Show or Tell? Modeling the evolution of request-making in Human-LLM conversations

Shengqi Zhu
Cornell University
sz595@cornell.edu

Jeffrey M. Rzeszotarski
Loyola University Maryland
jeff.rzeszotarski@gmail.com

David Mimno
Cornell University
mimno@cornell.edu

Abstract

Designing user-centered LLM systems requires understanding how people use them, but patterns of user behavior are often masked by the variability of queries. In this work, we introduce a new framework to describe request-making that segments user input into request content, roles assigned, query-specific context, and the remaining *task-independent* expressions. We apply the workflow to create and analyze a dataset of 211k real-world queries based on WildChat. Compared with similar human-human setups, we find significant differences in the language for request-making in the human-LLM scenario. Further, we introduce a novel and essential perspective of *diachronic analyses* with user expressions, which reveals fundamental and habitual user-LLM interaction patterns beyond individual task completion. We find that query patterns evolve from early ones emphasizing sole requests to combining more context later on, and individual users explore expression patterns but tend to converge with more experience. From there, we propose to understand communal trends of expressions underlying distinct tasks and discuss the preliminary findings. Finally, we discuss the key implications for user studies, computational pragmatics, and LLM alignment. Our data and code are available at https://github.com/CurlyZhu/ReCCRE.

1 Introduction

The versatile conversation format of chat-based LLM interfaces has created an unprecedented interaction paradigm by enabling open-ended user inputs (Zhu et al., 2025; Gao et al., 2024b). However, little attention has been given to the *language* aspects of human-LLM interaction, much less the formation and evolution of users' language patterns over many sessions: can we study *how* people ask, independent of *what* they are asking for? More often, NLP studies focus on the semantics and type of tasks conveyed (Tamkin et al., 2024; Cheng et al., 2025; Handa et al., 2025), while the text modality itself is simply assumed as the default, easy way. However, the natural language format is far more than a convenient form for tasks. How users organize and contextualize their requests directly reflects an LLM's user-perceived affordances, expectations, and social roles. Language does not just "help to show how LLMs are used"; it *is* how LLMs are used. Specifically, users' habitual expressions generalized across tasks, a fundamental linguistic feature of user-LLM interaction, remain largely understudied. Recent work observes real-world user-side nuances, such as users adapting their behaviors and expectations (Choi et al., 2024; Schroeder et al., 2025), or changes in utterances upon model updates (Ma et al., 2024). However, there is still yet to be a formal and generalizable framework to accommodate the linguistic analysis of input patterns beyond individual case studies.

In this work, we analyze how user inquiries are formed and explore the patterns in interactions over relative and absolute time. First, we perform a new segmentation task to tackle the challenge of users freely embedding *requests* in their *expressions*, together with other parts with substantial presence like *contexts* and assigned *roles*. Via an extendable, semi-automatic LLM annotation workflow, we present a dataset on top of WildChat (Zhao et al., 2024) with 211,414 parsed user utterances consisting of Request Content, Context, Roles, and Expressions (ReCCRE). The dataset features the clean separation of the parts specific to a request from the generic natural language expressions used to deliver the request.

Using the dataset, we make several key observations and conclusions. We show that the LLM chat modality is fundamentally different from natural request-making conversations, by comparing with the Stanford Politeness Datasets (Danescu-Niculescu-Mizil et al., 2013), a primary resource

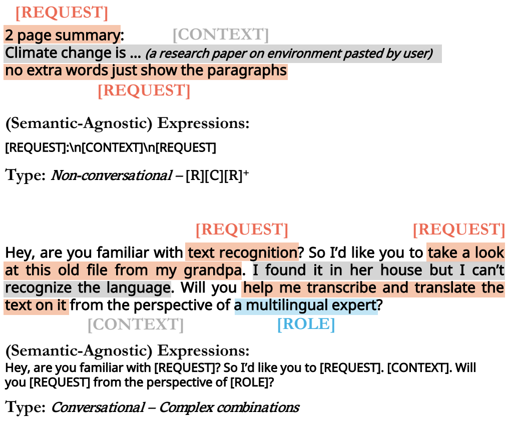
Findings of the Association for Computational Linguistics: EACL 2026, pages 5023–5034
March 24-29, 2026 ©2026 Association for Computational Linguistics

in computational pragmatics. We identify key repeating patterns in queries, which range on an axis between *request-centric* and *context-infused*. More importantly, we introduce how the perspective of expressions enables *diachronic* user modeling, with use records as a time-lapse data source for understanding user lifecycles and the community. We find clear traces that, as a user gains familiarity with the system, their expressions change less, and they migrate from simple chunks of requests to more context combinations. Further, we explore full-scale community analysis and discuss patterns over time supported by the data resource, showing that key trends are visible even with simple, non-parametric but indicative metrics like lexical diversity. We conclude with the tangible implications and future directions regarding understanding users, pragmatics, and LLM alignment.

2 Related Work

Interpreting Real-World Human-LLM Conversations has been a rising topic thanks to new data resources (Zhao et al., 2024; Zheng et al., 2024a). However, the major body of work has focused on the detection and categorization of what tasks users use LLMs for (Zhang et al., 2025b; Cheng et al., 2025; Mireshghallah et al., 2024) and their impacts (Handa et al., 2025; Tamkin et al., 2024; Kirk et al., 2024), as well as specific features such as values (Huang et al., 2025) and jailbreaking attempts (Jin et al., 2025). Beyond NLP, the formatting of prompts as well as the broader user experience with LLM input interfaces has also been core topics in Human-Computer Interaction (Zamfirescu-Pereira et al., 2023; He et al., 2025; Gao et al., 2024b; Zhang et al., 2025a).

