Feedback Effect in User Interaction with Intelligent Assistants: Delayed Engagement, Adaption and Drop-out

Zidi Xiu[1][†], Kai-Chen Cheng[1], David Q. Sun[1][†], Jiannan Lu[1], Hadas Kotek[1],
Yuhan Zhang[2][⋆], Paul McCarthy[1], Christopher Klein[1], Stephen Pulman[1],
Jason D. Williams[1]

[1] Apple, One Apple Park Way, Cupertino, CA 95014, USA
[2] Department of Linguistics, Harvard University, Cambridge, MA 02138
[†]{z_xiu,dqs}@apple.com

Abstract. With the growing popularity of intelligent assistants (IAs), evaluating IA quality becomes an increasingly active field of research. This paper identifies and quantifies the *feedback effect*, a novel component in IA-user interactions – how the capabilities and limitations of the IA influence user behavior over time. First, we demonstrate that unhelpful responses from the IA cause users to delay or reduce subsequent interactions in the short term via an observational study. Next, we expand the time horizon to examine behavior changes and show that as users discover the limitations of the IA's understanding and functional capabilities, they learn to adjust the scope and wording of their requests to increase the likelihood of receiving a helpful response from the IA. Our findings highlight the impact of the feedback effect at both the micro and meso levels. We further discuss its macro-level consequences: unsatisfactory interactions continuously reduce the likelihood and diversity of future user engagements in a feedback loop.

Keywords: Data Mining · Intelligent Assistant Evaluation

1 Introduction

Originated from spoken dialog systems (SDS), intelligent assistants (IAs) had rapid growth since the 1990s [9], with both research prototypes and industry applications. As their capabilities grow with recent advancements in machine learning and increased adoption of smart devices, IAs are becoming increasingly popular in daily life [3,14]. Such IAs often offer a voice user interface, allowing users to fulfill everyday tasks, get answers to knowledge queries, or start casual social conversations, by simply speaking to their device [17,26]; that is, they take human voice as input, which they process in order to provide an appropriate response [29]. The evolution of these hands-free human-device interaction systems brings new challenges and opportunities to the data mining community.

IA systems often consist of several interconnected modules: Automated Speech Recognition (ASR), Natural Language Understanding (NLU), Response Generation & Text-to-Speech (TTS), Task Execution (*e.g.*, sending emails, setting

---

⋆ Contributions made during the internship at Apple in the summer of 2022.

alarms and playing songs), and Question Answering [9,12,22]. Many of the active developments in the field are formulated as supervised learning problems where the model predicts a target from an input, e.g., a piece of text from a speech audio input (ASR), a predefined language representation from a piece of text (NLU), or a clip of audio from a string of text (TTS). Naturally, the evaluation of these models often involves comparing model predictions to some ground-truth datasets.

When building such an evaluation dataset from real-world usage, we inevitably introduce user behavior into the measurement. User interactions with IA are likely to be influenced by the their pre-existing perception of IA's capabilities and limitations, therefore introducing a bias in the distribution of "chances of success" in logged user interactions – users are more likely to ask what they know the IA can handle. This hypothesis makes intuitive sense and has been partly suggested by an earlier study on vocabulary convergence in users learning to speak to an SDS [19].

In this context, we define *feedback effect* as the behavior pattern changes in users of an interactive intelligent system (*e.g.*, IA) that are attributable to their cumulative experiences with said system. Our contributions can be summarized as follows. First, we establish a causal link between IA performance and immediate subsequent user activities, and quantify its impact on users of a real-world IA. Second, we identify distinct dynamics of behavior change for a cohort of new users over a set period of time, demonstrating how users first explore the IA's capabilities before eventually adapting or quitting. Third, having examined the *feedback effect* and its impact in detail, we provide generalizable recommendations to mitigate its bias in IA evaluation.

2   Related Work

IA evaluation methods and metrics. Many studies have been devoted to addressing the challenges in IA evaluation. Objective metrics like accuracy cannot present a comprehensive view of the system [8]. Human annotation is a crucial part of the process, but it incurs a high expense and is hard to scale [15]. Apart from human evaluation, *i.e.*, user self-reported scores or annotated scores, subjective metrics have been introduced. Jiang [12] designed a user satisfaction score prediction model based on user behavior patterns, ungrammatical sentence structures, and device features. Other implicit feedback from users (*e.g.*, acoustic features) are helpful to approximate success rates [16].

User adaptation and lexical convergence. *Adaptation* (or *entrainment*) describes the phenomenon whereby the vocabulary and syntax used by speakers converge as they engage in a conversation over time [27]. Convergence can be measured by observing repetitive use of tokens in users' requests [6] and high frequency words [24]. Adaptation happens subconsciously and leads to more successful conversations [7]. In SDS, the speakers in a dialogue are the IA and the user. When the IA actively adapts to the user in the conversation, the quality of the generated IA responses increases substantially [30,32]. The phenomenon of lexical adaptation of users to the IA system has been investigated as well [19,25]. Currently, most IAs are built upon a limited domain with restricted vocabulary

size [5]. Users' vocabulary variability tends to decrease as they engage with the IA over time. This naturally limits the linguistic diversity of user queries, although out-of-domain queries can happen from time to time [9].

3   Data Collection

We analyzed logged interactions (both user queries and the associated IA responses) from a real-world IA system. All data originate from users who have given consent to data collection, storage, and review. The data is associated with a *random, device-generated* identifier. Therefore, when we use the term 'user' in the context of the interaction data analysis, we are *actually* referring to this random identifier. While the identifier is a reasonable proxy of a user, we must recognize its limitations – in our analysis, we are unable to differentiate multiple users who share a single device to interact with the IA, nor to associate requests from a single user that were initiated on multiple devices.

The population of interest is US English-speaking smartphone users who interacted with the IA in 2021 and 2022. We randomly sampled interaction data from two distinct time periods *before* and *after* a special event in late 2021. This event entailed new software and hardware releases, potentially introducing nontrivial changes to user behavior and demographics, while simultaneously presenting unique opportunities for our particular investigation.

3.1   Study 1: Pre-event Control Period

To investigate the feedback effect on user engagement, we randomly sampled interaction data from a two-week period in August 2021. The choice of a relatively *short* time period *before* the special event helps us (i) directly control for seasonality and (ii) avoid the impact of the special event, where product releases and feature announcements usually stimulate user engagement and attract new users. (We return to a discussion of new users below.)

