn that users are preferring to issue their single-request learning requests through chat. As the information access systems with emerging technology such as generative AI and conversation-based modalities chart their ways to support users in new meaningful ways, they should consider their designs from user intent perspective.

### 4.3 Steps for Generating Intent Taxonomies

Now that we have described the methodology and demonstrated how it could be executed using an application, we summarize the lessons from these experiments and provide a guidance to anyone who wants to use LLMs for generating, validating, or applying user intent taxonomies.

- (1) *Identify Application and Data.* A taxonomy must fulfill the purpose for which it is built. That also means an existing taxonomy may not be right for your application. Assuming you want to build a purpose-driven taxonomy, prepare a detailed description of what user intent means for your application and how it should be used. For instance, in our case, it was important for us to stay focused on users' actions in a task rather than the objects when recognizing intents. This means we would not want intents that are tied to an object (e.g., 'finding information about tax') and stay close to a general action or objective (e.g., 'information retrieval'). It is also important to have as clean data as possible for an LLM to process it appropriately. Depending on which LLM you use, you may need to check for input requirements such as the size and language of input tokens.
- (2) *Build and Fine-tune Taxonomy.* Pay attention to the first prompt you prepare for building the taxonomy. Add details of your application/task, your criteria for a good taxonomy (see Table 1), and relevant constraints (e.g., number of levels, number of categories, length of a label). We recommend using different subsets of your training data to build different versions of the taxonomy to see how sensitive it is to the data used, like what we showed in Section 3.3. If a zero-shot taxonomy built with your LLM with this prompt is good enough for you, you can skip the following sub steps.
  - (a) *Check for Comprehensiveness.* Construct a prompt to annotate input data using the taxonomy built. Feed the training data to the LLM with this prompt to have it label the data. If what falls under 'Other' category is more than, say, 5% of your data, you may need to create additional categories or levels to make your taxonomy more complete.
  - (b) *Check for Consistency.* Assuming your training data is of a reasonable size, it may not be feasible to manually check labels for each of the samples, but you can take an appropriate random sample and see if the LLM consistently applied the definitions of various categories. You could also do multiple runs of Step-2 and see if a sample gets labeled the same way every time. Since this is an iterative and exploratory process, you can decide how far and deep you want to go before seeing good enough convergence and consistency.

- (c) *Improve Taxonomy's Clarity.* Once the above steps are done reasonably well or skipped as appropriate, your taxonomy is now fixed. At this point, you may ask the LLM to revise and expand the definition or description for each of your labels to improve its clarity. Often, feeding appropriate examples (positive and negative) can be useful – similar to how a human annotator is trained.
- (3) *Measure Accuracy and Conciseness.* As a final and another optional step, you can give the LLM test data, ensuring this data was not used before for any training purposes, for doing annotations using the final version of the taxonomy. Is there any category that does not get enough samples? If so, you may decide to remove that category to improve your taxonomy's conciseness. Note that if you do this, you may have to repeat some of the steps from before because now those samples will fall under other categories, which may affect some of the criteria evaluated before. Now take a random sample of labeled data and have a human annotator label it using the same instructions given to the LLM as prompt. Measure the ICR between human annotations and those from the LLM for the same data. This measurement will give you a sense of how accurate or valid your taxonomy is as well as your LLM's annotation capabilities. If at this point you have taken all the steps before (or skipped them as appropriate) and found a high enough ICR, your taxonomy and your LLM have been thoroughly tested.

## 5 CONCLUSION

Identifying user intents in online information access is highly crucial for most search and recommender systems. But doing so is often very challenging. Even if one has a pre-defined taxonomy of user intents, training a machine learning model or using such a model to annotate rapidly changing behavioral traits in new modalities such as AI chat can be expensive or infeasible. Large language models (LLMs) are shown to be effective at extracting concepts, descriptions or summaries, and examples from given set of text. This could be used for building and using taxonomies containing user intents, but there is a danger of creating a feedback loop without a clear evaluation.

In this paper we presented a novel methodology for using LLMs in generating, validating, and using taxonomies for identifying user intents in various applications. The methodology was demonstrated using an application of understanding user intents in AI chat logs. A case study was then presented with the application of contrasting user intents between search and chat. The results from both the applications are intriguing, presenting a set of new hypotheses and calling for further explorations. However, the primary contribution of this paper is the methodology for deploying LLMs in such research tasks.