Some existing work shares the focus on individual attributes relevant to our discussions. Lee et al. (2025) looks into evolving dialog patterns across time but focuses on inter-human and inter-LLM data synthesis instead of existing user-LLM documentations. For human-LLM, Huang et al. (2024) considers the "conversational tones" shared or (mis)aligned between humans and LLMs, and Zheng et al. (2024b) probes LLM performance with different role assignments. However, our work is fundamentally different: Prior work usually targets what a human-LLM conversation *should* look like, a dominant thread related to fine-tuning and deployment (Mott et al., 2024; Mishra et al., 2022; Ivey et al., 2024); Our work focuses on the mining of existing data and new analysis paradigms. More importantly, we present a new systematic view that is not covered by the studies on single attributes.



2 page summary:
Climate change is ... *(a research paper on environment pasted by user)*
no extra words just show the paragraphs

(Semantic-Agnostic) Expressions:

[REQUEST]:\n[CONTEXT]\n[REQUEST]

Type: *Non-conversational* – [R][C][R]$^+$

Hey, are you familiar with text recognition? So I'd like you to take a look at this old file from my grandpa. I found it in her house but I can't recognize the language. Will you help me transcribe and translate the text on it from the perspective of a multilingual expert?

(Semantic-Agnostic) Expressions:
Hey, are you familiar with [REQUEST]? So I'd like you to [REQUEST]. [CONTEXT]. Will you [REQUEST] from the perspective of [ROLE]?

Type: *Conversational – Complex combinations*

Figure 1: Two examples of user input annotated with request content, context, and roles. The precise definitions of the components are discussed in §3.1, and the expression types are elaborated in Table 1.

At a high level, our work is most related to Mysore et al. (2025) and Kolawole et al. (2025), which also seek to extract and describe latent patterns from massive user-LLM inputs. However, both works have distinct goals from this work: the former focuses on assisted writing and marks behavioral "PATHs" as a dialog proceeds, and the latter targets task taxonomies with semantic similarities. Both works resonate with key threads in conversation studies such as task-oriented dialogs and intent classification, which consider specific, pre-defined types of intents. In comparison, we target a more fundamental and unique level of user language in human-AI interaction. User-LLM conversations are extremely heterogeneous, covering any topic via any possible expressions. Our work models language use commonalities that define user-LLM interaction *despite* seemingly multifarious goals: Parallel to the completion of tasks, patterns of expressions and coordination of input components are directly related to user experience and require specific elicitation and understanding. In this work, we draw diachronic insights *from* the language composition and expression aspect *per se*, as opposed to filtering the "surface information" out to focus on key intents.

3 Annotating User Request-Making

We start by constructing the data infrastructure for modeling request-making behaviors by segmenting requests. Our goal is a generalizable annotation scheme that enables systematic quantitative and linguistic analysis of user requests, and allows adaptation to unseen data and to natural language conversations for comparisons.

3.1 Task Setup

Source Data We collect 317,373 initial user inputs (i.e., first turns) from the WildChat dataset (Zhao et al., 2024) as the base corpus.

Annotating User Queries As the first step, we distinguish between the making of an effortful request and the direct retrieval of answers (e.g., "who is the 3rd president of the U.S."). While the latter represents another common type of usage, the engagement level is low; it is less likely to involve a conversational scenario, and the dynamics are remarkably different from request-making. In practice, the annotator first reads the input text thoroughly and determines whether it involves a request or a direct question, or "both" or "neither".

Next, for the request-making cases, we outline the case-specific core semantics relevant to the request. This separates the framing templates used to deliver the request (expressions) across different dialogs. Specifically, we introduce the following annotations of three elements of requests plus the user expressions, on top of the full user input text:

- **Request Content** ($[R]$): A core span of text that specifies what task(s) exactly the user wants the LLM to perform, or what goal(s) the user wants to achieve;
- **Context** ($[C]$): A detailed span of context information that does not directly constitute the request, but provides support for neighboring requests. This includes the chunks of "target text" that the LLMs are requested to process (e.g., the pasted article for the request "2 page summary" in Fig. 1).
- **Role** ($[role]$): Any roles that the LLM is asked to take on to achieve the requests.
- **Expression**: The remaining text after extracting the above components, which represent the generic language templates used to embed and deliver the requests.

Each word in a request is assigned to precisely one of the four categories, and the labels thus form a non-overlapping full division of the user input. This forms the basis of the corpus annotation, and we discuss next the implementation in full scale.

3.2 Automating the annotation pipeline

To adapt the annotation to chat logs at the million-request scale, we seek to balance between automation and reliability (consistency). We implement a review-and-revise pipeline that simultaneously generates semi-supervised annotations and extends the available data for fine-tuning with consistent standards learned from human annotation.

The pipeline involves three contributors: a human annotator; an interim SotA LLM specialized in text understanding (L); and a smaller local LLM as full-scale annotator (l). We bootstrap from a small number of hand-labelled data, use L as pseudo-reference for fine-tuning l, and eventually automate annotations with the fine-tuned local l.

First, the human annotator manually labeled a small, random batch of "root" data. This small collection of references is consistent and we denote this *gold* dataset as D_0. Both LLM annotators are then provided with a detailed prompt of annotation rules and fine-tuned based on D_0. Next, both L and l are evaluated on a separate, randomly sampled set D_1^{raw}. We then convert this raw subset into a *silver* set D_1 as follows:

- If L and l *agree* on the instance (the total difference in labels sufficiently low), the annotation of L is accepted and added to D_1.
- Otherwise, if L and l *disagree*, the human annotator reviews and selects an output, and if both are incorrect, the instance is manually labeled. The reviewed/relabelled version is added to D_1.

D_1 as a silver set is then merged with D_0 to form an expanded annotated dataset for fine-tuning. Both L and l are then fine-tuned and evaluated on another new batch D_2^{raw}, and the process is repeated so on and so forth. Finally, after fine-tuning the models with incremental, semi-supervised data, we migrate the model prediction process from the black-box, high-cost L to our local model l to perform full-scale automated annotation.