We further control for software and hardware versions, before taking a random sample of approximately 14,000 users who had at least one interaction with the IA during the study period. We then randomly sampled *one* interaction per user and used human-label review to determine whether the IA response was helpful. We additionally analyzed the frequency of interactions for the user in the 2 weeks prior to and 2 weeks following our causal analysis. In our sample, approximately 80% of the interactions were labeled as *helpful* to the user.[3] The results of this study are presented in Section 4.
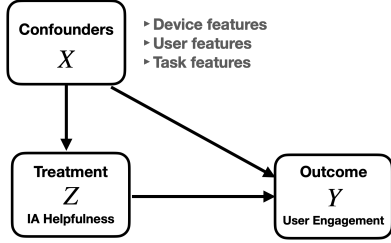
3.2   Study 2: Post-event New User Period

To investigate language convergence among a new user cohort, we randomly sampled data from a six-month period immediately *after* the special event. With our interest in analyzing long-term behavior changes of a new user cohort, this choice of sampling period has two interesting implications. First, special events often lead to a surge in new users of the IA. Second, feature announcements at the special event may cause some existing users to perceive the updated IA as "new" and explore it with a mindset akin to that of a new user. Given the

---

[3] This value is not necessarily a reflection of the aggregated or expected satisfaction metric, due to the sampling method and potential bias in the subpopulation of choice.

challenges inherent to determining new user cohorts (to be further discussed in Section 5), these two factors are valuable as they collectively increase the share of new users, thus boosting the observability of the cohort. From this six-month period, we took a random sample of 5,000 users who used the new software version of the IA. For each user, all interactions with the IA in the full study period were used in our analysis. The study is described in Section 5.

4   Feedback Effect on Engagement

Fig. 1. Causal graph illustrating the observational study of IA feedback effect accounting for the existence of confounding factors.

Intuitively, unhelpful responses from an IA may discourage users from future interactions with the IA. Our work aims to empirically shed light on the relation between IA helpfulness[4] (*helpful* or *unhelpful* on a single interaction) and users' subsequent engagement patterns with the IA, as illustrated in Figure 1. To establish such a causal relationship from IA performance to user engagement, we adopt an observational analysis framework with IA related features.

Given a dataset with N users, we denote unit i having (i) covariates $\mathbf{X}_i \in \mathbb{R}^p$, (ii) a treatment variable $Z_i \in \{0, 1\}$, indicating users experienced an *unhelpful* interaction with the IA or a *helpful* one respectively, and (iii) what would have happened if the unit is assigned to treatment and control, denoted by $Y_i(1)$ and $Y_i(0)$ respectively, according to the potential outcomes framework [10,28]. Consequently, the causal effect for unit i is defined as $\tau_i = Y_i(1) - Y_i(0)$, namely the difference between the outcomes if treated differently on the same user. However, the fundamental problem of causal inference is that only the potential outcome – the outcome in the group the subject was assigned to – can be observed, *i.e.*, $Y_i = Z_i Y(Z_i) + (1 - Z_i) Y(1 - Z_i)$. Individual level causal estimands, the contrast of values between the two potential outcomes, cannot be expressed with functions of observed data alone. Consequently, our primary focus is on population level causal effects like the Average Treatment Effect (ATE), $\tau = \mathbb{E}(\tau_i)$.

With a randomized controlled experiment, treated assignment mechanism is known and unconfounded, therefore we can directly and accurately estimate and infer causal effects (*e.g.*, ATE) from the observed data. However, in real-world scenarios, delicately designed experiments can be difficult or impossible to conduct. Instead, we must rely on observational techniques.

4.1   Covariates and Outcome Variables

Observational studies are susceptible to *selection bias* due to confounding factors, which affect both the treatment $\mathbf{Z}$ and the outcome $\mathbf{Y}$, as shown in Figure 1. To address any confoundness to the best of our ability, we have collected rich sets of the following IA related features, and assuming that there are no unobserved observed confounders. (i) Device features: The type of device used to interact

---

[4] The IA helpfulness of a given user request is defined as the user's satisfaction with the IA's response to the request, as determined by human annotators.

with the IA system, and the operating system version, (ii) Task features: The input sentence transcribed by the ASR system, the number of tokens in the input sentence, the word error rate (WER) of the transcription, and the domain that the IA executed with a confidence score provided by the NLU model (*e.g.*, weather, phone, etc.), (iii) User related features: Prior activity levels measured as the number of active days before the interaction, and temporal features including local day of the week and time of the day when the interaction happened.

To quantify user engagement after the annotated IA interaction, Section 4.3 focuses on time to next session ("immediate shock"), and Section 4.4 focuses on active day counts ("aftermath").

4.2 Observational Causal Methods

Matching methods. Matching is a non-parametric method to alleviate the effects of confounding factors in observational studies. The goal is to obtain well-matched samples from different treatment groups, hoping to replicate randomized trial settings. The popular Coarsened Exact Matching (CEM) model is based on a monotonic imbalance reducing matching method at a pre-defined granularity with no assumption on assignment mechanisms [11].
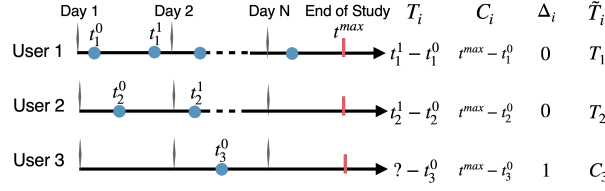
Weighting methods. Apart from matching covariates, weighting based methods use all of the high-dimensional data via summarizing scores, like the propensity score (PS). PS reflects the probability of being assigned to treatment based on user's background attributes [20,28], $e(x) = P(Z_i = 1|X_i = x) = \mathbb{E}(\mathbf{Z}|\mathbf{X})$ Since the true PS is unknown, we adopt generalized linear regression models (GLMs) to estimate it, which are widely adopted by the scientific community.

With PS estimates available, the next question is how to leverage them. Li [20] proposed a family of balancing weights which enjoys balanced weighted distributions of covariates among treatment groups. Inverse-Probability Weights (IPW) are a special case of this family, shown in Eq.(1). As the name suggests, the weight is the inverse of the probability that a unit is assigned to the observed group, and the corresponding estimand is the ATE. However, IPW is very sensitive to outliers, *i.e.*, when PS scores approach 0 or 1. To mitigate this challenge, Overlap Weights (OW) which emphasize a target population with the most covariate overlap [20], shown in Eq.(2).

$$\begin{cases} w_1^{\text{IPW}}(x) = \frac{1}{e(x)} \\ w_0^{\text{IPW}}(x) = \frac{1}{1-e(x)} \end{cases} \quad (1) \qquad \begin{cases} w_1^{\text{OW}}(x) = (1 - e(x)) \\ w_0^{\text{OW}}(x) = e(x), \end{cases} \quad (2)$$

where w_1 corresponds to the weight assigned to the treatment group, and w_0 to the control group, respectively. Then the population level causal estimands of interest, the Weighted Average Treatment Effect (WATE), are derived from the balanced weights. The target population varies with different weighting strategy. The causal estimand shown in Eq. (3) then becomes the average treatment effect for the overlap population.

$$\hat{\tau}^w = \frac{\sum_{i=1}^{N} w_1(\mathbf{x_i})\mathbf{Z_i}\mathbf{Y_i}}{\sum_{i=1}^{N} w_1(\mathbf{x_i})\mathbf{Z_i}} - \frac{\sum_{i=1}^{N} w_0(\mathbf{x_i})(\mathbf{1} - \mathbf{Z_i})\mathbf{Y_i}}{\sum_{i=1}^{N} w_0(\mathbf{x_i})(\mathbf{1} - \mathbf{Z_i})} \qquad (3)$$

Fig. 2. Illustration of the users' engagement (blue dots) after the request was annotated at time t_i^0 for user i. We observed the time-to-next-engagement for user 1 and 2, but user 3 was censored (the next engagement was not observed).