As a reference from our own experiments, building the full pipeline in that case study (Section 4) took less than half the time and effort compared to the process executed for developing the method (Section 3). The process described in Section 4.3 will take substantially less even if all the optional steps are executed. Such efficiency is more than simply reducing the effort for one set of experiments. Emerging technologies such as AI-driven chat are getting discovered and used by a large set of new users. As they

become more accustomed, we can expect to see the kind of tasks they do and the kind of intents they have with these modalities. The approach presented here will allow researchers to adapt to these evolving intents quickly and at lower cost and effort.

Through the development of this methodology, we learned that we could use GPT-4 for a zero-shot construction of a user intent taxonomy, given some log data with user requests in natural language. While this taxonomy is of reasonably high quality, we found the need to have human verification and fine tuning to ensure that such a taxonomy meets various criteria commonly expected in the literature and in practice, including comprehensiveness, consistency, clarity, accuracy, and conciseness. Through the development of this methodology and its subsequent application in a different case study, we showed how these criteria can be reliably fulfilled using GPT-4 and human-in-the-loop.

In this regard, we conclude that an LLM such as GPT-4 can serve as a collaborator or a copilot rather than a replacement for human researchers. This human-LLM collaboration can yield not only faster construction and validation of a new user intent taxonomy, but also higher quality outputs with crisply defined labels, descriptions, and examples. Once the phases of construction and validation are done, the LLM can very effectively and accurately perform the annotation task, turning from copilot to autopilot. This can allow us to analyze large-scale data and generate insights. Finally, we found that often GPT-4 not only made things go faster, but also better. In cases of disagreements with human annotators, we found that GPT-4 was producing user intent labels truer to the data given rather than extrapolating to situations for which we lacked evidence. In short, the work reported here charts a new territory for using LLMs as collaborators and consignors for user intent analysis in an effective, efficient, and responsible manner.