We use `gpt-4o-2024-08-06` as the intermediate L, and a flagship 10B-level open-source LLM at the time of the work, `Qwen2.5-14B-Instruct`, as the full-scale annotator l. The review-and-revise loop was repeated 3 times on a total of 762 instances, with the agreement rates between L and l validated as 57%, 79%, and 87%. A template-based verification showed that ill-formatted responses of l after the loops are $< 0.5\%$.

Type		Description	Examples
Non-conversational	$[R]$	A single *Request* component.	$[R]$ Give me $[R]$.
	$[R] * n$	Multiple *Request*s concatenated in simple ways, without conversational expressions.	$[R]$ and $[R]$. $[R]$. $[R]$. Also $[R]$.
	$[R][C]$	One *Request* followed by One *Context* component, without conversational expressions.	$[R]$: $[C]$ $[R]$ such as $[C]$.
	$[C][R]$	One *Context* followed by One *Request* component, without conversational expressions.	$[C]$. Now, $[R]$. $[C]$\n\n$[R]$
	$[R][C][R]^+$	A pair of *Request* and *Context*, followed by additional *Request* components.	$[R]$, $[C]$. $[R]$ and $[R]$. $[R]$: $[C]$. $[R]$. $[R]$. $[R]$.
	$[R][C][C]^+$	A pair of *Request* and *Context*, followed by additional *Context* components.	$[R]$ based of $[C]$: $[C]$. $[R]$. $[C]$. $[C]$. $[C]$.
	$[C]^+$	Concatenation of *Context* components only. Usually seen in early requests to complete writings.	$[C]$. $[C]$ $[C]$
	Other $[R]/[C]/[role]$ compositions	Other more complicated series of $[R]$, $[C]$, and $[role]$, with simple, non-conversational expressions.	$[R]$, $[C]$. Then, $[R]$, $[C]$. Finally, $[R]$, $[C]$. $[C]$. $[C]$. Given $[C]$, $[R]$.
Conversational	Single $[R]$	One single *Request* in a conversational expression.	Can you help me to $[R]$? Hi I wanna $[R]$.
	Simple $[R]/[C]/[role]$ combinations	Simple combinations of *Request*, *Context*, and *Role* using conversational expressions.	You are $[role]$. Now, $[R]$. I'm working on $[C]$ and I'd like you to $[R]$.
	Complex compositions	Other more complicated series of $[R]$, $[C]$, and $[role]$, with full, conversational expressions.	How can I $[R]$? $[R]$. Please be sure to $[R]$! Act as $[role]$ and $[R]$. You will $[R]$: $[C]$.

Table 1: Taxonomy of user expressions, with 8 non-conversational types and 3 conversational types.

4 The ReCCRE dataset

4.1 Basic information

After obtaining the valid annotations, we set up a spam filter to remove consecutive dialogs that are overly similar or created within a very short period, and also remove the instances that do not involve request-making. This results in a collection of 211,414 user request-making inputs from 18,964 users. The dataset covers the time window from April 2023 to May 2024, and spans 6 versions of the `gpt-3.5-turbo` and `gpt-4` models.[1]

Long-term Users To track the change of use patterns over time, we focus on the core users with sufficient experience and time to play with and adapt to the system. In practice, we seek users (1) whose use records (time between first and last dialog created) span more than 14 days, and (2) have started at least 10 dialogs. This yields a subset of 2,092 users with 59,175 request-making user inputs in total. We refer to this key group as *long-term* or *stable* users. Our analyses will primarily focus on this group, as it more reliably reflects use patterns and provides the legitimacy for diachronic observations. We present a full-scale case study comparing long-term users with all users and the full data in §5.2 under the Lexical Richness framework.

4.2 Categorization and Illustration

4.2.1 Taxonomy of Expressions

To understand and generalize the annotations, we start from a taxonomy of expressions in request-making shown in Table 1. We refer to an expression as *conversational* if it involves explicit signs of conversing with another party, such as person ("You" and "I"), politeness strategies, and greetings. Conversely, if an utterance is a mere imperative combination of $[R]$ and $[C]$ without signs of conversation, it is marked as *non-conversational*. We further break down the expressions based on their structure and complexity, e.g., whether multiple request or context components are involved.

Anchor Points To calibrate the varied user inputs, we collect 40 different instances of the most frequent expressions in the dataset, covering all 11 categories. We use them as *anchor points* in our analyses (including subsequent figures) to provide reference for visualizations, and more importantly, to allow categorization of arbitrary unseen expressions based on their closest anchor points.

4.2.2 Visualizing the data distribution

We vectorize user expressions with a SotA text-embedding model, `gte-large-en-v1.5` (Li et al., 2023), with $[R]$, $[C]$, and $[role]$ wrapped as formatted placeholder tokens (`__[REQUEST]__`, etc.) Each request-making utterance is encoded as a 1024-dim vector, and a user is represented by the average of their utterances. We then map all users as well as the anchor points to a 2-D space using PaCMAP (Wang et al., 2021), a SotA dimension reduction (DR) method. In this way, the collections of user expressions are depicted as a scatter plot in Figure 2, where each bubble represents a long-term user. The bubble size is in proportion to the number of dialogs a user created. Bubbles are colored based on the closest anchor point, or grey if not

[1] See Zhao et al. (2024) for the detailed documentation.