4.3   Time to Next Engagement

In this section, we establish causal links between interaction quality with the IA (as implied by the annotated helpfulness) and the user's time to next engagement. Specifically, our main hypothesis is that if a user has a helpful interaction with the IA, they are more likely to further engage with the IA in the future.
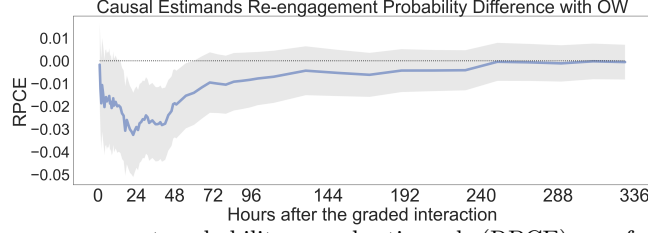
Unlike standard observational studies with well-defined and observable outcomes, time-to-event measures fall into the range of survival analyses, which focus on the length of time until the occurrence of a well-defined outcome [23,33]. A characteristic feature in the study of time-to-event distributions is the presence of *censored* instances: events that do *not* occur during the follow-up period of a subject. This can happen when the unit drops out during the study (right censoring), presenting challenges to standard statistical analysis tools.

As illustrated in Figure 2, assume for user i: the time-to-next engagement is T_i with censoring time C_i, the observed outcome $\tilde{T}_i = T_i \wedge C_i$, and the censoring indicator $\Delta_i = \mathbb{1}\{T_i \leq C_i\}$. Under time-to-event settings, we observe a quadruplet $\{Z_i, \mathbf{X}_i, \tilde{T}_i, \Delta_i\}$ for each sample. Each user also has a set of potential outcomes, $\{T_i(1), T_i(0)\}$. Users may use the IA system at some point in our research and be assigned a *helpfulness* score, but not show up again before the data collection period ends (*e.g.*, User 3 illustrated in Figure 2). This yields a censored time C_3 instead of a definite time-to-next-engagement outcome T_3 which is not observed within the study period.

Following Zeng [34], the causal estimand of interest is defined based on a function of the potential survival times, $\nu(T_i(z); t) = \mathbb{1}\{T_i(z) \geq t\}$. It can be interpreted as an at-risk function with the potential outcome $T_i(z)$. The expectation of the risk function corresponding to the potential survival function of user i, *i.e.*, the probability of no interaction with the IA until time t. Accordingly, the *re-engagement* probability for users in treatment group z within time t is therefore defined as Eq.(4).

$$\mathbb{E}[\nu(T_i(z); t)] = \mathbb{P}[T_i(z) \geq t] = \mathbb{S}_i(t; z) \quad (4) \qquad \mathbb{P}(t; z) = 1 - \mathbb{S}(t; z) \quad (5)$$

To properly apply balancing weights (2) with survival outcomes, right censoring needs to be accounted for. Pseudo-observation is therefore constructed based on re-sampling (a jack-knife statistic) and is interpreted as the individual contribution to the target estimate from a complete sample without censoring [2]. Given a time t, denote the expectation of the risk function at that time point , *i.e.*, $\mathbb{E}[\nu(T_i(z); t)]$ in Eq.(4), as $\theta(t)$, which is a population parameter. Without loss of generality, we discuss the pseudo observation omitting the potential

Fig. 3. The re-engagement probability causal estimands (RPCE) as a function of time after the annotated interaction, with associated 95% confidence interval (shaded gray).

outcome notations. The pseudo-observation for each unit i can be specified as, $\hat{\theta}_i(t) = N\hat{\theta}(t) - (N-1)\hat{\theta}_{-i}(t)$, where $\hat{\theta}(t)$ is the Kaplan-Meier estimator of the population risk at time t, which is based on Δ_i and T_i. $\hat{\theta}_{-i}(t)$ is calculated without unit i. In this way, classic propensity score methods become applicable. Then the conditional causal effect averaged over a target population at time t is:

$$\hat{\tau}^w(t) = \frac{\sum_{i=1}^{N} w_1(\mathbf{x_i})\mathbf{Z_i}\hat{\theta}_{\mathbf{i}}(\mathbf{t})}{\sum_{i=1}^{N} w_1(\mathbf{x_i})\mathbf{Z_i}} - \frac{\sum_{i=1}^{N} w_0(\mathbf{x_i})(\mathbf{1-Z_i})\hat{\theta}_{\mathbf{i}}(\mathbf{t})}{\sum_{i=1}^{N} w_0(\mathbf{x_i})(\mathbf{1-Z_i})}$$
$$= (1 - \hat{\mathbb{S}}^{w_0}(t;0)) - (1 - \hat{\mathbb{S}}^{w_1}(t;1)) = \hat{\mathbb{P}}^{w_0}(t;0) - \hat{\mathbb{P}}^{w_1}(t;1) \qquad (6)$$

The estimator in Eq.(6) represents the survival probability causal effect, *i.e.*, the difference of the weighted *re-engagement* probability in the *Unhelpful* group and the *Helpful* group, or the *Re-engagement* Probability Causal Effect (RPCE). The results are shown in Figure 3. The confidence interval is calculated based on the estimated standard error of SPCE [34].

The difference in estimated re-engagement probability is negative within the 336 hours (two weeks) following the initial interaction, with a maximum causal difference of 3.2% at around 24 hours (*p*-value is 0.007). The time window where the difference between the *Helpful* and *Unhelpful* groups is consistently statistically significant is between hours 8-65. The IPW result yields similar conclusions. Our main takeaway is that the inhibition effect of an unhelpful interaction reaches peak around 24 hours after the interaction and then gradually weakening.