## REFERENCES

- [1] Marissa D Abram, Karen T Mancini, and R David Parker. 2020. Methods to integrate natural language processing into qualitative research. *International Journal of Qualitative Methods* 19 (2020), 1609406920984608.
- [2] Eugene Agichtein, Ryan W White, Susan T Dumais, and Paul N Bennet. 2012. Search, interrupted: understanding and predicting search task continuation. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. 315–324.
- [3] Qingyao Ai, Ting Bai, Zhao Cao, Yi Chang, Jiawei Chen, Zhumin Chen, Zhiyong Cheng, Shoubin Dong, Zhicheng Dou, Fuli Feng, et al. 2023. Information Retrieval Meets Large Language Models: A Strategic Report from Chinese IR Community. *arXiv preprint arXiv:2307.09751* (2023).
- [4] Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063* (2019).
- [5] Muneera Bano, Didar Zowghi, and Jon Whittle. 2023. Exploring Qualitative Research Using LLMs. *arXiv preprint arXiv:2306.13298* (2023).
- [6] Garbiel Bénédicte, Ruqing Zhang, and Donald Metzler. 2023. Gen-IR@ SIGIR 2023: The First Workshop on Generative Information Retrieval. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 3460–3463.
- [7] Valeriia Bolotova, Vladislav Blinov, Falk Scholer, W Bruce Croft, and Mark Sanderson. 2022. A non-factoid question-answering taxonomy. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1196–1207.
- [8] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.
- [9] Andrei Broder. 2002. A taxonomy of web search. In *ACM Sigir forum*, Vol. 36. ACM New York, NY, USA, 3–10.
- [10] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.
- [11] Courtnei Byun, Piper Vasicek, and Kevin Seppi. 2023. Dispensing with Humans in Human-Computer Interaction Research. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–26.
- [12] Belen Carrion, Teresa Onorati, Paloma Diaz, and Vasiliki Triga. 2019. A taxonomy generation tool for semantic visual analysis of large corpus of documents. *Multimedia Tools and Applications* 78 (2019), 32919–32937.
- [13] Soumen Chakrabarti, Byron Dom, Rakesh Agrawal, and Prabhakar Raghavan. 1997. Using taxonomy, discriminants, and signatures for navigating in text databases. In *VLDB*, Vol. 97. 446–455.
- [14] Long Chen, Dell Zhang, and Levene Mark. 2012. Understanding user intent in community question answering. In *Proceedings of the 21st international conference on world wide web*. 823–828.
- [15] Pu-Jeng Cheng, Ching-Hsiang Tsai, Chen-Ming Hung, and Lee-Feng Chien. 2006. Query taxonomy generation for web search. In *Proceedings of the 15th ACM international conference on Information and knowledge management*. 862–863.
- [16] Shui-Lung Chuang and Lee-Feng Chien. 2003. Automatic query taxonomy generation for information retrieval applications. *Online Information Review* 27, 4 (2003), 243–255.
- [17] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20, 1 (1960), 37–46.
- [18] Nick Craswell, Onno Zoeter, Michael Taylor, and Bill Ramsey. 2008. An experimental comparison of click position-bias models. In *Proceedings of the 2008 international conference on web search and data mining*. 87–94.
- [19] Yogesh K Dwivedi, Nir Kshetri, Laurie Hughes, Emma Louise Slade, Anand Jayaraj, Arpan Kumar Kar, Abdullah M Baabdullah, Alex Koohang, Vishnupriya Raghavan, Manju Ahuja, et al. 2023. “So what if ChatGPT wrote it?” Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management* 71 (2023), 102642.
- [20] Guglielmo Faggioli, Laura Dietz, Charles Clarke, Gianluca Demartini, Matthias Hagen, Claudia Hauff, Noriko Kando, Evangelos Kanoulas, Martin Potthast, Benno Stein, et al. 2023. Perspectives on Large Language Models for Relevance Judgment. *arXiv preprint arXiv:2304.09161* (2023).
- [21] Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin* 76, 5 (1971), 378.
- [22] Jianfeng Gao, Chenyan Xiong, and Paul Bennett. 2020. Recent advances in conversational information retrieval. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2421–2424.
- [23] Philipp Grandeit, Carolyn Haberkern, Maximiliane Lang, Jens Albrecht, and Robert Lehmann. 2020. Using BERT for qualitative content analysis in psychosocial online counseling. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*. 11–23.
- [24] Jeffrey Heer. 2019. Agency plus automation: Designing artificial intelligence into interactive systems. *Proceedings of the National Academy of Sciences* 116, 6 (2019), 1844–1850.
- [25] Tom Hope, Doug Downey, Daniel S Weld, Oren Etzioni, and Eric Horvitz. 2023. A computational inflection for scientific discovery. *Commun. ACM* 66, 8 (2023), 62–73.
- [26] Bernard J Jansen, Danielle L Booth, and Amanda Spink. 2007. Determining the user intent of web search engine queries. In *Proceedings of the 16th international conference on World Wide Web*. 1149–1150.
- [27] Vitor Jeronimo, Luiz Bonifacio, Hugo Abonizio, Marzieh Fadaee, Roberto Lotufo, Jakub Zavrel, and Rodrigo Nogueira. 2023. InPars-v2: Large Language Models as Efficient Dataset Generators for Information Retrieval. *arXiv preprint arXiv:2301.01820* (2023).
- [28] Angelika Kaplan, Thomas Kühn, Sebastian Hahner, Niko Benkler, Jan Keim, Dominik Fuchß, Sophie Corallo, and Robert Heinrich. 2022. Introducing an Evaluation Method for Taxonomies. In *Proceedings of the 26th International Conference on Evaluation and Assessment in Software Engineering*. 311–316.
- [29] Melanie Kellar, Carolyn Watters, and Michael Shepherd. 2007. A field study characterizing Web-based information-seeking tasks. *Journal of the American Society for information science and technology* 58, 7 (2007), 999–1018.
- [30] Weize Kong, Rui Li, Jie Luo, Aston Zhang, Yi Chang, and James Allan. 2015. Predicting search intent based on pre-search context. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 503–512.
- [31] Dennis Kundsich, Jan Muntermann, Anna Maria Oberländer, Daniel Rau, Maximilian Röglinger, Thorsten Schoormann, and Daniel Szopinski. 2021. An update for taxonomy designers: methodological guidance from information systems research. *Business & Information Systems Engineering* (2021), 1–19.
- [32] Shuokai Li, Ruobing Xie, Yongchun Zhu, Xiang Ao, Fuzhen Zhuang, and Qing He. 2022. User-centric conversational recommendation with multi-aspect user modeling. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 223–233.
- [33] Helen Lippell. 2022. *Taxonomies: Practical Approaches to Developing and Managing Vocabularies for Digital Information*. Facet Publishing.