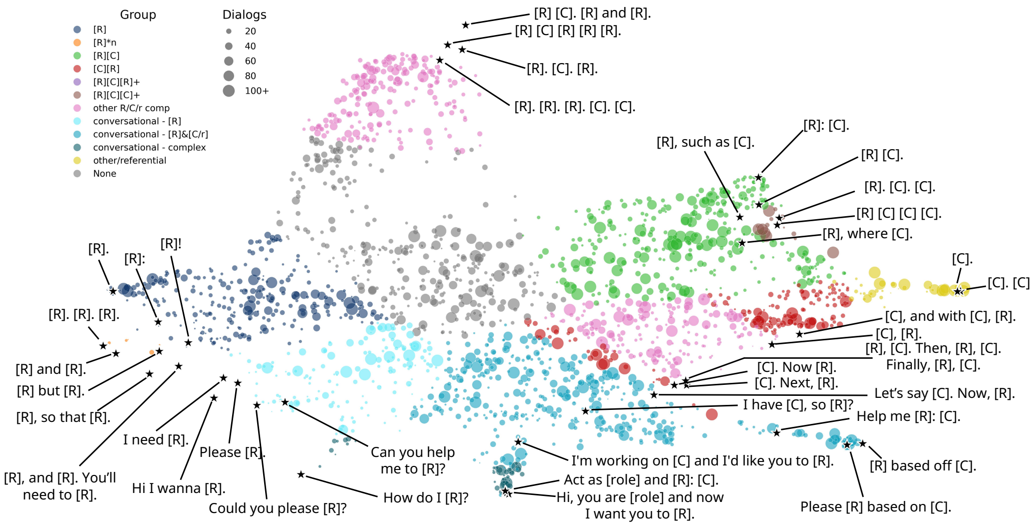


Figure 2: The overview of the ReCCRE dataset as a user-level 2-D plot. Each circle represents a user, with its size matching their total dialogs and color clustered based on the closest anchor point. In general, the horizontal axis displays the ratio of $[R]$ and $[C]$, and the vertical axis ranges from the most to least conversational.

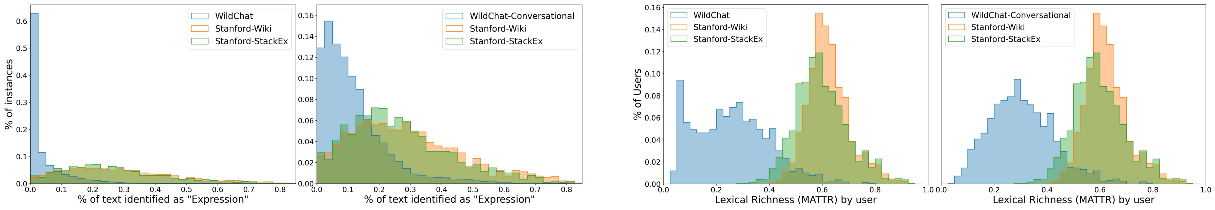


Figure 3: Comparing ReCCRE-WildChat and Stanford Politeness dataset: the amount (%) of task-independent expressions (left), and the distribution of speaker Lexical Richness (right; measured by Moving-Average Text-Token Ratio). For each metric, the first figure shows the full data and the second the conversational portion only.

close enough to any anchor points.

Note that the horizontal layout corresponds to the composition of Goals ($[R]$) and Context ($[C]$), where the leftmost represents the sole $[R]$ and the rightmost corresponds to $[C]$ only, and the combined forms are in between. Meanwhile, the vertical positions can be interpreted as the "conversationality": the bottom ones, featuring role assigning and politeness patterns like "please", are closest to the expressions and force of natural conversations; the topmost ones, with a repetitive sequence of $[R]$ and $[C]$ with no additional text, is most tool-like and unlikely in human-human request-making.

4.3 Human-LLM request-making differs from human-human counterparts

Factoring out case-specific contexts allows us to compare request-making language in LLM chats with human-human interaction. To evaluate this difference, we apply the same segmentation scheme to the Stanford Politeness Datasets (Danescu-Niculescu-Mizil et al., 2013), a prominent pragmatics corpus relevant to request-making. The corpus records two natural dialog sources, Wikipedia editor discussions (Stanford-Wiki) and the Stack-Exchange Q&A forum (Stanford-StackEx). Each utterance by design involves the speaker making requests to another member. We therefore compare how requests are delivered in the two human-human and the user-LLM scenarios.

We note the distinct composition of the utterances and the Lexical Richness of expressions (Laufer and Nation, 1995; Shen, 2022). Figure 3 compares (1) the percentage of input text as expression and (2) the user-level Moving-Average Text-Token Ratio (Covington and McFall, 2010) of ReCCRE and Stanford datasets. The two natural subsets share highly similar stats despite distinct sources and contexts; however, ReCCRE uses qualitatively fewer expressions to embed requests, and user language has much lower diversity. One major reason is the presence of non-conversational "imperative" spans (Table 1); e.g., "Write an article about jogging." is marked as a single $[R]$ com-

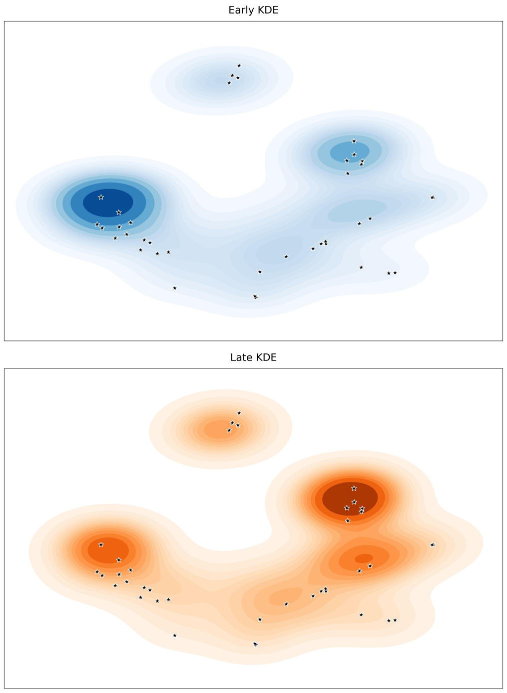


Figure 4: Distribution of the 1st (left) and 20th dialogs (right) of eligible long-term users as Kernel Density Estimation (KDE) plots in the 2D space from Fig. 2.

ponent with minimal or no formatting. This is a common paradigm different from natural conversations, as the latter would require appropriate social grounding. However, if confined to the *conversational* instances (the subplots on the right in both figures) by removing the simple combinations of request content and context, we see that the difference still holds. In other words, the existence of the non-conversational imperatives is not sufficient to account for the difference; there exist fundamental contrasts of language use specific to the human-LLM scenario, independent from request content.