Specifically, we conclude the following, (i) An unhelpful interaction tends to have a stronger effect on whether the user wants to use the assistant again around the same hour on the next few days, perhaps affecting daily tasks like starting navigation to work, (ii) About one week later, the re-engagement probability difference becomes insignificant, as users' recollections of the unhelpful interaction fade away.
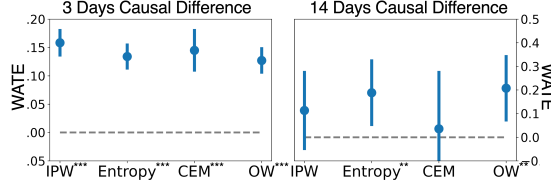
4.4   Number of Active Days

Section 4.3 established the immediate effect of IA helpfulness on time-to-next engagement. In this section, we widen the analysis window and focus on the number of active days after the annotated interaction. Let $A^{(k)}$ denote the number of active days within k-day window, $k \in \{3, 14\}$. The average treatment effect (ATE) is defined as $\mathbb{E}[A_i^{(k)}(1) - A_i^{(k)}(0)]$.

To estimate the causal effects with consistency, we applied four different statistical analysis tools at the two time windows respectively, belonging to two major branches of causal analysis. The first branch is weighting (IPW, entropy

Low Perplexity	High Perplexity: syntactically complex sentences	High Perplexity: lexically diverse and rare topics
What is the **weather**, 3.3	Show me hourly **weather** forecast, 17.1	What is the **UV index**, 11.8
	Could I have the **weather** for rest of the week in <Location>please, 20.8	Is there **tornado** nearby, 13.6
What is the **temperature**, 3.5	When is the **rain** supposed to start again, 19.5	How fast is the **wind** going, 15.6
Will it **rain** today, 5.9	When the **rain** going to stop, 21.8	When is the full **moon**, 15.7
Is it going to **snow** today, 7.6	How many inches of **snow** are we supposed to get, 20.4	What is the **barometric pressure** at <Location>, 27.1
	How tall will the **snow** get tonight, 27.6	

Table 1. Examples of low and high perplexity requests about weather.

weights, overlap weights).[5] The corresponding WATE function is defined similarly as Eq.(3). The second branch is matching. Considering the dimensionality, we used the CEM method [11].

Fig. 4. Causal effect of an unhelpful IA interaction on activity levels. Bar length indicates 95% CI

In line with our previous findings, we observe statistically significant causal impacts on the activity level 3 days after the annotated IA interaction, shown in Figure 4 (left). All four analysis tools yield p-values < 0.001. This also supports the finding that the inhibition effect of an unhelpful engagement fades in time. When we zoom out to a 14-day window, we observe that though the causal effects are not always significant, the directional consistency suggests a lessened effect of the unhelpful engagement compared to the 3-day window.

5   Language Convergence in New User Cohort

Having established the inhibition effect of an unhelpful interaction on a user's activity levels immediately following the interaction, we now expand both the scope and the time horizon of our analysis, to explore how prior engagements in turn shape users' linguistic choices over time.

5.1   New and Existing User Cohort Definition

Canonically, a *new user* to an IA is an individual who started using the IA for the first time in the observation window. As our data does not allow us to identify new users in this way, we rely instead on the following conservative, *necessary but insufficient*, condition for cohort determination: a user is assigned to the 'new user cohort' if they (i) had at least one interaction with the IA in the study period, and (ii) had no interaction with the IA in the first 60 days of the study period. By erring on the side of including existing users in the new user group, we can ensure that any patterns that remain are robust. Therefore, we argue that this determination method offers a reasonable (and likely inflated) approximation of the true new user cohort. In our dataset containing 6 months of interaction data, approximately 17% of all unique users were assigned to the new user cohort.

5.2   New user's self-Selection: Drop-out or adaption

We use a domain-specific language model-based perplexity (PP) score [18], which provides a comprehensive summary of the request's complexity characteristics
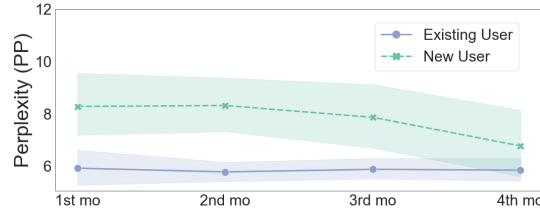
---

[5] Propensity weighting methods: https://cran.r-project.org/web/packages/PSweight

[1]. PP score is defined as the inverse joint probability that a sentence belongs to the trained language model normalized by the number of tokens in the sentence [13], $PP(W) = \sqrt[N]{\frac{1}{\mathbb{P}(w_1 w_2 ... w_N)}}$, where W is the target sentence, w_k is individual token and N is the token count of the sentence. In our analysis, we adopted a tri-gram language model [4]. Table 1 presents examples of requests with perplexity scores. Here we use paraphrased variants rather than actual user data for illustration purposes. Higher perplexity correlates with more complex sentence structures, more diverse language representations and broader topics.

Intuitively, new users tend to explore the limits of the IA system, with broader vocabulary and diverse paraphrases of their requests. In this study, we track the average PP scores of new and existing user cohorts over a six month period. First, we empirically show that the *existing user* cohort has a lower and more stable perplexity score over time compared to the *new user* cohort (Figure 5). This result suggests that requests from existing users are more likely to conform to the typical wording of requests within a certain domain.

Second, we discover that the perplexity score in the new user cohort is 30% higher than in the existing user group in the first month, but it gradually converges to that of the existing user cohort. This trend suggests that new users are less familiar with the IA's capabilities and are more exploratory when they are first introduced to the system. Over time, they gradually familiarize themselves with it. Eventually, they adopt similar sentence structures and other linguistic characteristics to those used by existing users when expressing similar intents.

Fig. 5. New user cohort has a higher average perplexity in the beginning and converges toward existing user cohort in language perplexity over time.

Next, within the new user cohort, we dive deeper into two subgroups: The *retained* group consists of users who were active for more than three out of the four month follow-up period. The *dropout* group includes users who were active for no more than 30 days within the study period. Based on these criteria, the retained group has an average perplexity score of 7.5, while the dropout group has an average perplexity score of 10.6 and the difference has p-value < 0.001. That is, users who stop using the IA within the first month tend to have substantially higher perplexity scores than users who are retained. Further, higher PP scores are closely related to higher unhelpful rates in the IA interactions. Following our findings from the previous sections, we expect unhelpful experiences to discourage users from continuing to engage with the IA.

In summary, we conclude that there are two plausible mechanisms that may explain the convergence of the perplexity score over time in the new user cohort:

1. **Dropout**: some new users who are either unfamiliar with the supported functionality or the language of the IA system suffer negative experiences. These high-perplexity language users stop using the system after a few tries.
2. **Adaptation**: despite some potential negative experiences in the beginning, some new users familiarize themselves with the system and adapt to its limitations. They continue to use the system after the first few months.