- [34] Jiqun Liu, Matthew Mitsui, Nicholas J Belkin, and Chirag Shah. 2019. Task, information seeking intentions, and user behavior: Toward a multi-level understanding of Web search. In *Proceedings of the 2019 conference on human information interaction and retrieval*. 123–132.
- [35] Matthew Mitsui, Chirag Shah, and Nicholas J Belkin. 2016. Extracting information seeking intentions for web search sessions. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. 841–844.
- [36] Steven Moore, Richard Tong, Anjali Singh, Zitao Liu, Xiangen Hu, Yu Lu, Joleen Liang, Chen Cao, Hassan Khosravi, Paul Denny, et al. 2023. Empowering Education with LLMs-The Next-Gen Interface and Content Generation. In *International Conference on Artificial Intelligence in Education*. Springer, 32–37.
- [37] Joe Raad and Christophe Cruz. 2015. A survey on ontology evaluation methods. In *Proceedings of the International Conference on Knowledge Engineering and Ontology Development, part of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*.
- [38] Filip Radlinski and Nick Craswell. 2017. A theoretical framework for conversational search. In *Proceedings of the 2017 conference on conference human information interaction and retrieval*. 117–126.
- [39] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research* 21, 1 (2020), 5485–5551.
- [40] Pengjie Ren, Zhongkun Liu, Xiaomeng Song, Hongtao Tian, Zhumin Chen, Zhaochun Ren, and Maarten de Rijke. 2021. Wizard of search engine: Access to information through conversations with search engines. In *Proceedings of the 44th International ACM SIGIR Conference on research and development in information retrieval*. 533–543.
- [41] Daniel E Rose and Danny Levinson. 2004. Understanding user goals in web search. In *Proceedings of the 13th international conference on World Wide Web*. 13–19.
- [42] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature* (2023), 1–9.
- [43] Scott Spangler and Jeffrey Kreulen. 2002. Interactive methods for taxonomy editing and validation. In *Proceedings of the eleventh international conference on Information and knowledge management*. 665–668.
- [44] Jaime Teevan, Eytan Adar, Rosie Jones, and Michael AS Potts. 2007. Information re-retrieval: Repeat queries in Yahoo’s logs. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. 151–158.
- [45] Petter Törnberg. 2023. How to use LLMs for Text Analysis. *arXiv preprint arXiv:2307.13106* (2023).
- [46] Pertti Vakkari. 2016. Searching as learning: A systematization based on literature. *Journal of Information Science* 42, 1 (2016), 7–18.
- [47] Alexandra Vtyurina, Charles LA Clarke, Edith Law, Johanne R Trippas, and Horatiu Bota. 2020. A mixed-method analysis of text and audio search interfaces with varying task complexity. In *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval*. 61–68.
- [48] Ryan Watkins. 2023. Guidance for researchers and peer-reviewers on the ethical use of Large Language Models (LLMs) in scientific research workflows. *AI and Ethics* (2023), 1–6.
- [49] Ryen W White and Steven M Drucker. 2007. Investigating behavioral variability in web search. In *Proceedings of the 16th international conference on World Wide Web*. 21–30.
- [50] Hong Xie. 2002. Patterns between interactive intentions and information-seeking strategies. *Information processing and Management* 38, 1 (2002), 55–77.
- [51] Hui Yang. 2012. Constructing task-specific taxonomies for document collection browsing. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. 1278–1289.
- [52] Oren Zamir and Oren Etzioni. 1998. Web document clustering: A feasibility demonstration. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. 46–54.
- [53] Hongfei Zhang, Xia Song, Chenyan Xiong, Corby Rosset, Paul N Bennett, Nick Craswell, and Saurabh Tiwary. 2019. Generic intent representation in web search. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 65–74.
- [54] Peiyan Zhang, Jiayan Guo, Chaozhao Li, Yueqi Xie, Jae Boum Kim, Yan Zhang, Xing Xie, Haohan Wang, and Sunghun Kim. 2023. Efficiently leveraging multi-level user intent for session-based recommendation via atten-mixer network. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*. 168–176.