5 Diachronic Analysis of Request-making

Enabled by the new data infrastructure, we move on to discuss a novel paradigm: modeling user behaviors and patterns in a *diachronic* manner, thereby understanding interaction as a systematic and dynamic process. We will first inspect the lifecycles of individual users, and discuss how the fundamental properties, such as the diversity of expression types, change across time. Next, we interpret the holistic evolution patterns formed by the community collectively, and extend further to compare the core user base and the lay public.

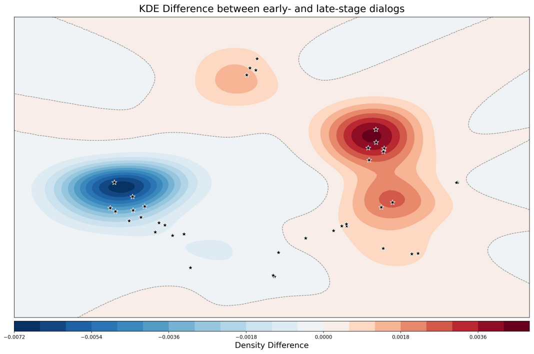


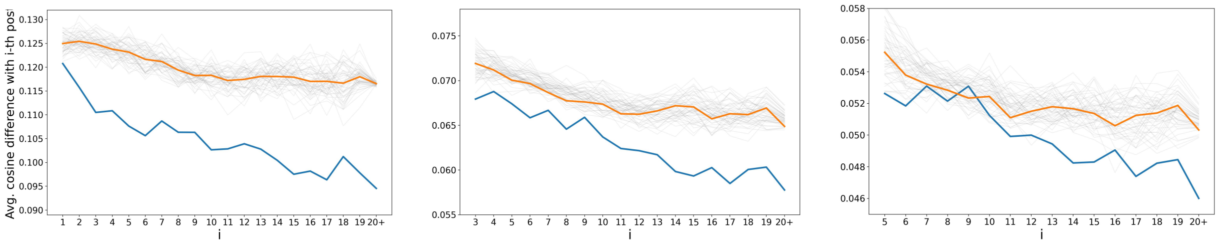
Figure 5: Difference of densities between the early and late inputs of long-term users (Fig. 4). In the later stage, the balanced combination of $[R]$ and $[C]$ (upper-right, red) sees major inflows, while the non-conversational stacks of $[R]$s (blue, left) significantly decreases.

5.1 Modeling User Lifecycle

Long-term, non-intrusive documentary of user-LLM interactions provides an interface for the user lifecycles (Zhu et al., 2025), i.e., the full course of actions and use history. Here, we discuss how a fully text-based analysis can help to model the change of use patterns across time.

5.1.1 What expressions were used and how are they structured?

As a natural continuation of Figure 2, we further add the dimension of time and inspect the most common patterns across individual dialogs at different stages. To model user lifecycles, we position a pair of early- and late-stage input data from long-term users in the same 2-D space from Fig. 2 and apply Kernel Density Estimation to model the frequency of user expressions. Figure 4 shows the comparison of the 1st (left) and the 20th (right) request-making dialog of all eligible long-term users, with the same anchor points from Fig. 2, and we directly depict the difference between the two KDEs in Figure 5. For the users' first explorations, most utterances consist of only one $[R]$ or the simple combination of two. This describes the exploratory stage of interaction with the system with more concise and generic requests. However, as users gain familiarity, this pattern drastly decreases (though still a major type), and the mass is significantly transitioned to the addition of more specific contexts on the right side, as well as the more complex type of multiple goals and context (top). This suggests a communal shift towards requests with higher specificity and complexity:



(a) $k = 1$ (b) $k = 3$ (c) $k = 5$

Figure 6: Convergence of user expressions over time under different window sizes k.

Users tend to tailor requests with more context details and fine-grained content later on. This "show or tell" transition reemphasizes that, whereas users may develop distinct use scenarios and tasks, there are indeed fundamental commonalities to be mined at the expression level.

5.1.2 User-level Evolutions

The user expression space like Fig. 2 enables the study focusing on individual users. Further, we also wonder if these paths of users can be collected across users to suggest deeper trends. We consider the effect of familiarity on a user's tendency to repeat the same kinds of expressions they have used previously. One hypothesis is that a user's expressions may converge with more experience, as they develop their "go-to" choices and stick to how that have worked in previous cases. Alternatively, as users gain familiarity, they may understand the capabilities and boundaries of the LLMs better, and thus rely less on rigorous prompt formats and interact in more casual, arbitrary ways.

To test the contrasting diachronic hypotheses, we examine the minimal difference between a long-term user's input and their most recent inputs, i.e., between their i-th and its previous k request utterances, $\min_{j=1,\ldots,k-1}[1 - sim(U_i, U_{i-j})]$, for all valid i given window size k. Then, we collect the minimal difference across all users and compute the average for each position i, to illustrate the averaged step-by-step convergence (or divergence) of expressions within a long-term user's lifecycle.

Figure 6 displays the results under different window sizes k. The actual chronological data is shown in blue, compared against the average of 50 random trials where the same user's dialogs are randomly shuffled (shown in orange), thus breaking diachronic ties. We observe strong evidence for the *convergence* hypothesis across time: the difference between a new input and its closest predecessors sees a drastic, continuous decline as users continue to interact with the system. This is quantitatively different from the baseline level of random shuffles, and the difference between the two gets more significant with more input requests. This suggests that users overall develop rather stable patterns of usage as they gain familiarity. Meanwhile, we also note that this pattern is most significant with window size $k = 1$ (Fig. 6a), i.e., when comparing each input with the exact one previous dialog. The gap between real and random situations is reduced with $k = 3$ (Fig. 6b) and further with $k = 5$ (Fig. 6c). This indicates that the most recent cases consistently serve as more significant references for a new user request, whereas the effect of earlier inputs decreases with their lower recency.