This represents a self-selection process among the users who choose to interact with the IA system: users adapt to the IA system in a way that lowers their language perplexity and consequently improves their experience, or they stop using it altogether. Crucially, as users adapt their behavior to the system over time, we expect to observe fewer and fewer requests that may lead to unhelpful interactions with the IA—as a result of the *feedback effect*. Consequently, we observe a bias that introduces a significant challenge to the meaningful offline evaluation of the IA system based on naive samples of the usage traffic.

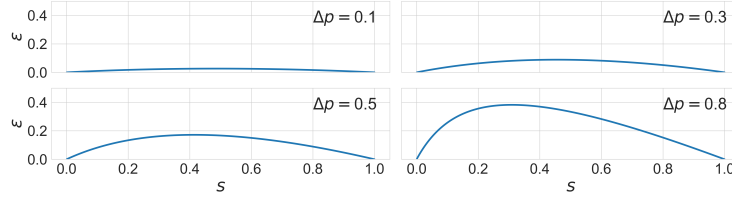## 6   The Feedback Effect: Challenges to Meaningful Metrics

6.1   User-based vs. Usage-based evaluations

Offline evaluation methodologies in the IA space mostly fall into two broad categories: *user-based* and *usage-based* approaches [12]. User-based approaches typically measure the overall satisfaction of a user, while usage-based approaches focus on success rates of the IA in correctly responding to a collection of requests. In this section, we discuss how the feedback effect introduces challenges to the construction of meaningful metrics for both types of approaches, informed by our findings in Sections 4 and 5.

6.2   Implications of the Inhibition Effect

As we established in Section 4, users who experience an unhelpful interaction with the IA are less likely to re-engage with it in the next few days. We may conclude, without loss of generality, that users who had unsatisfactory experiences in a preceding period are less likely to engage with the IA in the current period. Hence, if we are to survey *active* users in a fixed time period to measure their expected satisfaction levels, unsatisfied users would have a lower probability of being surveyed than their satisfied counterparts. This "heavy-user" bias is ubiquitous in data mining [31].

As an illustration, suppose that for the preceding period $T^{(k-1)}$ there were N active users within the period, and no new users joined. Further, let N' denote the number of re-engaged active users in the current period $T^{(k)}$ who were not active in $T^{(k-1)}$. Let s be the share of users who were satisfied with their experiences with the IA, hence $(1-s)$ represents the unsatisfied share. Let p be the probability of the satisfied users to re-engage in the current period (so that they may be surveyed), and Δp indicates the difference in re-engagement probability for the unsatisfied users. Among the re-engaged users in this period, there are sN' satisfied users and $(1-s)N'$ unsatisfied users.

Fig. 6. Estimated measurement error on user satisfaction rate for different Δp based on (8)

Accordingly, the total number of active users in the current period $T^{(k)}$ is $spN + sN'$, and the total number of users remaining in the study is $spN + (1 - s)(p - \Delta p)N + N'$, and the estimand of user satisfaction rate in the current period $\hat{s}$ is defined accordingly. However, this is not an unbiased estimator of s due to the feedback effect, shown in (7). We further assume that the system of interest has reached long-term equilibrium s.t. the number of active users in adjacent time periods is nearly identical, and empirically Δp is reasonably small s.t. $N' + pN \simeq N$. With these assumptions, we propose an estimator of the measurement error as in (8).

$$\epsilon = \hat{s} - s = \frac{s\Delta p(1 - s)}{p - \Delta p + s\Delta p + \frac{N'}{N}} \qquad (7) \qquad \hat{\epsilon} = \frac{s\Delta p(1 - s)}{1 - \Delta p(1 - s)} \qquad (8)$$

For an IA with an actual user satisfaction rate $s = 60\%$, $\Delta p = 0.3$, a simple survey of active users in the current period would yield a user satisfaction rate of 68%. A simulation study is presented in Figure 6. This error would be further amplified should the feedback effect (quantified by Δp) be stronger.

6.3   Implications of the Language Convergence

Language convergence has an equally profound impact on usage-based evaluation: as shown in Table S2, the IA is nearly 3 times more likely to return a unhelpful response to high perplexity requests – which account for a sizable share of the exploratory usage in the new user cohort – than to low perplexity ones.

Should the true purpose of evaluation be to understand how well the IA addresses what the users *truly want*, rather than a limited set of tasks that the users have *compromised on*, we should always give sufficient consideration to the exploratory usage in our evaluation datasets. For emergent IAs with fast-growing user bases, this means one should carefully analyze the exploratory usage and iterate on new functionalities so that new users are retained at a higher level of perplexity, a reasonable proxy for request diversity and activity levels. For more established IAs, one should make a proactive effort to identify new user cohorts and decide on appropriate sampling strategies that balance the new and old.

6.4   In Search of Meaningful Metrics: Some Recommendations

While much of our discussion thus far has been in the context of IAs, it is perhaps not too wild a conjecture that the same phenomena can be observed

more broadly in any intelligent systems that entail some form of an interactive user interface (e.g. dictation software, handwriting recognition, and etc.). We therefore provide some recommendations on how to construct more meaningful metrics in the presence the feedback effect:

1. **Error Estimation and Selection Bias Mitigation**: for user-based studies, one may build an estimator to correct for measurement error after validating and quantifying the impact of the feedback effect on engagement; to control for the selection bias, one may elect to sample from a more comprehensive list of users (or the true population if feasible) rather than a list of active users in some fixed period to form the cohort of analysis;
2. **Stratified Sampling along Multiple Dimensions**: one may identify key dimensions of interest to form stratified sampling strategies with enhanced coverage, such as system-designed function areas, user frequency, linguistic representations, and perplexities.
3. **Exploratory Usage Retention:** exploratory usage often contains a more complete set of requests that users wish to accomplish through the system, and/or a more diverse set of user request patterns (accents in speech recognition, handwriting styles); it is a highly informative to collect data points that are yet to be subjugated to the inevitable influences of the feedback effect.

7   Discussions

Evaluation of IA systems is an important yet challenging problem. On the one hand, the capabilities and limitations of IAs shape user behaviors (*e.g.*, delayed engagement, dropout, and adaption). On the other hand, these very user behavior shifts in turn influence data collection and consequently the assessment of the IAs' capabilities and limitations. To our knowledge, this two-sided problem has not been formally discussed in the literature, at least in the context of real-world IAs. To fill this gap, this paper empirically studied the "feedback effect" nature of IA evaluation. On the one hand, we demonstrated that unhelpful interactions with the IA led to delayed and reduced user engagements, both short-term and mid-term. On the other hand, we examined long-term user behaviors, which suggested that as users gradually learned the limitations of the IA, they either dropped out or adapted (*i.e.*, "gave in"), and consequently increased the likelihood of helpful interactions with the IA.