## A ETHICAL CONSIDERATIONS

The work described here heavily relies on LLMs, which are shown to have several issues in their training (e.g., bias in datasets used), application (e.g., toxicity and hallucination), as well as user perceptions (blindly believing all responses due to implicit trust in such systems). Our work is not associated with direct use of LLMs for end-users, but it does concern itself with interpreting what those users may be doing using LLMs. The main contribution in this paper is a methodology that can be used for developing and using user intent taxonomies. Given the efficiency and effectiveness of this method, one may be inclined to use it in a wide range of applications that may not be advisable. For example, this method should not be used on certain vulnerable populations such as children and people with different disabilities. The results – either in terms of constructing taxonomies to understand their behavior or in terms of applying taxonomies to generate insights or make recommendations – could be misleading and potentially harmful. We note that while we were using data that is not publicly available, it lacked any identifying information about the users. Abundance of caution was taken to ensure privacy and protection of this data and the users. However, someone else using our method on other datasets must do their own due diligence to protect the users from any potential harmful effects.

## B PROMPT FOR CHAT USER INTENT TAXONOMY GENERATION (SECTION 3.3)

### Context and data description

- Your primary goal is to generate an intent taxonomy from **the given data** and **the given existing taxonomy (if available)**. You can use the taxonomy to organize and understand your data.
- You will be given information about a list of human-AI conversations. For each conversation, you’ll be given a short summary about what the user task is performed in this conversation.
- You may also be given an existing taxonomy in the **table** format, where row in this taxonomy is a intent category. You can use this taxonomy to help you construct a new intent taxonomy. The schema of this intent taxonomy is as follows:
  - **title**: the title of the intent category
  - **description**: the description of the intent category
  - **examples**: a list of examples in the intent category, as well as a list of examples that should not be in that intent category to show the contrast

Your primary goal is to generate a taxonomy that can serve for the following use cases: The primary use case of this taxonomy is to help understand what users are doing in human-AI conversations. Entities in this taxonomy can be used to label **user intents** in human-AI conversations.

### Criteria of a generic taxonomy

- Accuracy: The definitions, descriptions of classes, properties, and individuals in a taxonomy should be correct.
- Completeness: All the data should be reliably classified using this taxonomy.

- **Conciseness:** The taxonomy should not include any irrelevant elements with regards to the user intents in AI Chat.
- **Clarity:** The taxonomy should communicate the intended meaning of the defined terms. Definitions should be objective and independent of the context.
- **Consistency:** The taxonomy does not include or allow for any contradictions.

### Requirements of your output taxonomy

- Your output **intent** taxonomy should focus on the user actions in a task, not the task objects. This is **different** from a **domain** taxonomy, which primarily describes the task objects.
- Your output taxonomy should match the existing taxonomy and the data as closely as possible, without leaving out important intent categories or including unnecessary ones. Please make sure there is no overlap or contradiction among the intent categories in your output taxonomy.
- Your output **title** of each category should be **no more than 5 words**. The title should be a concise and clear label for the intent category. It can be either verb phrases or noun phrases, whichever is more appropriate.
- Your output **description** of each category should be **no more than 30 words**. The description should explain the user's goal or purpose for the intent category, and should differentiate it from other intent categories.
- The number of examples for each intent category should be **no more than 3**. The examples should either come from the given taxonomy or the provided data with **exactly the same content**. Please do not invent new examples or intents that are not in the given taxonomy or the data.
- **Size limit of the output taxonomy**: The total number of intent categories should be **no more than 10**.
- Your output taxonomy and examples should be in **English** only.

### You are asked to answer the following questions

- Q1. Please check the above general criteria and the specific taxonomy requirements one-by-one. Does the provided taxonomy satisfy the above requirements, word limits and taxonomy size limit? Please answer "yes" or "no". If there is no given taxonomy, please answer "no".
- Q2. Please explain your answer to Q1. If your answer to Q1 is "no", please also describe if you'd like to construct the taxonomy structure from scratch or you plan to make changes on the given taxonomy. Your answer to this question should be **within 100 words**.
- Q3. If your answer to Q1 is "no", then generate a new intent taxonomy from the the given data and the given existing taxonomy (if available). Your output taxonomy should be in the **table** format with the same schema. If your answer to Q1 is "yes", please answer "N/A". Please make sure the new taxonomy satisfies **all of the above requirements**. Please **do not** invent new examples or new intents that are not in the existing taxonomy or the provided data.