5.2 Exploring and picturing the community

So far we have modeled the evolving request-making expressions from the perspectives of individual users and requests. We finally discuss a broader horizon of ReCCRE: Is it possible to draw a full landscape of evolving human-LLM interaction from observations, modeling the complete evolution trends in the user community as a whole?

We show that, with metrics as simple as Lexical Richness, we are poised to delineate the climate in great detail and discover the subtleties therein. Specifically, we use MTLD (McCarthy, 2005; McCarthy and Jarvis, 2010) as it's almost fully length-invariant and suitable for random collections. The intuitive interpretation of batched Lexical Richness is a measurement of expression diversity across all users: A higher richness indicates that there are more distinct presentations of requests, while a lower richness indicates converged, collective ways of interactions and clearer system affordances.

Comparing long-term and lay users While our analyses have focused on long-term users for the legitimate comparisons over time, the holistic view here enables us to compare the experience of the

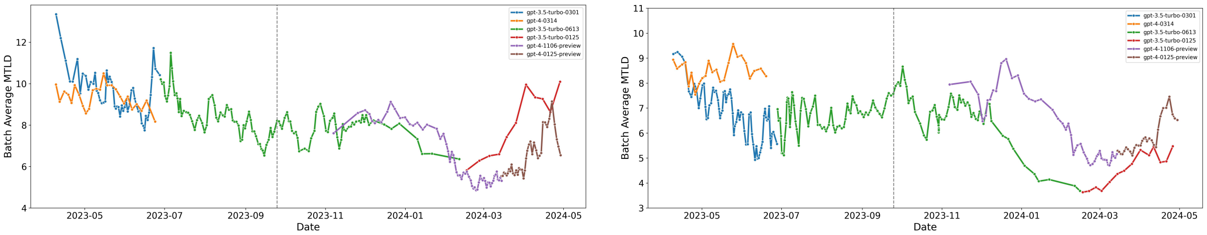


Figure 7: Lexical Richness of long-term users (left) and all users (right) across the full time span.

"regulars" and the entire public user base. The MTLD data for the full dataset with all 18,964 users is illustrated in Figure 7, where proportionally larger batch sizes are used to create comparable densities of data points in the two figures.

5.2.1 Observations

Both long-term and all users share a similar overall trend. The initial deployment of WildChat saw heavily heterogeneous attempts, especially in users who later became long-term. After initial oscillations, the diversity level of expressions stabilized in the extended period of `gpt-3.5-turbo-0613`. However, as the most advanced `gpt-4-1106-preview` was introduced, there seemed a paradigm shift towards a new low. The trend is further different with another major model replacement around March 2024, and both groups see drastically more diverse expressions.

New models, especially the ones from the same family (prefix), take on the exact use patterns of their predecessors. Simultaneous models also see shared evolutions: for instance, the emerging `gpt-4-1106-preview` leads users to also interact with `gpt-3.5-turbo-0613` with less diverse expressions, though the latter is no different from before. Interestingly, long-term users and the lay public seem to show flipped perceptions of model capabilities and usages, as seen in the earliest and latest parts of WildChat. These indicate further complexities yet to be explored, regarding the perception of LLMs in relation to familiarity.

6 Discussion

The ReCCRE framework enables us to move beyond surface-level task types to understand the deeper structure of *how* users formulate requests as LLM inquiries. Analyzing 211k real-world user inputs, we show that human-LLM interaction constitutes a distinct communicative domain with its own evolving linguistic conventions and implications for the broader NLP community.

Understanding the Language Use Dynamics of Human-LLM Pragmatics. Real-world user inputs are heterogeneous: filler language, role-play, and non-conversational inputs are common. We also show that users do not transfer existing conversational norms to LLM interactions, but instead develop new registers with distinct pragmatic features. Existing work has noted understudied features in user interaction such as underspecification (Mysore et al., 2025), as well as key contrasts in real-world users such as conversational and contextualized (Malaviya et al., 2025) vs. fragmented or task-oriented (Wang et al., 2025; Sarkar et al., 2025), and perceiving LLMs more as a tool vs. collaborator (Schroeder et al., 2025; Gao et al., 2024a). ReCCRE makes these fundamental user patterns visible and enables systematic quantitative analyses, connecting observations with qualitative studies and conversation studies. For instance, one important question is *why* the dynamics occur in LLM-based systems. Recent qualitative studies have described the specific difficulties and choices faced by users when prompting (Mahdavi Goloujeh et al., 2024; He et al., 2025) — does the observed "gradual addition of context", for instance, reflect the corresponding actions and strategies users develop? If so, how have they helped (if at all)? Connecting user-side feedback and quantitative data, future work can explore whether certain patterns correlate with more successful outcomes or fulfillment of actual task content, how these patterns vary across cultures or languages, and how systems can be designed to scaffold communication.

Modeling User-LLM Interactions Over Time. Our work is an NLP example of longitudinal user studies with LLMs in the wild (Long et al., 2025; Zhu et al., 2025; Chamberlain et al., 2012). We show that users individually employ converging expressions, and collectively move from request-centric to more context-rich interactions as their familiarity with the system grows. This differs from the majority of current usage of such data, where dialog sessions are introduced as individual cases and/or evaluated with overall and static metrics. Centering real-world language data over time, our work exemplifies the exploration towards systematic understanding of users' trajectories, revealing how people learn, adapt, and converge. We hope this work can expand the frameworks for computational linguistic analysis of user-LLM conversations, and connect with interdisciplinary work, especially Human-Computer Interaction studies with a rich background of longitudinal setups.

Rethinking Alignment via Natural Usage Patterns. One of the main goals of post-training and alignment procedures is to adapt language models to respond to user interactions, and they must account for the full spectrum of real-world user expression styles (Don-Yehiya et al., 2024; Malaviya et al., 2025; Gao et al., 2024a). Our decomposition provides a principled way to enhance training data to reflect this diversity by recombining different expression templates with various request types, or modeling the temporal progression from novice to experienced user patterns. The data supplies *authentic* ingredients for data augmentation under *fine-grained control* and for training reward models that prefer grounded, context-respecting responses. The ReCCRE components of a query may serve as independent supervision signals: requests signify intent, context supports retrieval-augmented grounding, and expressions show the "non-task" linguistic habits that models must identify and match. By separating the components of queries, we can create new training recipes and more informative evaluation, such as request-identification accuracy or robustness to noisy expressions.