Beside raising awareness within the data mining community, this paper aims to equip researchers and practitioners with tools for trustworthy IA evaluations. First, in cases where randomized controlled experiments are infeasible, we offered best practices on properly employing observational causal inference methods, and constructing offline metrics that take the censoring of user engagements into account. Second, to reduce the *feedback loop* problem in data collection and sampling, it is important to gauge users' experience with the IA and control for confounding factors if possible. When not possible, researchers should consider stratified sampling or boosting the signals from more complex intents, or creating synthetic test data that varies in complexity, especially targeting

more complex sentence structures and intent linguistic features which may be under-represented. Third, we have demonstrated that a key factor contributing to unsatisfactory IA experiences for new users is that the language they use is too complex in some way. We have also shown that users who fail to adapt by using simpler language often do not continue to use the IA. These insights immediately suggest growth opportunities to capitalize on. For example, multiple existing IAs offer a set of example conversations in different domains, in order to "train" new users to use the IA successfully right from the get go.

Our work implies multiple future directions, from both product and research perspectives. First, other than new user training (that might very well be skipped), what more can we do to convey the IA's capabilities and limitations, and help users engage more productively? Alternatively, how can we intervene early on and retain those "drop-outs," who provide invaluable feedback to help improve our system? Second, although we collected a rich set of covariates to ensure unconfoundedness, we can further assess the robustness of the established causal links, by leveraging classic sensitivity analysis techniques [21]. Third, while this paper focuses on off-line evaluation for IAs, it is possible to apply the proposed methodologies and recommendations in other settings (*e.g.*, on-line experimentation) and software products (*e.g.*, search engines).

Acknowledgements

References

1. Adiwardana, D., Luong, M.T., So, D.R., Hall, J., Fiedel, N., Thoppilan, R., Yang, Z., Kulshreshtha, A., Nemade, G., Lu, Y., et al.: Towards a human-like open-domain chatbot. arXiv preprint arXiv:2001.09977 (2020)
2. Andersen, P.K., Syriopoulou, E., Parner, E.T.: Causal inference in survival analysis using pseudo-observations. Statistics in Medicine **36**, 2669–2681 (2017)
3. de Barcelos Silva, A., Gomes, M.M., da Costa, C.A., da Rosa Righi, R., Barbosa, J.L.V., Pessin, G., De Doncker, G., Federizzi, G.: Intelligent personal assistants: A systematic literature review. Expert Systems with Applications **147**, 113–193 (2020)
4. Bird, S., Klein, E., Loper, E.: Natural language processing with Python: analyzing text with the natural language toolkit. O'Reilly Media, Inc (2009)
5. Chattaraman, V., Kwon, W.S., Gilbert, J.E., Ross, K.: Should AI-based, conversational digital assistants employ social-or task-oriented interaction style? A task-competency and reciprocity perspective for older adults. Computers in Human Behavior **90**, 315–330 (2019)
6. Duplessis, G., Clavel, C., Landragin, F.: Automatic measures to characterise verbal alignment in human-agent interaction. In: Proceedings of the 18th Annual Meeting of the Special Interest Group on Discourse and Dialogue. pp. 71–81 (2017)
7. Friedberg, H., Litman, D., Paletz, S.B.: Lexical entrainment and success in student engineering groups. In: Proceedings of the 2012 IEEE Spoken Language Technology Workshop (2012)

8. Gao, J., Galley, M., Li, L.: Neural approaches to conversational ai. Foundations and Trends in Information Retrieval **13**(2-3), 127–298 (2019)

9. Glass, J.: Challenges for spoken dialogue systems. In: Proceedings of the 1999 IEEE ASRU Workshop. vol. 696 (1999)

10. Holland, P.W.: Statistics and causal inference. Journal of the American Statistical Association **81**(396), 945–960 (1986)

11. Iacus, S.M., King, G., Porro, G.: Causal inference without balance checking: Coarsened exact matching. Political analysis **20**(1), 1–24 (2012)

12. Jiang, J., Hassan Awadallah, A., Jones, R., Ozertem, U., Zitouni, I., Gurunath Kulkarni, R., Khan, O.Z.: Automatic online evaluation of intelligent assistants. In: Proceedings of the 24th International Conference on World Wide Web. pp. 506–516 (2015)

13. Jurafsky, D.: Speech & language processing. Pearson Education India (2000)

14. Kepuska, V., Bohouta, G.: Next-generation of virtual personal assistants (Microsoft Cortana, Apple Siri, Amazon Alexa and Google Home). In: Proceedings of the 2018 IEEE 8th Annual Computing and Communication Workshop and Conference. pp. 99–103 (2018)

15. Kiseleva, J., Williams, K., Jiang, J., Hassan Awadallah, A., Crook, A.C., Zitouni, I., Anastasakos, T.: Understanding user satisfaction with intelligent assistants. In: Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval. pp. 121–130 (2016)

16. Komatani, K., Kawahara, T., Okuno, H.G.: Analyzing temporal transition of real user's behaviors in a spoken dialogue system. In: Proceedings of the 8th Annual Conference of the International Speech Communication Association. pp. 1837–1840 (2007)

17. Lee, G.G., Kim, H.K., Jeong, M., Kim, J.H.: Natural Language Dialog Systems and Intelligent Assistants. Springer (2015)

18. Lee, N., Bang, Y., Madotto, A., Khabsa, M., Fung, P.: Towards few-shot fact-checking via perplexity. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 1971–1981 (2021)

19. Levow, G.A.: Learning to speak to a spoken language system: Vocabulary convergence in novice users. In: SIGDIAL Workshop of Discourse and Dialogue (2003)

20. Li, F., Morgan, K.L., Zaslavsky, A.M.: Balancing covariates via propensity score weighting. Journal of the American Statistical Association **113**(521) (2018)

21. Liu, W., Kuramoto, S.J., Stuart, E.A.: An introduction to sensitivity analysis for unobserved confounding in nonexperimental prevention research. Prevention Science **14**, 570–580 (2013)

22. Lopatovska, I., Rink, K., Knight, I., Raines, K., Cosenza, K., Williams, H., Sorsche, P., Hirsch, D., Li, Q., Martinez, A.: Talk to me: Exploring user interactions with the amazon alexa. Journal of Librarianship and Information Science **51**(4) (2019)

23. Miller Jr, R.G.: Survival Analysis. John Wiley & Sons (2011)

24. Nenkova, A., Gravano, A., Hirschberg, J.: High frequency word entrainment in spoken dialogue. In: Proceedings of ACL-08: HLT, Short Papers (Companion Volume). pp. 169–172 (2008)

25. Parent, G., Eskenazi, M.: Lexical entrainment of real users in the let's go spoken dialog system. In: Proceedings of the 11th Annual Conference of the International Speech Communication Association. pp. 3018–3021 (2010)