### Tips

- If you're given an existing taxonomy, you can use the provided data to update this taxonomy. By incorporating the newly provided data, you can **add new categories**, **merge or generalize existing categories**, **split existing categories**, **reorganize the current tree structure**, **change titles and descriptions**, **swap examples** and do other operations if needed.
- If the intent category structure of the given taxonomy cannot be easily adjusted, then please construct a new structure of these intent categories based on their descriptions and the provided examples. Please make sure your new taxonomy covers the semantics of the existing taxonomy as thoroughly as possible. Please **do not** invent new intents that are not in the existing taxonomy or the provided data.
- You should carefully review the examples provided in each category and make sure they are correctly labeled. You can also reorganize the examples or create new categories from them when needed. You're allowed to have fewer than 3 for each category but your examples should only come from examples in **the given taxonomy or the provided data**. Please **do not** invent new examples that are not in the existing taxonomy or the provided data.
- Please make sure your new taxonomy satisfies the **word limits** and **taxonomy size limit**. You're allowed to have fewer than 10 categories in your final output. If you couldn't fit your new taxonomy into the limits, please consider merging or abstracting some specific categories into more general categories.
- Please make sure there is no overlaps or contradictions among the intent categories in your output taxonomy.

## C FINAL OUTLINE OF THE TAXONOMY FOR USER INTENTS IN CHAT (SECTION 3.4)

- **Information Retrieval:** Conversations where the user wants to find factual information or answers to specific questions. The agent's responses are typically direct, concise, and informative, providing the relevant information and/or links to the sources. This intent calls for retrieving or reconstructing factual information that already exists, rather than synthesizing or computing something new.
- **Problem Solving:** Conversations where the user wants to perform a mathematical or logical operation, such as a conversion, a percentage, a formula, or a function. The agent's responses are typically factual and computed or constructed based on available information and what the user provided. Unlike Information Retrieval intent, this intent calls for the agent to do some processing on top of simply retrieving or extracting information.
- **Learning:** Conversations where the user wants to understand a concept or acquire skills by getting detailed explanation, reasoning, or synthesis. The agent's responses are typically a synthesis of information based on several factual pieces of information, often from different sources. The Learning intent requests often involve questions like 'how', 'why', or requests like 'explain' – things that will indicate

asking for explanations or doing investigation. Also, while individual turns may be of information retrieval nature, if the user is asking multiple questions that drill into a topic, that's an indication of Learning intent.

- **Content Creation:** Conversations where the user asks the agent to either generate original content or translate existing content into new content based on specified criteria or constraints. In the case of generating original content, the user's questions require some degree of creativity, novelty, or innovation from the agent. The agent's responses contain original or translated outputs that match the user's specifications.
- **Leisure:** Conversations where the user wants to chat or play with the agent out of curiosity, boredom, or humor, or else explore broad ideas or areas of interest without a specific goal or information need in mind. There may not even be a specific question or a request. The agent's responses are typically suggestive and engaging. The agent may also encourage further inquiry or action from the user to deepen their discovery experience.

## D PROMPT FOR TAXONOMY APPLICATION (SECTION 3.5)

You will be given a conversation history between a User and an AI agent. Your task is to answer questions about the user's intent.