7 Conclusions

Our results show that there are consistent, large-scale patterns in how new and experienced users interact with LLM systems. These patterns only appear when we separate *how* people ask from *what* they ask for. While the WildChat dataset provides sufficient proof-of-concept, it also sets a protocol for further study across platforms, modalities, languages, communities, and cultures. Collectively, these directions position ReCCRE not only as an analytic lens but as a practical toolkit for advancing robust, user-centered NLP.

Limitations

While our work seeks to provide new resources and paradigms, the data and analysis work both have practical limitations. Due to the limited capabilities of the 14B LLM and the author as annotator and validator, the ReCCRE dataset is selected from the chat logs with English as the major language in the original dataset. This limits the reliable scope of the work, as cultural and language factors can lead to different interaction patterns. Further, our work is based fully on WildChat, which represents a specific type of data collection practice, a certain group of audience, and a specific time window. While there are no additional confounders introduced, it may still inherit biases present in the underlying WildChat logs, such as demographic skews or designs of its UI layout (HugginFace). It is also possible that findings are different in interesting ways in other documented chat logs, such as LMSys (Zheng et al., 2024a; Chiang et al., 2024) collected earlier with a different protocol. In our analyses, we largely utilize off-the-shelf metrics including the `gte` embeddings and Lexical Richness measurements. This is by design, as we would like to demonstrate the usability and low barrier of the dataset. However, we also note that this might limit the boundaries and depth of data analyses, and we encourage further explorations with the ReCCRE data, e.g., fine-tuning models with the resource.

References

Alan Chamberlain, Andy Crabtree, Tom Rodden, Matt Jones, and Yvonne Rogers. 2012. Research in the wild: understanding 'in the wild' approaches to design and development. In *Proceedings of the designing interactive systems conference*, pages 795–796.

Jingwen Cheng, Kshitish Ghate, Wenyue Hua, William Yang Wang, Hong Shen, and Fei Fang. 2025. REALM: A dataset of real-world LLM use cases. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8331–8341, Vienna, Austria. Association for Computational Linguistics.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, and 1 others. 2024. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning*.

Alexander S. Choi, Syeda Sabrina Akter, JP Singh, and Antonios Anastasopoulos. 2024. The LLM effect:

Are humans truly using LLMs, or are they being influenced by them instead? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22032–22054, Miami, Florida, USA. Association for Computational Linguistics.

Michael A Covington and Joe D McFall. 2010. Cutting the gordian knot: The moving-average type–token ratio (mattr). *Journal of quantitative linguistics*, 17(2):94–100.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 250–259, Sofia, Bulgaria. Association for Computational Linguistics.

Shachar Don-Yehiya, Leshem Choshen, and Omri Abend. 2024. Naturally occurring feedback is common, extractable and useful. *arXiv preprint arXiv:2407.10944*.

Ge Gao, Alexey Taymanov, Eduardo Salinas, Paul Mineiro, and Dipendra Misra. 2024a. Aligning LLM agents by learning latent preference from user edits. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Jie Gao, Simret Araya Gebreegziabher, Kenny Tsu Wei Choo, Toby Jia-Jun Li, Simon Tangi Perrault, and Thomas W Malone. 2024b. A taxonomy for human-llm interaction modes: An initial exploration. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '24, New York, NY, USA. Association for Computing Machinery.

Kunal Handa, Alex Tamkin, Miles McCain, Saffron Huang, Esin Durmus, Sarah Heck, Jared Mueller, Jerry Hong, Stuart Ritchie, Tim Belonax, and 1 others. 2025. Which economic tasks are performed with ai? evidence from millions of claude conversations. *arXiv preprint arXiv:2503.04761*.

Zeyu He, Saniya Naphade, and Ting-Hao Kenneth Huang. 2025. Prompting in the dark: Assessing human performance in prompt engineering for data labeling when gold labels are absent. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

Dun-Ming Huang, Pol Van Rijn, Ilia Sucholutsky, Raja Marjieh, and Nori Jacoby. 2024. Characterizing similarities and divergences in conversational tones in humans and LLMs by sampling with people. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10486–10512, Bangkok, Thailand. Association for Computational Linguistics.

Saffron Huang, Esin Durmus, Miles McCain, Kunal Handa, Alex Tamkin, Jerry Hong, Michael Stern, Arushi Somani, Xiuruo Zhang, and Deep Ganguli. 2025. Values in the wild: Discovering and analyzing values in real-world language model interactions. *arXiv preprint arXiv:2504.15236*.

Jonathan Ivey, Shivani Kumar, Jiayu Liu, Hua Shen, Sushrita Rakshit, Rohan Raju, Haotian Zhang, Aparna Ananthasubramaniam, Junghwan Kim, Bowen Yi, and 1 others. 2024. Real or robotic? assessing whether llms accurately simulate qualities of human responses in dialogue. *arXiv preprint arXiv:2409.08330*.

Zhihua Jin, Shiyi Liu, Haotian Li, Xun Zhao, and Huamin Qu. 2025. Jailbreakhunter: a visual analytics approach for jailbreak prompts discovery from large-scale human-llm conversational datasets. *IEEE Transactions on Visualization and Computer Graphics*.

Hannah Rose Kirk, Alexander Whitefield, Paul Rottger, Andrew M Bean, Katerina Margatina, Rafael Mosquera-Gomez, Juan Ciro, Max Bartolo, Adina Williams, He He, and 1 others. 2024. The prism alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. *Advances in Neural Information Processing Systems*, 37:105236–105344.