26. Purington, A., Taft, J.G., Sannon, S., Bazarova, N.N., Taylor, S.H.: "Alexa is my new BFF" social roles, user satisfaction, and personification of the Amazon

Echo. In: Proceedings of the 2017 CHI Conference: Extended Abstracts on Human Factors in Computing Systems. pp. 2853–2859 (2017)

27. Reitter, D., Keller, F., Moore, J.D.: Computational modelling of structural priming in dialogue. In: Proceedings of the 2006 Human Language Technology Conference of the NAACL, Companion Volume: Short Papers. pp. 121–124 (2006)

28. Rubin, D.B.: Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology **66**, 688 (1974)

29. Santos, J., Rodrigues, J., Casal, J., Saleem, K., Denisov, V.: Intelligent personal assistants based on internet of things approaches. IEEE Systems Journal **12**(2), 1793–1802 (2016)

30. Walker, M.A., Stent, A., Mairesse, F., Prasad, R.: Individual and domain adaptation in sentence planning for dialogue. Journal of Artificial Intelligence Research **30**, 413–456 (2007)

31. Wang, Y., Gupta, S., Lu, J., Mahmoudzadeh, A., Liu, S.: On heavy-user bias in A/B testing. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management. pp. 2425–2428 (2019)

32. Wen, T.H., Gasic, M., Mrksic, N., Su, P.H., Vandyke, D., Young, S.: Semantically conditioned lstm-based natural language generation for spoken dialogue systems. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. pp. 1711–1721 (2015)

33. Xiu, Z., Tao, C., Henao, R.: Variational learning of individual survival distributions. In: Proceedings of the ACM Conference on Health, Inference, and Learning. pp. 10–18 (2020)

34. Zeng, S., Li, F., Hu, L.: Propensity score weighting analysis of survival outcomes using pseudo-observations. Statistica Sinica (2021), DOI:10.5705/ss.202021.0175

Supplemental Material for "How the Feedback Effect Shapes User Behavior with Intelligent Assistants"

A Observational Study on IA's Helpfulness and users Engagement

A.1 Balancing Weights

Li [20] proposed a family of balancing weights which enjoys balanced weighted distributions of covariates among treatment groups. , which enjoys balanced weighted distributions of covariates among treatment groups. Inverse-probability weights (IPW) is a special case of this family. Let $f(x)$ denotes the covariates distribution of the population, and $f_0(x), f_1(x)$ as the control or treatment group distribution respectively.

With the balancing weights and tilting function $h(x)$, the weighted distributions for different treatment groups are evened out.

$$f_1(x)w_1(x) = f_0(x)w_0(x) = f(x)h(x),$$

The tilting function defines the target population and the estimands of interest, and also determined the weights accordingly.

$$\begin{cases} w_1^h(x) \propto \frac{h(x)}{e(x)}, & \text{for } Z = 1 \\ w_0^h(x) \propto \frac{h(x)}{1-e(x)}, & \text{for } Z = 0. \end{cases} \tag{9}$$

The population level causal estimands of interest, the weighted average treatment effect (WATE) shown in Eq.(10), is based on the balancing weights.

$$\hat{\tau} = \hat{\mu}_1 - \hat{\mu}_0 = \frac{\sum_{i=1}^N z_1^h(\mathbf{x_i})\mathbf{Z_i}\mathbf{Y_i}}{\sum_{i=1}^N z_1^h(\mathbf{x_i})\mathbf{Z_i}} - \frac{\sum_{i=1}^N z_0^h(\mathbf{x_i})\mathbf{Z_i}\mathbf{Y_i}}{\sum_{i=1}^N z_0^h(\mathbf{x_i})\mathbf{Z_i}}. \tag{10}$$

When $h(x) = 1$, it is the inverse-probability weights (IPW). As the name suggested, the inverse of the probability that a unit is assigned to the observed group is the weight, and the corresponding estimand is ATE.

In Table S1, it summarizes some weights from the balancing weights family.

Method	Target population	Tilting function $h(x)$
IPW	Combined	1
OW	Overlapped	$e(x)(1 - e(x))$
Entropy	Entropy based	$e(x)\log(e(x))$ $- (1 - e(x))\log(1 - e(x))$

Table S1. Target population, tilting functions comparison for the balancing weights methods

---

Algorithm 1: Causal discoveries of IA helpfulness on users time-to-next-engagement.

Input: Confounding variables $\mathbf{X}$, IA helpfulness indicator $\mathbf{Z}$, observed time-to-next-engagement or censored time $\tilde{T}$ with censoring indicator Δ.

for $t \in (0, t_{max}]$ **do**

    Step 1: Obtain pseudo-observations $\hat{\theta}_i(t)$

    Step 2: Estimate propensity score $\hat{e}(x)$

    Step 3: Calculate balancing weights $w_1(x), w_0(x)$ Eq. (2)

    Step 4: Derive the causal estimands of re-engagement probability difference (RPCE) $\hat{\tau}^w(t)$ Eq. (6)
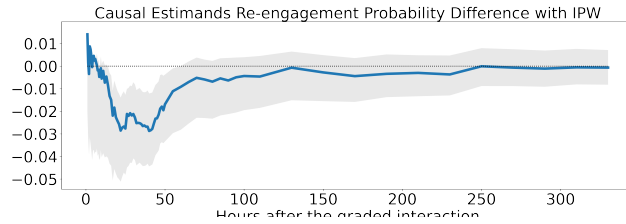
end

A.2   Outcome: time-to-next-engagement

We have summarized our analysis pipeline in Algorithm 1. Starting from estimating the pseudo-observations and propensity scores, then combining together to obtain the final weighted causal effect, re-engagement probability difference.

Apart from the OW results presented in the main context, the IPW-weighted RPCE implies similar conclusions, as presented in Figure S1. The IPW results yield a significant difference in the time window between hours 16-50 with p-values < 0.05.