**User Intent** A user intent is defined as the user's purpose for conversing with the AI agent. The categories of user intents are:

- **Information Retrieval:** Conversations where the user wants to find factual information or answers to specific questions. The agent's responses are typically direct, concise, and informative, providing the relevant information and/or links to the sources. This intent calls for retrieving or reconstructing factual information that already exists, rather than synthesizing or computing something new.
- **Problem Solving:** Conversations where the user wants to perform a mathematical or logical operation, such as a conversion, a percentage, a formula, or a function. The agent's responses are typically factual and computed or constructed based on available information and what the user provided. Unlike Information Retrieval intent, this intent calls for the agent to do some processing on top of simply retrieving or extracting information.
- **Learning:** Conversations where the user wants to understand a concept or acquire skills by getting detailed explanation, reasoning, or synthesis. The agent's responses are typically a synthesis of information based on several factual pieces of information, often from different sources. The Learning intent requests often involve questions like 'how', 'why', or requests like 'explain' – things that will indicate asking for explanations or doing investigation. Also, while individual turns may be of information retrieval nature, if the user is asking multiple questions that drill into a topic, that's an indication of Learning intent.
- **Content Creation:** Conversations where the user asks the agent to either generate original content or translate existing content into new content based on specified criteria or constraints. In the case of generating original content, the user's

questions require some degree of creativity, novelty, or innovation from the agent. The agent's responses contain original or translated outputs that match the user's specifications.

- **Leisure:** Conversations where the user wants to chat or play with the agent out of curiosity, boredom, or humor, or else explore broad ideas or areas of interest without a specific goal or information need in mind. There may not even be a specific question or a request. The agent's responses are typically suggestive and engaging. The agent may also encourage further inquiry or action from the user to deepen their discovery experience.
- **Other:** This intent label can be used if none of the above labels fit. Note that you should do your best to find an appropriate label from the list above and only in the rare circumstances when you have very little to no confidence in that ability, you can use 'Other' label.

### Examples

#### Information Retrieval

"User:where are the best places to watch the sun set in spring in Madison Wisconsin AI:There are many beautiful places to watch the sun set in spring in Madison Wisconsin. Here are some of the top-rated ones according to web sources: - The Edgewater Hotel: This historic hotel offers a stunning view of Lake Mendota and the Capitol. You can also enjoy movie nights, Brewgrass Fridays, and other evening events on the rooftop terrace. - Memorial Union Terrace: This classic spot overlooks Lake Mendota and is perfect for ice cream or a pitcher. You can also catch live music and entertainment on the stage. I hope this helps you find your ideal spot for a sunset experience." Problem Solving "User: The cost for the excel training will £15,000 per training. 60% will be at the start of the training and 40% in the following month. so if the months for year were july to december interpret that information AI:Based on the information you provided, the cost for the Excel training will be £15,000 per training. 60% of this amount, which is £9,000, will be due at the start of the training. The remaining 40%, which is £6,000, will be due in the following month. If the training takes place between July and December of a given year, then these payments would be made during those months."

#### Learning

"User:why do the gummies not work for me AI:There are several reasons why THC gummies might not work for some people. One reason could be due to their **THC tolerance** and physical factors like **body size**. Another reason could be due to **gastrointestinal disorders** that affect how the small intestine absorbs nutrients, water, and medicine. Additionally, some people may have a **slow metabolism** due to their genetics. Would you like more information on any of these reasons?"

#### Content Creation

"User:create a resume tagline for this profile: Graduate in Business and Information Systems Engineering with a strong interest in Data Science, Machine Learning, and AI.

AI:How about this tagline: "Data-driven Business Engineer with a passion for Machine Learning and AI?"

### Leisure

"User: i challenge you to a pokemon battle

AI: I'm sorry, but I'm just a search engine and I don't have the ability to participate in Pokémon battles. However, you can try one of the Pokémon battle simulators I mentioned earlier to find other trainers to battle with."

### Tips

- The following intentions indicate seeking information: Information Retrieval, Problem Solving, and Learning. Here is how we can differentiate among them. Information Retrieval is often going to be the default as it involves retrieving what typically already exists. Problem Solving will involve doing something more on top of retrieval information. So, knowing the tax brackets is Information Retrieval intent, but having the agent calculate one's taxes is a Problem Solving intent. Similarly, asking what function in Python does a certain operation has Information Retrieval intent, but asking the agent to write a function in Python has Problem Solving intent. Learning typically calls for an explanation. If the user asks the agent to describe how state income tax and estate tax differ, that has Learning intent. Similarly, if the user asks the agent to explain how a Python function works, that is Learning. Also, if a conversation involves multiple (3 or more) turns that shows the user is drilling more into a given topic, that is an indication of Learning.
- The following intentions are beyond information retrieval: Transaction, Creation, and Leisure. Transaction is where the user is looking to buy, plan, or execute something and needs help or recommendations. Creation is where the user wants the agent to generate something new instead of retrieving it. Leisure is where the user is using the agent as a substitute for a human to have natural language chitchat, explore a topic for fun. The user may not even provide a question, and simply have a statement to initiate an interesting conversation.