Steven Kolawole, Keshav Santhanam, Virginia Smith, and Pratiksha Thaker. 2025. Parallelprompt: Extracting parallelism from large language model queries. *Proceedings of the 39th Conference on Neural Information Processing Systems (NeurIPS 2025) Datasets and Benchmarks Track*.

Batia Laufer and Paul Nation. 1995. Vocabulary size and use: Lexical richness in l2 written production. *Applied linguistics*, 16(3):307–322.

Dong-Ho Lee, Adyasha Maharana, Jay Pujara, Xiang Ren, and Francesco Barbieri. 2025. Realtalk: A 21-day real-world dataset for long-term conversation. *arXiv preprint arXiv:2502.13270*.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.

Tao Long, Sitong Wang, Émilie Fabre, Tony Wang, Anup Sathya, Jason Wu, Savvas Dimitrios Petridis, Ding Li, Tuhin Chakrabarty, Yue Jiang, and 1 others. 2025. Facilitating longitudinal interaction studies of ai systems. In *Adjunct Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology*, pages 1–5.

Zilin Ma, Yiyang Mei, Krzysztof Z. Gajos, and Ian Arawjo. 2024. Schrödinger's update: User perceptions of uncertainties in proprietary large language model updates. In *Extended Abstracts of the CHI*

Conference on Human Factors in Computing Systems, CHI EA '24, New York, NY, USA. Association for Computing Machinery.

Atefeh Mahdavi Goloujeh, Anne Sullivan, and Brian Magerko. 2024. Is it AI or is it me? understanding users' prompt journey with text-to-image generative ai tools. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Chaitanya Malaviya, Joseph Chee Chang, Dan Roth, Mohit Iyyer, Mark Yatskar, and Kyle Lo. 2025. Contextualized evaluations: Judging language model responses to underspecified queries. *Transactions of the Association for Computational Linguistics*, 13:878–900.

Philip M McCarthy. 2005. *An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD)*. Ph.D. thesis, The University of Memphis.

Philip M McCarthy and Scott Jarvis. 2010. Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2):381–392.

Niloofar Mireshghallah, Maria Antoniak, Yash More, Yejin Choi, and Golnoosh Farnadi. 2024. Trust no bot: Discovering personal disclosures in human-llm conversations in the wild. In *First Conference on Language Modeling*.

Kshitij Mishra, Mauajama Firdaus, and Asif Ekbal. 2022. Please be polite: Towards building a politeness adaptive dialogue system for goal-oriented conversations. *Neurocomputing*, 494:242–254.

Terran Mott, Aaron Fanganello, and Tom Williams. 2024. What a thing to say! which linguistic politeness strategies should robots use in noncompliance interactions? In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, pages 501–510.

Sheshera Mysore, Debarati Das, Hancheng Cao, and Bahareh Sarrafzadeh. 2025. Prototypical human-AI collaboration behaviors from LLM-assisted writing in the wild. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 16819–16846, Suzhou, China. Association for Computational Linguistics.

Rupak Sarkar, Bahareh Sarrafzadeh, Nirupama Chandrasekaran, Nagu Rangan, Philip Resnik, Longqi Yang, and Sujay Kumar Jauhar. 2025. Conversational user-ai intervention: A study on prompt rewriting for improved llm response generation. *arXiv preprint arXiv:2503.16789*.

Hope Schroeder, Marianne Aubin Le Quéré, Casey Randazzo, David Mimno, and Sarita Schoenebeck. 2025. Large language models in qualitative research: Uses, tensions, and intentions. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

Lucas Shen. 2022. LexicalRichness: A small module to compute textual lexical richness.

Alex Tamkin, Miles McCain, Kunal Handa, Esin Durmus, Liane Lovitt, Ankur Rathi, Saffron Huang, Alfred Mountfield, Jerry Hong, Stuart Ritchie, and 1 others. 2024. Clio: Privacy-preserving insights into real-world ai use. *arXiv preprint arXiv:2412.13678*.

Xiaoyi Wang, Yuran Wang, and Xingyi Qiu. 2025. How to talk to ai: The role of preset prompt language styles in shaping conversational experience. *International Journal of Human–Computer Interaction*, 41(12):7763–7778.

Yingfan Wang, Haiyang Huang, Cynthia Rudin, and Yaron Shaposhnik. 2021. Understanding how dimension reduction tools work: An empirical approach to deciphering t-sne, umap, trimap, and pacmap for data visualization. *Journal of Machine Learning Research*, 22(201):1–73.

J.D. Zamfirescu-Pereira, Richmond Y. Wong, Bjoern Hartmann, and Qian Yang. 2023. Why johnny can't prompt: How non-ai experts try (and fail) to design llm prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.

Chao Zhang, Shengqi Zhu, Xinyu Yang, Yu-Chia Tseng, Shenrong Jiang, and Jeffrey M Rzeszotarski. 2025a. Navigating the fog: How university students recalibrate sensemaking practices to address plausible falsehoods in llm outputs. In *Proceedings of the 7th ACM Conference on Conversational User Interfaces*, pages 1–15.

Zhouqing Zhang, Kongmeng Liew, and Tham Piumsomboon. 2025b. What one million prompts tells us about ai usage, topics, and preferences. In *2025 IEEE Conference on Artificial Intelligence (CAI)*, pages 174–179. IEEE.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024. Wildchat: 1m chatgpt interaction logs in the wild. In *The Twelfth International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric Xing, and 1 others. 2024a. Lmsys-chat-1m: A large-scale real-world llm conversation dataset. In *The Twelfth International Conference on Learning Representations*.

Mingqian Zheng, Jiaxin Pei, Lajanugen Logeswaran, Moontae Lee, and David Jurgens. 2024b. When "a helpful assistant" is not really helpful: Personas in system prompts do not improve performances of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15126–15154, Miami, Florida, USA. Association for Computational Linguistics.

Shengqi Zhu, Jeffrey M. Rzeszotarski, and David Mimno. 2025. Data paradigms in the era of llms: On the opportunities and challenges of qualitative data in the wild. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '25, New York, NY, USA. Association for Computing Machinery.