Fig. S1. The re-engagement probability causal estimands (RPCE) as a function of time after the annotated interaction, with associated 95% confidence interval (shaded gray). The dotted horizontal line represents the difference is 0. The average re-engagement probability in *Unhelpful* cohort is lower than *Helpful* cohort in the following 336 hours (2 weeks) period. The significant gap happens around the $8 \sim 65$ hours with p-value < 0.05.
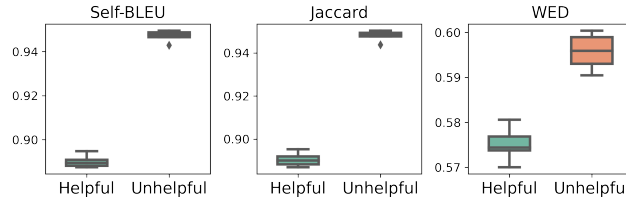
B   Language Convergence

B.1   Discussions on Different Language Complexity Metrics

Sentence complexity provides a linguistic measurement of user requests. Some simple metrics of sentence complexity are the number of tokens and the number of distinct N-grams per request [40]. These are quite intuitive but hardly reveal deep insights into

	High Perplexity	Low Perplexity	Total
Helpful	1245	11592	12837
Unhelpful	308	839	1147
Total	1553	12431	13984

Table S2. Contingency table of IA Quality (helpful vs. unhelpful) and PP score, with a statistically significant correlation.
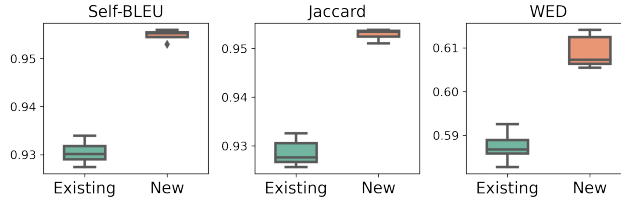
the requests [36]. To compare the sentence complexity of new requests with existing ones, there are three common metrics. Two of them, the selfBlEU score and the Jaccard similarity, are pairwise metrics based on overlapping N-grams between target and reference sentences [41, 39]. The third one, word embedding diversity (WED) measures cosine distances of word embeddings of sentences in comparison. Yet, these metrics are computationally inefficient owning to the enumeration of sentences in large datasets. On the other hand, the perplexity score (PP) is a scalable and informative metric that reveals both the syntactic complexity and the lexical diversity of a new request within a given text based on a trained language model [37]. Researches have also shown that the PP score of user requests inversely correlates with the success of the subsequent IA response: the higher the PP score, the more complex the request, the less likely the subsequent IA response is going to be successful [1].

Fig. S2. Language diversity for helpful utterances and unhelpful utterances. The bootstrapping results have shown that the unhelpful group has greater diversity with all three diversity metrics.

We first quantify the diversity with three widely used pairwise-based metrics without considering domain specific PP scores, i) selfBLEU: measures the diversity within the utterances cluster, i.e., average (1-pairwise BLEU score) ii) Jaccard similarity: average (1- pairwise word overlap) within the group iii) Word Embedding Diversity (WED): average (1- pairwise cosine distance) among embeddings of vectors in the utterances set. From Table S2, the unhelpful interactions have significantly larger diversities than the helpful group in general. Also, to check the language diversity changing over time, we have run the above analysis on a group of likely new users (non-habitual) against a group of habitual users (in Figure S3). From the exploratory analysis, the unhelpful interactions cohort and new users cohort share a high diversity in common.

Revisiting the datasets described in Section 3.1, we further studied the correlation between language perplexity and the human-label review of helpfulness. As illustrated in the 2 × 2 contingency table (Table S2), roughly 20% of high perplexity requests are unhelpful, on the contrary, only 6% of low perplexity ones are unhelpful. The association between perplexity score and IA helpfulness is statistically significant with p-value less than 0.0001.

Fig. S3. Language diversity for new and existing users cohort. Obviously the new group has greater diversity.

Revisiting the datasets described in Section 3.1, we studied the correlation between language perplexity and the human-label review of helpfulness. About 20% of high perplexity requests are unhelpful, compared to 6% of low perplexity ones. The association between perplexity score and IA helpfulness is statistically significant with p-value $< 1e-4$ with the Chi-squared test (Table S2).

B.2   Evaluation Metrics for Request Complexity

We evaluate user requests complexity along the following dimensions:

1. **Syntactic complexity.** The same intent can be expressed in numerous ways, using varying levels of complexity of sentence structure. For example, "call mom" and "place a telephone call to my mom please" express the same intent, but the latter is more complex structurally.
2. **Sub-intents entanglement within a request.** The level of detail required to address a user request may vary, with more complex intents requiring more detailed answers. For example, a simple request could be "Is it going to snow today?" and a more complex one could be "How many inches of snow are we expecting over the next 7 days?".
3. **Lexical and semantic diversity of the request.** Unlike common topics within a domain, infrequent topics are more difficult for the IA to handle. For example, within the weather domain, users commonly ask about the *weather*, *temperature*, and *rain*, while diverse items include topics like *wind*, *tide*, *barometric*, *moon*, and *tornado*.

Task-oriented IA systems often have a list of predefined domains: *Weather*, *Payment*, *Phone*, *Music*, *LocalBusiness*, etc. Consequently, requests in a specific domain often share typical recurrent linguistic patterns. As a consequence, it is important to examine language diversity on a by-domain level, rather than on the dataset as a whole. This is because different domains may have different vocabulary sizes and high frequency tokens. To evaluate the complexity and diversity of a request, a simple metric like vocabulary size may be helpful in some utility domains (*e.g.*, timer, alarm), but it could be misleading in communication domains (*e.g.*,, phone call, SMS). Specifically, in communication domains, vocabulary size may be biased by the request's content (or payload) rather than the sophistication with which users express their intent. For example, if a user sends a longer message, the vocabulary size will increase accordingly, but the way they ask the IA to send the message may remain the same. Pairwise comparison metrics such as the previously mentioned selfBLEU, Jaccard, and WED are sensitive to keywords and topics, but they do not distinguish payload from non-payload content.

SM References

35. Adiwardana, D., Luong, M.T., So, D.R., Hall, J., Fiedel, N., Thoppilan, R., Yang, Z., Kulshreshtha, A., Nemade, G., Lu, Y., et al.: Towards a human-like open-domain chatbot. arXiv preprint arXiv:2001.09977 (2020)
36. Evans, N., Levinson, S.C.: The myth of language universals: Language diversity and its importance for cognitive science. Behavioral and Brain Sciences **32**, 429–448 (2009)
37. Holtzman, A., Buys, J., Du, L., Forbes, M., Choi, Y.: The curious case of neural text degeneration. In: Proceedings of International Conference on Learning Representations (2020)
38. Li, F., Morgan, K.L., Zaslavsky, A.M.: Balancing covariates via propensity score weighting. Journal of the American Statistical Association **113**(521) (2018)
39. Niwattanakul, S., Singthongchai, J., Naenudorn, E., Wanapu, S.: Using of jaccard coefficient for keywords similarity. In: IMECS (2013)
40. Xu, J., Ren, X., Lin, J., Sun, X.: Diversity-promoting gan: A cross-entropy based generative adversarial network for diversified text generation. In: EMNLP (2018)
41. Zhu, Y., Lu, S., Zheng, L., Guo, J., Zhang, W., Wang, J., Yu, Y.: Texygen: A benchmarking platform for text generation models. In: SIGIR (2018)