## E PROMPT FOR COMPARING SEARCH AND CHAT USER INTENT TAXONOMY GENERATION (SECTION 4.1)

### Context and data description

- Your primary goal is to generate an intent taxonomy from **the given data**. You can use this taxonomy to organize and understand your data.
- You will be given data that contain user requests to a search system.
- Ignore the requests that are not in English.
- Ignore data with multiple intents.
- The schema of this intent taxonomy is as follows:
  - **title**: the title of the intent category
  - **description**: the description of the intent category
  - **examples**: a list of examples in the intent category, as well as a list of examples that should not be in that intent category to show the contrast

Here is an example of a taxonomy:

```
“markdown
|title|description|examples| | | |”
```

```
|title of the category|description of the category|example
1; example 2; example 3|
“
```

### Your primary goal is to generate a taxonomy that can serve for the following use cases

The primary use case of this taxonomy is to help understand what users are doing when they submit a request to a search system. Entities in this taxonomy can be used to label **user intents** in human-system interactions.

### Here are the criteria of a good taxonomy

- Accuracy: The definitions, descriptions of classes, properties, and individuals in a taxonomy should be correct.
- Completeness: All the data should be reliably classified using this taxonomy.
- Conciseness: The taxonomy should not include any irrelevant elements with regards to the user intents in AI Chat.
- Clarity: The taxonomy should communicate the intended meaning of the defined terms. Definitions should be objective and independent of the context.
- Consistency: The taxonomy does not include or allow for any contradictions.

### Requirements of your output taxonomy

- Your output **intent** taxonomy should focus on the user actions in a task, not the task objects. This is **different** from a **domain** taxonomy, which primarily describes the task objects.
- Your output taxonomy should match the data as closely as possible, without leaving out important intent categories or including unnecessary ones. Please make sure there is no overlap or contradiction among the intent categories in your output taxonomy.
- Your output **title** of each category should be **no more than 3 words**. The title should be a concise and clear label for the intent category. It can be either verb phrases or noun phrases, whichever is more appropriate.
- Your output **description** of each category should be **no more than 30 words**. The description should explain the user's goal or purpose for the intent category, and should differentiate it from other intent categories.
- The number of examples for each intent category should be **no more than 3**. The examples should either come from the given taxonomy or the provided data with **exactly the same content**. Please do not invent new examples or intents that are not in the given taxonomy or the data.
- **Size limit of the output taxonomy**: The total number of intent categories should be **no more than 5**.
- Your output taxonomy and examples should be in **English** only.

## F FINAL OUTLINE OF THE TAXONOMY FOR USER INTENTS IN SEARCH VS. CHAT (SECTION 4.1)

- **Ask for Advice or Recommendation**: The intent to seek suggestions, opinions, or guidance from others on a specific topic or situation. Examples: recommend me a good book; how can I improve my sleep quality. On the other hand,

if a request is for factual information, it should not be in this category. Example: how many countries does the UN recognize?

- **Create:** The intent to use AI tools or platforms to generate, edit, or manipulate information objects. Example: write me a poem about breakup; create a short report on effects of glaciers melting. On the other hand, if a request is for an existing item, it should be in Information Retrieval, and not in this category. Example: show me the most popular Picasso painting.
- **Information Retrieval:** The intent to find existing information or answers on the internet. Examples: what is the capital of France; where is biggest deposit of oil. On the other hand,

if the request clearly indicates that the user is further interested in learning about a topic, it should be in 'Learn' and not here. Example: explain how total solar eclipse happens.

- **Learn:** The intent to acquire new knowledge or skills on a subject of interest. Examples: teach full stack web development; explain market cap rate in commercial real estate. On the other hand, if the request is looking for retrieving existing factual information from the web, it should be marked as 'Information Retrieval'. Example: how many states are there in India?
- **Leisure:** The intent to enjoy oneself by engaging in amusing activities such as games, jokes, stories, etc. Examples: play tic tac toe; make me laugh.